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

Adaptive Blind Audio Source Extraction Supervised by Dominant Speaker Identification using X-vectors

ICASSP 2020 - virtual conference

Jakub Janský **Jiří Málek**

in cooperation with

Jaroslav Čmejla **Tomáš Kounovský** **Zbyněk Koldovský** **Jindřich Žd'ánský**

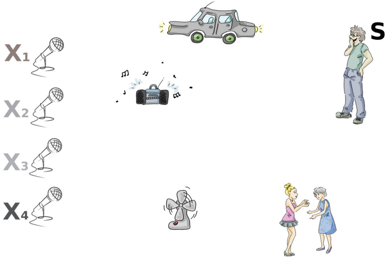
A.S.A.P. Group & Speechlab, Institute of Information Technology and Electronics (ITE)

Part I

Introduction

What do we want to do . . .

- **We have:** a mixture of moving audio sources
- **We want:** to extract the desired source of interest (SOI) with almost no assumptions
    - Blind Source Separation/ Extraction (BSS/BSE)
- **Why:**
    - Speech enhancement of noisy recordings
    - Speech separation during cross-talk
- **In this paper:**
    - Adaptive fast converging algorithm for BSE
    - Piloting towards SOI using dominant speaker identification



Blind source separation/extraction

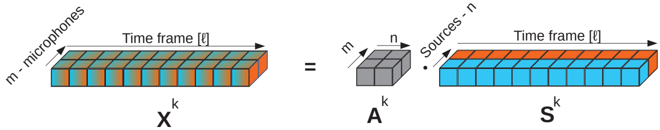
- **BSS:** Attempts to extract all sources contained in the mixture
- **BSE:** Extracts only single desired source from the mixture
- Often based on maximizing the independence of the sources
- BSE methods:
  - Independent component extraction (ICE)
    - Maximize independence between SOI and rest of the sources
    - Requires additional solution to component permutation problem
  - Independent vector extraction (IVE)
    - Maximizes independence between SOI and rest of the sources
    - Preserves dependence of frequency components within SOI
    - SOI selected randomly by initialization
  - (Semi-)supervised IVE
    - Introduction of prior information about SOI (e.g., by pilot signal)
    - Allows selection of SOI

Blind Source Separation

Mixing model

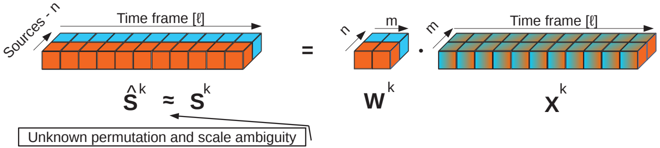
Instantaneous mixing model in time-frequency domain with k frequency bin index.

$$\mathbf{X}^k = \mathbf{A}^k\mathbf{S}^k \qquad \mathbf{S}^k = (\mathbf{A}^k)^{-1}\mathbf{X}^k$$



Goal: finding demixing matrix

$$\mathbf{W}^k \approx (\mathbf{A}^k)^{-1}$$



BSS model reduction

Blind Source Extraction

Source Extraction

We can decompose $\mathbf{A}^k$, $\mathbf{W}^k$, $\mathbf{S}^k$ as follows:

$$\mathbf{W}_k = \begin{pmatrix} (\mathbf{w}_k)^H \\ \mathbf{B}_k \end{pmatrix} \qquad \mathbf{A}_k = \left(\mathbf{a}_k, \mathbf{Q}_k\right) \qquad \mathbf{S}_k = \begin{pmatrix} \mathbf{s}_k \\ \mathbf{Z}_k \end{pmatrix}$$

- $\mathbf{s}_k = (\mathbf{w}_k)^H \mathbf{X}_k$ is an extracted SOI
- $\mathbf{Z}_k = (\mathbf{B}_k)^H \mathbf{X}_k$ are rest of signals

Goal is estimation of the extraction vector $\mathbf{w}^k$ without computing full $\mathbf{W}^k$

Part II

AuxIVE

IVE

Orthogonal Constraint Independent Vector Extraction

Assumptions:

