Target Speech Extraction: Independent Vector Extraction Guided by Supervised Speaker Identification

Jiri Malek, *Member, IEEE*, Jakub Jansky, Zbynek Koldovsky, *Senior Member, IEEE*, Tomas Kounovsky, Jaroslav Cmejla, and Jindrich Zdansky

***Abstract*—This manuscript proposes a novel robust procedure for the extraction of a speaker of interest (SOI) from a mixture of audio sources. The estimation of the SOI is performed via independent vector extraction (IVE). Since the blind IVE cannot distinguish the target source by itself, it is guided towards the SOI via frame-wise speaker identification based on deep learning. Still, an incorrect speaker can be extracted due to guidance failings, especially when processing challenging data. To identify such cases, we propose a criterion for non-intrusively assessing the estimated speaker. It utilizes the same model as the speaker identification, so no additional training is required. When incorrect extraction is detected, we propose a "deflation" step in which the incorrect source is subtracted from the mixture and, subsequently, another attempt to extract the SOI is performed. The process is repeated until successful extraction is achieved. The proposed procedure is experimentally tested on artificial and real-world datasets containing challenging phenomena: source movements, reverberation, transient noise, or microphone failures. The method is compared with state-of-the-art blind algorithms as well as with current fully supervised deep learning-based methods.**

***Index Terms*—Blind extraction, supervised speaker identification, target speech extraction.**

I. Introduction

A FREQUENT goal of speech processing is to recover a speaker of interest (SOI) from a mixture of speech sources and environmental noise. This task is often solved using *speech separation* methods; all sources present in the mixture are estimated and subsequently the SOI is identified among them. Separation can be performed either via *data-driven techniques* deriving their models from large sets of training signals [1]–[5] or via *model-based techniques* utilizing general statistical assumptions about the sources and the mixing [6]–[12].

Both approaches have distinct advantages. The data-driven techniques [1] employ principles of *machine/deep learning* to estimate the separating models. If provided with relevant scenario-specific training data, they achieve high separation quality. The machine learning-based approaches primarily perform spectral filtering on single-channel data, which can be supplemented by additional spatial filtering if additional channels are available. In contrast, the model-based approaches [6] employ only general statistical models of the sources/mixing. These approaches do not need any training data and require only minimum information about the target scenario, i.e., they perform the *blind separation*. The blind approaches are, in theory, applicable to a wide range of tasks without any need for adaptation. Arguably, this freedom is achieved at the cost of lower separation accuracy since the employed statistical models only approximate the real conditions. Many blind techniques aim to estimate spatial filters and thus require multi-channel mixtures.

Focusing on blind methods, separation of speech usually proceeds in the time-frequency domain. Independent component analysis (ICA, [13]) separates the sources based on their statistical independence. For a wide-band signal, ICA is separately applied to each frequency bin, which leads to the so called *permutation ambiguity* [14]. The recovered frequency components have a random order and all components corresponding to the wide-band source need to be identified in order to reconstruct it in the time-domain. To alleviate this drawback, the independent vector analysis (IVA, [7], [8]) has been proposed. It binds together the frequency components corresponding to a single source using higher-order dependencies among them. Non-negative matrix factorization (NMF, [15]) attempts to factorize spectrogram of a single-channel mixture as a product of two non-negative components, recurring patterns and their activations. Multichannel NMF (MNMF, [9]) extends this concept for analysis of multi-channel mixtures. Independent low-rank matrix analysis (ILRMA, [10]–[12]) unifies the principles of IVA and NMF. Spectral masking methods [16], [17] use the assumption that only one source is dominant at each time-frequency point.

The full separation attempts to estimate all sources in the mixture, which usually requires the knowledge of the number of sources. This arguably limits the practicality/flexibility of the separation (see the discussion in [18]–[20]). To alleviate,

Manuscript received 5 November 2021; revised 13 May 2022 and 28 June 2022; accepted 30 June 2022. Date of publication 14 July 2022; date of current version 25 July 2022. This work was supported in part by the Technology Agency of the Czech Republic under Grant TO01000027, in part by the Czech Science Foundation under Grant 20-17720S, and in part by the Student Grant Scheme 2021 Project of the Technical University in Liberec. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Timo Gerkmann. *(Corresponding author: Jiri Malek.)*

The authors are with the Faculty of Mechatronics, Informatics and Interdisciplinary Studies, Technical University of Liberec, 461 17 Liberec, Czech Republic (e-mail: jiri.malek@tul.cz; jakub.jansky@tul.cz; zbynek.koldovsky@tul.cz; tomas.kounovsky@tul.cz; jaroslav.cmejla@tul.cz; jindrich.zdansky@tul.cz).

Digital Object Identifier 10.1109/TASLP.2022.3190739

the recovery can be focused exclusively on the SOI, which is referred to as *target speech extraction*. The extraction can thus be interpreted as simultaneous identification and estimation of the SOI. Again, this task can be solved using machine learning-based approach [18], [19], [21]–[24] or blind source extraction (BSE, [25], [26]).

The independent vector extraction (IVE) is a sub-problem of IVA focusing on the SOI [25]–[28]. By definition, IVE methods extract an arbitrary source depending on (often random) initialization. To extract the SOI, information identifying this source is required. Such information can be provided via an initialization focused on the SOI. The utilization of a video stream was proposed to this end in [29]. Such initialization, however, does not guarantee that the method will remain focused on the SOI during the updates. Alternatively, the extraction can be limited to a direction containing the SOI via the geometric constraint [30], [31], which requires information about SOI location. A practical alternative is to introduce a *pilot signal* that is related to the SOI and directs the convergence towards it. For example, piloting using voice activity detection was proposed for mixtures containing a single speaker in [32]. For mixtures of multiple speakers, pilots using supervised speaker identification via embeddings [33], [34] have recently proven to be very effective in [35]. The embeddings were used to tackle the problem concerning the ambiguity of the SOI in the deep learning-based extractors as well [18], [21].

Speaker embeddings are Deep-Neural-Network-based (DNN) features encoding the characteristics of a speaker. Several variants have recently been introduced, differing mainly in the architecture of the extracting DNN. Embeddings derived from fully-connected feed-forward DNN were proposed in [36]. Approaches utilizing the context of the data via recursive long short-term memory (LSTM) networks were presented in [37]. The recursive modeling allows for more precise classification, however, the training is data-demanding and time-consuming.

To alleviate those demands, non-recursive architectures capturing the context have been proposed, such as time-delayed neural networks [38] (TDNN) or feed-forward sequential memory networks [39] (FSMN). The "context layers" within these networks process a set of frames (or feature vectors produced by their previous layer) centered around the current frame. The processing of the context significantly increases the number of learnable parameters. To reduce this number, TDNN subsamples the set of frames, because the neighboring frames are assumed to be correlated. In contrast, FSMN weights all frames at an input of a layer by a trainable matrix and performs mean time-pooling. Thus FSMN can be seen as a generalization of TDNN in which the importance of frames is learned during training rather than selected during design.

The traditional model of IVE is time invariant, i.e., it is suitable for separation of immobile sources (*static approach*). To extract moving sources, the time invariant methods are consecutively applied to short intervals of data where the sources are approximately static, and their parameters are recursively updated. The drawback of this *block-wise static approach* [35] lies in difficult tuning of the block length and the recursion weight. Recently, an alternative approach based on the constant separating vector (CSV) model [40] has been proposed. It allows for changes of mixing parameters within the processed interval of data. Compared to the block-wise static approach, the CSV-based method exploits longer intervals of signals. Consequently, this allows us to achieve a higher extraction accuracy. This higher accuracy was proven theoretically in [41] and demonstrated experimentally in [42].

In this manuscript, we propose an improvement of a guided blind IVE-based method involving the above described advancements introduced in our previous works [35], [40], [42]. It consists of a combination of model-based extraction and data-driven frame-wise identification of the SOI. The extraction is based on an IVE algorithm endowed with the CSV mixing model. The IVE algorithm is guided towards the SOI using piloting exploiting speaker embeddings computed via an FSMN model. This combination simplifies/reduces the amount of training compared to fully data-driven techniques. The proposed method needs only to learn how to identify the SOI while its extraction is free of training, which makes it applicable to a wide variety of realistic mixtures. Thanks to this decoupling of the identification and the extraction, the identification can be trained generally, independent of a specific mixing scenario.

The contribution of this manuscript to the basic concept is threefold. 1) An improvement for piloting is introduced by incorporating a non-intrusive criterion for the assessment of the extraction performance. The assessment allows for the detection of the cases in which an incorrect source is being extracted. 2) These incorrect cases are treated using a deflation approach: the unwanted source is subtracted from the mixture, and the extraction is attempted again. This cycle continues until the SOI is extracted. The pilot signal and the criterion share the same pretrained FSMN model, i.e., no additional training is required. 3) Compared to our previous works, we perform a more detailed experimental analysis of the properties and limitations of piloting. The proposed extractor is verified using two widely analyzed datasets (CHiME-4 [43], the spatialized version of wsj0-2mix [2]) and an ad-hoc dataset featuring source movements. The benefits of the deflation step are demonstrated, and the results are compared to the state-of-the-art deep-learning-based and blind methods.

This article is organized as follows. The blind extraction algorithm is described in Section II-B. Section II-C2 provides the principles of piloting. The non-intrusive criterion for assessment of extraction quality is proposed in Section II-D. The deflation is presented in Section II-E. The proposed method is experimentally evaluated in Section III, while Section IV concludes the manuscript.

II. Algorithm Description

A. Problem Definition

A time varying mixture of d 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 $k = 1, \dots, K$ is the frequency and $\ell = 1, \dots, L$ is the frame index. $\mathbf{x}_\ell^k \in \mathbb{C}^d$ denotes a vector of the mixed signals recorded on d microphones, $\mathbf{y}_\ell^k \in \mathbb{C}^d$ is a vector whose ith component corresponds to the ith original signal and $\mathbf{A}_\ell^k \in \mathbb{C}^{d\times d}$ is the mixing matrix. For practical reasons, it is often assumed that the mixing is approximately static over a small number of subsequent frames. Let this interval be referred to as *block* in this manuscript. The mixture is thus divided into $t = 1, \dots, T$ equally long blocks of length 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 [35] is given by

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

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

In IVA, a complete de-mixing matrix $\mathbf{W}_t^k \in \mathbb{C}^{d\times d}$ 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. In contrast, IVE seeks only one row of $\mathbf{W}_t^k$, denoted by $\mathbf{w}_t^k$, such that it specifically extracts 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 $\mathbf{A}_t^k = [\mathbf{a}_t^k \ \mathbf{Q}_t^k]$. Then, the (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 conjugate transpose.

