EXTRACTION OF INDEPENDENT VECTOR COMPONENT FROM UNDERDETERMINED MIXTURES THROUGH BLOCK-WISE DETERMINED MODELING

Zbyněk Koldovský, Jiří Málek, and Jakub Janský

Acoustic Signal Analysis and Processing Group,
Faculty of Mechatronics, Informatics and Interdisciplinary Studies,
Technical University of Liberec, Liberec, Czech Republic

ABSTRACT

We propose a new model for blind source extraction where the source of interest is assumed to be static while the background noise is dynamic. The model is determined within short blocks (the same number of sources as that of sensors), however, the noise subspace can be changing from block to block. We propose a gradient-based algorithm that jointly extracts an independent vector component from a set of mixtures obeying the model based on maximum quasi-likelihood principle. Simulations confirm the validity of the approach, and experiments with real-world recordings show promising results.

Index Terms— Blind Source Extraction, Underdetermined Mixing, Independent Vector Analysis, Speech Enhancement

1. INTRODUCTION

Frequency-domain Independent Component Analysis (FDICA) has been a popular method for blind separation of audio signals that were recorded in a reverberated room [1]. FDICA operates in a Short-term Fourier Transform (STFT) domain by applying ICA to each frequency bin [2]. The observed data in the kth frequency bin are described as a determined instantaneous mixture

$$\mathbf{x}^k = \mathbf{A}^k \mathbf{s}^k, \qquad k = 1, \ldots, K, \tag{1}$$

where $\mathbf{s}^k$ and $\mathbf{x}^k$ denote the $d \times 1$ vector of random variables representing the original and mixed signals (their STFT values in frames), respectively; the signals are modeled as i.i.d. sequences. $\mathbf{A}^k$ denotes the $d \times d$ nonsingular mixing matrix; K is the number of frequency bins in the STFT.

In ICA, $\mathbf{s}^k$ are assumed to be mutually independent and non-Gaussian. When the frequency bins are treated separately as in FDICA, there is the permutation problem (separated signals have random orders) [3]. Therefore, Independent Vector Analysis (IVA) has been proposed that separates K mixtures jointly. The separated frequency components belonging to the same source, represented by a vector component, are assumed to be mutually dependent [4]. The jth vector component is denoted as $\mathbf{s}_j = [s_j^1, \ldots, s_j^K]^T$. Similarly to ICA, $\mathbf{s}_1, \ldots, \mathbf{s}_d$ are assumed to be independent.

In ICA and IVA, the same number of sources as is the number of microphones is assumed (determined mixing), which causes certain limitations, especially, in highly dynamic and underdetermined mixing environments. For example, in CHiME-3 and CHiME-4, multi-channel noisy recordings of a speaker were considered [5], where the position of the speaker is almost static (small movements with head) while the background is dynamic and quickly changing. The background involves various sounds such as cars passing by, environmental noises (street, cafeteria, bus), other speakers, etc. Motivated by these situations, we propose an advanced mixing model for situations like these.

In the proposed model, we assume that the observed signals can be described as the determined mixture (1) during short intervals (blocks). The mixing matrix and the distributions of sources can be varying from block to block, up to the first column of the mixing matrix, which is assumed to be constant. The first source can thus represent a static speaker. The proposed model is described through

$$\mathbf{x}^{k,m} = \mathbf{A}^{k,m} \mathbf{s}^{k,m}, \tag{2}$$

where m is the block index, $m = 1, \ldots, M$. The first column of $\mathbf{A}^{k,m}$, denoted by $\mathbf{a}_1^{k,m}$, is independent of m, so it can be denoted as $\mathbf{a}^k = \mathbf{a}_1^{k,m}$. The first source will be denoted by $s^{k,m} = s_1^{k,m}$ and its corresponding vector component, for the mth block, as $\mathbf{s}^m = [s^{1,m}, \ldots, s^{K,m}]^T$. The mixing matrices $\mathbf{A}^{k,m}$ are assumed to be square and non-singular, so the model can be referred to as *block-wise determined*.

