TECHNICAL UNIVERSITY OF LIBEREC
Faculty of Mechatronics, Informatics and Interdisciplinary Studies

SKFM 2020
Studentská konference Fakulty mechatroniky, informatiky a mezioborových studií

Online Implementation of Independent Vector Extraction based on Auxiliary function (AuxIVE)

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

Abstract and motivation

- Blind separation of audio sources based on Independent Vector Analysis
- **IVA:** Sources mutually independent while separation proceeds jointly for all frequency bins by exploiting inter-channel higher-order dependencies
- **Speech enhancement:** Only single source of interest (SOI) needs to be retrieved, i.e., independent vector extraction (IVE)[1]
  - **Our contribution:** Modification of IVA for extraction of single source only; auxiliary-function-based optimization ensures high convergence speed of the proposed algorithm
- **Indeterminacy:** Without knowledge about SOI, any source can be extracted
  - **Prior information using pilot:** Additional signal in the cost function related to SOI ensures convergence to the desired source [2]
  - **Our contribution:** Proposal of pilot based on x-vectors; can be utilized in the presence of both non-speech noise and speech interference

Source extraction model

- **IVA model:** for a d signal mixture observed on d microphones for kth frequency bin

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

  - $\mathbf{X}$ denotes mixture, $\mathbf{S}$ denotes original signals and $\mathbf{A}$ is a mixing matrix,
  - Looking for demixing matrix $\mathbf{W}_k \mathbf{X}_k = \mathbf{W}_k \mathbf{A}_k \mathbf{S}_k \approx \mathbf{S}_k$

- **IVE model:**

$$\begin{pmatrix} \mathbf{x}_k^1 \\ \mathbf{x}_{k,}^2 \\ \vdots \\ \mathbf{x}_k^d \end{pmatrix} = [\mathbf{a}_k, \mathbf{Q}_k] \begin{pmatrix} \mathbf{s}_k \\ \mathbf{z}_k^1 \\ \vdots \\ \mathbf{z}_k^{d-1} \end{pmatrix}, \tag{2}$$

  - $\mathbf{s}$ denotes SOI, $\mathbf{z}$ denotes rest of signals and $\mathbf{a}$ is a mixing vector of SOI,
  - Looking only first row of $\mathbf{W}_k = [(\mathbf{w}_k)^H; \mathbf{B}_k]$

- **Contrast function:** by assuming independence of signals we obtain contrast function

$$\mathcal{Q}(\{\mathbf{w}_k\}_{k=1}^K, \{\mathbf{a}\}_{k=1}^K, r_\ell) = -\frac{1}{2}\sum_{k=1}^K (\mathbf{w}_k)^H \mathbf{V}_k \mathbf{w}_k - \frac{1}{L}\sum_{k=1}^K \sum_{\ell=1}^L \mathbf{x}_{k,\ell}^H \mathbf{B}_k^H \mathbf{C}_{\mathbf{z}_k}^{-1} \mathbf{B}_k \mathbf{x}_{k,\ell} + \sum_{k=1}^K \log|\det \mathbf{W}_k|^2 + R, \tag{3}$$

- Where $\mathbf{V}_k = \frac{1}{L}\sum_{\ell=1}^L \varphi(r_\ell)\mathbf{x}_{k,\ell}\mathbf{x}_{k,\ell}^H$ and $r_\ell = \sqrt{\sum_{k=1}^K |(\mathbf{w}_k)^H \mathbf{x}_{k,\ell}|^2}$
- **Update rules:** by solving (3) for $\mathbf{w}$ we obtain

$$\begin{aligned} r_{\ell,i} &= \sqrt{\sum_{k=1}^K |\mathbf{w}_{k,i-1}^H \mathbf{x}_{k,\ell} + \mathbf{P}_\ell|^2} \\ \mathbf{V}_{k,i} &= \alpha \mathbf{V}_{k,i-1} + (1-\alpha)\frac{1}{L_b}\sum_{\ell=\ell_s}^{\ell_e} [\varphi(r_\ell)\mathbf{x}_{k,\ell}\mathbf{x}_{k,\ell}^H], \\ \hat{\mathbf{C}}_{k,i} &= \alpha \hat{\mathbf{C}}_{k,i-1} + (1-\alpha)\frac{1}{L_b}\sum_{\ell=\ell_s}^{\ell_e} \mathbf{x}_{k,\ell}\mathbf{x}_{k,\ell}^H \\ \mathbf{a}_{k,i} &= \frac{\hat{\mathbf{C}}_{k,i}\mathbf{w}_{k,i-1}}{\mathbf{w}_{k,i-1}^H \hat{\mathbf{C}}_{k,i}\mathbf{w}_{k,i-1}}, \\ \mathbf{w}_{k,i} &= \mathbf{V}_{k,i}^{-1}\mathbf{a}_{k,i} \end{aligned}$$

- α is forgetting factor, L_b is block size and $\mathbf{P}$ denotes pilot signal

The x-vector deep neural network (DNN)

