ESTIMATION OF TIME-VARYING ROOM IMPULSE RESPONSES OF MULTIPLE SOUND SOURCES FROM OBSERVED MIXTURE AND ISOLATED SOURCE SIGNALS

Joonas Nikunen, Tuomas Virtanen

Tampere University of Technology
Korkeakoulunkatu 1, 33720 Tampere, Finland
joonas.nikunen@tut.fi, tuomas.virtanen@tut.fi

ABSTRACT

This paper proposes a method for online estimation of time-varying room impulse responses (RIR) between multiple isolated sound sources and a far-field mixture. The algorithm is formulated as adaptive convolutive filtering in short-time Fourier transform (STFT) domain. We use the recursive least squares (RLS) algorithm for estimating the filter parameters due to its fast convergence rate, which is required for modeling rapidly changing RIRs of moving sound sources. The proposed method allows separation of reverberated sources from the far-field mixture given that their close-field signals are available. The evaluation is based on measuring unmixing performance (removal of reverberated source) using objective separation criteria calculated between the ground truth recording of the preserved sources and the unmixing result obtained with the proposed algorithm. We compare online and offline formulations for the RIR estimation and also provide evaluation with blind source separation algorithm only operating on the mixture signal.

Index Terms— Online room impulse response estimation, informed source separation, source unmixing, adaptive filtering

1. INTRODUCTION

In this paper we propose an online method for estimating room impulse responses (RIR) for multiple moving sources by observing their noisy and reverberated mixture and assuming availability of one or more source signals. The source signals can be obtained by close-field microphones (voice and acoustic instruments) or from playback material outputted through loudspeakers in a live performance recorded for 3D spatial audio. The estimated RIRs are used to obtain isolated reverberated source signals (as captured by one or more far-field microphones), which allows individual 3D audio reconstruction of each reverberated source and unmixing of the sources from mixture to obtain the ambient background.

The problem of time-varying RIR estimation from live mixtures has previously not been widely studied in the setting where the dry source signals are available. It can be thought as a special case of informed source separation where the unknown parameter is the source mixing filters. In an offline scenario where block-wise stationarity of the mixing process is assumed, a least squares (LS) optimal solution of the RIRs can be obtained as in the preparation of the material for CHiME-3 [1] where it was used for removing a single source from a noisy recording. The online setting is related to acoustic echo cancellation (AEC) [2, 3] where the goal is to subtract and suppress the double talked speech. The differences of the proposed task to AEC are the following: 1) source-to-receiver distance can be significantly larger (e.g. up to tens of meters) causing long initial acoustic delay, 2) instead of a single source, there can be multiple close-miked sources and the level of the sources within the mixture can be significantly more varying.

Additionally, a link between the proposed work and research on oracle source separation performance [4, 5] can be made. The widely used BSS eval toolkit [5] finds a single time-invariant projection between the reference and estimated source signal to account for the acoustic delay and reverberation. The evaluation paradigm fails in case of moving sources and using close-field capture as a reference since the single projection cannot account for the time-varying RIRs. The time-varying RIR estimation can perform the projection operation for moving sound sources.

We propose to extend the STFT domain RIR estimation framework [2, 3, 6] for highly time-varying RIRs of moving sound sources with large source-to-receiver distances and high amount of reverberation. Robust operation is achieved by introduction of several novel extensions: source activity based regularization, short-term spectrum based regularization, and frequency-dependent RIR lengths and recursion factor. This paper addresses the joint estimation of RIRs of multiple sound sources, which has not been investigated in previous studies and it is shown to significantly increase the performance.

For algorithm evaluation we use isolated recordings of speech with various types of movement and mix the isolated source signals to obtain test mixtures. The evaluation is based on using the estimated RIR for unmixing a source from the mixture and comparing the result to the ground truth recording of the preserved sources by objective separation criteria [5, 7, 8]. We compare the performance of the proposed online RIR estimation to offline formulation [1]. As a blind baseline assuming that the source signals are not available we use multichannel NMF-based method which has been shown to obtain state-of-the-art results in separation of moving sources [9].

