BLIND EXTRACTION OF MOVING AUDIO SOURCE IN A CHALLENGING ENVIRONMENT SUPPORTED BY SPEAKER IDENTIFICATION VIA X-VECTORS

Jiri Malek, Jakub Jansky, Tomas Kounovsky, Zbynek Koldovsky and Jindrich Zdansky

Acoustic Signal Analysis and Processing Group
Faculty of Mechatronics, Informatics, and Interdisciplinary Studies, Technical University of Liberec
Studentská 2, 461 17 Liberec, Czech Republic

ABSTRACT

We propose a novel approach for semi-supervised extraction of a moving audio source of interest (SOI) applicable in reverberant and noisy environments. The blind part of the method is based on independent vector extraction (IVE) and uses the recently proposed constant separating vector (CSV) mixing model. This model allows for changes of mixing parameters within the processed interval of the mixture, which potentially leads to higher accuracy of SOI estimation. The supervised part of the method concerns a pilot signal, which is related to the SOI and ensures the convergence of the blind method towards the SOI. The pilot is based on robust detection of frames where SOI is dominant via speaker embeddings called X-vectors. Robustness of the detection is achieved through augmentation of the data for the supervised training of the X-vectors. The pilot-supported extraction yields significantly better performance compared to its unsupervised counterpart identifying SOI solely using the initialization.

Index Terms— Independent vector extraction, moving sources, block-wise processing, x-vector, speaker identification.

1. INTRODUCTION

In speech enhancement scenarios, a source of interest (SOI) should be recovered from a mixture of other sources and environmental noise. Without any prior information about mixing conditions, the task can be solved in an unsupervised manner using Blind Source Extraction (BSE) applied in the frequency-domain. Independent vector extraction (IVE) is a BSE approach assuming that signals of different sources are mutually independent while there exist higher-order dependencies between frequency components corresponding to the same source. Conventional IVE model, suitable for separation of static sources, is time invariant. A gradient descent algorithm for IVE was proposed in [1], fast converging methods based on optimization of auxiliary function were proposed in [2, 3].

In practice, the extraction should be time-varying due to movements of the sources. Conventionally, time invariant methods are consecutively applied to short intervals with approximately static sources, and their parameters are recursively updated. The drawback of this *block-wise approach* lies in difficult tuning of the interval length or the recursion weight. An adaptive fast converging IVE algorithm for simple acoustic conditions was proposed in [4]. Recently, an alternative approach was proposed based on the constant separating vector (CSV) model [5], which allows for changes of mixing parameters within the processed block of data. The CSV-based method can thus operate with longer intervals and achieves, in theory [6], higher accuracy compared the block-wise approach.

With no prior information identifying the SOI, the BSE extracts an arbitrary source. The identification can be provided via suitable initialization [7] or by a geometric constraint [8] focusing the extraction towards the desired direction. Alternatively, a pilot signal can be introduced, which is related to the SOI and controls the convergence. A pilot based on the constrained location of the SOI has been exploited in [9], additional audio measurements in the area of SOI were used in [10]. Acquisition of the above described types of prior information is, however, challenging in practice.

Useful alternatives provide pilots based on machine learning principles, which do not require any additional assumptions or measurements. Piloting using voice activity detection was proposed in [11] for mixtures of a single speaker and background noise. For mixtures of multiple speakers, detection of SOI dominance relying on X-vectors [12] was presented in [4].

X-vectors are Deep Neural Network-based (DNN) features extracted from an utterance, which aim to encode the characteristics of an active speaker. When utilized in speaker identification, it is usually assumed that the analyzed signal contains single speaker only. However, it was shown in [4] that, when two speakers are simultaneously active, the dominant one is identified reliably. This phenomenon can be used to control the IVE convergence towards SOI.

The current paper extends the work [4] concerning the extraction of a moving SOI. Two contributions are discussed. 1) The extraction is performed using BSE method based on the novel CSV mixing model. Its applicability to longer mixture intervals results into more accurate SOI extraction compared to the block-wise approach from [4]. 2) The computation of reverberation/noise-robust X-vectors is discussed. The robustness is achieved through the augmentation of the training data for the X-vector network. This is shown beneficial for the detection of intervals with dominant SOI within multi-speaker mixtures and for the computation of X-vector-based pilot. The combination of both contributions allows the application of the proposed approach in scenarios, where SOI is moving in significantly reverberant and noisy environment.