- Probability distributions of $\mathbf{s}$ and $\mathbf{Z}$ are independent ,
- Probability distribution of $\mathbf{Z}$ is Gaussian,
- Matrix $\mathbf{B}$ is chosen to be orthogonal to $\mathbf{a}$,
- Speech have laplacian distribution - $f(\cdot)$
- The relation between $\mathbf{w}$ and $\mathbf{a}$ is $\mathbf{w}_k^H \mathbf{a}_k = 1$

Relation $\mathbf{w}^k$ and $\mathbf{a}^k$ is too weak and need to be make stronger by orthogonal constraint (OG):

$$\mathbf{a}_k = \frac{\widehat{\mathbf{C}}_{\mathbf{x}_k} \mathbf{w}_k}{\mathbf{w}_k^H \widehat{\mathbf{C}}_{\mathbf{x}_k} \mathbf{w}_k}, \quad k = 1, \dots, K,$$

AuxIVE

Optimization by auxiliary function

By including all assumption we can obtain contrast function for estimation extraction vector with respect to OG

$$\mathcal{J}(\mathbf{w}_k) = \mathrm{E}[\log f(\widehat{\mathbf{s}}_1, \dots, \widehat{\mathbf{s}}_K)] - \sum_{k=1}^{K} \mathrm{E}[\mathbf{x_k^H B_k^H C_{z_k}^{-1} B_k x_k}] + \log|\det \mathbf{W_k}|^2,$$

We choose to minimize the contrast function using auxiliary function technique. By rewriting $\mathcal{J}$ we obtain the form

$$\mathcal{Q}(\mathbf{w}_k, \mathbf{V}_k) = -\frac{1}{2}\sum_{k=1}^{K}(\mathbf{w}_k)^H \mathbf{V}_k \mathbf{w}_k - \mathrm{E}[\mathbf{x_k^H B_k^H C_{z_k}^{-1} B_k x_k}] + \log|\det \mathbf{W_k}|^2,$$

which leads to fast and stable converging algorithm.

AuxIVE

block-by-block processing

We assume a **moving SOI**

- $\mathbf{w}_k$ and $\mathbf{V}_k$ needs to be time-varying

We propose:

- Adaptation using short block of data L_b
- Initialization of $\mathbf{w}_k$ and $\mathbf{V}_k$ by data from previous block $k-1$
- Utilization of forgetting factor α to control adaptation speed

Two approaches:

- **Block online AuxIVE:** $\alpha = 0$; $L_b > 1$
- **Online AuxIVE:** $\alpha \in (0, 1]$; $L_b = 1$

AuxIVE

Update Rules (block-by-block processing)

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

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

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

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

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

where α is a forgetting factor; ℓ_s and ℓ_e denote the beginning and the end of the i-th block, respectively.

Supervised source extraction

- Due to time-varying activity of the sources, blind AuxIVE extracts arbitrary independent source from the mixture
- **To mitigate:** introduction of **a pilot signal** dependent on SOI
  - Has the effect of soft bounding the solution space in order to guide the separation towards SOI
  - **Advantage:** Easy introduction into AuxIVE algorithm

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

  - **Disadvantage:** Difficult estimation for realistic acoustic scenarios
  - Several previously proposed approaches (for specific scenarios)
    - *Voice activity detection* - single speaker with background noise
    - *Mouth movement detection using video* - possible for cross-talk, cumbersome to obtain

Part III

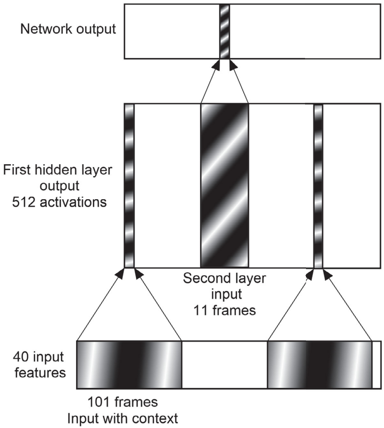
Pilot Signal Design using X-vectors

Pilot design using X-vectors

- **Advantage:** Pilot is usable for both noisy and cross-talk scenario
- **Principle:** Pilot is highly active in frames, which are assigned to the SOI by speaker identification system
- Based on concept of **speaker embeddings:**
    - Embeddings map utterances to fixed-dimensional vectors which encode characteristics of the given speaker
    - Single active speaker usually assumed
    - **In this work:** identification in the presence of cross-talk is shown possible for dominant speaker (in the sense of energy)
- **X-vector:** Variant of speaker embeddings extracted by Time-delayed neural network (TDNN)
    - Topology for sequence classification, less complicated training compared to recurrent neural networks