B. Blind Extraction: CSV-AuxIVE Algorithm

The extraction part of the proposed procedure is a blind algorithm from [40]. Here, we overview its most important ideas and provide some intuition on how the final update rules were obtained. The CSV model is based on the assumption that the separating vector $\mathbf{w}_t^k$ is constant within all T blocks ($\mathbf{w}_t^k = \mathbf{w}^k, t = 1 \dots T$). 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 still assumed to vary with respect to t.

The estimation of $\mathbf{w}^k$ stems from the following *log-likelihood function*. Let $\mathbf{s}_{\ell_t} = [s_{\ell_t}^1 \dots s_{\ell_t}^K]$ be a vector of all frequency components corresponding to the SOI. The elements of $\mathbf{s}_{\ell_t}$ are assumed to be dependent; they thus need to be modeled by a joint pdf $p_s(\mathbf{s}_{\ell_t})$. The background signals $\mathbf{z}_{\ell_t}^1 \dots \mathbf{z}_{\ell_t}^K$ are Gaussian and their frequency components are assumed to be uncorrelated. Consequently, their higher order dependencies are zero and they can be modeled as independent; let their density be denoted $p_\mathbf{z}(\mathbf{z}_{\ell_t}^k)$. The log-likelihood function is then

$$\mathcal{L}(\{\mathbf{w}^k\}_{k\le K}, \{\mathbf{a}_t^k\}_{k\le K} | \{\mathbf{x}_{\ell_t}^k\}_{k\le K}) = \log p_s(\{\hat{s}_{\ell_t}^k\}_{k\le K}) + \sum_{k=1}^K \log p_\mathbf{z}(\hat{\mathbf{z}}_{\ell_t}^k) + \log|\det \mathbf{W}_t^k|^2, \tag{4}$$

where $\hat{\mathbf{z}}_{\ell_t}^k$ is the estimate of the background signals. The notation $\{\cdot\}_{k\le K}$ describes a variable with all values of index k, e.g., $\{\mathbf{w}^k\}_{k\le K} = \mathbf{w}^1, \dots, \mathbf{w}^K$.

Subsequently, *a contrast function* is formulated using the assumption that all samples are independently distributed and the log-likelihood function (4) thus can be averaged over all blocks and samples. Optimization of the contrast function is performed using the *auxiliary function optimization* technique [44]. 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. Then the new auxiliary function is alternately optimized in the original and the auxiliary variables. Moreover, since the true model of the $p_s(\mathbf{s}_{\ell_t})$ is unknown, a surrogate density function suitable for speech signals is chosen in the form $f(x) \propto \exp\{-\|x\|\}$.

The *update rules* for finding the optimum point of the auxiliary contrast function are obtained in the form:

$$r_{\ell_t} = \sqrt{\sum_{k=1}^K |(\mathbf{w}^k)^H \mathbf{x}_{\ell_t}^k|^2} \qquad \text{for all } \ell_t, \tag{5}$$

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

$$\widehat{\mathbf{C}}_t^k = \hat{\mathrm{E}}_t\left[\mathbf{x}_{\ell_t}^k(\mathbf{x}_{\ell_t}^k)^H\right], \tag{7}$$

$$\mathbf{a}_t^k = \frac{\widehat{\mathbf{C}}_t^k \mathbf{w}^k}{(\mathbf{w}^k)^H \widehat{\mathbf{C}}_t^k \mathbf{w}^k}, \tag{8}$$

$$\hat{\sigma}_{k,t} = \sqrt{(\mathbf{w}^k)^H \widehat{\mathbf{C}}_t^k \mathbf{w}^k}, \tag{9}$$

$$\mathbf{w}^k \leftarrow \left(\sum_{t=1}^T \frac{\mathbf{V}_t^k}{(\hat{\sigma}_t^k)^2}\right)^{-1} \sum_{t=1}^T \frac{(\mathbf{w}^k)^H \mathbf{V}_t^k \mathbf{w}^k}{(\hat{\sigma}_t^k)^2} \mathbf{a}_t^k, \tag{10}$$

where r_{ℓ_t}, $\mathbf{V}_t^k$ are the auxiliary variables, $\varphi(r_{\ell_t}) = r_{\ell_t}^{-1}$ 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 expectation over the frames in block t. Equation (8) 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. A normalization $\mathbf{w}^k \leftarrow \mathbf{w}^k / \sqrt{\sum_{t=1}^T (\mathbf{w}^k)^H \mathbf{V}_t^k \mathbf{w}^k}$ is performed after each iteration (i.e., sequence of update rules (5)–(10)) to enable stable convergence.

When $T = 1$, CSV-AuxIVE corresponds to the auxiliary function-based IVE for static sources from [27], which is denoted as FS-IVE in the experiments within Section III. Successive application of the static IVE to blocks $t = 1 \dots T$ gives the block-wise static IVE approach from [42] (BS-IVE) that allows for dynamic mixing. To cope with the lack of data in short blocks, BS-IVE performs the following two steps. First, the extraction on the block t is initialized by the de-mixing vectors achieved on the block $t-1$, and, second, the statistics required in the update rules (5)–(10) are computed in a recursive manner.

TABLE I
DESCRIPTION OF THE FSMN PRODUCING THE X-VECTORS

Layer	**Layer**	**Total**	**Input**
	context	**context**	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 input sizes for the context layers are stated after the mean pooling operation.

C. Extraction Guided Towards the SOI: Piloting Using the Supervised Speaker Identification

CSV-AuxIVE extracts an arbitrary source from the mixture if no prior information concerning the SOI is available. This section discusses how the supervised speaker identification via embeddings is used to focus the extraction on the SOI. First, our implementation of the FSMN network for computation of the conventional *sentence-wise embeddings* is described. Subsequently, a general concept of a pilot signal is introduced. The pilot signal is statistically dependent on the SOI. It is submitted to the CSV-AuxIVE with the mixture and forces the blind algorithm to converge towards the SOI. Finally, modifications to the FSMN network are proposed, which allow computation of *frame-wise embeddings* and the design of a practically usable pilot.

1) Network Producing the Embeddings, X-Vectors: Our implementation of the embedding network stems from the FSMN[1] architecture [39] and is summarized in Table I. 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 a length of 400 and a frame-shift of 200 samples. Subsequently, six Context layers are present, i.e., 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. Overall, the size of the model is 1.8 million parameters. The network is trained to classify N speakers via minimization of the cross-entropy loss function.

After training, the two latest classification layers are removed and the embeddings are extracted from the Pooling layer. This is done to allow for classification of the speakers absent in the training set. In the test phase, an embedding of an 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, [45]). PLDA is a machine learning approach that tests a hypothesis that a an enrollment vector $\boldsymbol{\xi}$ and test vector $\boldsymbol{\chi}$ corresponds to a single speaker. The statistical distributions necessary for this testing are derived from a training dataset of precomputed embeddings. PLDA returns a score $M(\boldsymbol{\xi}, \boldsymbol{\chi})$, which is high if the hypothesis is correct.

The training data for the FSMN and PLDA originate from the development part of the Voxceleb1 database [46] and the training part of the LibriSpeech corpus [47]. The recording of Voxceleb1 (149 k utterances, about 340 hours) contain real-world reverberation and noise. Librispeech (part train-360-clean, 104 k utterances, 360 hours) is free of distortions. It was subjected to augmentations discussed below, in order to train X-vectors robust with respect to environmental distortions. The environmental noise was taken from the simulated part of the CHiME-4 training dataset [43] and the development dataset available in Task 1 of the DCASE2018 challenge [48].

The augmented X-vectors were trained on one unchanged instance of Voxceleb1/Librispeech and three augmented instances of the Librispeech dataset, where the following augmentations were applied:

1) Reverberation: The utterances were convolved with artificial room impulse responses (RIRs) generated by [49]. The artificial RIRs originated from a shoe-box room of size $8 \times 7 \times 3$ m; four different reverberation times T_{60}, ranging from $175 - 650$ ms, were considered. The source-microphone distance was $1 - 2$ m.
2) Noise: The environmental noise was summed with the original Librispeech utterances at signal-to-noise-ratio (SNR) equal to 10 dB.
3) Reverberation+noise: The noise was added to the reverberated Librispeech dataset with SNR$=$ 10 dB.

The PLDA was trained using the three augmented variants of the Librispeech dataset.

In this manuscript, we denote the extracted embeddings as X-vectors. In a narrow sense, this term is reserved for features estimated by the TDNN [33]. However, since both topologies are closely related, we believe such naming can be used without ambiguity.

2) The Concept of Piloting: The pilot signal represents an information identifying the SOI for the CSV-AuxIVE. It forces the blind algorithm to converge towards the SOI. The pilot signal is introduced through modification of the update step in (5). This equation corresponds to a factor that binds together all frequency components belonging to a single source. Without this factor, the independence of the outputs would be achieved in each frequency bin k separately, and the reconstruction of the wide-band SOI would suffer the permutation problem described in the Introduction. Modification of the equation (5) into the form

$$r_{\ell_t} = \sqrt{\sum\nolimits_{k=1}^{K} |(\mathbf{w}^k)^H \mathbf{x}_{\ell_t}^k|^2 + g_{\ell_t}}. \tag{11}$$

adds the dependency of all the frequency components on the pilot signal $\mathbf{g}$ and consequently also the SOI. The pilot $\mathbf{g}$ is independent of the mixing model parameters and thus does not change the remaining update rules of the CSV-AuxIVE.

The signal $\mathbf{g}$ needs to be designed as statistically dependent on the SOI. The term under the square root of (5) describes the

[1] The utilized network architecture does not differ from our previous works in [35], [42]. There we described the embedding network as TDNN with modifications. A more detailed research of literature revealed that it is more accurate to label the network as FSMN.

total energy of the extracted components. Thus the frame-wise energy of the SOI appears to be a suitable choice. Since the actual energy is unknown and difficult to estimate, a reasonable approximation is given by the frames of the mixture, where the energy of the unwanted sources is low (the energy of the SOI is *dominant*). We propose to compute the pilot signal $\mathbf{g}$ for the ℓth frame (note that $\mathbf{g}$ is independent of CSV blocks) as

$$g_\ell = \begin{cases} \sum_{k=1}^{K} |x_\ell^k(1)|^2 & \text{the SOI is dominant,} \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

where $x_\ell^k(1)$ is the mixture on the first microphone. Specific pilot signals (and their respective ways how to determine the dominance of the SOI) are introduced in Section II-C4.