- **Input:** Single channel audio, 40 filter bank coefficients
- **Target:** Labels to classify N speakers
- **Topology:** Time-delayed DNN (TDNN) from [3]
  - The TDNN layers originating in [4] operate on context of frames centered on the current frame ℓ
- **Some of our modifications to the TDNN :**
  - All frames are used in the context without sub-sampling
  - Time-pooling at the output of TDNN layers (reduces number of trainable parameters)
- **Speaker identification during cross-talk:**
  - Closed set of small number of speakers
  - Assignment of test recording to enrollment speakers using PLDA
  - Dominant speaker from the perspective of energy is often identified
- **Pilot:** $O(\mathbf{s}_\ell)$, $O(\mathbf{y}_\ell^m)$ - PLDA score of the SOI and other enrolllment speakers

$$\mathbf{P}_\ell^{\text{XVEC}} = \begin{cases} \sum_{k=1}^K |\mathbf{X}_{k,\ell}|^2 & \frac{O(\mathbf{s}_\ell)}{\max(O(\mathbf{y}_\ell^1),\ldots,O(\mathbf{y}_\ell^M))} \geq \eta, \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

Tato prace byla podpořena z projektu Studentksé grantové soutěže (SGS) na Technické univerzitě v Liberci v roce 2020

Table: 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	–	$\max(151, L_c)$	$128 \times N$

Experiments

- **Setup:**
  - Simulated speaker movements in room with $T_{60} = 100$ ms
  - Speech: CHiME-4 simulated test/development,
  - Enrollment: four considered speakers (1 minute of speech): F01, F06, M04, M05
- **Case study: dominant speaker identification**
  - Speech of F01 and M04 summed (F01 energy higher by 2 dB)
  - Highest PLDA score - speaker with the highest energy (accuracy for context $L_c = 151$ is 79.8%, for $L_c = 10$ is 62.4%)
  - Second highest PLDA score does not reflect the other active speaker

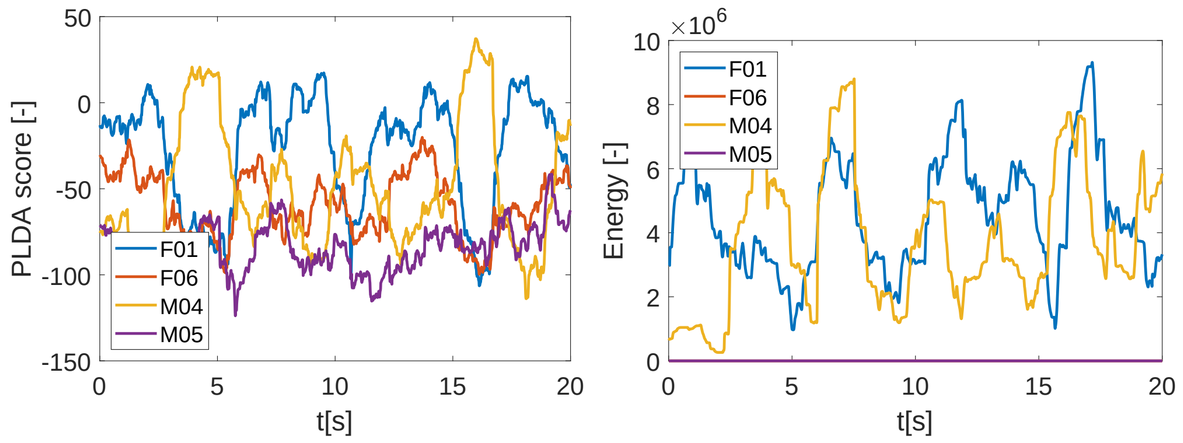

Figure: PLDA score and the corresponding utterance energy

- **Extraction in noisy environment:**
  - 600 mixtures
  - In each mixture:
    - Moving Signal of Interest (SOI) in semicircle (40 cm/s)
    - Interference speech signal (IS)
    - Pedestrian area noise (-10dB)
  - Two possible IS position
  - Evaluation by improvement in Signal-to-Interference-and-Noise-Ratio (SINR) and fail rate when SINR< 1 dB

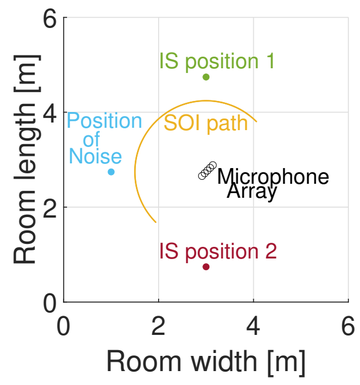

Figure: Room setup

Table: Result of noisy enviroment experiment

		Block online AuxIVE			Online AuxIVE		
		Blind	$\mathbf{P}_\ell^{\text{XVEC}}$	$\mathbf{P}_\ell^{\text{ORAC}}$	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
Average time per mixture [s]		4.55	11.8	4.65	15.24	24.42	15.32

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, 67(4):1050–1064, Feb 2019.

[2] F. Nesta and Z. Koldovský.
Supervised independent vector analysis through pilot dependent components.
In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 536–540, March 2017.

[3] 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)*, pages 5329–5333. IEEE, 2018.

[4] 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.