X-Vector neural network

- Time-delayed neural network
- Each layer considers context of frames around current frame ℓ
- **Input:** single-channel, 40 filter bank coefficients (25 ms length, 10 ms shift)
- **Target:** Classification - 1 of N speakers
- Voxceleb2 training dataset ($N \sim 6000$ speakers, ~ 1000000 utterances)



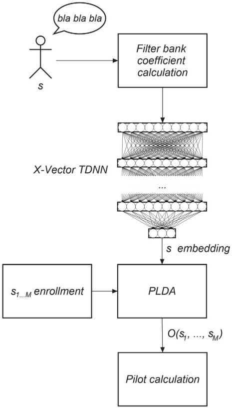
X-Vector neural network

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	—	$\max(151, L_c)$	$128 \times N$

- No frame sub-sampling within contexts
- Mean time-pooling of frames within context
- $L_c = 151$ during the training phase
- X-vectors extracted at the pooling layer

Speaker identification using x-vectors

- **Classifier:** Probabilistic Linear Discriminant Analysis (PLDA)
    - Allows classification of speakers, which are not part of the training set
    - **Enrollment set:** x-vectors representing speakers, which will be classified during testing
    - Tests a hypothesis that an unknown x-vector corresponds to each of the speakers in the enrollment set
    - **Output:** log-likelihood PLDA score $O(s_i)$ for speaker s_i
    - **In this work:** During cross-talk, the dominant speaker (from the perspective of energy) corresponds to the highest PLDA score



Pilot signal design

- **Oracle:** energy-based pilot signal

$$\mathbf{P}_\ell^{\mathrm{ORAC}} = \begin{cases} \sum_{k=1}^{K} |\mathbf{X}_{k,\ell}|^2 & \frac{\sum_{k=1}^{K} |\mathbf{s}_{k,\ell}|^2}{\sum_{j=2}^{d} \sum_{k=1}^{K} |\mathbf{z}_{k,\ell}^{j}|^2} \geq \nu, \\ 0 & \text{otherwise} \end{cases}$$

- **Proposed:** PLDA-score-based pilot signal

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

Part IV

Experiments

X-vectors during cross-talk: case study

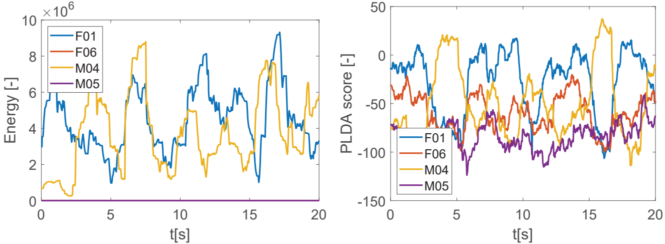
Comparison of PLDA score(s) and true energy of the active speaker(s)

Setup:

- **Enrollment:** four speakers from the CHiME-4 database (two male, two female)
- **Analyzed signal:** Noiseless mixture of female F01 and male M04
- Mild reverberation added ($T_{60} = 100$ ms)
- Global energy of F01 about 2 dB higher
- **TDNN and oracle energy context:** 151 frames (1.5s)

X-vectors during cross-talk: case study

- Dominant speaker is correctly marked by PLDA score in 79.8% of frames
- Second highest PLDA score does not correspond to the second active speaker



- Piloting needs more time-localized information, the TDNN context shortens to 10 frames
- Dominant speaker is correctly marked by PLDA score in 62.4% of frames

