TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií

SKFM 2020
Studentská konference Fakulty mechatroniky,
informatiky a mezioborových studií

Adaptive blind audio source extraction supervised by dominant speaker identification using x-vectors

Jakub Janský, Jiří Málek, Jaroslav Čmejla, Tomáš Kounovský, Zbyněk Koldovský and Jindřich Žďánský

We propose a novel algorithm for adaptive blind audio source extraction. The proposed method is based on independent vector analysis and utilizes the auxiliary function optimization to achieve high convergence speed. The algorithm is partially supervised by a pilot signal related to the source of interest (SOI), which ensures that the method correctly extracts the utterance of the desired speaker. The pilot is based on the identification of a dominant speaker in the mixture using x-vectors. The proposed approach is verified in a scenario with a moving SOI, static interfering speaker and environmental noise.

Key words: Independent vector extraction, adaptive processing, auxiliary function, x-vector, speaker identification

Introduction

Independent Vector Analysis (IVA) performed in the frequency-domain is a popular approach to Blind Source Separation (BSS) of audio sources. It assumes that sources are mutually independent while separation proceeds jointly for all frequency bins by exploiting inter-channel higher-order dependencies.

Typically, only a desired source should be extracted from the mixture, which is the goal of Blind Source Extraction (BSE). A modification of IVA for the BSE problem, relating mixing and de-mixing vectors corresponding to the SOI through the orthogonal constraint (OG), was proposed in [1]. In this work, a novel BSE algorithm referred to as AuxIVE (Auxiliary-function-based Independent Vector Extraction) is derived.

In BSE, there is the so-called global permutation problem that means that a different Interfering Source (IS) can be extracted instead of the SOI when no information about the SOI is available. In order to avoid this problem, a piloted IVA was introduced in [2] where the pilot signal related to the SOI is used for influencing the convergence to the SOI. However, the acquisition of a proper pilot signal poses a challenge. In this work, we utilize a pilot that is based on the identification of a dominant speaker using speaker embeddings based on Deep Neural Network-based (DNN) x-vectors [3]. With a pretrained x-vector DNN, speaker identification is often performed using probabilistic linear discriminant analysis (PLDA). Here, a hypothesis is tested whether the embedding of an unknown speaker is produced by any of known speakers computed from short clean utterances called enrollments.

Adaptive supervised AuxIVE algorithm

In the time-frequency domain, a mixture of d original signals observed by d microphones can be, within the kth frequency bin, approximated by the instantaneous mixing model

$$\mathbf{X}_k = \mathbf{A}_k\mathbf{S}_k \tag{1}$$

where $\mathbf{S}_k$ and $\mathbf{X}_k$ denote the original and mixed signals, respectively. In IVA, we are looking for a de-mixing matrix $\mathbf{W}_k$ that fulfills $\mathbf{W}_k\mathbf{X}_k = \mathbf{W}_k\mathbf{A}_k\mathbf{S}_k \approx \mathbf{S}_k$, which means $\mathbf{W}_k^{-1} \approx \mathbf{A}_k$ up to the global permutation and scales of the separated signals. In IVE, only one row of $\mathbf{W}_k$ is sought such that it extracts the SOI from the mixture. Without any loss on generality, let the SOI be the first signal in $\mathbf{S}_k$ and $\mathbf{A}_k$ be partitioned as $\mathbf{A}_k = [\mathbf{a}_k, \mathbf{Q}_k]$ and $\mathbf{W}_k$ can be partitioned as $\mathbf{W}_k = [(\mathbf{w}_k)^H; \mathbf{B}_k]$, where $\mathbf{w}_k$ is a vector extracting the SOI.

Now, we apply the IVE statistical model from [1] and the auxiliary function technique in a similar way to [4]. The auxiliary contrast function have the form

$$\mathcal{Q}(\{\mathbf{w}_k\}_{k=1}^K, \{\mathbf{a}\}_{k=1}^K, r_\ell) = -\frac{1}{2}\sum_{k=1}^K(\mathbf{w}_k)^H\mathbf{V}_k\mathbf{w}_k \\ -\frac{1}{L}\sum_{k=1}^K\sum_{\ell=1}^L \mathbf{x}_{k,\ell}^H\mathbf{B}_k^H\mathbf{C}_{\mathbf{z}_k}^{-1}\mathbf{B}_k\mathbf{x}_{k,\ell} + \sum_{k=1}^K \log|\det\mathbf{W}_k|^2 + R, \tag{2}$$

