BLIND EXTRACTION OF TARGET SPEECH SOURCE: THREE WAYS OF GUIDANCE EXPLOITING SUPERVISED SPEAKER EMBEDDINGS

Jiri Malek, Jaroslav Cmejla, Zbynek Koldovsky

Faculty of Mechatronics, Informatics, and Interdisciplinary Studies,
Technical University of Liberec, Studentská 2, 461 17, Liberec, Czech Republic.

ABSTRACT

The manuscript deals with the robust extraction of a speaker of interest (SOI) from a mixture of audio sources. A blind algorithm based on independent vector extraction (IVE) is used, which, by definition, extracts an arbitrary source. To focus the extraction towards the SOI, a prior knowledge identifying the target source is required. To this end, the manuscript exploits speaker-identification based on embedding features computed via a pretrained forward sequential memory network (FSMN). We introduce and experimentally validate three ways how this prior knowledge can be employed in a blind algorithm, namely, speaker-specific initialization, pilot signal, and supervised deflation of the mixture. The experiments show that the proposed techniques complement each other and lead to robust identification/extraction of the SOI in difficult mixtures of three speakers.

Index Terms— Target speech extraction, blind extraction, supervised speaker identification, initialization, deflation, pilot signal.

1. INTRODUCTION

An important task in speech processing is to recover a speaker of interest (SOI) from a mixture of speakers and noises. This task is conventionally solved using a *full separation* of all speakers (their number must be usually known) and subsequent identification of the SOI among them. The separation can be based on *data-driven* principles of machine/deep learning [1–4] or on *model-based* blind source separation [5–9]. Data-driven techniques achieve high separation quality provided that scenario-specific training data are available. Blind techniques do not require any data and can be, in theory, applied in an arbitrary scenario. This freedom is, arguably, achieved at a cost of lower separation quality, because the employed statistical models only approximate the real conditions.

This work considers blind methods where the separation of speech proceeds in the time-frequency domain. Independent component analysis (ICA, [10]) estimates the sources separately at each frequency bin based on their statistical independence. This leads to the *permutation ambiguity* [11], and the affiliation of each frequency component to one of the sources must be identified. Independent vector analysis (IVA, [6, 7]) alleviates this drawback. It binds together the frequency components corresponding to a single source using higher-order dependencies among them. Non-negative matrix factorization (NMF, [8]) attempts to factorize the spectrogram of the mixture into a product of two non-negative components, recurring patterns and their activations. Independent low-rank matrix analysis (ILRMA, [9]) unifies the principles of IVA and NMF.

The number of speakers is often unknown and can be time-varying. It is therefore practical to limit the recovery exclusively to the SOI while leaving the unwanted speakers and other sources mixed. This task, referred to as *target speech extraction* [12–14], is the focus of this manuscript. To estimate the SOI, we utilize a variant of IVA called Independent vector extraction (IVE, [15]). By definition, IVE extracts an arbitrary independent source (based on often random/uniform initialization), so it must be guided towards the SOI by some prior information identifying the target source. To this end, we exploit the identification via embedding features encoding the characteristics of a speaker. The embeddings are computed using a pre-trained neural network; our implementation uses the feed-forward sequential memory network (FSMN, [16]).

This manuscript discusses three ways, how the embedding-based information can be introduced into the IVE algorithm. 1) *SOI specific initialization* uses the embeddings to find intervals in the mixture where the SOI is highly active. These intervals are used to obtain a rough estimate of the covariance matrix of the SOI and subsequently a multi-channel Wiener filter (MCWF, [2]) directed towards the SOI. The filter serves as starting point for IVE iterations. 2) *Pilot signal* [14] is a reference signal derived using the intervals of the mixture where the SOI is dominant. It is injected into the IVE to bind together frequency components that show dependence with the pilot, and, consequently, also with the SOI. 3) *Supervised deflation of the mixture* [17] is a correction mechanism which is applied when the initialization and the piloting fail. Using the embeddings, the extracted source is assessed whether it corresponds to the SOI. In case an unwanted source is identified, it is subtracted from the mixture, and the extraction is attempted again on the reduced mixture.