3) Frame-Wise Speaker Identification for Piloting: The utilization of X-vectors and PLDA for piloting differs from the conventional speaker identification in several aspects.

a) Conventionally, speaker identification operates on long intervals/sentence-wise. However, the pilot signal in (12) requires frame-wise information about the dominance of the SOI, i.e., *a frame-wise sequence of X-vectors*. Each X-vector then describes the identity of a speaker in a short interval centered around the current frame. To obtain such a sequence, the input context of the FSMN is gradually shifted by a single frame. For each shift, an X-vector is computed based on pooling with a shortened context (e.g., $L_c = 11$).

b) The identification is performed in the presence of cross-talk. However, only the identity of *the dominant speaker* is required (due to the definition of the pilot in (12)).

c) Only the *identity of the SOI is of interest*; it must not be confused with any interfering speaker. The substitution within the set of interferers is irrelevant because the pilot in (12) is set to zero when any unwanted source is assumed dominant.

d) The set of the potential speakers (enrollment set) is significantly smaller. Conventionally, hundreds of speakers need to be distinguished. For the purposes of piloting, the enrollment set contains X-vector for each speaker, which can be active in the processed dataset (e.g., 18 vectors for the wsj0-2mix dataset [2]).

The aspects a) and b) complicate the identification task, whereas the aspects c) and d) simplify it.

To perform the frame-wise identification, PLDA scores $M_\ell(\boldsymbol{\xi}, \boldsymbol{\chi}_{x_\ell})$ are computed. Here, $\boldsymbol{\xi}$ 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 around the ℓth frame within the first channel of the mixture. The speaker with the highest $M_\ell(\boldsymbol{\xi}, \boldsymbol{\chi}_{x_\ell})$ is the most distinctive from the perspective of the X-vectors and is also *assumed to be dominant* in the mixture. Validity of this assumption is experimentally verified in Section III-D1.

4) Specific Pilot Variants: Two variants of a pilot signal are considered in this manuscript. The properties and limitations of the proposed pilots are demonstrated experimentally in Section III-D.

The realizable *X-vector-based pilot* $\mathbf{g}^{\text{XVEC}}$ is computed according to (12), where the SOI is considered dominant in the ℓth frame if

$$\begin{aligned} &M_\ell(\boldsymbol{\xi}_s, \boldsymbol{\chi}_{x_\ell}) > \max\{M_\ell(\boldsymbol{\xi}_{z_j}, \boldsymbol{\chi}_{x_\ell}), j = 1 \dots J\} \text{ and} \\ &M_\ell(\boldsymbol{\xi}_s, \boldsymbol{\chi}_{x_\ell}) > \mu_{\text{PLDA}}(\boldsymbol{\xi}_s), \end{aligned} \tag{13}$$

where $\boldsymbol{\xi}_s$ denotes the X-vector corresponding to the SOI and $\mu_{\text{PLDA}}(\boldsymbol{\xi}_s)$ is the lowest PLDA score, where the SOI is still considered active. The variable $\boldsymbol{\xi}_{z_j}$ denotes the X-vector corresponding to the jth potential interfering speaker from the enrollment set containing the SOI and J other speakers. To compute the $\mathbf{g}^{\text{XVEC}}$, the following two pieces of information are *needed:* the identity of SOI (we need to know which source we want to extract) and the enrollment set containing the X-vector for each speaker present in the processed dataset. On the other hand, the number of sources in the mixture or the identities of the active interferers are *not required*.

An *oracle pilot* $\mathbf{g}^{\text{ORAC}}$ is used to analyze the possibilities of the piloting proposed via (12). The dominance of the SOI is always determined correctly using true unobservable energies of the sources. Due to the use of unavailable information, it cannot be used in practice. $\mathbf{g}^{\text{ORAC}}$ is computed using (12), where SOI is considered dominant within the ℓth frame if

$$\sum_{k=1}^{K} |s_\ell^k|^2 > \mu_{\text{ORAC}} \sum_{k=1}^{K} ||\mathbf{z}_\ell^k||^2, \tag{14}$$

where μ_{ORAC} is a free parameter reflecting the desired level of dominance.

D. Non-Intrusive Assessment of Extraction Quality

This Section proposes a non-intrusive criterion to assess whether the extraction of the SOI was successful. This criterion is based on the same X-vectors and PLDA as the piloting, i.e., no additional training is required.

The assessment represents the entire signal through a single PLDA score $M(\boldsymbol{\xi}_s, \boldsymbol{\chi}_{\hat{s}})$, where $\boldsymbol{\chi}_{\hat{s}}$ is the X-vector independent of ℓ computed from an estimate of the SOI (FSMN pooling context is set $L_c = L$). As in the conventional speaker identification, this score can be seen as a measure of similarity between the X-vector computed from the enrollment utterance of the SOI and an unknown test X-vector. The two following observations concerning the score hold. 1) When the SOI is truly active in the test utterance, its PLDA score is higher than the non-active speakers' scores. 2) Interferences decrease the similarity/score compared to values observed on undistorted test signals. Based on these observations, the *extraction assessment* is proposed:

Assessment (of extraction quality): *Having X-vectors for two signals containing the same component corresponding to the SOI*[2] *denoted by $\boldsymbol{\chi}_{\hat{s}}, \boldsymbol{\chi}_{\dot{s}}$; if $M(\boldsymbol{\xi}_s, \boldsymbol{\chi}_{\hat{s}}) > M(\boldsymbol{\xi}_s, \boldsymbol{\chi}_{\dot{s}})$ then $\boldsymbol{\chi}_{\hat{s}}$ corresponds to a superior estimate of this SOI component in the sense of speech quality.*

The extraction assessment is experimentally validated in Section III-E. The Section shows a strong linear dependence between increments in criteria measuring quality of speech and the

[2] For example, the original mixture and the extracted signal.

Algorithm 1: Deflation mechanism for CSV-AuxIVE using the extraction assessment. Variables $\chi^i_{\hat{s}}$ and χ^i_x denote X-vectors corresponding to the SOI estimate and the first channel of the mixture after i deflation steps.

```
Require: Multi-channel mixture $\mathbf{x}^k_\ell$, X-vector FSMN,
  enrollment set including the SOI, PLDA model
  for $i \leftarrow 0, i < I$ do
    Extract $\hat{\mathbf{s}}^{k,i}_{\ell_t}$ from $\mathbf{x}^{k,i}_{\ell_t}$ using piloted CSV-AuxIVE
    if $M(\boldsymbol{\xi}_s, \boldsymbol{\chi}^i_{\hat{s}}) > M(\boldsymbol{\xi}_s, \boldsymbol{\chi}^i_x)$ then
      return $\hat{\mathbf{s}}^{k,i}_{\ell_t}$ {Extracted source is the SOI estimate}
    else
      $\mathbf{x}^{k,i+1}_{\ell_t} \leftarrow$ Subtract $\hat{\mathbf{s}}^{k,i}_{\ell_t}$ from $\mathbf{x}^{k,i}_{\ell_t}$ using (15)
      if $M(\boldsymbol{\xi}_s, \boldsymbol{\chi}^i_x) > M(\boldsymbol{\xi}_s, \boldsymbol{\chi}^{i+1}_x)$ then
        return $\mathbf{x}^{k,i}_{\ell_t}$ {Reduced mixture is not closer to the
          SOI, end the deflation}
      else
        {Continue the deflation}
      end if
    end if
  end for
  return $\mathbf{x}^{k,i+1}_{\ell_t}$ {Maximum number of steps reached}
```

PLDA score. The extraction assessment is used in the deflation process as a *decision mechanism*. It determines whether the extracted source is an estimate of the SOI or of an unwanted source (and the deflation should be applied).

E. Re-Estimation of the SOI on Extraction Failure: Deflation.

The deflation provides a mechanism to extract the SOI from mixtures in which the desired source is difficult to identify via pilot alone. This may happen, e.g., when the SOI is the weaker source and only a small number of frames with dominant SOI exist to form an efficient pilot.

The deflation is summarized in Algorithm 1 and proceeds as follows. The first signal is extracted using the piloted CSV-AuxIVE. The extraction assessment 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 (on each CSV block) using least square projection. 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 a predefined number I of deflation steps has been performed. It is reasonable to select I close to the assumed number of speakers active in the mixture. Owing to the utilization of the pilot signal, the CSV-AuxIVE is forced to converge towards speech signals. Thus, the active speakers are the first extracted sources in most cases.

Let $\mathbf{x}^{k,i}_{\ell_t} \in \mathbb{C}^{d-i}$ and $\mathbf{w}^{k,i} \in \mathbb{C}^{d-i}$ denote the input mixture and the separating vector after i deflation steps, respectively. The reduced mixture $\mathbf{x}^{k,i+1}_{\ell_t} \in \mathbb{C}^{d-i-1}$ is obtained by the least square subtraction of the extracted signal $\hat{\mathbf{s}}^{k,i}_{\ell_t} = (\mathbf{w}^{k,i})^H \mathbf{x}^{k,i}_{\ell_t}$ from $\mathbf{x}^{k,i}_{\ell_t}$. Let $\mathbf{a}^{k,i}_t$ be the mixing vector after i deflation steps computed on the tth block via (8). Due to the orthogonality of $\mathbf{w}^{k,i}$ and $\mathbf{a}^{k,i}_t$, the subtraction is achieved through

$$\mathbf{x}^{k,i+1}_{\ell_t} = \mathbf{D}^{k,i}(\mathbf{x}^{k,i}_{\ell_t} - \mathbf{a}^{k,i}_t (\mathbf{w}^{k,i})^H \mathbf{x}^{k,i}_{\ell_t}), \qquad (15)$$

where $\mathbf{D}^{k,i}$ is a $(d-i-1) \times (d-i)$ full row-rank matrix, reducing the dimension of $\mathbf{x}^{k,i+1}_{\ell_t}$ by one compared to $\mathbf{x}^{k,i}_{\ell_t}$. This reduction needs to be applied to avoid rank deficiency of the "deflated" mixture. Matrix $\mathbf{D}^{k,i}$ can be found via principal component analysis [50] or can simply omit one element of $\mathbf{x}^{k,i}_{\ell_t}$.

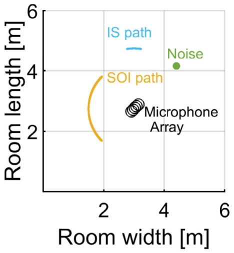

Fig. 1. Source trajectories and locations for the Dynamic dataset.

III. Experiments

The following experiments pursue three goals. 1) The possibilities and limitations of piloting are investigated as a motivation for the proposed deflation. 2) The functionality of the proposed extraction assessment is analyzed. 3) The benefits of the deflation are demonstrated and the performance of the proposed extractor is compared to results published in the literature.

A. Datasets

The experiments are performed on the following three datasets, which contain various detrimental phenomena such as source movements, high reverberation and noise activity, transients or low energy of the SOI.