The rest of the paper is organized as follows. In Section 2 we introduce the STFT domain mixing and introduce the joint RIR estimation of multiple source by recursive least squares (RLS) algorithm in Section 3. We introduce the extensions to the RLS based RIR estimation in Section 4. Evaluation of the algorithm performance for reverberated source unmixing is given in Section 5 with conclusions in Section 6.

2. CONVOLUTIVE MIXING IN STFT DOMAIN

The proposed algorithm operates independently on each far-field signal and thus for the algorithm derivation we omit the channel index of the possible microphone array used for spatial audio capture.

A far field microphone observes a mixture of $p = 1, \ldots, P$ source signals $x^{(p)}(n)$ sampled at discrete time instances indexed

This research was supported by Nokia Technologies.

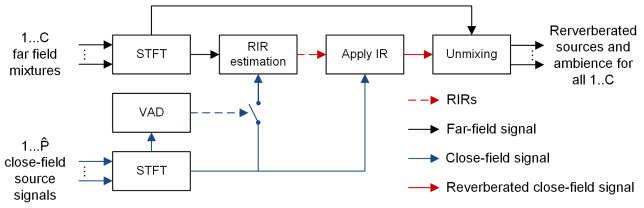

Fig. 1. The block diagram of the proposed processing.

by n and convolved with their RIRs $h_n^{(p)}(\tau)$. The sources are moving and have time-varying mixing defined for each time index n. The resulting mixture signal can be given as

$$y(n) = \sum_{p=1}^{P} \sum_{\tau} x^{(p)}(n-\tau) h_n^{(p)}(\tau) + s(n), \tag{1}$$

where $s(n)$ is additive uncorrelated noise.

Applying the short time Fourier transform (STFT) to the time-domain array signal $y(n)$ and assuming RIRs being stationary within a short time frame allows expressing the source mixing by frame-wise convolution across frequencies defined as

$$y_{ft} \approx \sum_{p=1}^{P} \sum_{d=0}^{D-1} x_{ft-d}^{(p)} h_{ftd}^{(p)} + s_{ft} = \sum_{p=1}^{P} \hat{x}_{ft}^{(p)} + s_{ft}. \tag{2}$$

The STFT of the far-field signal is denoted by y_{ft} where f and t are frequency and frame index, respectively. The source signal as captured by the far-field microphone is modeled by frame-wise convolution between the source STFT $x_{ft}^{(p)}$ and its STFT domain RIR $h_{ftd}^{(p)}$ with frame delays $d = 0, \dots, D-1$. Noise is denoted by its STFT s_{ft} and reverberated source signals are denoted by $\hat{x}_{ft}^{(p)}$.

The model in Equation (2) with convolution using $D-1$ previous frames at each frequency is known in the literature as subband filtering model [10]. It is only an approximation of the convolutive time domain mixing in Equation (1) because of omitting the effect of energy spreading into adjacent frequency bins by FFT, which would require also considering intra-frequency (2-D) convolution [6].

3. ONLINE RIR ESTIMATION IN STFT DOMAIN

The block-diagram of the proposed method is given in Figure 1 and it consists of following steps. We assume availability of $p = 1, \dots, \hat{P}$ close-field source signals ($\hat{P} \leq P$). First, the STFT is applied to both inputs, the far field signal $y(n)$ and close-field source captures $x^{(p)}(n)$. A voice activity detection (VAD) is estimated from the close-field signal in order to determine when the RIR estimate can be updated, i.e., if the source does not emit any signal its RIR cannot be updated. Both STFTs y_{ft} and $x_{ft}^{(p)}$ are inputs to the RIR estimation by the RLS algorithm [11]. As a result, a set of RIRs in the STFT domain are obtained. The estimated RIRs are applied to the original close-field signals to obtain estimate of $\hat{x}_{ft}^{(p)}$ and unmixing of one or more sources can be done by subtraction.

