Estimation of Time-varying Room Impulse Responses of Multiple Sound Sources from Observed Mixture and Isolated Source Signals

Joonas Nikunen and Tuomas Virtanen
Laboratory of Signal Processing, Tampere University of Technology, Finland

Introduction

Online estimation of room impulse responses (RIR) between multiple isolated source signals and far-field mixtures

- **Motivation:** Obtain isolated spatial audio tracks from a live mixture recording
- **Application:** Spatial audio remixing and editing for live broadcast and post-production
- **Method:** RIR estimation in short time Fourier transform (STFT) domain by adaptive filtering [1]
- **Evaluation:** Unmixing moving speakers from far-field mixture and use of objective separation metrics for evaluation

Processing overview

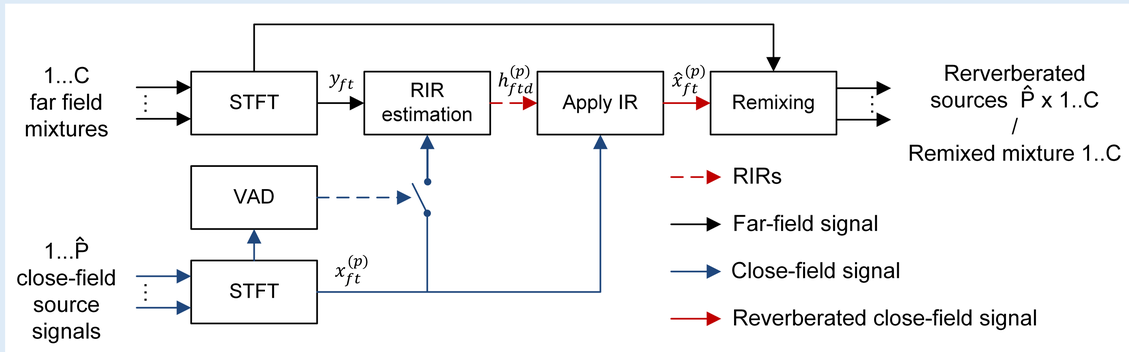

Figure: The block diagram of the proposed processing.

Extending the STFT domain RIR estimation framework [1]

- Joint estimation of RIRs for multiple available dry source signals
- Highly time-varying RIRs due to moving sound sources
- Large source-to-receiver distances and high amount of reverberation

Proposed method

STFT subband filtering model: $y_{ft} = \sum_{p=1}^{P} \sum_{d=0}^{D-1} x_{ft-d}^{(p)} h_{ftd}^{(p)}$

- Assuming availability of $\hat{P} \leq P$ close-field source signals
- Estimate $h_{ftd}^{(p)}$ independently at each frequency f and jointly $\forall\ \hat{P}$

RIR estimation by Recursive Least Squares (RLS) algorithm

- Stacking sources and RIR coefficients: $\mathbf{x}_{ft} \in \mathbb{C}^{\hat{P}D \times 1}$, $\mathbf{h}_{ft} \in \mathbb{C}^{\hat{P}D \times 1}$
- Filtering operation for multiple sources: $\hat{x}_{ft} = \mathbf{x}_{ft}^T \mathbf{h}_{ft}$
- Estimating $\mathbf{h}_{ft}$ minimizing $C(\mathbf{h}_{ft}) = \sum_{i=0}^{t} \lambda^{t-i} (y_{ft} - \hat{x}_{ft})^2$

Levenberg-Marquardt regularized RLS: Used for controlling when specific elements $h_{ftd}^{(p)}$ within $\mathbf{h}_{ft}$ are updated

1. **Source activity based regularization:** Large regularization weight halts the update of the filter weights when source is inactive
2. **RMS-ratio based regularization:** Controls how much attenuation or amplification on average is allowed between close-field and mixture signals
3. **Relative spectrum based regularization:** prevents excess amplification at frequencies where a close-field signal has no energy
4. **Frequency-dependent RIR length and recursion factor:** RT60 generally lower at high frequencies and small changes in source position can cause large changes in RIR at high frequencies $\rightarrow$ shorter RIR length and less error contribution from past frames

Practical considerations

1. Source signals by close miking (voice and acoustic instruments) or use of any other playback material
2. Mixture by microphone array (spatial audio capture)
3. Method applied on all channels independently
4. Does not need exact synchronization of the source and mixture signals