2. BLIND SOURCE EXTRACTION

A time varying mixture of d original signals observed by d microphones can be approximated in the short-time frequency domain by the mixing model

$$\mathbf{x}_{k,\ell} = \mathbf{A}_{k,\ell}\mathbf{y}_{k,\ell}, \qquad (1)$$

This work was supported by the Technology Agency of the Czech Republic (Project No. TH03010018), by The Czech Science Foundation (Project No. 20-17720S), by the department of the Navy, Office of Naval Research Global, through Project No. N62909-19-1-2105, and by the Student Grant Scheme 2020 project of the Technical University in Liberec.

where $k = 1, \dots, K$ denotes frequency, $\ell = 1, \dots, L$ denotes frame and $\mathbf{y}_{k,\ell}$ and $\mathbf{x}_{k,\ell}$ denote the original and mixed signals, respectively. In independent component analysis (IVA), a de-mixing matrix $\mathbf{W}_{k,\ell}$ is sought such that it fulfills $\mathbf{W}_{k,\ell}\mathbf{x}_{k,\ell} = \mathbf{W}_{k,\ell}\mathbf{A}_{k,\ell}\mathbf{y}_{k,\ell} = \hat{\mathbf{y}}_{k,\ell} \approx \mathbf{y}_{k,\ell}$.

In IVE, only one row of $\mathbf{W}_{k,\ell}$ is sought such that it extracts the SOI from the mixture. Without any loss on generality, let the SOI be the first signal in $\mathbf{y}_{k,\ell}$ and $\mathbf{A}_{k,\ell}$ be partitioned as $\mathbf{A}_{k,\ell} = \begin{bmatrix}\mathbf{a}_{k,\ell} & \mathbf{Q}_{k,\ell}\end{bmatrix}$. Then, (1) can be expressed in the form

$$\mathbf{x}_{k,\ell} = [\mathbf{a}_{k,\ell} \quad \mathbf{Q}_{k,\ell}] \begin{bmatrix} s_{k,\ell} \\ \mathbf{z}_{k,\ell} \end{bmatrix}, \tag{2}$$

where $s_{k,\ell}$ represents the SOI, $\mathbf{z}_{k,\ell}$ are the other $d-1$ signals in the mixture. Using the parameterization from [13], $\mathbf{W}_{k,\ell}$ can be partitioned as $\mathbf{W}_{k,\ell} = [\mathbf{w}_{k,\ell} \quad \mathbf{B}_{k,\ell}^H]^H$, where $\mathbf{w}_{k,\ell}$ is a vector extracting the SOI, $\mathbf{B}_{k,\ell}$ is called a blocking matrix which satisfies $\mathbf{B}_{k,\ell}\mathbf{a}_{k,\ell} = \mathbf{0}$, and $\cdot^H$ denotes conjugate transpose.

The estimation of $\mathbf{w}_{k,\ell}$ for each frame ℓ is not accurate when there is lack of available samples. Thus, it is often assumed that the mixing is approximately static over a small number of subsequent frames. The mixture is divided into $t = 1, \dots, T$ blocks of L_t frames, with a block-constant separating vector $\mathbf{w}_{k,t}$; this approach will be referred to as *block-wise* [4].

2.1. CSV AuxIVE algorithm

The following text describes an algorithm based on the CSV mixing model, which was derived in [5] and which constitutes the blind part of our proposed extraction method. In the CSV model, further reduction of estimable parameters is done by assuming that the separating vector $\mathbf{w}_{k,t}$ is constant over all T blocks ($\mathbf{w}_{k,t} = \mathbf{w}_k, t = 1 \dots T$). This means that the separating vector must obey condition $\mathbf{w}_k^H\mathbf{x}_{k,\ell_t} = \hat{s}_{k,\ell_t} \approx s_{k,\ell_t}$ for each block t, where ℓ_t are frame indices corresponding to the tth block.