Piloted AuxIVE

Experimental setup

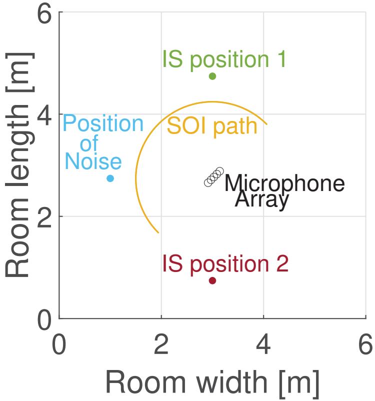
For the experimental evaluation we use

- Closed set of four speakers (2 male and 2 female), two simultaneously active at a time
- One speaker as SOI and different speaker as interference signal (IS) located at two possible different locations
- 5 unique 1 minute utterances for each speaker
- Pedestrian area noise (-10 dB with respect to speech mixture)
- We operate in STFT domain with 512 frequency bins and 160 shift (10ms)
- SOI was moving in semicircle by 40cm/s
- Simulated 5-channel signal in room with $T_{60} = 100$ ms
- 600 mixtures with mean input SNR 1.35 dB
  - 6 speaker combinations $\times$ 2 speaker roles $\times$ 25 utterance combinations $\times$ 2 IS positions.

Piloted AuxIVE

Room setup

- When IS is in position 1, SOI and IS become aligned during movement
- Without pilot, there is a strong possibility that IS is extracted instead of SOI



Piloted AuxIVE

Methods setup

Tested methods:

- Block online AuxIVE with 100 frames block size and 75 frames block shift
- Online AuxIVE with 1 frame block size and forgetting factor $\alpha = 0.97$

Pilot setup:

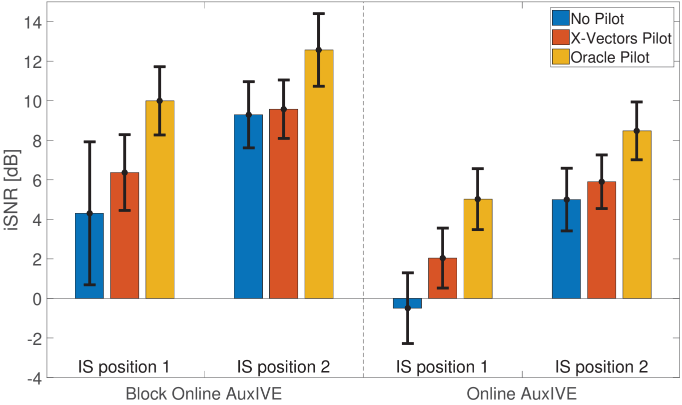
- Oracle with $\nu = 0.5$
- X-Vectors pilot with $\eta = \exp(-5)$

Evaluation:

- Improvement in Signal-to-Noise ratio
- Fail cases when iSNR < 1 dB

Piloted AuxIVE

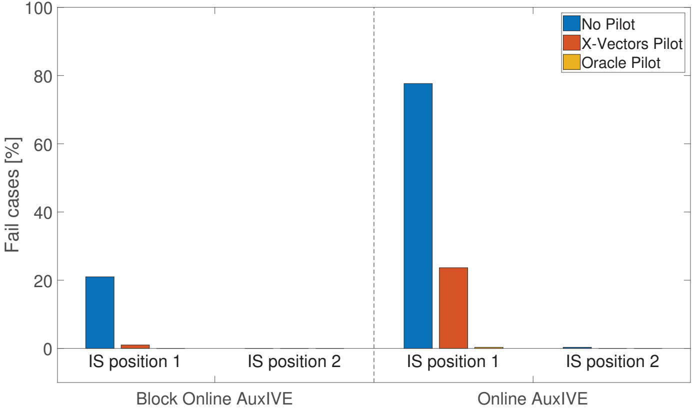
Improvement in SNR



X-vector-based pilot prevents the unwanted extraction of the IS, especially for the IS located at position 1

Piloted AuxIVE

Failed cases



X-vector-based pilot prevents the unwanted extraction of the IS, especially for the IS located at position 1

Conclusion

- **AuxIVE:** blind adaptive and fast converging method for BSE proposed
- Suitable for extraction of moving SOI in the noisy cross-talk scenario
- Extraction of SOI ensured by supervision via pilot signal
- **Pilot:** using x-vectors, frames where SOI is dominant (from the perspective of energy) are identified
- Functionality of the pilot verified for low-reverberation scenarios
- **Future work:** Investigation of x-vectors in the presence of cross-talk for more reverberant and noisy environments

Thank you for your attention