This model is suitable when only the static source should be extracted (Blind Source Extraction - BSE). By rewriting $\mathbf{x}^{k,m}$ in the form

$$\mathbf{x}^{k,m} = \mathbf{A}^{k,m} \mathbf{s}^{k,m} = \mathbf{a}^k s^{k,m} + \mathbf{y}^{k,m}, \tag{3}$$

where $\mathbf{y}^{k,m}$ denotes the dynamic part of the mixture, we can

This work was supported by the United States Department of the Navy, Office of Naval Research Global, through Project No. N62909-18-1-2040, by The Czech Science Foundation through Project No. 17-00902S, and by the Student Grant Scheme 2018 project of the Technical University in Liberec.

formulate the main assumption: $s^{k,m}$ and $\mathbf{y}^{k,m}$ are independent. For the joint extraction, we will also assume that the elements of $\mathbf{s}^m$ are mutually dependent as in IVA [4, 6].

It should be noted that if $\mathbf{A}^{k,m}$, $m = 1, \ldots, M$, share any other constant column, the corresponding (static) source can also play the role of the source to be extracted. To prevent from extracting a different static source, we assume that an initial (approximate) value of $\mathbf{a}^k$, $k = 1, \ldots, K$, is given.

In the following, we derive a gradient algorithm for the proposed joint extraction problem based on the maximum quasi-likelihood approach. Section 3 is devoted to simulations and Section 4 describes an application to the CHiME-4 recordings.

2. BLOCK-WISE INDEPENDENT VECTOR EXTRACTION

To extract the signal of interest (SOI) from K block-wise determined mixtures, we will modify the approach recently proposed in [6] referred to as Independent Vector Extraction (IVE). IVE is designed for extracting the vector component $\mathbf{s}^m$ jointly from the K mixtures $\mathbf{x}^m = [\mathbf{x}^{1,m}; \ldots; \mathbf{x}^{K,m}]$, that is, considering one block only. For $K = 1$, IVE coincides with Independent Component Extraction (ICE) [7, 8].

2.1. Determined mixtures

For now, let us consider only one block. In IVE, the de-mixing matrix is parametrized to separate $\mathbf{x}^{k,m}$ into two independent components: the extracted source $s^{k,m}$ and the other background signals $\mathbf{z}^{k,m}$ (spanning the same $(d-1)$-dimensional subspace as $\mathbf{y}^{k,m}$). Owing to (3), the de-mixing matrix can have the following structure:

$$\mathbf{W}^{k,m} = \begin{pmatrix} \mathbf{w}^{k,m\,H} \\ \mathbf{B}^k \end{pmatrix} = \begin{pmatrix} \beta^{k,m\,*} & \mathbf{h}^{k,m\,H} \\ \mathbf{g}^k & -\gamma^k \mathbf{I}_{d-1} \end{pmatrix}, \quad (4)$$

where $\mathbf{w}^{k,m} = [\beta^{k,m}; \mathbf{h}^{k,m}]$ is the separating vector such that $s^{k,m} = (\mathbf{w}^{k,m})^H \mathbf{x}^{k,m}$, and $\mathbf{a}^k = [\gamma^k; \mathbf{g}^k]$. The structure of $\mathbf{B}^k$, which depends purely on $\mathbf{a}^k$, guarantees that $\mathbf{z}^{k,m} = \mathbf{B}^k \mathbf{x}^{k,m}$ does not contain any contribution of $s^{k,m}$ when $\mathbf{a}^k$ is the true mixing vector, because $\mathbf{B}^k \mathbf{a}^k = \mathbf{0}$.

The free parameters in (4) are represented by $\mathbf{w}^{k,m}$ and $\mathbf{a}^k$. These vectors should be both related to the same source $s^{k,m}$ (the separating and the mixing vector, respectively). However, they are linked only through the distortionless condition $(\mathbf{w}^{k,m})^H \mathbf{a}^k = 1$. It might thus happen that the estimated values of $\mathbf{w}^{k,m}$ and $\mathbf{a}^k$ eventually do not correspond to the same source. To prevent from this unwanted result, the orthogonality constraint (OG) can be applied; see, e.g., [9].