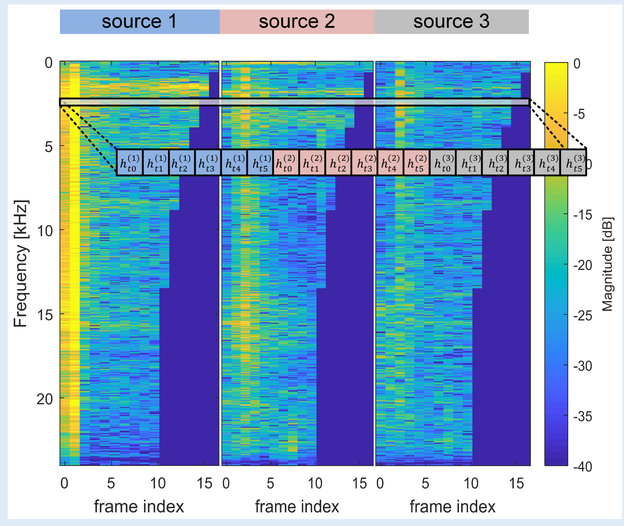

Figure: Magnitudes of the estimated RIRs.

Material

- Moving speakers recorded in reverberant open space
- Source-to-receiver distance varying from 0.5 m to 3 m
- Capture with spherical microphone array (8 channels, $F_s = 48$ kHz)
- 2 speakers mixed together (AA, AB, AS, AT, BS, BT and ST)
- 78 test mixtures each with 30-second duration (22 test / 56 eval)
- Evaluation by objective source separation metrics

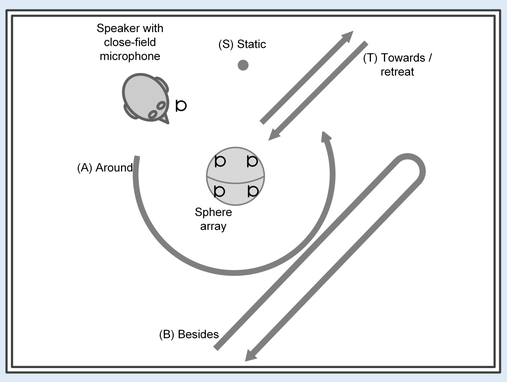

Figure: Recording setup.

Results

- **OL-RLS:** Proposed online RLS based RIR estimation
- **OF-LS:** Offline LS regression based RIR estimation
- **OF-BSS:** Blind source separation algorithm [2]

Results with various $\hat{P}$ and $D = 8$

Method	($\hat{\mathbf{P}}$)	SDR	SIR	SAR	STOI	fwSNR
OL-RLS	(2)	7.38 dB	11.96 dB	9.60 dB	0.7285	30.32 dB
OF-LS	(2)	8.79 dB	13.54 dB	10.82 dB	0.7782	30.30 dB
OL-RLS	(1)	5.35 dB	10.80 dB	7.24 dB	0.6896	29.82 dB
OF-LS	(1)	6.09 dB	12.56 dB	7.51 dB	0.7324	28.86 dB
OF-BSS	(-)	3.59 dB	4.84 dB	11.31 dB	0.6505	29.17 dB

Results of OL-RLS with various D

RIR length	SDR	SIR	SAR	STOI	fwSNR
$D = 4$	6.53 dB	10.59 dB	9.16 dB	0.7002	30.32 dB
$D = 16$	7.24 dB	12.75 dB	8.97 dB	0.7253	30.96 dB
$D = 12...6$	7.44 dB	12.41 dB	9.44 dB	0.7311	30.55 dB

Conclusions

- **Proposed method enables live remixing of spatial audio captures for sources with close-field signal available**
- **Proposed online algorithm performs comparable to offline formulation**
- **Future work: objective evaluation with music content and listening tests of the unmixing application**

[1] Carlos Avendano, "Acoustic echo suppression in the STFT domain," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2001, pp. 175–178.

[2] Joonas Nikunen, Aleksandr Diment, and Tuomas Virtanen, "Separation of moving sound sources using multichannel NMF and acoustic tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 281–295, 2018.

http://www.cs.tut.fi/sgn/arg/nikunen/ firstname.lastname@tut.fi