Assuming that the mixing model in Equation (2) is uncorrelated across frequencies then the filter weights can be estimated independently for each frequency. The filtering equation for the $\hat{P}$ known signals at frequency f and frame t is specified as

$$\hat{x}_{ft} = \sum_{p=1}^{\hat{P}} \sum_{d=0}^{D-1} x_{ft-d}^{(p)} h_{ftd}^{(p)} = \mathbf{x}_{ft}^T \mathbf{h}_{ft}, \tag{3}$$

where the vector variables $\mathbf{x}_{ft} \in \mathbb{C}^{\hat{P}D \times 1}$ and $\mathbf{h}_{ft} \in \mathbb{C}^{\hat{P}D \times 1}$ contain the source signals and filter coefficients as stacked:

$$\mathbf{x}_{ft} = [x_{ft}^{(1)}, x_{ft-1}^{(1)}, \dots, x_{ft-D-1}^{(1)}, \dots, x_{ft}^{(\hat{P})}, x_{ft-1}^{(\hat{P})}, \dots, x_{ft-D-1}^{(\hat{P})}]^T,$$
$$\mathbf{h}_{ft} = [h_{ft0}^{(1)}, h_{ft1}^{(1)}, \dots, h_{ftD-1}^{(1)}, \dots, h_{ft0}^{(\hat{P})}, h_{ft1}^{(\hat{P})}, \dots, h_{ftD-1}^{(\hat{P})}]^T.$$

Online estimation of the filter weights $\mathbf{h}_{ft}$ in the least squares sense can be obtained by formulating the problem as system identification in adaptive filtering framework. We use the RLS algorithm [11, 12] where the modeling error at time step t is specified as $e_{ft} = y_{ft} - \hat{x}_{ft}$ and y_{ft} is the observed mixture signal. The cost function to be minimized with respect to filter weights at each frequency f is

$$C(\mathbf{h}_{ft}) = \sum_{i=0}^{t} \lambda^{t-i} e_{fi}^2, \quad 0 < \lambda \leq 1. \tag{4}$$

The exponentially decaying weight λ^{t-i} in the cost function is considered as forgetting factor which determines how much error in past frames contribute to the estimation of the filter weights at current frame. The formulation corresponds to assuming stationarity of RIRs over several time frames controlled by the forgetting factor.

The RLS algorithm minimizing Equation (4) applied individually for each frequency f can be summarized as follows:

Initialization: $\mathbf{h}_{f0} = \mathbf{0}$, $\mathbf{R}_{f0} = \delta \mathbf{I}$

Repeat for $t = 1, 2, \dots$

$$\alpha_{ft} = y_t - \mathbf{x}_{ft}^T \mathbf{h}_{ft-1}$$
$$\mathbf{R}_{ft} = \lambda \mathbf{R}_{ft-1} + \mathbf{x}_{ft}^* \mathbf{x}_{ft}^T \tag{5}$$
$$\mathbf{h}_{ft} = \mathbf{h}_{ft-1} + \mathbf{R}_{ft}^{-1} \mathbf{x}_{ft}^* \alpha_{ft}, \tag{6}$$

where * denotes complex conjugate and $\mathbf{R}_{ft}$ is the autocorrelation matrix of $\mathbf{x}_{ft}$ and it is initialized with identity matrix scaled by δ.

With the above definitions the RLS algorithm can be used to jointly estimate all close-field signal RIRs simultaneously. The algorithm is applied independently for all frequencies to obtain $h_{ftd}^{(p)}$ and the reverberated sources can be obtained as,

$$\hat{x}_{ft}^{(p)} = \sum_{d=0}^{D-1} x_{ft-d}^{(p)} h_{ftd}^{(p)}, \quad p \in [1, \dots, \hat{P}]. \tag{7}$$

Time-domain signals can be reconstructed by inverse FFT and overlap-add sythesis. The modifications of the mixture signal using the reverberated sources is linear additive operation and can be done in either STFT or time-domain.