1) Dynamic Dataset: The first dataset is an ad-hoc simulated one containing noisy recordings of two simultaneously active moving speakers (SOI and an interfering source (IS)). The sources are located in a room of dimensions $6 \times 6 \times 3$ m; reverberation times $T_{60} \in \{100, 300, 600\}$ ms are considered. A linear array of five omni-directional microphones with spacing of 8 cm is placed close to the center of the room and rotated counter-clockwise by $45°$. Both sources move on a half-circle around the array, the radius is 1.5 m for the SOI and 2 m for the IS. SOI performs a large angular movement in the left-hand half-plane and IS a small one in the right-hand half-plane. A static directional noise source is located perpendicular to the microphone array axis to the right. The situation is depicted in Fig. 1.

The speech (sampled at 16 kHz) originates from the test/development sets of CHiME-4; four potential speakers (F01, F06, M04, M05) are considered. The cafeteria sounds used for a directional noise originates from the QUT corpus [51]. Different utterances are concatenated to form 5 unique test signals of length 25 s for each speaker. The movements of SOI and positions of the static sources are simulated using the RIR generator [49]. One instance of the experiment (for one T_{60} value) thus consists of 300 mixtures (6 speaker combinations × 2 speaker roles × 25 utterance combinations). The sources are mixed at an input signal-to-interference-ratio of 0 dB (SIR, ratio of energy of SOI and IS) and an input signal-to-noise-ratio of 10 dB (SNR, ratio of all speech to noise energy).

2) CHiME-4 Dataset: CHiME-4 dataset [43] contains six-channel real-world and simulated recordings of a single speaker active in a highly noisy environment. The dataset does not contain cross-talk; however, the real-world part contains a lot of microphone failures and transient noises. These non-speech signals are occasionally extracted instead of SOI.

3) Multi-Channel Wall Street Journal - 2mix Dataset: The multi-channel version of the Wall Street Journal - 2mix dataset (MC-WSJ0-2mix, [2]) is currently often used to compare speaker separation and extraction algorithms. The MC-WSJ0-2mix dataset contains 3,000 simulated mixtures recorded in a reverberant environment using a microphone array containing eight microphones. Each mixture contains two active speakers, i.e., there is 6,000 extraction experiments in total. The sources are mixed with SIR between $\langle -5, +5\rangle$ dB. Some of the recordings are very short; their durations range from 1.6 s to 13.9 s. The recordings are highly reverberant ($T_{60} \in \langle 200, 600\rangle$ ms), and captured in rooms with variable dimensions. The geometry 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 dataset does not contain environmental noise or source movements. The 8 kHz variant of the mixtures is used.[3]

B. Evaluation Measures and Common Settings

The extraction is evaluated in terms of the following metrics. SIR and SDR are computed using BSS_EVAL [52]. The perceptual quality of the extracted sources is quantified using the "perceptual evaluation of speech quality" (PESQ [53]) or "short-time objective intelligibility measure" (STOI, [54]). These metrics are evaluated over the entire signal lengths with the exception of the Dynamic dataset, for which (due to source movements) the measures are evaluated within intervals of length 1 s each and subsequently averaged. The metrics are either stated as values or as improvements with respect to the mixture (iSIR, iSDR, iPESQ, iSTOI).

When the extraction algorithm fails to track a moving SOI (the SOI moves out of the spatial focus of the method), the desired speech vanishes from the estimated signal. To measure this phenomenon, we also provide *the standard deviation* of the "SOI Attenuation" metric, defined as $\sum_k |\hat{s}_\ell^k|^2 / \sum_k |s_\ell^k|^2$, where $\hat{s}_\ell^k$ is the estimate of s_ℓ^k. For a properly extracted moving SOI, this deviation should be close to zero and it increases if the gain of the desired speech fluctuates.

All the experiments have been performed without any adaptation of the algorithm or the FSMN network to a specific scenario. The enrollment set always consists of 1 minute of speech for each target speaker considered in the given scenario, augmented by reverberation as described in Section II-C2. The FSMN pooling context length is $L_c = 11$.

C. CSV Model for Extraction of a Moving SOI

This experiment is performed on the Dynamic dataset. It demonstrates the benefits of the CSV-model on mixtures with moving sources and the ability of $\mathbf{g}^{\text{XVEC}}$ to direct the extraction towards a moving SOI. The deflation is not applied in these experiments, since the mixtures are 25 s long and $\mathbf{g}^{\text{XVEC}}$ founds sufficient number of frames to successfully guide the extraction. The results of CSV-AuxIVE are compared to the fully static (FS-IVE, [27]) and the block-wise static (BS-IVE, [42]) variants of AuxIVE. The name of a method followed by subscript L_T (e.g. BS-IVE$_{200}$) denotes the number of frames within the analyzed block.

CSV-AuxIVE and FS-IVE process the entire mixture as a whole using 50 iterations. BS-IVE processes each block independently and applies 5 iterations to each block of length L_T and shift $L_T/4$ frames. The inner statistics in BS-IVE are accumulated using recursive forgetting with $\alpha = 0.3$ (see [35]). All these methods are initialized using the location of the SOI at the beginning of the recording; BS-IVE initializes the extraction at each block by the solution from the previous one. The NFFT length is 1,024 and shift 200 samples. The threshold $\mu_{\text{ORAC}} = 2$.

All criteria in Table II indicate that the pilot-guided methods extract the SOI more precisely than the methods relying on initialization (without any pilot). Due to the limited identification accuracy, the performance with $\mathbf{g}^{\text{XVEC}}$ is inferior to that with $\mathbf{g}^{\text{ORAC}}$ (by $1.6 - 2.8$ dB of iSIR). The CSV-AuxIVE achieves superior (or at least comparable) performance compared to its static or block-wise static counterparts. This is notable especially when $\mathbf{g}^{\text{XVEC}}$ is used. CSV-AuxIVE appears to be more robust than BS-IVE with respect to pilot inaccuracies. The performance of the method decreases with increasing reverberation. However, even when $T_{60} = 600$ ms, the CSV-AuxIVE + $\mathbf{g}^{\text{XVEC}}$ is able to achieve iSIR 6.9 dB. The iSDR is low in this case, which means that the suppression of interference/noise introduces some distortions into the estimated SOI. However, this scenario is very challenging for spatial filtering due to the low direct to reverberation ratio (the SOI distance is 1.5 m) and rather high movement speed of the sources.

The important parameter of CSV-AuxIVE/BS-IVE is the length of block L_T, which influences the compromise between adaptivity to movement and the amount of available data. Excessively long blocks (FS-IVE or BS-IVE$_{800}$) yield high iSIR and iPESQ but also increase Attenuation compared to the suitable block length (BS-IVE$_{200}$). Using long blocks, the methods are unable to adapt well to the source movements and the SOI moves out of their spatial focus (the sound vanishes for certain time intervals).[4] The increased Attenuation is observable for the CSV$_{800}$ as well; the increase of iSIR/iPESQ is, however, not

[3] We interpolate the mixtures to 16 kHz in order to be able to process it via the FSMN network. We found that this approach gives comparable results to retraining the network on training datasets down-sampled to 8 kHz.

[4] Note that the Attenuation describes the vanishing of the SOI well for $\text{T}_{60} \leq$ 300 ms but fails to capture this phenomenon for more reverberant scenario. We can observe that this fact is due to the reverberation of the SOI, which is still present in the estimate even when the location (direct path) of the SOI lies outside of the spatial focus of the methods.

TABLE II
DYNAMIC DATASET: THE EXTRACTION PERFORMANCE OF THE CSV-AUXIVE (CSV), THE STATIC (FS-IVE) AND THE BLOCK-WISE STATIC (BS-IVE) IVE TECHNIQUES

Unprocessed mixture										
	input PESQ [-]			**input SDR [dB]**			**input SIR [dB]**			
	100ms	300ms	600ms	100ms	300ms	600ms	100ms	300ms	600ms	
	1.15	1.20	1.27	1.10	1.12	1.14	1.10	1.12	1.14	

Processed using suitable block length														
Method	L_T	**Pilot**	**iPESQ [-]**			**iSDR [dB]**			**iSIR [dB]**			**Attenuation**		
			100ms	300ms	600ms	100ms	300ms	600ms	100ms	300ms	600ms	100ms	300ms	600ms
FS-IVE	2000	-	0.32	0.05	-0.01	6.84	-0.95	-3.73	15.33	6.97	4.04	0.42	0.17	0.10
FS-IVE	2000	$\mathbf{g}^{\rm ORAC}$	0.61	0.22	0.11	8.05	2.74	-0.17	19.72	13.31	9.72	0.32	0.19	0.12
FS-IVE	2000	$\mathbf{g}^{\rm XVEC}$	0.51	0.14	0.04	7.29	1.31	-2.37	18.46	11.16	6.61	0.33	0.18	0.10
Proposed CSV	200	-	0.49	0.11	0.02	7.89	0.21	-2.88	17.03	7.60	4.60	0.36	0.15	0.10
Proposed CSV	200	$\mathbf{g}^{\rm ORAC}$	0.87	0.27	0.13	11.52	3.58	0.16	22.01	13.29	9.64	0.25	0.13	0.10
Proposed CSV	200	$\mathbf{g}^{\rm XVEC}$	0.76	0.20	0.06	10.24	2.23	-1.80	20.46	11.38	6.85	0.28	0.14	0.10
BS-IVE	200	-	0.09	0.00	-0.07	2.99	-1.07	-3.39	12.50	6.53	4.17	0.32	0.18	0.14
BS-IVE	200	$\mathbf{g}^{\rm ORAC}$	0.39	0.16	0.05	8.11	3.54	0.47	19.34	13.54	9.92	0.26	0.17	0.14
BS-IVE	200	$\mathbf{g}^{\rm XVEC}$	0.22	0.04	-0.05	5.33	0.61	-2.49	15.69	9.25	5.69	0.28	0.17	0.13
Processed using excessively long/short blocks														
Proposed CSV	800	$\mathbf{g}^{\rm XVEC}$	0.52	0.15	0.03	7.05	1.11	-2.53	18.11	10.68	6.17	0.32	0.17	0.10
BS-IVE	800	$\mathbf{g}^{\rm XVEC}$	0.42	0.11	0.00	6.67	1.24	-2.22	18.11	10.35	5.91	0.31	0.19	0.13
Proposed CSV	50	$\mathbf{g}^{\rm XVEC}$	0.60	0.14	0.02	10.47	1.80	-1.83	17.63	8.47	4.45	0.16	0.11	0.10
BS-IVE	50	$\mathbf{g}^{\rm XVEC}$	0.04	-0.03	-0.10	2.77	-0.47	-3.01	11.83	7.63	4.85	0.23	0.15	0.12

present. The prolongation of inner blocks does not bring the advantage of more available data. Application of an insufficiently short block (50 frames) allows for good adaptation (low SOI Attenuation), but the overall IS suppression is deteriorating (low iSIR).

D. Properties and Limitations of Piloting

This Section analyzes the accuracy of the frame-wise speaker identification. Subsequently, the influence of inaccurate pilot on the extraction accuracy is investigated and the causes of pilot failures are discussed.