Similarly to IVE, such separating vector $\mathbf{w}_k$ is sought which maximizes the independence between SOI $\mathbf{s}$ and the other signals $\mathbf{z}$. Based on the maximum-likelihood approach, a contrast function is derived as described in [5]. To find its optimum points, the auxiliary function technique is applied in a similar way as in [14]. This leads to the following update rules:

$$r_{\ell_t} = \sqrt{\sum_{k=1}^{K} |\mathbf{w}_k^H \mathbf{x}_{k,\ell_t}|^2} \qquad \text{for all } \ell_t, \tag{3}$$

$$\mathbf{V}_{k,t} = \mathrm{E}_t\left[\varphi(r_{\ell_t})\mathbf{x}_{k,\ell_t}\mathbf{x}_{k,\ell_t}^H\right], \tag{4}$$

$$\mathbf{a}_{k,t} = \frac{\widehat{\mathbf{C}}_{k,t}\mathbf{w}_k}{\mathbf{w}_k^H\widehat{\mathbf{C}}_{k,t}\mathbf{w}_k}, \tag{5}$$

$$\hat{\sigma}_{k,t} = \sqrt{\mathbf{w}_k^H\widehat{\mathbf{C}}_{k,t}\mathbf{w}_k} \tag{6}$$

$$\mathbf{w}_k = \left(\sum_{t=1}^{T} \frac{\mathbf{V}_{k,t}}{\hat{\sigma}_{k,t}^2}\right)^{-1} \sum_{t=1}^{T} \frac{\mathbf{w}_k^H\mathbf{V}_{k,t}\mathbf{w}_k}{\hat{\sigma}_{k,t}^2}\mathbf{a}_{k,t}, \tag{7}$$

where r_{ℓ_t} and $\mathbf{V}_{k,t}$ are auxiliary variables, $\varphi(\cdot)$ is a suitable nonlinear function [14], E_t denotes the expectation operator over the frames in block t, $\widehat{\mathbf{C}}_{k,t}$ is sample estimate of covariance matrix of mixture for the kth frequency and the tth block. The equation (5) realizes the orthogonal constraint, and $\hat{\sigma}_{k,t}$ is the variance estimate of the SOI on the tth block and kth frequency. To ensure the stability of the convergence, the normalization after each update of the separation vector $\mathbf{w}_k \leftarrow \mathbf{w}_k / \sqrt{\sum_{t=1}^{T} \mathbf{w}_k^H \mathbf{V}_{k,t}\mathbf{w}_k}$ is proposed. Note that for $T = 1$ the algorithm coincides with the auxiliary function-based IVE for static sources from [2].

To ensure the extraction of the SOI, we propose to employ a pilot component as in [15]. Let $\mathbf{P} = [P_1, \dots, P_{\ell_t}, \dots, P_L]$ be a pilot signal that is SOI-dependent and independent with the other sources in the mixture. $\mathbf{P}$ is independent of the mixing model parameters and thus does not change the analytic learning rules of AuxIVE up to the difference that the non-linearity $\varphi(\cdot)$ depends on $\mathbf{P}$. Consequently, the step given by (3), for the ℓ_tth frame, is modified to

$$r_{\ell_t} = \sqrt{\sum\nolimits_{k=1}^{K} |\mathbf{w}_k^H \mathbf{x}_{k,\ell_t}|^2 + P_{\ell_t}}, \tag{8}$$

The rest of the algorithm remains unchanged.

3. PILOTING USING FRAMES WITH DOMINANT SOI

A pilot signal related to SOI guides the convergence of IVE towards the desired source. The pilot requires knowledge of frames, where the SOI is dominant. These frames are determined using speaker identification focused on SOI via pretrained features called X-vectors.

3.1. X-vector DNN

The implementation of the X-vector DNN, described in Table 1, comes from [16]. The DNN is trained to classify N speakers and possibly a non-speech class. Its input consists of a single-channel audio signal sampled at 16 kHz. The input features are 40 filter bank coefficients computed from frames of length of 400 and frame-shift 200 samples. The TDNN (time-delayed DNN) layers introduced in [17] operate on frames with a temporal context centered on the current frame ℓ. The TDNN layers build on top of the context of the earlier layers, thus the final context is a sum of the partial ones.

Our implementation of TDNN contains the following differences compared to [16]: 1) Longer context is used without any frame sub-sampling. 2) The omission of sub-sampling increases the number of trainable parameters. To reduce it, all frames in the context are weighted by a trainable matrix at the input of each TDNN layer and mean time-pooling is performed. 3) The pooling layer computes only variances of frames (means are omitted), the context length is $L_c = 101$ during training. 4) The rectified linear units at the output of layers are replaced by exponential linear units (ELU), which speeds up convergence in our case.