OG means that $\mathbf{w}^{k,m}$ and $\mathbf{a}^k$ are linked so that the sample-based correlations between the estimates of $s^{k,m}$ and $\mathbf{z}^{k,m}$ are zero. Specifically, OG requires that

$$(\mathbf{w}^{k,m})^H \hat{\mathrm{E}}[\mathbf{x}^{k,m}(\mathbf{x}^{k,m})^H] \mathbf{B}^k = (\mathbf{w}^{k,m})^H \widehat{\mathbf{C}}_{\mathbf{x}}^{k,m} \mathbf{B}^k = \mathbf{0}, \quad (5)$$

where $\hat{\mathrm{E}}[\cdot]$ denotes the sample averaging operator, and $\widehat{\mathbf{C}}_{\mathbf{x}}^{k,m} = \hat{\mathrm{E}}[\mathbf{x}^{k,m}(\mathbf{x}^{k,m})^H]$ denotes the sample-based correlation of $\mathbf{x}^{k,m}$. In [9] it is derived that, together with the distortionless condition $(\mathbf{w}^{k,m})^H \mathbf{a}^k = 1$, the OG means

$$\mathbf{w}^{k,m} = \frac{(\widehat{\mathbf{C}}_{\mathbf{x}}^{k,m})^{-1} \mathbf{a}^k}{(\mathbf{a}^k)^H (\widehat{\mathbf{C}}_{\mathbf{x}}^{k,m})^{-1} \mathbf{a}^k}. \quad (6)$$

In the statistical model, we assume that the joint pdf of $\mathbf{s}^m$, denoted by $p_m(\mathbf{s}^m)$, is non-Gaussian. Since each $\mathbf{z}^{k,m}$ is a mixture of unknown latent variables, and IVE does not aim at further analysis of its components (as opposed to ICA/IVA), it is reasonable to exploit only its second-order statistics and assume that the pdf of $\mathbf{z}^m = [\mathbf{z}^{1,m}; \ldots; \mathbf{z}^{K,m}]$, denoted by $p_{\mathbf{z}^m}$, is Gaussian.

From the assumption that $\mathbf{s}^m$ is independent of $\mathbf{z}^m$, the log-likelihood function for one sample of $\mathbf{x}^m$ reads [6]

$$\mathcal{L}_m(\mathbf{x}^m | \{\mathbf{a}^k, \mathbf{w}^{k,m}\}_{k=1}^K) = \log p_m(\mathbf{s}^m) - \\ - \sum_{k=1}^K \Big\{ (\mathbf{z}^{k,m})^H \mathbf{R}^{k,m} \mathbf{z}^{k,m} + \log |\det \mathbf{W}^{k,m}|^2 \Big\}, \quad (7)$$

where $\mathbf{R}^{k,m}$ is the inverse of the covariance matrix of $\mathbf{z}^{k,m}$. Since this matrix is not known, it is later replaced by the inverse of the sample-based covariance of the current estimate of $\mathbf{z}^{k,m}$, i.e., by $\hat{\mathrm{E}}[\mathbf{z}^{k,m}(\mathbf{z}^{k,m})^H]^{-1}$ (the quasi-likelihood approach).

By taking the average of (7) over the samples available for the mth block, and by setting $\mathbf{R}^{k,m} = \hat{\mathrm{E}}[\mathbf{z}^{k,m}(\mathbf{z}^{k,m})^H]^{-1}$, it can be shown as in [6] that the gradient of (7) subject to $\mathbf{a}^k$ under the constraint (6) is

$$\frac{\partial \mathcal{L}_m}{\partial \mathbf{a}^k} = \mathbf{w}^{k,m} - \lambda^{k,m} (\widehat{\mathbf{C}}_{\mathbf{x}}^{k,m})^{-1} \hat{\mathrm{E}}[\mathbf{x}^{k,m} \psi^{k,m}(\mathbf{s}^m)], \quad (8)$$

where $\lambda^{k,m} = [(\mathbf{a}^k)^H (\widehat{\mathbf{C}}_{\mathbf{x}}^{k,m})^{-1} \mathbf{a}^k]^{-1}$, and $\psi^{k,m}(\mathbf{s}^m) = -\partial / (\partial s^{k,m}) \log p_m(\mathbf{s}^m)$ is the kth score function of p_m. Since these score functions are not known, they are replaced by an appropriate nonlinearity $\phi^k(\mathbf{s}^m)$ (for simplicity, let us select ϕ^k which is independent of m). Then, ϕ^k must be normalized so that $\hat{\mathrm{E}} = [s^{k,m} \phi^k(\mathbf{s}^m)] = 1$[1], so the modified gradient reads

$$\Delta^{k,m} = \mathbf{w}^{k,m} - \lambda^{k,m} (\widehat{\mathbf{C}}_{\mathbf{x}}^{k,m})^{-1} \hat{\mathrm{E}}[\mathbf{x}^{k,m} \phi^k(\mathbf{s}^m)] / \nu^{k,m}, \quad (9)$$

where $\nu^{k,m} = \hat{\mathrm{E}}[s^{k,m} \phi^k(\mathbf{s}^m)]$.