1) The Frame-Wise Dominant Speaker Identification: As a ground truth in this task, we use the true identity of the speaker with the highest energy in the mixture. This energy is computed using the same context of frames as the pooling context of FSMN ($L_c \in \{7, 11, 21\}$, i.e., $\{9, 14, 26\}$ ms). The accuracy of the identification is thus computed by a comparison of regions determined by the X-vectors and the oracle information obtained using the true energies. The most reverberant part ($T_{60} = 600$ ms) of the Dynamic dataset is revisited. Multiple variants of this dataset are considered, each changing the input SIR $\in \{-5, 0, 5, 10, 20\}$ dB and the input SNR $\in \{0, 10, \infty\}$ dB. Markers in Figs. 2 and 3 correspond to the averaged accuracy over all mixtures in one such variant.

Let us first verify the assumption that the source with the highest PLDA score is also the dominant one in the mixture. Considering $L_c = 11$ and the noiseless case, Fig. 2 confirms our assumption with the accuracy ranging from 59%–77%. By definition of the pilot in (12), the identity of the interfering speaker is irrelevant for $\mathbf{g}^{\rm XVEC}$. The classification is thus simplified to a binary task whether the SOI or an arbitrary other source is dominant. Fig. 3 shows that the accuracy of SOI dominance identification is 69%–77%. The presence of noise decreases the accuracy to 63%–73%. The results of the extraction in the previous Section indicate that such accuracy leads to a functional $\mathbf{g}^{\rm XVEC}$, which improves the performance of CSV-AuxIVE by iSIR $= 2.3$ dB over its non-piloted counterpart. Utilization of $\mathbf{g}^{\rm ORAC}$ leads to another increase by 2.8 dB. The influence of inaccurate pilot on extraction performance is further investigated in Section III-D3.

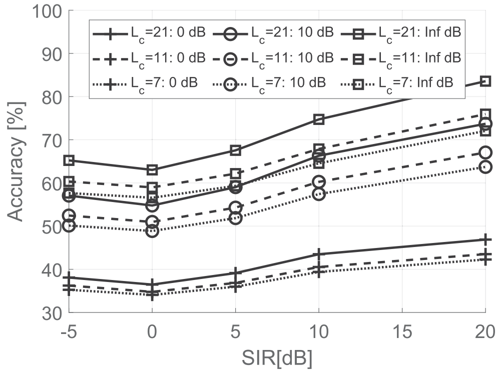

Fig. 2. Accuracy in the task of the dominant speaker identification; each marker corresponds to a different SNR.

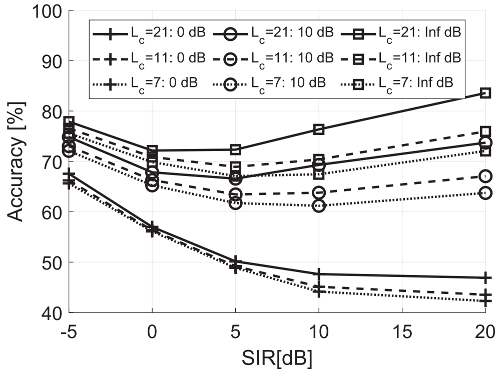

Fig. 3. Accuracy in the task of the SOI dominance identification; each marker corresponds to a different SNR.

TABLE III
Dynamic Dataset: The Extraction Performance of Piloted CSV-AuxIVE ($L_T = 200$) With Respect to X-Vector Context L_c

Context	**iSDR [dB]**			**iSIR [dB]**		
L_c	100ms	300ms	600ms	100ms	300ms	600ms
7	10.25	2.24	-1.83	20.48	11.40	6.71
11	10.24	2.23	-1.80	20.46	11.38	6.85
21	10.07	2.04	-1.84	20.14	10.87	6.64

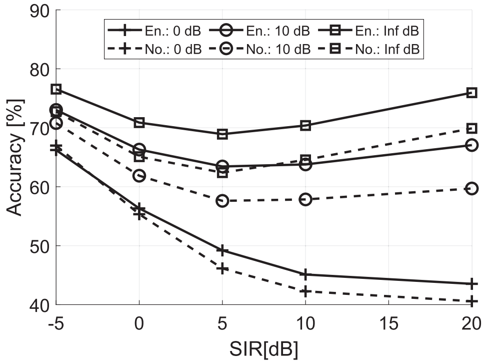

Fig. 4. Accuracy in the task of the SOI dominance identification with respect to language of speakers in the enrollment set; $L_c = 11$ and each marker corresponds to a different SNR.

TABLE IV
Dynamic Dataset: The Extraction Performance With Respect to Spoken Language (English or Norwegian); the Language Dependence of $\mathbf{g}^{\mathrm{XVEC}}$

Method	**Pilot**	**Lang.**	**iPESQ**		**iSDR [dB]**		**iSIR [dB]**	
			100ms	600ms	100ms	600ms	100ms	600ms
CSV_{200}	-	Eng.	0.49	0.02	7.89	-2.88	17.03	4.60
CSV_{200}	$\mathbf{g}^{\mathrm{ORAC}}$	Eng.	0.87	0.13	11.52	0.16	22.01	9.64
CSV_{200}	$\mathbf{g}^{\mathrm{XVEC}}$	Eng.	0.76	0.06	10.24	-1.80	20.46	6.85
CSV_{200}	-	Nor.	0.42	-0.01	7.37	-3.12	15.05	2.62
CSV_{200}	$\mathbf{g}^{\mathrm{ORAC}}$	Nor.	0.81	0.11	12.54	0.80	22.42	9.70
CSV_{200}	$\mathbf{g}^{\mathrm{XVEC}}$	Nor.	0.74	0.02	11.14	-1.52	20.86	6.02

The subscript L_T denotes the number of frames within the analyzed block.

TABLE V
MC-WSJ0-2mix, 4 Channels: The Number of Cases, When CSV-AuxIVE: 1) Extracts an Unwanted Source Due to Insufficient Piloting (iSDR < -2 dB), 2) Extracts No Source (2 dB $<$ iSDR $<$ -2 dB), 3) Successfully Extracts the SOI (iSDR > 2 dB)

Pilot/deflation	**Unwanted source extracted**	**No source extracted**	**SOI extracted**
No pilot	2986	616	2398
$\mathbf{g}^{\mathrm{XVEC}}$	697	753	4550
$\mathbf{g}^{\mathrm{XVEC}}$ + deflation	58	1016	4926
$\mathbf{g}^{\mathrm{ORAC}}$	24	253	5723

It might seem surprising that accuracy of SOI dominance identification is high despite the low SIR. This is caused by a low occurrence of frames with a dominant SOI (for SIR$= -5$ dB, only 28.5% of frames). The classifier is thus often correct when it assigns the frame to the easily classifiable interfering source with high energy.

A short context of the pooling layer L_c is required for the frame-wise identification. However, it deteriorates the accuracy due to the increased variability of the X-vectors (less data is available for the pooling). Figs. 2 and 3 indicate that this accuracy is, as expected, highest for $L_c = 21$ and monotonically deteriorates with decreasing L_c. On the other hand, Table III shows that the long context $L_c = 21$ achieves the worst extraction performance; the piloting is no longer well localized in time. As a compromise, context $L_c = 11$ is utilized throughout this manuscript.

2) Language Dependence of the SOI Identification: The blind CSV-AuxIVE algorithm is language independent. However, piloting using $\mathbf{g}^{\mathrm{XVEC}}$ is based on deep-learning and thus is designed to work on English language present in the training dataset. Its accuracy might deteriorate if applied to an unseen language. To quantify, this scenario compares the SOI identification/extraction achieved on English with results yielded on unseen Norwegian. It analyzes a slightly modified version of the Dynamic dataset. The original English speakers are replaced by four Norwegian (2 male and 2 female) originating in the NST speech database [55].

The results in Fig. 4 corroborate that X-vectors are slightly language dependent; the accuracy for Norwegian speakers is lower by about 4.5%. However, this does not influence the extraction performance much. The metrics in Table IV indicate that the non-piloted extraction is slightly less accurate for the Norwegian dataset. This decrease does not stem from the language as such but it is caused by longer silences between Norwegian sentences. When the SOI is quiet, the non-piloted extractor tends to converge to an arbitrary active source. The utilization of a pilot completely removes this difference. The results for CSV-AuxIVE piloted via $\mathbf{g}^{\mathrm{XVEC}}$ are comparable for both datasets (difference is maximally 1 dB in iSIR and iSDR); the proposed method can thus be considered language independent in this experiment.

3) Limitations of the Embedding-Based Piloting in Low SIR Scenarios: By definition in (12), the pilot is non-zero/active when the SOI is dominant in a subset of frames. This condition becomes difficult to fulfill when SIR is low. Let us demonstrate using mixtures in the Dynamic dataset. Considering three levels of SIR=$\{20, 0, -5\}$ dB; the SOI is dominant in $93.2\%, 48.3\%$ and 28.5% of frames, respectively. For a low SIR, the potential support is limited, which weakens the guidance provided by the pilot. $\mathbf{g}^{\mathrm{XVEC}}$ suffers from a further reduction of the support, because it incorrectly identifies a subset of the dominant frames. An extreme case of the pilot being equal to zero for all frames leads to non-piloted extraction (which, moreover, tends to extract the dominant interfering source).

The deflation approach provides a mechanism to alleviate these limitations. Let us demonstrate via an extraction experiment on MC-WSJ0-2mix dataset [2]. The dataset contains 3,000 mixtures of two active speakers. Since each speaker can assume the role of the SOI, 6,000 independent extractions can be performed. Let us observe in Table V the number of cases when CSV-AuxIVE successfully extracts a source, but it is an unwanted source due to insufficient guidance. We assume this happens when the iSDR is less than -2 dB. The non-piloted

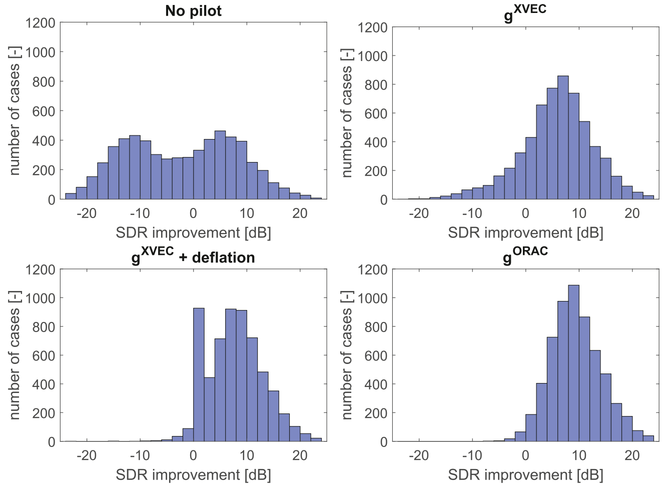

Fig. 5. MC-WSJ0-2mix, 4 channels: iSDR distributions achieved by CSV-AuxIVE endowed with various forms of guidance.