3.2. Training datasets and their augmentation

The training data for the TDNN originate from the development part of the Voxceleb database [18] and the training part of the LibriSpeech corpus [19]. The Voxceleb utterances contain real-world reverberation and noise. Librispeech (part train-360-clean) is free of distortions and is subjected to augmentations discussed further.

Environmental noise was taken from the simulated part of the CHiME-4 training dataset [20] and the development dataset available in Task 1 of the DCASE2018 challenge [21]. This data were also added to the training set without speech, in order to create samples for the *non-speech class*.

Two variants of the X-vector TDNN were trained, differing by the datasets included within the training set. 1) *Baseline X-vectors* were trained on one instance of Voxceleb and unmodified instance

Table 1: Description of the DNN producing the X-vectors. The input size for the TDNN layers is stated after the mean pooling operation.

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

of Librispeech. 2) *Augmented X-vectors* were trained on one unchanged instance of Voxceleb/Librispeech and three augmented instances of the Librispeech dataset, where the following augmentations were applied: a) Reverberation: The utterances are convolved with artificial room impulse responses (RIRs) generated by [22]. The artificial RIRs originate from a shoe-box room of size $8 \times 7 \times 3$ m. We generate RIRs corresponding to four different rooms with T_{60} ranging from $175 - 650$ ms. The source-microphone distance is $1 - 2$ m. b) Noise: The environmental noise was summed with the original Librispeech utterances at signal-to-noise-ratio (SNR) equal to 10 dB. c) Reverberation+noise: The environmental noise was added to the reverberated Librispeech dataset with SNR$=$ 10 dB.

3.3. Speaker identification in cross-talk using X-vectors

During the test phase, the X-vectors are extracted at the output of the pooling layer; pooling context is small ($L_c = 11$) in order to obtain time-localized speaker info. The sequence of X-vectors describing the time-dependent activity of speakers within the mixture is obtained by shifting the input context of the TDNN by a single frame at the time.

The speaker identification is performed via Probabilistic Linear Discriminant Analysis (PLDA, [23]). It requires short utterances of the considered speakers; these are used to compute reference X-vectors called enrollment set. Using a pretrained PLDA model, a hypothesis is tested whether the current X-vector computed from a mixture was produced by speakers in the enrollment set. The result is a PLDA score for each of the enrollment speakers, the speaker with the highest score is assumed dominant (see [4] for detailed discussion and a case study).

The PLDA model for the Baseline X-vectors was trained using the original Librispeech data; PLDA for the Augmented X-vectors was trained using three (augmented) instances of the Librispeech described above (original, reverberation, reverberation + noise).

3.4. Pilot signal

We propose to construct the pilot signal $\mathbf{P}$ for the ℓ_tth frame as

$$P_{\ell_t} = \begin{cases} \sum_{k=1}^{K} |x^1_{k,\ell_t}|^2 & \frac{M(s_{k,\ell_t})}{M(\mathbf{z}_{k,\ell_t})} \geq \eta, \\ 0 & \text{otherwise}, \end{cases} \tag{9}$$

where x^1_{k,ℓ_t} is the signal on the first microphone, s_{k,ℓ_t} is the SOI, $\mathbf{z}_{k,\ell_t}$ represents the rest of the considered speech sources, $M(\cdot)$ is a function estimating activity of the source and η is a free threshold parameter. This manuscript presents two pilot variants: 1) The oracle $\mathbf{P}^{\text{ORAC}}$, where $M(\cdot)$ represents the sum of energy of the respective true sources, 2) The X-vector based $\mathbf{P}^{\text{XVEC}}$ where $M(\cdot)$ represents the maximum PLDA score of the respective sources within the enrollment set.

4. EXPERIMENTAL EVALUATION

The experiments simulate mixtures of two active speech sources (moving SOI and static interfering source (IS)) and static directional noise. We consider room in Fig. 1 of dimensions $6 \times 6 \times 3$ m and three reverberation times $T_{60} \in \{100, 300, 600\}$ ms. 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°. The SOI is moving around the array at speed 40 cm/s on a quarter circle with diameter 1.5 m. IS is placed behind the trajectory of SOI at coordinates $(3, 4.74)$. This is a difficult scenario for a pure BSE, SOI and IS can be interchanged due to their close proximity (see permutation problem in [4]). The directional noise is situated at $(4.41, 4.16)$. The static sources are 2 m distant from the array.