The gradient algorithm, referred to as OGIVE$_{\mathbf{a}}$, being initialized by an approximate values of the true $\mathbf{a}^1, \ldots, \mathbf{a}^K$, iterates by performing small updates in the directions given by $\Delta^{k,m}$, that is,

$$\mathbf{a}^k \leftarrow \mathbf{a}^k + \mu \Delta^{k,m}, \quad (10)$$

where μ is a step size constant. After each update, the separating vector $\mathbf{w}^{k,m}$ is recomputed according to (6). The optimization is stopped when the norm of $\Delta^{k,m}$ is smaller than a selected threshold.

[1] The nonlinearity must be normalized so that the true value of $\mathbf{a}^k$ is the stationary point of the contrast (for infinite number of samples) [10, 6].

The observed signals can be whitened before OGIVE$_\mathbf{a}$ is applied so that $\widehat{\mathbf{C}}_\mathbf{x}^{k,m} = \mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix. It is an often used pre-processing step in ICA/IVA. In that case, (6) simplifies to $\mathbf{w}^{k,m} = \mathbf{a}^k / \|\mathbf{a}^k\|^2$.

2.2. Algorithm for block-wise IVE

Now, we start considering all M blocks, and we take the advantage of the fact that $\mathbf{a}_k$ is constant over the blocks for all $k = 1, \ldots, K$. The contrast function is obtained through averaging (7) over the blocks, because, in the model, samples are assumed to be independently distributed. Hence, also the gradient of the quasi-log-likelihood contrast (9) can be averaged over the blocks, by which we obtain $\Delta^k = \frac{1}{M}\sum_{m=1}^{M} \Delta^{k,m}$.

Then, the gradient update is, similarly to (10), $\mathbf{a}^k \leftarrow \mathbf{a}^k + \mu\Delta^k$. Owing to the scaling ambiguity, it is useful to normalize $\mathbf{a}^k$ after each update so that its first element is equal to one[2]. The block-dependent separating vectors are computed using (6). A pseudocode of the proposed method, from here referred to as BOGIVE$_\mathbf{a}$ (Block-wise OGIVE$_\mathbf{a}$), is described in Algorithm 1.

Algorithm 1: BOGIVE$_\mathbf{a}$: Block-wise orthogonally constrained independent vector extraction

Input: $\mathbf{x}^{k,m}, \mathbf{a}_{\text{ini}}^k$ $(k, m = 1, 2, \ldots)$, μ, `tol`
Output: $\mathbf{a}^k, \mathbf{w}^{k,m}$

1 **foreach** $k = 1, \ldots, K$ **do**
2 $\quad \widehat{\mathbf{C}}_\mathbf{x}^{k,m} = \hat{\mathrm{E}}[\mathbf{x}^{k,m}(\mathbf{x}^{k,m})^H]$;
3 $\quad \mathbf{a}^k = \mathbf{a}_{\text{ini}}^k / (\mathbf{a}_{\text{ini}}^k)_1$;
4 **end**
5 **repeat**
6 $\quad$ **foreach** $k = 1, \ldots, K$, $m = 1, \ldots, M$ **do**
7 $\quad\quad \mathbf{w}^{k,m} \leftarrow ((\mathbf{a}^k)^H(\widehat{\mathbf{C}}_\mathbf{x}^{k,m})^{-1}\mathbf{a}^k)^{-1}(\widehat{\mathbf{C}}_\mathbf{x}^{k,m})^{-1}\mathbf{a}^k$;
8 $\quad\quad \mathbf{s}^{k,m} \leftarrow (\mathbf{w}^{k,m})^H\mathbf{x}^{k,m}$;
9 $\quad$ **end**
10 $\quad$ **foreach** $k = 1, \ldots, K$, $m = 1, \ldots, M$ **do**
11 $\quad\quad \nu^{k,m} \leftarrow \hat{\mathrm{E}}[s^{k,m}\phi^k(\mathbf{s}^m)]$;
12 $\quad\quad$ Compute $\Delta^{k,m}$ according to (9);
13 $\quad$ **end**
14 $\quad$ **foreach** $k = 1, \ldots, K$ **do**
15 $\quad\quad \mathbf{a}^k \leftarrow \mathbf{a}^k + \mu\Delta^k$;
16 $\quad\quad \mathbf{a}^k \leftarrow \mathbf{a}^k / (\mathbf{a}^k)_1$;
17 $\quad$ **end**
18 **until** $\max\{\|\Delta^1\|, \ldots, \|\Delta^K\|\} < \texttt{tol}$;