CSV-AuxIVE fails in 2,986 cases. It has no way to focus on a specific SOI and, in addition, fails to process some of the mixtures (output SDR is close to 0 dB). $\mathbf{g}^{\mathrm{XVEC}}$ reduces the fail rate by about 77 % to 680 cases; in 440 of these mixtures, the SOI is the weaker source (input SDR$<$ 0 dB). The deflation significantly reduces the number to 58 cases, which is comparable to the utilization of $\mathbf{g}^{\mathrm{ORAC}}$ (which does not suffer from the erroneous classification of frames).

The distributions of iSDR achieved in this task are shown in Fig. 5. The CSV-AuxIVE without a pilot has a symmetric distribution of iSDR because it cannot focus on specific SOI. Utilization of $\mathbf{g}^{\mathrm{XVEC}}$ shifts the distribution towards the positive iSDR. However, many cases of negative iSDR remain, corresponding to unsuccessful piloting. For some cases, the piloting prevents extraction of an unwanted source, but fails to guide the extraction towards the SOI. Therefore, CSV-AuxIVE+$\mathbf{g}^{\mathrm{XVEC}}$ yields a slightly increased number of cases with no extracted source compared to CSV-AuxIVE without a pilot (see Table V). The deflation manages to remedy almost all failed piloting cases and further shifts the distribution to the positive values. However, part of these remedied cases does not lead to successful extraction of the SOI; their output iSDR is equal to 0 dB. This effect is caused by an overly conservative behavior in the assessment of the extraction quality (see Section III-E for further discussion). It recognizes that an unwanted source was extracted and performs the deflation of the mixture. However, the reduced mixture is not recognized as a better estimate of the SOI than the original mixture. Consequently, the original mixture is returned as the SOI estimate. The utilization of an accurate $\mathbf{g}^{\mathrm{ORAC}}$ causes a successful extraction of the SOI for most of the mixtures.

The influence of the incorrectly classified frames in the pilot on the final SDR is shown in Fig. 6. In this experiment, we pilot the extraction on the MC-WSJ0-2mix dataset by $\mathbf{g}^{\mathrm{ORAC}}$. We gradually replace 10% of frames with dominant SOI by 10% of frames corresponding to the unwanted source. The frames with comparable energy are swapped first; the frames with a highly dominant source are swapped as the last ones. It can be seen that the substitution of about 20% of frames does not significantly influence the performance. When all frames are substituted, i.e., the pilot contains only frames corresponding to the interfering source, CSV-AuxIVE achieves highly negative SDR because it is in all cases guided towards the interfering source.

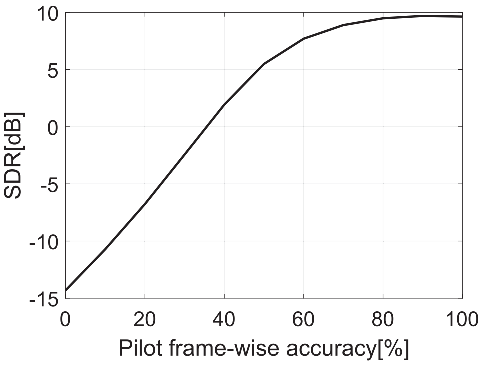

Fig. 6. MC-WSJ0-2mix, 4 channels: SDR achieved by the CSV-AuxIVE using oracle pilot, whose frames corresponding to SOI are gradually interchanged with frames dominated by the interfering source.

The accuracy of SOI dominance identification in $\mathbf{g}^{\mathrm{XVEC}}$ is 66.3% on the MC-WSJ0-2mix dataset, which yields an SDR of 6 dB. Comparing the results with Fig. 6, such accuracy should yield an SDR of about 8 dB. The modeling of errors by distorting the $\mathbf{g}^{\mathrm{ORAC}}$ thus appears to be slightly more optimistic than the results achieved using the realizable $\mathbf{g}^{\mathrm{XVEC}}$.

E. Non-Intrusive Assessment of Extraction Quality

This section verifies whether the PLDA score can be used to select a superior SOI estimate within several available variants. The superiority is measured using the standard objective and perceptual metrics (SIR, SDR, PESQ, STOI).

We use two datasets: the simulated development part of the CHiME-4 dataset contains 1,640 mixtures of speech (produced by 4 speakers) and noise, whereas MC-WSJ0-2mix contains 3000 mixtures of two utterances (produced by 18 speakers). The non-piloted CSV-AuxIVE with uniform initialization is applied to these recordings and stopped consecutively after $\{0, 5, 10, 15, 20, 25\}$ iterations for the CHiME-4 and $\{0, 15, 30, 50\}$ iterations for MC-WSJ0-2mix. For each utterance and each stop, the PLDA score $M(\boldsymbol{\xi}_s, \boldsymbol{\chi}_{\hat{s}})$ and the metrics are evaluated. Subsequently, the differences with respect to the previous stop are computed because the goal is to find the relationship between the change of $M(\boldsymbol{\xi}_s, \boldsymbol{\chi}_{\hat{s}})$ and the change in the metrics.

The differences plotted in Figs. 7 and 8 indicate the existence of a linear dependence. The Pearson correlation coefficient reaches a value of 0.83 for STOI. From another perspective, the assessment can also be perceived as a binary classifier: given the increase/decrease of $M(\boldsymbol{\xi}_s, \boldsymbol{\chi}_{\hat{s}})$, we want to predict the respective change in the objective criterion. Tables VI and VII show that the classification accuracy is 72.7 % and 75.5 % for SIR on speech-noise and speech-speech mixtures, respectively.

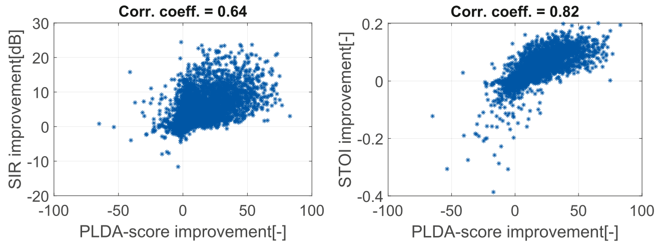

Fig. 7. CHiME-4 (simulated development part): dependency between the improvements of the objective criteria and the improvements of PLDA score on speech-noise mixtures.

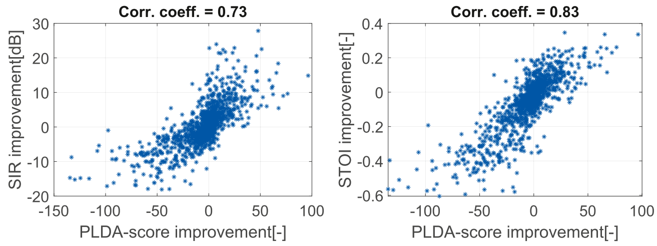

Fig. 8. MC-WSJ0-2mix: dependency between the improvements of the objective criteria and the improvements of PLDA score on speech-speech mixtures.

TABLE VI
CHiME-4 (Simulated Development Part, Speech-Noise Mixtures): The Evaluation of the Proposed Extraction Assessment Serving as a Binary Classifier of the Speech-Quality Metrics

	SIR	**SDR**	**PESQ**	**STOI**
Correlation coefficient [-]	0.64	0.74	0.58	0.82
Accuracy[%]	72.7	70.4	70.4	69.2
False positives [%]	5.7	15.3	16.3	20.2
Significant false positives [%]	0.3	1.2	0.8	1.6
False negatives [%]	21.6	14.2	13.3	10.5
Significant false negatives [%]	7.9	2.7	1.7	0.8

Significant cases denote samples in which the erroneous increase/decrease in SIR/SDR is larger than 1 dB, 0.01 in STOI or 0.05 in PESQ.

TABLE VII
MC-WSJ0-2mix (Speech-Speech Mixtures): The Evaluation of the Proposed Extraction Assessment Serving as a Binary Classifier of the Speech-Quality Metrics

	SIR	**SDR**	**PESQ**	**STOI**
Correlation coefficient [-]	0.73	0.77	0.63	0.83
Accuracy[%]	75.5	75.6	71.7	76.5
False positives [%]	10.3	12.5	12.5	14.3
Significant false positives [%]	5.5	6.4	5.8	11.0
False negatives [%]	14.3	11.9	15.8	9.1
Significant false negatives [%]	8.4	4.9	7.9	6.1

Significant cases denote samples in which the erroneous increase/decrease in SIR/SDR is larger than 1 dB, 0.01 in STOI or 0.05 in PESQ.

There are two types of error: 1) False positives ($M(\boldsymbol{\xi}_s, \boldsymbol{\chi}_{\hat{s}})$ increases, but the metrics decrease) are more severe and potentially lead us to select an interfering source. Fortunately, the number of cases with significant deterioration is not very high. A decrease worse than 1 dB in SIR happens only in 5.5% of cases for speech-speech mixtures. 2) False negatives ($M(\boldsymbol{\xi}_s, \boldsymbol{\chi}_{\hat{s}})$ decreases, but criteria increase) potentially lead us to a selection of an inferior estimate. An 8.4% proportion of the cases exhibits a significant decrease in SIR for speech-speech mixtures. These cases cause the overly conservative behavior of the deflation described in Section III-D3.

The proposed assessment is functional in both speech-noise and speech-speech mixtures. However, the number of significant incorrect cases is larger for the speech-speech mixtures, where the active sources are more similar and can be confused more easily.

F. *Extraction Via Deflation on Public Datasets*

The following experiments provide comparison between results achieved by the proposed method and the results reported in the literature. The experiments also show benefits brought by deflation.

1) Extraction of the SOI From Noisy Recordings With Transients and Microphone Failures: Piloting and deflation should not be necessary on CHiME-4 data since the recordings contain only one active speaker. However, the real-world part of CHiME-4 is sometimes distorted by transients and microphone failures. These signals behave like sources that are strongly non-Gaussian, which have wide areas of attraction in contrast functions of blind algorithms. Therefore, they can be extracted instead of speech. The piloting and deflation used in our method provide effective solutions for this phenomenon.

The enhancement via piloted CSV-AuxIVE is compared with two enhancers known to be very successful on the CHIME-4 data: BeamformIt [56], a weighted delay-and-sum beamformer, which is used as a front-end algorithm in the original CHiME-4 baseline system. The Generalized Eigenvalue Beamformer (GEV) is a front-end solution proposed in [57], [58]. The latter represents one of the most successful enhancers for CHiME-4. It relies on voice activity detection (VAD) via deep networks trained specifically for the CHiME-4 data. We utilize the feed-forward topology of the VAD (the training procedure was kindly provided to us by the authors of [57]) and re-train the network using the training part of the CHiME-4 data.