The identification of SOI using embeddings is also essential for deep-learning extractors; see, e.g., [12]. A video stream can be an alternative source of prior information for guidance. An initialization based on video was used to focus a blind algorithm towards the SOI in [18] and a video-based pilot signal was proposed in [19].

The current manuscript extends our previous works on blind extraction guided via speaker identification [14, 17], where the pilot signal and the deflation were introduced. The novel contribution of this work is the SOI-specific initialization exploiting embeddings and the experimental analysis of the combination of all three techniques. The SOI-specific initialization is computationally cheap and, when combined with piloting, reduces the number of cases when an unwanted source is extracted. Consequently, the costly deflation needs to be applied to a smaller number of mixtures, which reduces the running time of the extractor. The experimental analysis was performed on the spatialized multi-speaker Wall Street Journal datasets (MC-WSJ0-2mix, MC-WSJ0-3mix, [2, 20]). The results indicate that the three guiding techniques complement each other. Their simultaneous use improves the ability of IVE to extract SOI from difficult mixtures compared to the application of a subset of techniques.

This work was supported by the Technology Agency of the Czech Republic (Project No. TO01000027) and by The Czech Science Foundation (Project No. 20-17720S).

978-1-6654-6867-1/22/$31.00 ©2022 IEEE

2. PROBLEM FORMULATION

A potentially time varying mixture of D sources emitting *original signals* observed by D microphones can, in the short-time frequency domain, be approximated by the mixing model

$$\mathbf{x}_\ell^k = \mathbf{A}_\ell^k \mathbf{y}_\ell^k, \tag{1}$$

where $k = 1, \ldots, K$ is the frequency and $\ell = 1, \ldots, L$ is the frame index, $\mathbf{y}_\ell^k$ and $\mathbf{x}_\ell^k$ denote the original and mixed signals, respectively and $\mathbf{A}_\ell^k$ denotes the mixing matrix.

The mixing is assumed approximately static over a small number of subsequent frames, which is referred to as *block* in this manuscript. The mixture is thus divided into $t = 1, \ldots, T$ equally long blocks of $\ell_t = 1, \ldots, L_T$ frames, with a block-constant mixing matrix $\mathbf{A}_t^k$; the index of the ℓth frame within the tth block is denoted by ℓ_t. The mixing model for this *block-wise static* approach [14] is given by

$$\mathbf{x}_{\ell_t}^k = \mathbf{A}_t^k \mathbf{y}_{\ell_t}^k \qquad \text{for } \ell = 1, \ldots, L_T;\ t = 1, \ldots, T. \tag{2}$$

Note that this model becomes fully static when $T = 1$ or maximally time-varying when $T = L$.

In IVA, a complete de-mixing matrix $\mathbf{W}_t^k$ is sought such that it fulfills $\mathbf{W}_t^k \mathbf{x}_{\ell_t}^k = \mathbf{W}_t^k \mathbf{A}_t^k \mathbf{y}_{\ell_t}^k = \hat{\mathbf{y}}_{\ell_t}^k \approx \mathbf{y}_{\ell_t}^k$, i.e., it recovers all the sources present in the mixture assuming their independence. IVE seeks only one row of $\mathbf{W}_t^k$, denoted by $\mathbf{w}_t^k$, such that it extracts a single independent source (the SOI). The SOI ambiguity stems from the fact that any independent source can play the role of the SOI. Without any loss on generality, let the SOI be the first signal in $\mathbf{y}_{\ell_t}^k$ and $\mathbf{A}_t^k$ be partitioned as $\mathbf{A}_t^k = \begin{bmatrix}\mathbf{a}_t^k & \mathbf{Q}_t^k\end{bmatrix}$. Then, the equation (2) can be expressed in the form

$$\mathbf{x}_{\ell_t}^k = [\mathbf{a}_t^k \quad \mathbf{Q}_t^k] \begin{bmatrix} s_{\ell_t}^k \\ \mathbf{z}_{\ell_t}^k \end{bmatrix}, \tag{3}$$

where $s_{\ell_t}^k$ represents the SOI, and $\mathbf{z}_{\ell_t}^k$ are the other $d-1$ signals in the mixture. Subsequently, $\mathbf{W}_t^k$ can be partitioned as $[\mathbf{w}_t^k \quad (\mathbf{B}_t^k)^H]^H$, where, $\mathbf{B}_t^k$ is called a blocking matrix and $\cdot^H$ denotes the conjugate transpose.