2.3. OGIVE$_\mathbf{a}$ vs. BOGIVE$_\mathbf{a}$, and BOGIVE$_\mathbf{w}$

To see the main differences between OGIVE$_\mathbf{a}$ and BOGIVE$_\mathbf{a}$, consider these algorithms if they were applied to the entire batch of data (considering all blocks) obeying the model (2). While OGIVE$_\mathbf{a}$ optimizes mixing and separating vectors that are constant over blocks, BOGIVE$_\mathbf{a}$ assumes different separating vector in each block as follows from lines 7-8 in Algorithm 1. From line 7 it follows, that the outputs of the algorithms are different, especially, when $\widehat{\mathbf{C}}_\mathbf{x}^{k,m}$ is significantly block-dependent, and vice versa.

The whitening of input signals can be done also before BOGIVE$_\mathbf{a}$ is applied, which means that $M^{-1}\sum_{m=1}^{M} \widehat{\mathbf{C}}_\mathbf{x}^{k,m} = \mathbf{I}$. However, compared to OGIVE$_\mathbf{a}$, it is *not* possible to simplify (6) to $\mathbf{w}^{k,m} = \mathbf{a}^k / \|\mathbf{a}^k\|^2$ in BOGIVE$_\mathbf{a}$, because the sample-covariances on blocks remain different from the identity matrix after the whitening.

The idea of the block-determined modeling can be realized also in a variant where the mixing vectors $\mathbf{a}^{k,m}$ related to the SOI are varying from block to block while the separating vector $\mathbf{w}^k$ is constant. A gradient algorithm derived based on this model will be denoted BOGIVE$_\mathbf{w}$, however, details are not provided here due to lack of space.

3. SIMULATIONS

As a proof of concept, we conduct a simulated experiment with $d = 6$, $M = 5$, and $K = 3$. Signals are mixed according to (2). The first signals in the K mixtures are drawn independently from the circular Laplacean distribution and mixed by a random unitary matrix, so they are dependent (but uncorrelated). The other signals are circular Gaussian. The length of each block is $N_b = 1000$ samples; the entire batch has $N = 3000$ samples. Each signal in each block is multiplied by a random factor from $[0.01, 1]$ (random variance). Then, the global power (block-averaged variance) of the first signal is re-scaled so that it is equal to the average global power of the other signals. The mixing matrices are drawn from the uniform distribution; the real part in $[1; 2]$ and the imaginary part in $[0, 1]$; their first columns are constant over the blocks.

To extract the SOI from the kth mixture, the compared algorithms are initialized by a perturbed value of the true mixing vector, that is, by $\mathbf{a}_{\text{ini}}^k = \mathbf{a}^k + \mathbf{e}_{\text{ini}}^k$, where $\mathbf{e}_{\text{ini}}^k$ is random and orthogonal to $\mathbf{a}^k$, and $\|\mathbf{e}_{\text{ini}}^k\|^2 = \epsilon^2$. Then, the extracted signal is obtained using the separating vectors computed according to (6) where $\mathbf{a}^k$ is the estimated mixing vector. Signal-to-Interference Ratio (SIR) is computed as the ratio between the block-averaged powers of the SOI and of the background signals in the extracted signal.

We have compared several BSE algorithms in 1000 independent trials. It is not surprising that methods assuming the mixing model (1) such as OGIVE$_\mathbf{a}$, One-unit FastICA [12] or Natural Gradient [13] failed to estimate the mixing vector in this experiment, because the mixtures are block-wise determined (the resulting SIR is typically below or near 0dB). Therefore, we do not show results of these algorithms.