4. ROBUST RIR ESTIMATION BY RLS

The RLS algorithm introduced in Section 3 can be used as is for RIR estimation, however usual capturing scenarios involve challenging properties that require addressing the robustness of the algorithm. For example, multiple sources can be simultaneously active with very different relative loudness while some sources can be silent for long periods of time. The source spectrum can be sparse (only few harmonic spectral components) and amount of reverberation varies over frequency. In this section we propose novel extensions to the STFT domain RIR estimation in order to make it robust in all operation environments and source types.

4.1. Activity detection, source spectrum and RLS regularization

The source activity detection can be used for controlling when the RIRs are updated, but since the RIR estimation of multiple sources

is formulated as joint optimization problem, there is need to control the update of specific elements $h_{ftd}^{(p)}$ within $\mathbf{h}_{ft}$. For this we propose to use Levenberg-Marquardt regularized RLS algorithm [13] with autocorrelation matrix update in Equation (5) replaced with

$$\mathbf{R}_{ft} = \lambda \mathbf{R}_{ft-1} + \mathbf{x}_{ft}^{*} \mathbf{x}_{ft}^{T} + (1-\lambda)\mathrm{diag}(\mathbf{b}_{ft}), \quad (8)$$

where $\mathrm{diag}(\mathbf{b})$ denotes a diagonal matrix with vector $\mathbf{b}$ on its main diagonal. The regularization weights $\mathbf{b}_{ft} \in \mathbb{R}^{\hat{P}D \times 1}$ are defined as

$$\mathbf{b}_{ft} = [\underbrace{b_{ft}^{(1)}, \ldots, b_{ft}^{(1)}}_{D}, \ldots, \underbrace{b_{ft}^{(\hat{P})}, \ldots, b_{ft}^{(\hat{P})}}_{D}], \quad (9)$$

where each set of D weights corresponds to one source. In order to avoid updating RIR of inactive source p at time step t the respective regularization weights $b_{ft}^{(p)}$ are set to very high values. This effectively halts the update of the filter weights when the second term in Equation (8) is very large and the inverse of $\mathbf{R}_{ft}$ ends up having very small effect in filter weights update in Equation (6) leading to $\mathbf{h}_{ft} \approx \mathbf{h}_{ft-1}$. In the following, we will break down the regularization weight into signal level dependent part $a_t^{(p)}$ and close-field relative spectrum dependent part $c_{ft}^{(p)}$ so that $b_{ft}^{(p)} = a_t^{(p)} c_{ft}^{(p)}$.

4.1.1. Signal RMS level -based regularization

The amount of regularization needed is dependent on how much attenuation or amplification on average is required between close-field and far-field signals. For this we use the overall signal RMS level ratio between the close-field signal $x_{ft}^{(p)}$ and the far-field signal y_{ft} estimated recursively as,

$$L_t^{(p)} = \gamma L_{t-1}^{(p)} + (1-\gamma)\mathrm{RMS}[x_{ft}^{(p)}]/\mathrm{RMS}[y_{ft}], \quad (10)$$

where $\mathrm{RMS}[x_f] = (1/F \sum_f |x_f|^2)^{1/2}$ and γ controls the amount of recursion, i.e. that the RMS estimate does not react too fast for rapid changes in RMS ratio. The amount of regularization for active source p is set to $a_t^{(p)} = \sigma \max_{0<t'<t}[L_{t'}^{p}]$ which denotes maximum observed RMS ratio since from the start of the processing and scaled with global constant σ. For example, if $L_t^{(p)} = 1$ (0 dB) it indicates that the signals have the same overall RMS level. The details of the VAD implementation are explained in Section 5.