3. BLIND EXTRACTION: CSV-AUXIVE ALGORITHM

The extractor part of the proposed procedure is the blind algorithm derived in [21], named CSV-AuxIVE, assuming constant separating vector (CSV) mixing model; here, we overview its most important ideas. The CSV model is based on an assumption that the separating vector $\mathbf{w}_t^k$ is constant within all T blocks ($\mathbf{w}_t^k = \mathbf{w}^k, t = 1 \ldots T$). This means that the separating vector obeys $(\mathbf{w}^k)^H \mathbf{x}_{\ell_t}^k = \hat{s}_{\ell_t}^k \approx s_{\ell_t}^k$ for each block t, where $\hat{s}_{\ell_t}^k$ is the SOI estimate. The mixing vector $\mathbf{a}_t^k$ and the blocking matrix $\mathbf{B}_t^k$ are allowed to depend on t.

The estimation of $\mathbf{w}^k$ stems from a formulation of a log-likelihood function describing the mixing, where the frequency components of SOI are modeled as dependent drawn from Laplace distribution and the unwanted background signals are assumed to be uncorrelated Gaussian variables. Following computations from [21], a contrast function is derived from the log-likelihood function, which is subsequently optimized using the *auxiliary function optimization* technique. The main idea is to replace the nonlinear contrast function with an auxiliary function, which is easier to optimize and retains the same optimal solution. The following update rules are obtained by optimizing the auxiliary function alternately in original ($\mathbf{w}^k$) and auxiliary variables (r_{ℓ_t}, $\mathbf{V}_t^k$):

$$r_{\ell_t} = \sqrt{\sum_{k=1}^{K} |(\mathbf{w}^k)^H \mathbf{x}_{\ell_t}^k|^2} \qquad \text{for all } \ell_t, \tag{4}$$

$$\mathbf{V}_t^k = \hat{\mathrm{E}}_t \left[\varphi(r_{\ell_t}) \mathbf{x}_{\ell_t}^k (\mathbf{x}_{\ell_t}^k)^H \right], \tag{5}$$

$$\widehat{\mathbf{C}}_t^k = \hat{\mathrm{E}}_t \left[\mathbf{x}_{\ell_t}^k (\mathbf{x}_{\ell_t}^k)^H \right], \tag{6}$$

$$\mathbf{a}_t^k = \frac{\widehat{\mathbf{C}}_t^k \mathbf{w}^k}{(\mathbf{w}^k)^H \widehat{\mathbf{C}}_t^k \mathbf{w}^k}, \tag{7}$$

$$\hat{\sigma}_{k,t} = \sqrt{(\mathbf{w}^k)^H \widehat{\mathbf{C}}_t^k \mathbf{w}^k}, \tag{8}$$

$$\mathbf{w}^k = \left(\sum_{t=1}^{T} \frac{\mathbf{V}_t^k}{(\hat{\sigma}_t^k)^2} \right)^{-1} \sum_{t=1}^{T} \frac{(\mathbf{w}^k)^H \mathbf{V}_t^k \mathbf{w}^k}{(\hat{\sigma}_t^k)^2} \mathbf{a}_t^k, \tag{9}$$

where $\varphi(r_{\ell_t}) = r_{\ell_t}^{-1}$ is a nonlinearity suitable for super-Gaussian signals such as speech, $\widehat{\mathbf{C}}_t^k$ is the sample-based covariance matrix of the mixture on the tth block, and $\hat{\mathrm{E}}_t$ denotes the sample-based average over the frames in block t. Equation (7) is the *orthogonal constraint* (OGC) ensuring mutual orthogonality of subspaces generated by the SOI and the other signals, and $\hat{\sigma}_t^k$ is the sample-based variance of the SOI.

4. PRIOR KNOWLEDGE BASED ON SUPERVISED SPEAKER IDENTIFICATION

Without any prior knowledge about the SOI, CSV-AuxIVE extracts an arbitrary source from the mixture. Which source is actually extracted usually depends on the initialization, which is due to lack of any relevant information often random or uniform. In this section, we describe the three ways for employing the SOI-specific prior knowledge to extract the target source.