[2]This normalization of $\mathbf{a}^k$ such that $(\mathbf{a}^k)_1 = 1$ means that the scale of the extracted signal corresponds to that of its contribution (image) on the first input channel [11].

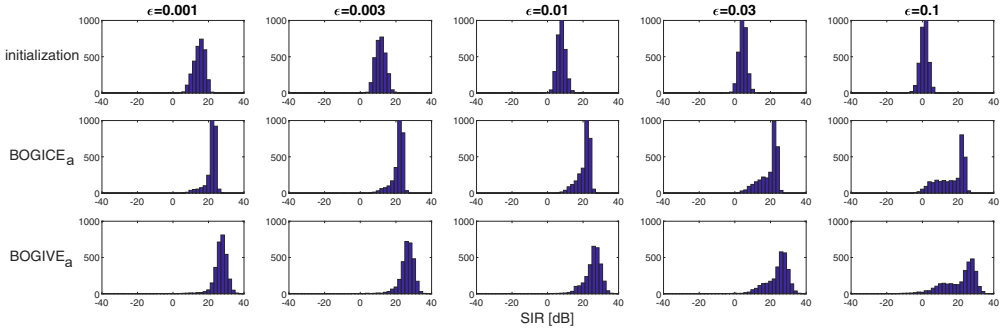

Fig. 1. Histograms of the output SIR (in dB) achieved by BOGICE$_\mathbf{a}$ and BOGIVE$_\mathbf{a}$ in 1000 trials ($K \times 1000 = 3000$ mixtures) for various levels of the initial perturbation ϵ. To compare, the SIR obtained with the initial value of the mixing vector is shown.

BOGIVE$_\mathbf{a}$, assuming the true number (and length) of the blocks, has been applied either to each of K mixtures separately as in ICE (the approach referred to as BOGICE$_\mathbf{a}$) or jointly (BOGIVE$_\mathbf{a}$). The nonlinearity used by the algorithm is $\phi^k(\mathbf{s}^m) = \tanh^*(s^{k,m})/\sqrt{\sum_{\ell=1}^K |s^{k,\ell}|^2}$. The problem of choosing an appropriate nonlinearity is beyond the scope of this paper; see, e.g., [4, 14, 15].

The histograms in Fig. 1 show that the achieved SIRs are often higher than the SIRs obtained with the initial mixing vector values, which proves the efficiency of the proposed algorithms. The number of successful extractions (SIR significantly higher than 0 dB) is decreasing with growing ϵ^2, which follows the fact that $\mathbf{a}_{\text{ini}}^k$ must lie in the region of convergence of the BSE algorithm to extract the SOI [6]. Finally, BOGIVE$_\mathbf{a}$ outperforms BOGICE$_\mathbf{a}$ on average as it takes the advantage of the joint extraction.

4. EXPERIMENTAL RESULTS

Here, we show results achieved on the CHiME-4 dataset of six-channel[3] recordings [5]. The dataset contains simulated (SIMU) and real-world (REAL) noisy utterances. Some recordings involve microphone failures. A method from [16] is used to detect these failures, and the malfunctioning channels are excluded.

The compared methods are used to extract speech from noisy recordings, and the enhanced signals are forwarded to the baseline speech recognizer from [5]; the achieved Word Error Rate (WER) is the final criterion. We operate in the STFT domain with the FFT length 512 and hop-size 128 using the Hamming window; the sampling frequency is 16 kHz.

Interestingly, BOGIVE$_\mathbf{a}$ did not succeed compared to BOGIVE$_\mathbf{w}$ in this experiment. This could be explained by the fact that speakers are performing small movements (varying mixing vector). Therefore, we apply BOGIVE$_\mathbf{w}$ (with $N_b = 170 \approx 1.3$ s) instead. It is initialized by the Relative Transfer Function (RTF) estimator from [17] which is (or is not) improved by Voice Activity Detector (VAD) as in [16]. The VAD is a feed-forward neural net (DNN) trained on the training set of CHiME-4. It estimates speech presence probability in the STFT domain.

[3]Microphone two is not used in the experiments as it is oriented away from the speaker.