4.1.2. Relative spectrum based regularization

The close-field signal $x_{ft}^{(p)}$ can have very low energy at certain frequencies and practically no evidence of it can be observed in the mixture y_{ft}. This applies especially to musical instruments. In order to avoid updating the filter coefficients with no relevant observations, we propose a source spectrum based regularization. We keep short-term average statistics of the close-field signal magnitude spectrum $m_{ft}^{(p)} = \sum_{t'=t-M}^{t} |x_{ft'}^{(p)}|$, where M denotes the number of averaged frames. The spectrum based regularization given the current processed frequency f is defined as

$$c_{ft}^{(p)} = 1 - \log_{10}(m_{ft}^{(p)}/\max_f[m_{ft}^{(p)}]). \quad (11)$$

The frequency index with most energy in the short-term average spectrum results to $c_{ft}^{(p)} = 1$ whereas frequencies with lower energy have $c_{ft}^{(p)} > 1$ in logarithmic relation.

4.2. Variable forgetting factor and RIR length

The contribution of error from past frames to the RIR filter estimate at current frame t is controlled by the forgetting factor λ, which can be varied over frequency f. Small changes in source position can cause substantially large changes in the RIRs at high frequencies due to highly reflected and diffuse sound propagation path. Therefore the contribution of past frames at high frequencies needs to be lower than at low frequencies. It is assumed that the RIR changes slowly at lower frequencies and observations can be integrated over longer periods. The details of used forgetting factor are given in Section 5.

The length of the STFT domain RIR can vary from few frames to several tens of frames, for example a 10 meter distance between close and far-field microphones results to $\tau_{dir} = 29$ ms direct path delay (speed of sound $c = 345$ m/s). Assuming STFT window size of $N = 1024$ samples with 50% overlap, the direct path peak occurs at frame $d_{dir} = \tau_{dir} F_s/(N/2) = 2.7$. If we want to model τ_{rev} ms of reverberation after the direct path, we need to use $D = d_{dir} + \tau_{rev} F_s/(N/2)$ amount of previous frames for the RIRs $h_{ftd}^{(p)}$.

The RIR lengths D in the proposed method can be different for each frequency. Typical rooms have shorter reverberation time at high frequencies than in low frequencies. This is due to high frequencies becoming more easily absorbed by porous materials, whereas lower frequencies interact with low order room modes and have very long reverberation time. Thus the higher frequencies require generally less amount of frames after the direct path d_{dir} frame for accurate modeling of the RIR. Additionally, different sources can have different RIR lengths at the same frequency, which is useful if the direct path delay differs across sources, but all are subject to same amount of reverberation. This requires estimation or prior knowledge about the source-to-receiver distance and this extension is not used in the evaluation. The detailed choice of RIR lengths at different frequencies is given in Section 5

5. ALGORITHM EVALUATION

In this section we evaluate the performance of the proposed algorithm in an unmixing scenario, i.e. removal of one of the reverberated sources from the mixture.

5.1. Material and evaluation procedure

The test material was collected with a 3D printed spherical microphone array (r = 7.5 cm) embodying 8 miniature omnidirectional microphones (DPA 4060). The place of recording was a typical office building coffee lounge with irregular walls and furnishing ($T_{60} \approx 400$ ms). Isolated recordings of human speakers moving around the array or being stationary were recorded and the movement paths (A/B/S/T) are illustrated in Figure 2. The maximum distance from source-to-receiver was approximately 3 meters. The close-field source signal was captured using a head-worn wireless microphone. All signals were recorded using same audio interface with sampling rate of $F_s = 48$ kHz.