4.1. Speaker identification using embedding features (X-vectors)

Our implementation of the network producing the embeddings stems from the FSMN architecture [16] and is summarized in Table 1. Its input consists of a single-channel audio signal sampled at 16 kHz. The input features are 40 filter bank coefficients computed from frames of length 400 samples with the frame-shift of 200 samples. Subsequently, six context layers follow. In each layer, the context of frames is weighted by a trainable matrix; mean time-pooling is performed; and a linear transformation is applied. The output of each layer is weighted by the exponential linear unit (ELU). The Pooling layer computes variances of frames. Its context length is $L_c = 101$ during training. The network is trained to classify N speakers via minimization of the cross-entropy loss function.

After the training, the two latest classification layers are removed and the embeddings are extracted from the Pooling layer. In the test phase, an embedding of the unknown speaker is compared to the set of embeddings (called *enrollment*) corresponding to the potential speakers. This comparison is performed by Probabilistic Linear Discriminant Analysis (PLDA, [22]). PLDA is a machine learning approach that tests a hypothesis that an enrollment vector $\boldsymbol{\xi}$ and the test vector $\boldsymbol{\chi}$ correspond to the same speaker. PLDA returns a score $M(\boldsymbol{\xi}, \boldsymbol{\chi})$, whose ”higher” values signify correct hypothesis. In this manuscript, we denote the extracted embeddings as *X-vectors*.

Table 1. Description of the FSMN producing the X-vectors. The input sizes for the context layers are stated after the mean pooling.

Layer	Layer context	Total context	Input x output
Context 1	$\ell \pm 80$	161	40×1024
Context 2	$\ell \pm 4$	169	1024×768
Context 3	$\ell \pm 4$	177	768×512
Context 4	$\ell \pm 4$	185	512×384
Context 5	$\ell \pm 4$	193	384×256
Context 6	$\ell \pm 4$	201	256×128
Fully-conn. 1	ℓ	201	128×128
Pooling	$\ell \pm \frac{L_c-1}{2}$	$201 + L_c$	$(L_c \cdot 128) \times 128$
Fully-conn. 2	ℓ	$201 + L_c$	128×128
Softmax	$-$	$201 + L_c$	$128 \times N$

The training data for the FSMN and PLDA originate from the development part of the Voxceleb database [23] and from the training part of the LibriSpeech corpus [24]. Since the Librispeech (part train-360-clean) is free of distortions, it was subjected to robustness introducing augmentations focused on environmental noise and reverberation. Details about the augmentation process can be found in [17]. The environmental noise was taken from the simulated part of the CHiME-4 training dataset [25] and from the development dataset available in Task 1 of the DCASE2018 challenge [26].

4.2. Non-intrusive assessment of the extraction quality

The X-vectors and PLDA are used to non-intrusively assess the extracted signal. This allows detection of cases when an unwanted source is extracted and the SOI needs to be re-estimated via deflation. The assessment represents the entire signal through a single PLDA score $M(\boldsymbol{\xi}_s, \boldsymbol{\chi})$. It measures the similarity between the X-vector computed from the enrollment utterance of the SOI $\boldsymbol{\xi}_s$ and the unknown test X-vector $\boldsymbol{\chi}$. When the SOI is active in a clean test signal, the score tends to be high. The value decreases with the presence of distortions. The *extraction assessment* compares two signals containing the same SOI component (e.g., the original and the extracted signal with X-vectors $\boldsymbol{\chi}_{\hat{s}}$ and $\boldsymbol{\chi}_{\dot{s}}$). If $M(\boldsymbol{\xi}_s, \boldsymbol{\chi}_{\hat{s}}) > M(\boldsymbol{\xi}_s, \boldsymbol{\chi}_{\dot{s}})$ then $\boldsymbol{\chi}_{\hat{s}}$ is assumed to be a better SOI estimate.