Since the true references of the sources are not available for the real-world part of CHiME-4, the experiments are evaluated using the WER of the original baseline recognizer from [59]. All of the proposed methods are initialized by the relative transfer function estimator from [60]. CSV-AuxIVE performs 5 iterations in the STFT domain with an FFT length of 512, hop-size of 200 (the shift of the FSMN network) and applied Hamming window; the sampling frequency is 16 kHz. The length of the CSV-AuxIVE block is 2 seconds ($L_T = 160$ frames). The enrollment set for piloting contains 8 speakers; respective speech signals originate from the simulated development part of CHiME-4.

The results in Table VIII indicate that the WER of CSV-AuxIVE[5] is lowered by using piloting and further using deflation. This is in agreement with the discussion presented in Section III-D3: namely, the piloting significantly reduces the number of diverged cases and the deflation allows for re-estimation of the SOI when the piloting fails. The proposed

[5] Slightly different WER of CSV-AuxIVE was reported in [40]; it is caused by a different FFT frame-shift and the number of performed iterations.

TABLE VIII
WER [%] YIELDED ON THE REAL-WORLD PART OF THE CHiME-4 DATASETS

	Mix. ch.5.	**Beam-form-It**	**GEV**	**CSV**	**CSV** +pilot	**CSV** +pilot +defl.
Dev.	9.8	5.8	4.6	5.8	5.4	5.4
Test	19.9	11.5	8.1	9.9	9.5	9.3

Mixture results are achieved using data from channel 5.

method yields lower WER values compared to BeamformIt but is still outperformed by GEV. Nevertheless, GEV is a technique specifically tailored to CHiME-4 due to dataset-specific VAD and is limited to enhancement of recordings without cross-talk. In contrast, the proposed technique is, without adaptation, applicable to both speech enhancement and extraction. Even without piloting, CSV-AuxIVE achieves results approaching those of GEV without a need for training.

2) Extraction of the SOI From Cross-Talk in a Reverberant Environment: The following experiment compares the performance of the proposed method on the MC-WSJ0-2mix dataset to the results reported in the literature. The competing methods can be divided into three groups: 1) Oracle methods representing ideal extractors. These methods cannot be used in practice as they utilize information that is normally not available. 2) Methods based on machine learning (ML), which rely on the existence of a scenario-specific training dataset. 3) Blind source separation/extraction methods, which exploit spatial information extracted from the multi-channel mixture.

For ML-based methods, we consider extraction approaches that identify the SOI and solely recover this source from the mixture. For blind approaches, the literature usually presents methods performing the complete separation (BSS). Here, all sources in the mixture are estimated (the number of sources must be known), and the SOI is subsequently identified among them. This can be done either in an oracle manner using the true reference during evaluation or using ML-based speaker identification (to this end, we use the same FSMN network as for piloting; the X-vector pooling context $L_c = L$). In contrast, the piloted CSV-AuxIVE extracts (BSE) only the SOI and does not require the number of interfering sources.

The oracle approaches are represented by the 1) multi-channel Wiener filter (MCWF), which uses the oracle covariance matrix of the target speech and constitutes the upper boundary for the extraction based on spatial filtering. The machine learning-based separation is represented by: 2) TasNet from [3], which is based on a convolutional topology performing full separation in the time domain; subsequently, the SOI is selected via speaker identification. 3) The frequency (FD) and time domain (TD) variants of SpeakerBeam [18], [22], which perform speaker extraction based on an enrollment utterance. Blind methods are represented by 4) masking-based binaural MESSL [16], 5) binaural GCC-NMF [15] based on non-negative matrix factorization, 6) consistent ILRMA from [11], 7) GLOSS [17] using sparsity-based spectral masking and single-channel post-filter and 8) static auxiliary function-based independent vector extraction FS-IVE [27].

TABLE IX
MC-WSJ0-2MIX: SDR [DB] YIELDED USING MACHINE-LEARNING (ML), BLIND SOURCE SEPARATION (BSS) AND BLIND SOURCE EXTRACTION (BSE)

Approach	**Chan. num.**	**Tr. data [hrs.]**	**Sepa-ration**	**Spk. id.**	**SDR [dB]**
Mixture	-	-	-	-	0.2
MCWF	2	-	Orac.	Orac.	9.0
MCWF	4	-	Orac.	Orac.	13.4
TasNet [3]	2	50	ML	ML	8.4
FD-SpkBeam [18]	2	50	ML	ML	7.9
TD-SpkBeam-Orig. [18]	2	50	ML	ML	11.5
TD-SpkBeam-Ext. [22]	2	50	ML	ML	12.9
ILRMA [11]	2	-	BSS	Orac.	5.9
GCC-NMF [15]	2	-	BSS	Orac.	2.7
MESSL [16]	2	-	BSS	Orac.	3.3
Prop. CSV+$\mathbf{g}^{\text{ORAC}}$	2	-	BSE	Orac.	5.4
ILRMA + ML spk. ident.	2	-	BSS	ML	5.4
FS-IVE+$\mathbf{g}^{\text{XVEC}}$+defl.	2	-	BSE	ML	4.5
Prop. CSV+$\mathbf{g}^{\text{XVEC}}$+defl.	2	-	BSE	ML	4.1
ILRMA [11]	4	-	BSS	Orac.	7.6
GLOSS [17]	4	-	BSS	Orac.	9.3
Prop. CSV+$\mathbf{g}^{\text{ORAC}}$	4	-	BSE	Orac.	9.6
ILRMA +ML spk. ident.	4	-	BSS	ML	7.2
FS-IVE+$\mathbf{g}^{\text{XVEC}}$+defl.	4	-	BSE	ML	7.7
Prop. CSV+$\mathbf{g}^{\text{XVEC}}$+defl.	4	-	BSE	ML	7.8
Prop. CSV+$\mathbf{g}^{\text{XVEC}}$	4	-	BSE	ML	6.0

The column "Tr. data" quantifies the volume of the required scenario-specific training data.

These methods are evaluated in terms of SDR implemented in the BSS_EVAL toolbox [52]. CSV-AuxIVE operates in the STFT domain with an FFT length of 1,000, hop-size of 100 (the shift of the FSMN network), and an applied Hamming window; the sampling frequency is 8 kHz. The length of the CSV-AuxIVE block is 2 seconds ($L_T = 160$ frames). The demixing filters are initialized with a vector of ones, because the locations of the sources and the topology of the microphone array are unknown. The enrollment set contains 18 speakers; the X-vectors are computed using unused sentences from the original WSJ0 dataset. The publicly available implementation[6] of consistent ILRMA [11] is used. ILRMA (using 100 iterations) and MCWF[7] were applied in the STFT domain with window length of 1024 and hop-size 512 samples. The results for TasNet were taken over from [18]; for MESSL and GCC-NMF, they were found in [2]. The other results originate from their respective references.

Restricting the methods to two channels, the results presented in Table IX show that the ML-based spatial+spectral filtering outperforms the blind spatial filtering by a large margin. The supervised methods are even comparable to oracle MCWF using two/four microphones. This is possible due to the existence of a strictly matching training part of the MC-WSJ0-2mix dataset. In this setting, CSV-AuxIVE with $\mathbf{g}^{\text{XVEC}}$ and deflation outperforms MESSL and GCC-NMF, but is outperformed by ILRMA.

The two-channel setting is, however, arguably unfair for blind methods relying solely on spatial diversity of the sources. Utilization of four channels increases the SDR for all blind methods.

[6] [Online]. Available: https://github.com/d-kitamura/ILRMA

[7] Different results of MCWF reported in [2] are caused by a short STFT length of 256 used there.

Using $\mathbf{g}^{\mathrm{XVEC}}$, deflation and 4 microphones, CSV-AuxIVE[8] is comparable to ML-based FD-SpeakerBeam and outperforms blind ILRMA performing full separation followed by ML-based speaker identification. Using $\mathbf{g}^{\mathrm{ORAC}}$, CSV-AuxIVE achieves results comparable to GLOSS. The results confirm that CSV-AuxIVE coincides with FS-IVE if the mixed sources are static. The best performance overall is achieved by the variants of TD-SpeakerBeam, which approach the oracle MCWF using 4 channels.

Concerning the benefits of deflation, the failures of $\mathbf{g}^{\mathrm{XVEC}}$ (discussed in Section III-D3), caused by a weak SOI activity and limited classification accuracy, deteriorate significantly the average SDR. Considering 4 microphones, CSV-AuxIVE using $\mathbf{g}^{\mathrm{XVEC}}$ yields SDR lower by 3.6 dB compared to CSV-AuxIVE using $\mathbf{g}^{\mathrm{ORAC}}$. The deflation partly alleviates this issue and increases the average SDR by 1.8 dB.

IV. Conclusion

This manuscript presents a novel method for target speech extraction from realistic mixtures. It consists of a combination of blind extraction using CSV-AuxIVE method and data-driven identification of the SOI. Due to decoupling of the extraction and the identification, the training required by the method is simpler compared to fully data-driven approaches. Moreover, the proposed method is applicable to a wide variety of realistic extraction scenarios without any adaptation. The guidance of the blind technique towards the SOI is ensured through two techniques: the piloting and the successive deflation of the multi-source mixture. Evaluation of the proposed approach leads to the following conclusions: 1) The presented frame-wise SOI identification applied to mixtures exhibits accuracy of 67% in highly reverberated and noisy scenarios ($T_{60} = 600$ ms and $\mathrm{SIR} = 0$ dB). 2) This accuracy is sufficient to form an efficient pilot able to guide the extraction in most scenarios. However, the embedding-based piloting fails when the mixture contains a small number of frames where the SOI is dominant, such as when the activity of the SOI is short and has a low energy level. These cases can be remedied using successive deflation of the mixture along with the re-estimation of the SOI. 3) The proposed non-intrusive assessment of extraction quality can successfully be used as a decision mechanism to determine whether the deflation should be applied. It is strongly correlated with the objective/perceptual criteria used to evaluate quality of speech; the Pearson coefficient between PLDA score and STOI improvements reaches a value of 0.83. 4) The procedure as a whole is language independent. The accuracy of the speaker identification deteriorates slightly for an unseen language, but this has negligible effect on the extraction. 5) The CSV-AuxIVE achieves more precise extraction compared to a blind block-wise static approach for mixtures of moving sources. On mixtures of static sources, the piloted CSV-AuxIVE is comparably accurate to competing blind approaches performing full separation followed by the ML-based/oracle speaker identification. In contrast to full separation approaches, CSV-AuxIVE does not require to know the number of speakers. 6) The proposed approach achieves a lower performance compared to the state-of-the-art machine learning-based algorithms, as observed on widely known CHiME-4 and MC-WSJ0-2mix datasets. On the other hand, it does not require any scenario-specific training data.