where $\mathbf{C}_{\mathbf{z}_k}$ is the covariance of the background signals (modeled as Gaussian),

$$\mathbf{V}_k = \frac{1}{L}\sum_{\ell=1}^L \varphi(r_\ell)\mathbf{x}_{k,\ell}\mathbf{x}_{k,\ell}^H, \tag{3}$$

and r is the auxiliary variable, R is a constant term, and $\varphi(\cdot)$ is related to pdf of the signal according to Theorem 1 from [4].

Considering the sequential (block-by-block) adaptive processing of data blocks with the length of L_b frames with shift L_{shift}, the update rules for the ith block have a form

$$r_{\ell,i} = \sqrt{\sum\nolimits_{k=1}^K |\mathbf{w}_{k,i-1}^H\mathbf{x}_{k,\ell}|^2} \qquad \text{for } \ell = \ell_s, \dots, \ell_e \tag{4}$$

$$\mathbf{V}_{k,i} = \alpha\mathbf{V}_{k,i-1} + (1-\alpha)\frac{1}{L_b}\sum\nolimits_{\ell=\ell_s}^{\ell_e}[\varphi(r_\ell)\mathbf{x}_{k,\ell}\mathbf{x}_{k,\ell}^H], \tag{5}$$

$$\hat{\mathbf{C}}_{k,i} = \alpha\hat{\mathbf{C}}_{k,i-1} + (1-\alpha)\frac{1}{L_b}\sum\nolimits_{\ell=\ell_s}^{\ell_e}\mathbf{x}_{k,\ell}\mathbf{x}_{k,\ell}^H \tag{6}$$

$$\mathbf{a}_{k,i} = \frac{\hat{\mathbf{C}}_{k,i}\mathbf{w}_{k,i-1}}{\mathbf{w}_{k,i-1}^H\hat{\mathbf{C}}_{k,i}\mathbf{w}_{k,i-1}}, \tag{7}$$

$$\mathbf{w}_{k,i} = \mathbf{V}_{k,i}^{-1}\mathbf{a}_{k,i}, \tag{8}$$

where α is a forgetting factor; $\ell_s = (i-1)L_{\text{shift}} + 1$ and $\ell_e = (i-1)L_{\text{shift}} + L_b$ denote the beginning and the end of the ith block, respectively. For the case $L_b = 1$ and $\alpha \in (0, 1\rangle$, we refer to the proposed method as to "Online AuxIVE", and, for $L_b > 1$ and $\alpha = 0$, we call it "Block Online AuxIVE".

To ensure the extraction of the desired source, we propose to employ a pilot component. Let $\mathbf{P}$ be a pilot signal that is SOI-dependent and independent with the other signals in the mixture. Consequently, the step given by (4), for the ith block, is modified to

$$r_{\ell,i} = \sqrt{\sum\nolimits_{k=1}^K |\mathbf{w}_{k,i-1}^H\mathbf{x}_{k,\ell}|^2 + \mathbf{P}_\ell}, \quad \ell = \ell_s, \dots, \ell_e. \tag{9}$$

The rest of the algorithm remains unchanged.

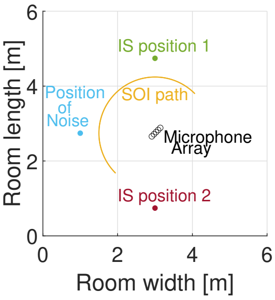

Fig. 1: Setup of the simulated room scenario

TABLE I: Mean iSINR [dB] with standard deviation and percentage of failed extractions (iSINR < 1 dB) for experiments

		Block online AuxIVE			Online AuxIVE		
		Blind		$\mathbf{P}_\ell^{\text{ORAC}}$	Blind	$\mathbf{P}_\ell^{\text{XVEC}}$	$\mathbf{P}_\ell^{\text{ORAC}}$
IS position 1	iSINR [dB]	4.3 ± 3.6	6.4 ± 1.9	10.0 ± 1.7	-0.5 ± 1.8	2.0 ± 1.5	5.0 ± 1.5
	fail cases [%]	24.67	2	0	77.67	23.67	1.67
IS position 2	iSINR [dB]	9.3 ± 1.7	9.6 ± 1.5	12.6 ± 1.8	5.0 ± 1.6	5.9 ± 1.4	8.5 ± 1.5
	fail cases [%]	0	0	0	0.34	0	0
Time per mixture[s]		4.55	11.8	4.65	15.24	24.42	15.32