4.3. Guidance through the pilot signal

The pilot signal $\mathbf{g}$ is dependent on the SOI and is introduced into CSV-AuxIVE via modification of the update step (4) into

$$r_{\ell_t} = \sqrt{\sum_{k=1}^{K} |(\mathbf{w}^k)^H \mathbf{x}_{\ell_t}^k|^2 + g_{\ell_t}}. \tag{10}$$

The right-hand side of (4) corresponds to a factor that binds together components corresponding to the same source in IVE. Without this factor, the independence would be achieved in each frequency bin k separately and the reconstruction of the SOI would suffer the permutation problem. Equation (10) introduces the dependence of all components on $\mathbf{g}$ and consequently on the SOI.

The pilot signal $\mathbf{g}$ for the ℓth frame is computed as

$$g_\ell = \begin{cases} \sum_{k=1}^{K} |x_\ell^k(1)|^2 & \text{the SOI is dominant,} \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

where $x_\ell^k(1)$ is the mixture on the first microphone. The *realizable pilot* $\mathbf{g}^{\text{XVEC}}$ is computed using the frame-wise X-vectors. To obtain them, the context of the FSMN network is shifted by a single frame at the time and the context of the Pooling layers is reduced to $L_c = 10$ frames. Subsequently, the frame-wise PLDA scores $M_\ell(\boldsymbol{\xi}, \boldsymbol{\chi}_{x_\ell})$ are computed, where $\boldsymbol{\xi}$ is the enrollment X-vector corresponding to one of the potential speakers and $\boldsymbol{\chi}_{x_\ell}$ is the X-vector computed using the context at the ℓth frame of the first mixture channel. The SOI is considered *dominant* when its score $M_\ell(\boldsymbol{\xi}_s, \boldsymbol{\chi}_{x_\ell})$ is the highest compared to scores of the other enrollment speakers.

To compare, we consider an *oracle pilot* $\mathbf{g}^{\text{ORAC}}$ to explore the potential of an ideal piloting. $\mathbf{g}^{\text{ORAC}}$ is computed using (11) where the SOI is considered dominant if

$$\sum_{k=1}^{K} |s_\ell^k|^2 > \mu_{\text{ORAC}} \sum_{k=1}^{K} ||\mathbf{z}_\ell^k||^2 \text{ and } \sum_{k=1}^{K} |s_\ell^k|^2 > \mu_{\text{ENE}}(s), \tag{12}$$

where μ_{ORAC} is a free parameter reflecting the desired level of dominance, and $\mu_{\text{ENE}}(s)$ is the minimum energy for which the SOI is considered as active. More detailed description of the pilots can be found in [17].

4.4. Guidance through the speaker-specific initialization

The frame-wise PLDA score $M_\ell(\boldsymbol{\xi}_s, \boldsymbol{\chi}_{x_\ell})$ corresponding to the SOI is higher compared to scores of other enrollment speakers when the SOI is assumed to be the dominant speaker in the mixture. Using these frames, a multi-channel Wiener filter (MCWF) can be estimated, which roughly points towards the SOI and initializes the CSV-AuxIVE closer to the desired solution.

Let $\mathcal{S} = \{\ell_1, \dots, \ell_{|\mathcal{S}|}\} \subset 1, \dots, L$ denote the set of indices where the SOI is assumed dominant by the X-vectors, and $|\mathcal{S}|$ is the cardinality of this set. Let us define an $L \times 1$ masking vector

$$h_\ell = \begin{cases} 1, & \ell \in \mathcal{S} \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Then, the covariance matrix of the SOI is estimated by

$$(\widehat{\mathbf{C}}_s)^k = \frac{1}{|\mathcal{S}|} \sum_{\ell} \left[h_\ell \mathbf{x}_{\ell_t}^k (\mathbf{x}_{\ell_t}^k)^H \right], \tag{14}$$

and the SOI-specific MCWF is given by

$$\mathbf{w}^k = (\widehat{\mathbf{C}}^k)^{-1} (\widehat{\mathbf{C}}_s)^k \mathbf{u}, \tag{15}$$

where $\widehat{\mathbf{C}}^k = \frac{1}{L} \sum_\ell \left[\mathbf{x}_{\ell_t}^k (\mathbf{x}_{\ell_t}^k)^H\right]$, and $\mathbf{u}$ is a one-hot vector of size $D \times 1$ denoting the reference microphone. In our implementation, we estimate the SOI as it is recorded on the first microphone.