[8] We placed examples of the extraction on our web-page: https://asap.ite.tul.cz/demos/blind-extraction-of-target-speech-source-guided-by-piloting-and-deflation/

References

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speec, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[2] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 1–5.

[3] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech,Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[4] M. Togami, "End to end learning for convolutive multi-channel wiener filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 8032–8036.

[5] C. Boeddeker *et al.*, "Convolutive transfer function invariant SDR training criteria for multi-channel reverberant speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 8428–8432.

[6] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*. Hoboken, NJ, USA: Wiley, 2018.

[7] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 70–79, Jan. 2007.

[8] R. Scheibler and M. Togami, "Surrogate source model learning for determined source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 176–180.

[9] K. Sekiguchi, Y. Bando, A. A. Nugraha, M. Fontaine, and K. Yoshii, "Autoregressive fast multichannel nonnegative matrix factorization for joint blind source separation and dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 511–515.

[10] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1626–1641, Sep. 2016.

[11] D. Kitamura and K. Yatabe, "Consistent independent low-rank matrix analysis for determined blind source separation," *EURASIP J. Adv. Signal Process.*, vol. 2020, no. 1, pp. 1–35, 2020.

[12] T. Nakashima, R. Scheibler, M. Togami, and N. Ono, "Joint dereverberation and separation with iterative source steering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 216–220.

[13] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Hoboken, NJ, USA: Wiley, 2001.

[14] 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 Trans. Trans. Speech, Audio Process.*, vol. 12, no. 5, pp. 530–538, Sep. 2004.

[15] S. U. Wood, J. Rouat, S. Dupont, and G. Pironkov, "Blind speech separation and enhancement with GCC-NMF," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 745–755, Apr. 2017.

[16] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[17] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Global and local simplex representations for multichannel source separation," *IEEE/ACM Trans. Audio, Speech,Lang. Process.*, vol. 28, pp. 914–928, 2020.

[18] M. Delcroix *et al.*, "Improving speaker discrimination of target speech extraction with time-domain speakerbeam," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 691–695.

[19] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Multi-stage speaker extraction with utterance and frame-level reference signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6109–6113.

[20] M. Delcroix, K. Zmolikova, T. Ochiai, K. Kinoshita, and T. Nakatani, "Speaker activity driven neural speech extraction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6099–6103.

[21] K. Žmolíková *et al.*, "Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 4, pp. 800–814, Aug. 2019.

[22] J. Han, X. Zhou, Y. Long, and Y. Li, "Multi-channel target speech extraction with channel decorrelation and target speaker adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6094–6098.

[23] Q. Wang *et al.*, "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proc. Interspeech*, 2019, pp. 2728–2732.

[24] C. Xu, W. Rao, E. S. Chng, and H. Li, "SpEx: Multi-scale time domain speaker extraction network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1370–1384, 2020.

[25] Z. Koldovsky and P. Tichavsky, "Gradient algorithms for complex non-Gaussian independent component/vector extraction, question of convergence," *IEEE Trans. Signal Process.*, vol. 67, no. 4, pp. 1050–1064, Feb. 2019.

[26] R. Ikeshita and T. Nakatani, "Independent vector extraction for fast joint blind source separation and dereverberation," *IEEE Signal Process. Lett.*, vol. 28, pp. 972–976, 2021.

[27] R. Scheibler and N. Ono, "Independent vector analysis with more microphones than sources," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 185–189.

[28] R. Ikeshita, T. Nakatani, and S. Araki, "Overdetermined independent vector analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 591–595.

[29] Y. Liang, S. M. Naqvi, and J. A. Chambers, "Audio video based fast fixed-point independent vector analysis for multisource separation in a room environment," *EURASIP J. Adv. Signal Process.*, vol. 2012, no. 1, 2012, Art. no. 183.

[30] A. Brendel, T. Haubner, and W. Kellermann, "A unified probabilistic view on spatially informed source separation and extraction based on independent vector analysis," *IEEE Trans. Signal Process.*, vol. 68, pp. 3545–3558, 2020.

[31] A. Brendel, T. Haubner, and W. Kellermann, "Spatially guided independent vector analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 596–600.

[32] F. Nesta, S. Mosayyebpour, Z. Koldovsky, and K. Palecek, "Audio/video supervised independent vector analysis through multimodal pilot dependent components," in *Proc. 25th Eur. Signal Process. Conf.*, 2017, pp. 1150–1164.

[33] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5329–5333.

[34] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 4930–4934.

[35] J. Jansky, J. Malek, J. Cmejla, T. Kounovsky, Z. Koldovsky, and J. Zdansky, "Adaptive blind audio source extraction supervised by dominant speaker identification using X-vectors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 676–680.

[36] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 4052–4056.

[37] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 5115–5119.

[38] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 3214–3218.

[39] S. Zhang, C. Liu, H. Jiang, S. Wei, L. Dai, and Y. Hu, "Feedforward sequential memory networks: A new structure to learn long-term dependency," 2015, *arXiv:1512.08301v2*.

[40] J. Janskỳ, Z. Koldovskỳ, J. Málek, T. Kounovskỳ, and J. Čmejla, "Auxiliary function-based algorithm for blind extraction of a moving speaker," *EURASIP J. Audio, Speech,Music Process.*, vol. 2022, no. 1, pp. 1–16, 2022.

[41] V. Kautský, Z. Koldovský, P. Tichavský, and V. Zarzoso, "Cramér-Rao bounds for complex-valued independent component extraction: Determined and piecewise determined mixing models," *IEEE Trans. Signal Process.*, vol. 68, pp. 5230–5243, 2020.

[42] J. Malek, J. Jansky, T. Kounovsky, Z. Koldovsky, and J. Zdansky, "Blind extraction of moving audio source in a challenging environment supported by speaker identification via x-vectors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 226–230.

[43] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "The 4th CHiME speech separation and recognition challenge," Accessed on: Sep. 9, 2021. [Online]. Available: http://spandh.dcs.shef.ac.uk/chime_challenge/chime2016/

[44] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2011, pp. 189–192.

[45] S. Ioffe, "Probabilistic linear discriminant analysis," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 531–542.

[46] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," 2017, *arXiv:1706.08612v2*.

[47] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.

[48] "DCASE 2018 challenge," Accessed on: Sep. 9, 2021. [Online]. Available: http://dcase.community/challenge2018/index

[49] E. A. Habets, "Room impulse response generator," vol. 2, Technische Universiteit Eindhoven, Eindhoven, Netherlands, Tech. Rep. 2.4, 2006.

[50] I. Jolliffe, "Principal component analysis," in *Encyclopedia of Statistics in Behavioral Science*. New York, NY, USA: Springer, 2005.

[51] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc.* 2010, pp. 3110–3113.

[52] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[53] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, vol. 2, pp. 749–752.

[54] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[55] G. Andersen, "NST norwegian ASR database," Accessed on: Sep. 29, 2021. [Online]. Available: https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-13/

[56] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2011–2022, Sep. 2007.

[57] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 196–200.

[58] J. Heymann, L. Drude, and R. Haeb-Umbach, "Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition," in *Proc. 4th Int. Workshop Speech Process. Everyday Environ.*, 2016, Art. no. 79.

[59] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Comput. Speech Lang.*, vol. 46, pp. 535–557, Nov. 2017.

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

Jiri Malek (Member, IEEE) was born in Czechia in 1983. He received the Ph.D. degree in technical cybernetics from the Technical University of Liberec, Liberec, Czechia, in 2011. Since 2011, he has been an Assistant Professor with the Faculty of Mechatronics, Technical University of Liberec, Liberec, Czechia. His main research interests include enhancement/separation of audio signals and robust automatic speech recognition. He is a Reviewer for journals and conferences focused on digital signal processing, including the IEEE Transactions on Audio, Speech And Language Processing, *IET Signal Processing,* or ICASSP.

Jakub Jansky was born in Czechia, in 1989. He received the M.S. degree in application of software engineering from the Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Prague, Czechia, in 2014. Since 2014, he has been working toward the Ph.D. degree with the Faculty of Mechatronics, Technical University of Liberec, Liberec, Czechiam. Since 2014, he has been also a Research Assistant with the Faculty of Mechatronics, Technical University of Liberec. His main research interests include blind source separation, independent vector analysis, and sparse reconstruction.

Zbynek Koldovsky (Senior Member, IEEE) received the M.S. and Ph.D. degrees in mathematical modeling from the Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University, Prague, Czech Republic, in 2002 and 2006, respectively. Since 2020, he has been a Full Professor with the Institute of Information Technology and Electronics,Technical University of Liberec, Liberec, Czech Republic, and the Leader of Acoustic Signal Analysis and Processing Group. He is currently the Associated Dean for Science, Research and Doctoral Studies with the Faculty of Mechatronics, Informatics and Interdisciplinary Studies. His main research interests inlclude blind source separation based on advanced mixing models applied in independent component/vector analysis and extraction. He was the General Co-Chair of the 12th Conference on Latent Variable Analysis and Signal Separation, Liberec, Czech Republic, and as a Technical Co-Chair of the 16th International Workshop on Acoustic Signal Enhancement, Tokyo, Japan. Since 2019, he has been a Member of the IEEE SPS Committee Audio and Acoustic Signal Processing. He was the Area Chair for the area of Analysis of Speech and Audio Signals of Interspeech 2021 and 2022.

Tomas Kounovsky was born in Czechia, in 1991. He received the M.S. degree in Information technology in 2016 from the Faculty of Mechatronics, Informatics, and Interdisciplinary Studies, Technical University of Liberec, Liberec, Czechia, where he is has been working toward the Ph.D. degree since 2016. He is a Member of the Acoustic Signal Analysis and Processing Group led by Prof. Zbynek Koldovsky. His research interests include audio signal processing, mainly speech enhancement, and source separation.

Jaroslav Cmejla received the M.S. degree in information technology in 2016 from the Faculty of Mechatronics, Informatics and Interdisciplinary Studies, Technical University of Liberec, Liberec, Czechia, where he has been working toward the Ph.D. degree. He is a Member of the Acoustic Signal Analysis and Processing Group led by Prof. Zbynek Koldovsky. His research interests include audio signal processing and blind source separation. His current works are related to the blind source extraction problem.

Jindrich Zdansky was born in Ceska Lipa, Czechia, in 1978. He received the M.S. degree in applied electronics from the Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czechia, in 2002, and the Ph.D. degree in applied cybernetics from the Institute of Information Technology and Electronics, Technical University of Liberec, Liberec, Czechia 2006. Since 2005, he has been a Member of the Speech Processing Group with the Technical University of Liberec, Liberec, Czechia. His main research interests include audio signal processing, voice-to-text, and speaker diarization technologies.