System	Development		Test	
	REAL	SIMU	REAL	SIMU
Unprocessed	9.83	8.86	19.90	10.79
BeamformIt	5.77	6.76	11.52	10.91
GEV (VAD)	**4.61**	4.65	**8.10**	**5.99**
OGIVE$_\mathbf{w}$	5.59	4.96	9.51	6.34
BOGIVE$_\mathbf{w}$	5.64	4.84	8.98	6.21
BOGIVE$_\mathbf{w}$ (VAD)	5.39	**4.62**	8.54	6.30

Table 1. WERs [%] achieved in the CHiME-4 challenge.

The results are shown in Table 1. BeamformIt [18] is the baseline preprocessor of CHiME-4. The Generalized Eigenvalue Beamformer (GEV) from [19, 20] is one of the most successful enhancers in CHiME-4. In these experiments, it is endowed with a feed-forward DNN-based VAD. All the methods significantly improve the WER compared to the unprocessed case. The best results are achieved by GEV and by BOGIVE$_\mathbf{w}$ (VAD). BOGIVE$_\mathbf{w}$ without the DNN-supported initialization and BeamFormIt stand for fully unsupervised approaches. BOGIVE$_\mathbf{w}$ significantly outperforms BeamFormIt and its results are comparable with that of GEV.

5. CONCLUSIONS

The block-wise determined model was shown to be useful alternative when modeling underdetermined mixtures where only a target source should be extracted. The model can involve up to $(d-1)M+1$ sources. BOGIVE$_\mathbf{w}$ derived based on this model has been successfully applied in CHiME-4, showing competitive results to supervised DNN-based approaches. Simulations show that there is room for improvements in terms of the global convergence of the algorithm, which will be subject of our future research.

6. REFERENCES

[1] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Independent Component Analysis and Applications Series. Elsevier Science, 2010.

[2] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[3] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1592–1604, July 2007.

[4] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 70–79, Jan. 2007.

[5] Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, 2016.

[6] Z. Koldovský and P. Tichavský, "Gradient algorithms for complex non-gaussian independent component/vector extraction, question of convergence," *IEEE Transactions on Signal Processing*, vol. 67, no. 4, pp. 1050–1064, Feb 2019.

[7] Zbyněk Koldovský and Francesco Nesta, "Performance analysis of source image estimators in blind source separation," *IEEE Transactions on Signal Processing*, vol. 65, no. 16, pp. 4166–4176, Aug. 2017.

[8] V. Kautský, Z. Koldovský, and P. Tichavský, "Cramér-Rao-induced bound for interference-to-signal ratio achievable through non-Gaussian independent component extraction," in *2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Dec 2017, pp. 1–4.

[9] Zbyněk Koldovský, Petr Tichavský, and Václav Kautský, "Orthogonally constrained independent component extraction: Blind MPDR beamforming," in *Proceedings of European Signal Processing Conference*, Sept. 2017, pp. 1195–1199.

[10] Dinh-Tuan Antoine Pham, "Blind partial separation of instantaneous mixtures of sources," in *Proceedings of International Conference on Independent Component Analysis and Signal Separation*. 2006, pp. 868–875, Springer Berlin Heidelberg.

[11] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proceedings of International Conference on Independent Component Analysis and Signal Separation*, Dec. 2001, pp. 722–727.

[12] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1483–1492, July 1997.

[13] S. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," in *Proceedings of Neural Information Processing Systems*, 1996, pp. 757–763.

[14] Dinh Tuan Pham and P. Garat, "Blind separation of mixture of independent sources through a quasi-maximum likelihood approach," *IEEE Transactions on Signal Processing*, vol. 45, no. 7, pp. 1712–1725, Jul 1997.

[15] M. Novey and T. Adali, "Adaptable nonlinearity for complex maximization of nongaussianity and a fixed-point algorithm," in *2006 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, Sept 2006, pp. 79–84.

[16] Jiri Malek, Zbyněk Koldovský, and Marek Bohac, "Block-online multi-channel speech enhancement using DNN-supported relative transfer function estimates," *submitted*, 2018, available at: https://gitlab.tul.cz/jiri.malek/multichannel-enhancement.

[17] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug 2001.

[18] Xavier Anguera, Chuck Wooters, and Javier Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[19] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 196–200.

[20] Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach, "Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition," in *Proc. of the 4th Intl. Workshop on Speech Processing in Everyday Environments, CHiME-4*, 2016.