4.5. Guidance through the supervised deflation of the mixture

The deflation provides a correction mechanism to extract the SOI from mixtures where the guidance through the pilot and the initialization fails. The deflation starts with the extraction of the first signal via CSV-AuxIVE. The assessment of extraction quality is used to determine whether this signal represents a better estimate of the SOI than the original mixture. If so, the first signal is returned and the extraction ends. Otherwise, the first signal is subtracted from the mixture using least squares. Using the assessment, the reduced mixture is compared to the original one. If the original mixture is chosen, the extraction ends (the deflation did not bring the mixture closer to the SOI). If the reduced mixture is selected, the piloted CSV-AuxIVE is applied to it and the second signal is extracted. This process is repeated until an estimate of the SOI is found or until the number of deflation steps exceeds selected maximum I. More details concerning deflation can be found in [17].

5. EXPERIMENTAL VERIFICATION

The experimental evaluation of the proposed guided extractor is performed using reverberated mixtures of two or three speakers originating in the multi-channel Wall Street Journal datasets (MC-WSJ0-2mix, MC-WSJ0-3mix, [2,20]). These are spatialized versions of the single-channel datasets described in [27]. Each of the datasets contains $3,000$ simulated mixtures recorded in a reverberant environment using a microphone array containing eight microphones. Each mixture contains two/three active speakers, i.e., there is $6,000/9000$ extraction experiments in total. The sources are mixed at SIR between $\langle -5, +5\rangle$ dB. The recordings are highly reverberant ($T_{60} \in \langle 200, 600\rangle$ ms), and captured in rooms with variable dimensions. The topology of the microphone array is varying, as well as the source-microphone distance, which is 1.3 m with 0.4 m standard deviation. The sampling frequency is 8 kHz.

Table 2 and Table 3 contain results achieved by CSV-AuxIVE endowed with various combinations of the prior knowledge. In addition, the achieved results are compared to 1) the oracle multi-channel Wiener filter implemented according to formulas in [2] which utilizes an oracle covariance matrix of the SOI and represents an upper boundary on for the extraction based on the spatial filtering. 2) The X-vector-based MCWF computed via (15), which is used as a SOI-specific initialization for CSV-AuxIVE. 3) The state-of-the-art ILRMA algorithm from [9], which performs a full blind separation (BSS, requires the true number of sources) and is followed by an oracle matching of the estimated and true sources for the evaluation. CSV-AuxIVE for extracting the SOI is applied with blocks of length $L_T = 160$ frames. The experiments were computed in Matlab 2021a on a desktop computer with Intel i7-6800K processor at 3.4GHz and 64GB of RAM. The experiments were evaluated using BSS_EVAL toolbox [28]. The presented measures are signal-to-interference-ratio (SIR), which quantifies suppression of the unwanted sources and signal-to-distortion-ratio (SDR), which measures both the suppression and the distortion of the SOI.

Focusing on MC-WSJ0-2mix in Table 2, the oracle MCWF using four microphones achieves the highest SDR of 13.4 dB. Using a strong oracle information in the form of $\mathbf{g}^{\mathrm{ORAC}}$, CSV-AuxIVE achieves comparable SIR but SDR is lower by 3.8 dB. The X-vector-based MCWF produces an approximate estimate of the SOI achieving SDR of 4.8 dB. Using the MCWF as an initialization, the CSV-AuxIVE improves the SDR only slightly to 5.6 dB. Detailed analysis of the results indicates that CSV-AuxIVE + initialization or CSV-AuxIVE + $\mathbf{g}^{\mathrm{XVEC}}$ often converge to an unwanted source when these types of prior knowledge are exploited individually. This is caused by a limited accuracy of the frame-wise speaker identification provided by the X-vectors. The situation is significantly remedied using $\mathbf{g}^{\mathrm{XVEC}}$ with the deflation, which yields SDR of 7.8 dB. The disadvantage of the deflation is its high computation complexity; it requires analysis of several signals (the mixture, the extracted signal, the deflated mixture) via the FSMN embedding network. However, the same SDR of 7.8 dB can be achieved using piloting together with initialization, which reduces the time of computation to 149 minutes from 350 using pilot with the deflation. The best SDR of CSV-AuxIVE (8 dB) is achieved using all three techniques introducing the prior knowledge. Surprisingly, this variant also slightly decreases the computational time compared to the piloting + deflation. Here, the utilization of a fine initialization prevents the need for deflation for some of the mixtures. The fully guided extractor achieves comparable performance to full separation provided by ILRMA in the terms of SDR and outperforms it when observing SIR. This indicates that CSV-AuxIVE is able to suppress the unwanted source more but introduces additional distortions into the estimated SOI. The ILRMA is less computationally demanding compared to fully guided CSV-AuxIVE. Note, however, that the matching of sources to the speakers is not included in the time required by ILRMA, since it is performed in an oracle way as a part of the evaluation.