Three male speakers spoke the Harvard sentences [14] separately with the 4 different types of movement illustrated in Figure 2 resulting in 12 recordings each with 60-second duration. The recorded signals were split to 30-second segments and two speakers ($P = 2$) were mixed together with the movement combinations AA, AB, AS, AT, BS, BT and ST, also in reversed permutation (AB $\rightarrow$ BA, except for AA), resulting in 13 different speaker/movement combinations. Each 30-second segment from each speaker was used once for each combination, resulting in 6 mixtures per condition and in total $13 \times 6 = 78$ test mixtures each with 30-second duration.

Evaluation is based on measuring the unmixing performance, i.e. subtracting one reverberated source from the mixture and comparing the result to the recording of the remaining source. The mixture signal without *pth* source is denoted by $y_{ft}^{(p)}$ and the correspond-

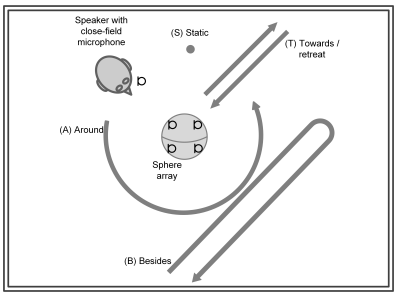

Fig. 2. Recording setup and source movement patterns.

ing estimate by the algorithm is obtained as $\hat{y}_{ft}^{(p)} = y_{ft} - \hat{x}_{ft}^{(p)}$. We use the conventional BSS evaluation scores (SDR, SIR and SAR) [5], frequency-weighted SNR (fwSNR) [7] and short-time objective intelligibility measures (STOI) [8]. We measure the unmixing for both sources and report the average.

5.2. Tested methods and implementation details

For comparison we consider two other methods: offline RIR estimation from the mixture and offline blind source separation (BSS) algorithm. The offline block-wise LS optimal RIR estimation from [1] was modified to produce joint estimates for multiple sources and is referred to as OF-LS and acts as an upper reference. We used block size of 80 frames with 75% overlap between the blocks, resulting into algorithmic delay of $\sim$850 ms as opposed to one frame $\sim$20 ms delay with the proposed online method. The proposed method is referred to as OL-RLS.

The offline BSS algorithm proposed in [9] is used for comparing how well a blind method with assumption of instantaneous mixing in STFT domain performs in the same unmixing task. The method from [9] is based on source spectrogram estimation by multichannel non-negative matrix factorization parametrized by source direction of arrival (DOA) trajectory. Annotated ground truth source DOA trajectories were used to obtain best achievable result and more details of the annotations can be found from [15]. The method is referred to as OF-BSS and its parameters are the same as reported in [9].

All of the parameters were experimentally optimized using a subset of the dataset (22 out of 78 mixtures) and the rest (56 mixtures) was used for evaluation. The STFT window length is 1024 samples with 50% frame overlap. The forgetting factor was set to $\lambda = 0.98$ for 0 Hz and it linearly decreases to 0.95 for $F_s/2 = 24$ kHz. The chosen values correspond to error accumulation extending to past 1.5 seconds for 0 Hz and past 0.8 seconds for 24 kHz. Recursion factor for RMS level ratio was set to $\gamma = 0.97$ and the global constant $\sigma = 10^{-4}$. If the source is inactive regularization is set as $b_{ft}^{(p)} = 100\, a_t^{(p)} c_{ft}^{(p)}$. The base results are with $D = 8\ \forall f$ but we also report and analyze results with different lengths for the RIRs.

The source activity detection was implemented by recursively estimating the RMS level of each close-field signal (as in Equation (10) but without division by mixture signal RMS). We store the minimum RMS value observed as from the beginning of processing which acts as noise floor estimate for each close-field microphone, assuming that source is momentarily silent. We use 3 dB detection threshold above the noise floor to set the source active.

5.3. Results and discussion

The results of each tested method are reported in Table 1. The scores are averaged over the 8 channels and the 56 mixtures and column

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
OF-BSS	(-)	3.59 dB	4.84 dB	11.31 dB	0.6505	29.17 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

Table 1. Unmixing results with the different tested methods.