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 noise originates from the QUT corpus [24]. The enrollment set consists of 1 minute of speech for each considered speaker, augmented by reverberation as in Section 3.2. Different utterances are concatenated to form 5 unique test signals of length 25 s for each speaker. The SOI movements and positions of the static sources are simulated using the RIR generator [22]. The input signal-to-interference-ratio (SIR, ratio of energy of SOI and IS) is 0 dB and the input signal-to-noise-ratio (SNR, ratio of all speech energy to noise energy) is 10 dB. 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 extraction is evaluated in terms of the improvement of SIR (iSIR, all undesired sources are included in the interference term) and signal-to-distortion-ratio (SDR) as defined in BSS_EVAL [25]. We also provide the improvement of PESQ score [26] (iPESQ) and the standard deviation of "SOI Attenuation" defined as $\sum_k |\hat{s}_{k,\ell}|^2 / \sum_k |s_{k,\ell}|^2$, where $\hat{s}_{k,\ell}$ is the estimate of $s_{k,\ell}$. For a well extracted SOI, this criterion should be close to zero. When the SOI moves out of the focus of the algorithm (is vanishing within the extracted signal), the deviation increases.

4.1. Accuracy of the SOI dominance detection

The pilot $\mathbf{P}^{\text{XVEC}}$ is based on the information whether SOI is currently dominant in the mixture or not. To measure the accuracy of such classification, we compare the frames where SOI has the highest PLDA score with frames where SOI has the highest energy. The context of TDNN and the energy-based reference is $L_c = 11$.

Fig. 2 shows the classification accuracy achieved in reverberant environment ($T_{60} = 600$ ms) using Baseline (XVEC$_B$) and Augmented (XVEC$_A$) X-vectors. Each point in the graph corresponds to the averaged accuracy over all mixtures in one instance of the experiment described in Section 4. Various points/instances differ by input SIR $\in \{-5, 0, 5, 10, 20\}$ dB and input SNR $\in \{0, 10, \infty\}$ dB.

The highest accuracy is achieved when SIR is low ($<$ 5 dB), i.e., when the SOI is mostly not dominant. For a higher SIR ($>$ 5 dB), the SOI can be easily recognized as being dominant, however, only when the SNR is high enough. For low SNR scenarios, the classification becomes challenging. In all SNR and SIR settings, the

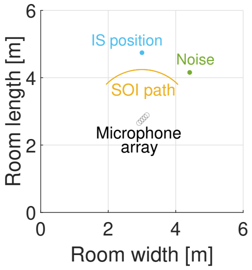

Fig. 1: Setup of the simulated room scenario

Table 2: The extraction performance for the CSV AuxIVE (CSV$_{L_t}$) and Block-Online AuxIVE (BO$_{L_t}$) techniques. The subscript L_t denotes the number of frames within the analyzed block. The results are averaged over all respective mixtures with specific T_{60} level. Methods without pilot identify SOI solely using the initialization.

Method	**Pilot**	**iPESQ**			**SDR [dB]**			**iSIR [dB]**			**SOI Attenuation**		
		100ms	300ms	600ms	100ms	300ms	600ms	100ms	300ms	600ms	100ms	300ms	600ms
CSV$_{200}$	-	0.28	0.07	0.01	5.91	0.45	-2.03	12.64	5.32	3.30	0.37	0.14	0.09
CSV$_{200}$	$\mathbf{P}^{\text{ORAC}}$	0.75	0.31	0.17	10.93	5.31	1.87	20.67	14.00	10.53	0.27	0.12	0.09
CSV$_{200}$	$\mathbf{P}^{\text{XVEC}_A}$	0.36	0.18	0.04	7.13	3.04	-0.96	14.72	10.14	5.29	0.34	0.14	0.09
CSV$_{200}$	$\mathbf{P}^{\text{XVEC}_B}$	0.39	0.21	0.05	7.45	3.59	-0.47	13.40	9.27	4.25	0.34	0.13	0.09
BO$_{200}$	-	0.13	0.01	-0.04	4.38	0.34	-1.73	11.66	6.33	4.51	0.34	0.19	0.14
BO$_{200}$	$\mathbf{P}^{\text{ORAC}}$	0.35	0.15	0.06	8.65	4.74	1.85	17.03	12.96	10.25	0.29	0.17	0.14
BO$_{200}$	$\mathbf{P}^{\text{XVEC}_A}$	0.13	0.03	-0.04	4.37	1.32	-1.41	11.60	8.03	5.11	0.33	0.18	0.14
BO$_{200}$	$\mathbf{P}^{\text{XVEC}_B}$	0.12	0.04	-0.04	4.17	1.72	-1.15	9.29	6.42	3.32	0.33	0.17	0.13
CSV$_{50}$	$\mathbf{P}^{\text{ORAC}}$	0.57	0.21	0.15	11.00	4.49	1.52	18.09	10.99	10.70	0.19	0.13	0.11
BO$_{50}$	$\mathbf{P}^{\text{ORAC}}$	0.12	0.04	-0.03	6.83	3.70	1.15	14.38	11.59	9.15	0.19	0.14	0.11
CSV$_{800}$	$\mathbf{P}^{\text{ORAC}}$	0.57	0.27	0.15	8.02	4.52	1.52	18.55	13.76	10.70	0.51	0.14	0.11
BO$_{800}$	$\mathbf{P}^{\text{ORAC}}$	0.44	0.19	0.10	6.99	4.20	1.45	17.23	12.93	10.16	0.45	0.17	0.14