Table 2. MC-WSJ0-2mix: SIR/SDR [dB] averaged over all 6000 extractions using 4 microphones. The column "Time" reports time needed to extract all speakers (dataset duration 296 minutes).

Approach	**Time [min.]**	**Sepa-ration**	**Spk. id.**	**SIR [dB]**	**SDR [dB]**
Mixture	-	-	-	0.2	0.2
MCWF - Oracle	-	Orac.	Orac.	17.3	13.4
CSV + $\mathbf{g}^{\mathrm{ORAC}}$	-	BSE	Orac.	17.3	9.6
ILRMA	191	BSS	Orac.	11.3	7.6
MCWF - X-vector	36	ML	ML	8.0	4.8
CSV + Init.	119	BSE	ML	11.1	5.6
CSV + $\mathbf{g}^{\mathrm{XVEC}}$	113	BSE	ML	11.7	6.0
CSV + $\mathbf{g}^{\mathrm{XVEC}}$ + Init.	149	BSE	ML	14.3	7.8
CSV + $\mathbf{g}^{\mathrm{XVEC}}$ + Defl.	350	BSE	ML	13.4	7.8
CSV + all three	338	BSE	ML	13.8	8.0

Table 3. MC-WSJ0-3mix: SIR/SDR [dB] averaged over all 9000 extractions using 8 microphones. The column "Time" reports time needed to extract all speakers (dataset duration 261 minutes).

Approach	**Time [min.]**	**Sepa-ration**	**Spk. id.**	**SIR [dB]**	**SDR [dB]**
Mixture	-	-	-	-3.0	-3.0
MCWF - Oracle	-	Orac.	Orac.	15.2	10.8
CSV + $\mathbf{g}^{\mathrm{ORAC}}$	-	BSE	Orac.	11.6	6.1
ILRMA [9]	309	BSS	Orac.	6.8	3.8
MCWF - X-vector	58	ML	ML	2.6	0.4
CSV + Init.	219	BSE	ML	4.3	0.5
CSV + $\mathbf{g}^{\mathrm{XVEC}}$	202	BSE	ML	4.1	0.5
CSV + $\mathbf{g}^{\mathrm{XVEC}}$ + Init.	261	BSE	ML	6.7	2.4
CSV + $\mathbf{g}^{\mathrm{XVEC}}$ + Defl.	597	BSE	ML	6.3	2.7
CSV + all three	586	BSE	ML	7.6	3.7

Table 3 confirms that all the phenomenons observed on mixtures of two speakers remain valid for three speakers as well. The performance of all algorithms slightly deteriorates on this more difficult dataset. Each mixture contains two unwanted sources with comparable energy, the input SIR is thus -3 dB. Also here, the combination of the three approaches brings the highest performance of CSV-AuxIVE.

6. CONCLUSION

We have presented the efficient combination of three approaches utilizing the speaker-identification-based prior knowledge to guide the blind extraction algorithm CSV-AuxIVE towards the desired speaker. The experiments have shown that the guided extraction with CSV-AuxIVE achieves higher or, at least, comparable performance to ILRMA, where full separation is performed, and the SOI is selected among the separated sources in an oracle way.

7. REFERENCES

[1] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[2] Zhong-Qiu Wang, Jonathan Le Roux, and John R Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *ICASSP 2018*. IEEE, 2018, pp. 1–5.

[3] Masahito Togami, "End to end learning for convolutive multichannel wiener filtering," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 8032–8036.