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

Table 2. Performance of OL-RLS with different RIR lengths.

($\hat{P}$) indicates the number available close-field signals. The unmixing performance of the proposed OL-RLS is 1.5 to 2.0 dB lower in BSS eval scores and 0.05 lower in STOI in comparison to offline processing by OF-LS, whereas the frequency weighted SNR is slightly better. The compromise required for the online operation is thus considered to be small in terms of objective quality. The results with the blind offline method has substantially lower performance in all metrics but especially in SIR and thus cannot be considered to perform complete unmixing of the sources whereas the RIR estimation based informed methods almost completely unmix the source based on informal subjective listening of the results. The last two rows with $\hat{P} = 1$ indicate the algorithm performance when RIRs of the two sources are estimated independently leading to decreased performance for both OL-RLS and OF-LS.

Additionally we have included results of studying the effect of RIR length to the unmixing performance and have reported a few different configurations of OL-RLS in Table 2. The notation $D = 12...6$ denotes linearly decreasing RIR length from 0 Hz to 24 kHz. The unmixing quality is decreased with too short RIR length ($D = 4$), since it does not model all the reverberation. Also too long filters ($D = 16$) lead to lower performance due to overfitting, the RIRs start modeling unwanted correlations between close and far-field signal. The results indicate that using $\tau_{rev} \approx 1/4\, T_{60}$ ms leads to best results in the dataset. Using the variable RIR length leads to slightly better SDR, SIR and STOI for the proposed method.

The algorithm has been tested using up to $\hat{P} = 6$ musical instrument sources played back simultaneously from set of loudspeakers and up to source-receiver distances of 15 meters. The preliminary results were subjectively evaluated to be very promising regarding the task of unmixing and the proposed extensions had greater impact on algorithm performance with musical sources. Future work will consist of reporting the algorithm results with music content and evaluating the unmixing quality by listening tests.

6. CONCLUSION

We presented a method for online estimation of RIRs in STFT domain between a mixture signal and close-field captures of multiple moving sound sources. We proposed novel extensions to the filter parameter estimation by the RLS algorithm and showed that the algorithm performs comparable to the equivalent offline formulation. The application novelty of the proposed algorithm is that it allows separation of reverberated sources from a far-field array capture and preserves the spatial properties of the sources allowing 3D audio reconstruction of each isolated source.

7. REFERENCES

[1] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe, "The third CHiME speech separation and recognition challenge: dataset, task and baselines," in *in Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 504–511.

[2] 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.

[3] Carlos Avendano and Guillermo Garcia, "STFT-based multichannel acoustic interference suppressor," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2001, vol. 1, pp. 625–628.

[4] Emmanuel Vincent, Rémi Gribonval, and Mark D Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933–1950, 2007.

[5] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[6] Yekutiel Avargel and Israel Cohen, "System identification in the short-time fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[7] Yi Hu and Philipos C Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[8] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[9] 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.

[10] Andre Gilloire and Martin Vetterli, "Adaptive filtering in subbands with critical sampling: analysis, experiments, and application to acoustic echo cancellation," *IEEE transactions on signal processing*, vol. 40, no. 8, pp. 1862–1875, 1992.

[11] Simon S Haykin, *Adaptive filter theory*, Pearson Education India, 2008.

[12] Jin Jiang and Youmin Zhang, "A revisit to block and recursive least squares for parameter estimation," *Computers & Electrical Engineering*, vol. 30, no. 5, pp. 403–416, 2004.

[13] Steven L Gay, "Dynamically regularized fast rls with application to echo cancellation," in *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 1996, vol. 2, pp. 957–960.

[14] IEEE Subcommittee, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, pp. 225–246, 1969.

[15] Joonas Nikunen and Tuomas Virtanen, "Time-difference of arrival model for spherical microphone arrays and application to direction of arrival estimation," in *in Proceedings of 25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1295–1299.