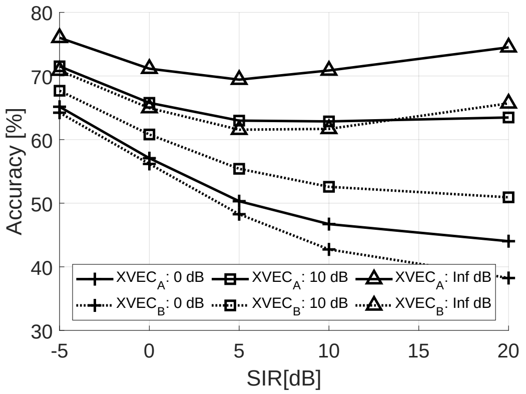

Fig. 2: Accuracy for the Baseline (XVEC$_B$) and Augmented (XVEC$_A$) X-vector variants in the task of the SOI dominance detection; each line corresponds to different SNR.

classification based on XVEC$_A$ significantly outperforms the one based on XVEC$_B$.

4.2. Extraction of SOI

Here, the extraction performed by the proposed CSV AuxIVE (abbreviated by CSV) and the Block-Online AuxIVE (BO) from [4] are compared. CSV processes each mixture as one batch, the number of iterations is set to 50. BO analyzes the recording in a block-wise manner, performing 5 iterations in each block. The methods are initialized by the location of the SOI at the beginning of the recording; BO initializes the extraction at each block by the solution from the previous one. The NFFT length is 1024 with shift 200 samples; BO blocks have 75% overlap. The threshold $\eta = 2$ for $\mathbf{P}^{\text{ORAC}}$ and $\eta = 1$ for $\mathbf{P}^{\text{XVEC}}$; the suitable values are determined based on preliminary experiments. The thresholds are distinct, because both pilots stem from different underlying principles (signal energy for $\mathbf{P}^{\text{ORAC}}$ and PLDA score for $\mathbf{P}^{\text{XVEC}}$).

All criteria in Table 2 indicate that the CSV achieves more precise extraction compared to its block-wise counterpart, especially for $T_{60} = 100$ ms. All pilot-supported methods are more successful in extraction of SOI compared to methods relying solely on initialization (without any pilot). This indicates that the piloted extractors are robust to permutation problem, when sources are shortly in cover. The performance of $\mathbf{P}^{\text{XVEC}}$ is lower than that of $\mathbf{P}^{\text{ORAC}}$. Comparing the two variants of X-vector-based pilots, methods endowed with the augmented XVEC$_A$ yield higher iSIR (by $1-2$ dB); the other criteria are comparable. To summarize, the currently proposed method (CSV$_{200}$ with $\mathbf{P}^{\text{XVEC}_A}$) outperforms the previously presented approach [4] (BO$_{200}$ with $\mathbf{P}^{\text{XVEC}_B}$) for all T_{60} levels and almost all criteria (both methods achieve comparable SOI Attenuation).