[4] Christoph Boeddeker et al., "Convolutive transfer function invariant sdr training criteria for multi-channel reverberant speech separation," in *ICASSP 2021*. IEEE, 2021, pp. 8428–8432.

[5] Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot, *Audio Source Separation and Speech Enhancement*, Wiley Publishing, 2018.

[6] Taesu Kim, Hagai T Attias, Soo-Young Lee, and Te-Won Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 1, pp. 70–79, 2006.

[7] Robin Scheibler and Masahito Togami, "Surrogate source model learning for determined source separation," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 176–180.

[8] Kouhei Sekiguchi et al., "Autoregressive fast multichannel nonnegative matrix factorization for joint blind source separation and dereverberation," in *ICASSP 2021*. IEEE, 2021, pp. 511–515.

[9] Daichi Kitamura and Kohei Yatabe, "Consistent independent low-rank matrix analysis for determined blind source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2020, no. 1, pp. 1–35, 2020.

[10] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[11] 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," in *IEEE transactions on speech and audio processing*, Apr. 2003, pp. 505–510.

[12] Marc Delcroix, Tsubasa Ochiai, Katerina Zmolikova, Keisuke Kinoshita, Naohiro Tawara, Tomohiro Nakatani, and Shoko Araki, "Improving speaker discrimination of target speech extraction with time-domain speakerbeam," in *ICASSP 2020*. IEEE, 2020, pp. 691–695.

[13] Jiangyu Han, Xinyuan Zhou, Yanhua Long, and Yijie Li, "Multi-channel target speech extraction with channel decorrelation and target speaker adaptation," in *ICASSP 2021*. IEEE, 2021, pp. 6094–6098.

[14] Jakub Jansky, Jiri Malek, Jaroslav Cmejla, Tomas Kounovsky, Zbynek Koldovsky, and Jindrich Zdansky, "Adaptive blind audio source extraction supervised by dominant speaker identification using X-vectors," in *ICASSP 2020*. IEEE, 2020, pp. 676–680.

[15] Robin Scheibler and Nobutaka Ono, "Independent vector analysis with more microphones than sources," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 185–189.

[16] Shiliang Zhang, Cong Liu, Hui Jiang, Si Wei, Lirong Dai, and Yu Hu, "Feedforward sequential memory networks: A new structure to learn long-term dependency," *arXiv preprint arXiv:1512.08301*, 2015.

[17] Jiri Malek, Jakub Jansky, Zbynek Koldovsky, Tomas Kounovsky, Jaroslav Cmejla, and Jindrich Zdansky, "Target speech extraction: Independent vector extraction guided by supervised speaker identification," *arXiv preprint arXiv:2111.03482*, 2021.

[18] Yanfeng Liang, Syed Mohsen Naqvi, and Jonathon A Chambers, "Audio video based fast fixed-point independent vector analysis for multisource separation in a room environment," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, pp. 183, 2012.

[19] Francesco Nesta, Saeed Mosayyebpour, Zbynek Koldovsky, and Karel Palecek, "Audio/video supervised independent vector analysis through multimodal pilot dependent components," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 1150–1164.

[20] "Scripts to generate the wsj0-mix multi-speaker dataset [online]," `https://www.merl.com/demos/deep-clustering`, Accessed: 01.03.2022.

[21] Jakub Janský, Zbyněk Koldovský, Jiří Málek, Tomáš Kounovský, and Jaroslav Čmejla, "Auxiliary function-based algorithm for blind extraction of a moving speaker," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, pp. 1–16, 2022.

[22] Sergey Ioffe, "Probabilistic linear discriminant analysis," in *European Conference on Computer Vision*. Springer, 2006, pp. 531–542.

[23] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[24] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *ICASSP 2015*. IEEE, 2015, pp. 5206–5210.

[25] Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer, "The 4th CHiME speech separation and recognition challenge [online]," Accessed: 29.9.2021.

[26] "DCASE 2018 challenge [online]," `http://dcase.community/challenge2018/index`, Accessed: 29.9.2021.

[27] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *ICASSP 2016*. IEEE, 2016, pp. 31–35.

[28] 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, July 2006.