X-vector computation

Our implementation of the x-vector DNN, described in Table III. Its input consists of a single-channel audio signal, i.e., no spatial information is used. The input features are 40 filter bank coefficients computed from frames of length of 25 ms and frame-shift of 10 ms. The TDNN (time-delayed DNN) layers 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. The DNN was trained to classify N speakers. The training examples consisted of 151 frames of features and the speaker label.

TABLE III: Description of the DNN producing the x-vectors

Layer	Layer context	Total context	Input x output
TDNN 1	$\ell \pm 50$	101	40×512
TDNN 2-6	$\ell \pm 5$	151	512×512
Fully-conn. 1	ℓ	151	512×128
Pooling	$\ell \pm \frac{L_c - 1}{2}$	$\max(151, L_c)$	$(L_c \cdot 128) \times 128$
Fully-conn. 2	ℓ	$\max(151, L_c)$	128×128
Softmax	–	$\max(151, L_c)$	$128 \times N$

The pilot signal needs to be dependent on the SOI, e.g., it should exhibit high values when the SOI is active and vice versa. The highest PLDA score corresponding to the SOI is usually related to the frames where it is dominant.

Experimental evaluation

The experiments simulate speaker movements in a reverberant room, see Fig. 1. We utilized 4 speakers originating from the CHiME-4 database. We obtained 5 unique one minute long test signals for each speaker sampled at 16 kHz. Noise signal consists of one minute of pedestrian area sounds. The experiments are evaluated using Signal-To-Interference-And-Noise ratio improvement (iSINR), where all undesired sources are included in the noise term. Parameters was set for Block online AuxIVE, $L_b = 100$, $L_{\text{shift}} = 75$ and $\alpha = 0$. For Online AuxIVE, $L_b = 1$, $L_{\text{shift}} = 1$ and $\alpha = 0.97$. Both methods used $K = 512$ with frame shift 160 samples. The de-mixing vector updates were computed using single iteration in every block.

We consider 600 simulated mixtures. Each mixture consists of a moving SOI, a fixed IS and noise. The global energy of the noise was 10 dB. The IS was situated either in position 1, where the methods are prone to permutation problem due to close proximity to the SOI path or position 2, which should not have such problem. We consider the SOI extraction as failed, when iSINR < 1 dB. We compare oracle pilot $\mathbf{P}_\ell^{\text{ORAC}}$ with x-vector pilot $\mathbf{P}_\ell^{\text{ORAC}}$.

The results in Table 1 indicate that the blind variants of AuxIVE yield positive iSINR for most of the considered mixtures. The introduction of a pilot improves the extraction especially for cases with IS in position 1. The higher iSINR along with its lower variance and lower fail-rate show that the utilization of pilot successfully prevents from the source permutation.

In comparison to $\mathbf{P}_\ell^{\text{ORAC}}$, the supervision by $\mathbf{P}_\ell^{\text{XVEC}}$ brings lower improvements in average iSINR. Compared to the blind method, $\mathbf{P}_\ell^{\text{XVEC}}$ improves the iSINR in 83% of cases for Block online AuxIVE and in 100% cases for the online version, considering IS in position 1. In the remaining cases, when $\mathbf{P}_\ell^{\text{XVEC}}$ lowers the performance, the deterioration is always lower than 1 dB.

Conclusion

A blind adaptive and fast converging method for BSE was proposed, suitable for SOI extraction in the noisy cross-talk scenario. To avoid extraction of an incorrect source, a pilot based on x-vectors related to SOI activity was presented. For environments with low reverberation, it was shown that the x-vectors can identify the energy-dominant speaker even in the presence of cross-talk.

The future research can be directed towards improvement of the pilot. The x-vector DNN can be adapted to function even for higher reverberation levels (e.g. through augmentation of the training data) or to produce more time-localized estimates of the active speaker.

Thanks

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.

References

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

[2] Francesco Nesta and Zbyněk Koldovský, “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.

[3] 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.

[4] Nobutaka Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2011, pp. 189–192.