Let us focus on the selection of the block-length and restrict the discussion to the methods using the most accurate $\mathbf{P}^{\text{ORAC}}$. Methods using longer block (800 frames) yield high values of iSIR (IS is well suppressed) but low SDR and increased Attenuation; the techniques are slow to adapt to source movements and the SOI moves out of their spatial focus. Observing the criteria, this phenomenon is clearly noticeable for $T_{60} = 100$ ms. However, it is to a certain degree present in the extracted signals for all T_{60} levels, but the considered criteria do not reflect it well [1]. Methods using short blocks (50 frames) are able to adapt well to SOI movements (low SOI Attenuation), but achieve lower values in other criteria due to small quantity of samples for estimation.

5. CONCLUSION

A blind algorithm for robust extraction of a moving audio source of interest (SOI) was proposed. The extractor utilized the novel *constant separating vector* mixing model (CSV AuxIVE), which allows, in theory, more accurate estimation of moving SOI compared to conventional block-wise approach. A variant of X-vector speaker embeddings robust to environmental conditions was presented and used to guide the extraction towards SOI.

The joint utilization of both principles resulted in an extraction algorithm applicable to mixtures of moving sources originating in challenging acoustic conditions. In the presented experiments, the piloted CSV AuxIVE was applied in a batch manner to mixtures of length 25 s and still, due to its mixing model, was able to adapt to the SOI movements better than its block-wise counterpart.

[1] To demonstrate, samples of the extracted signals are available at `https://asap.ite.tul.cz/demos/`

6. REFERENCES

[1] Zbynek Koldovsky, Petr Tichavsky, and Vaclav Kautsky, “Orthogonally constrained independent component extraction: Blind MPDR beamforming,” Sept. 2017, pp. 1195–1199.

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

[3] Rintaro Ikeshita, Tomohiro Nakatani, and Shoko Araki, “Overdetermined independent vector analysis,” in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*. IEEE, 2020, pp. 591–595.

[4] Jakub Jansky, Jiri Malek, Jaroslav Cmejla, Tomas Kounovsky, Zbynek Koldovsky, and Jindrich Zdansky, “Adaptive blind audio source extraction supervised by dominant speaker identification using X-vectors,” in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*. IEEE, 2020, pp. 676–680.

[5] Jakub Jansky, Zbynek Koldovsky, Jiri Malek, Tomas Kounovsky, and Jaroslav Cmejla, “Auxiliary function-based algorithm for blind extraction of a moving speaker,” *arXiv preprint arXiv:2002.12619v2*, 2020.

[6] Vaclav Kautsky, Zbynek Koldovsky, Petr Tichavsky, and Vicente Zarzoso, “Cramer-rao bounds for complex-valued independent component extraction: Determined and piecewise determined mixing models,” *arXiv preprint arXiv:1907.08790*, 2019.

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

[8] Andreas Brendel, Thomas Haubner, and Walter Kellermann, “A unified probabilistic view on spatially informed source separation and extraction based on independent vector analysis,” *IEEE Transactions on Signal Processing*, vol. 68, pp. 3545–3558, 2020.

[9] Francesco Nesta and Zbynek Koldovsky, “Supervised independent vector analysis through pilot dependent components,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 536–540.

[10] Jaroslav Cmejla, Tomas Kounovsky, Jiri Malek, and Zbynek Koldovsky, “Independent vector analysis exploiting prelearned banks of relative transfer functions for assumed target’s positions,” in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2018, pp. 270–279.

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

[12] Daniel Garcia-Romero, David Snyder, Gregory Sell, Daniel Povey, and Alan McCree, “Speaker diarization using deep neural network embeddings,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4930–4934.

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

[14] Nobutaka Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” in *WASPAA 2011*, 2011, pp. 189–192.

[15] F. Nesta and Z. Koldovsky, “Supervised independent vector analysis through pilot dependent components,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, March 2017, pp. 536–540.

[16] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, “X-vectors: Robust DNN embeddings for speaker recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[17] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

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

[19] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “Librispeech: an ASR corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[20] Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer, “The 4th CHiME speech separation and recognition challenge [online],” Accessed: 19.8.2020.

[21] “DCASE 2018 challenge [online],” Accessed: 19.8.2020.

[22] Emanuel AP Habets, “Room impulse response generator,” *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, pp. 1, 2006.

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

[24] David B Dean, Sridha Sridharan, Robert J Vogt, and Michael W Mason, “The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms,” *Proceedings of Interspeech 2010*, 2010.

[25] E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation,” vol. 14, no. 4, pp. 1462–1469, July 2006.

[26] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in *2001 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2001)*. IEEE, 2001, vol. 2, pp. 749–752.