Network Architectures for Manifold Learning in MIMO Acoustic System Identification

Till Hardenbicker, Johannes Hahn, Peter Jax
Institute of Communication Systems, RWTH Aachen University, Germany
{hardenbicker, hahn, jax}@iks.rwth-aachen.de

***Abstract*—Identification of linear time-varying acoustic systems with multiple inputs and outputs is required in signal processing tasks like echo cancellation or crosstalk cancellation. When all inputs are excited simultaneously, identification is difficult because each output is a superposition of the influence of all inputs. If the inputs are correlated, identification is even more difficult due to the so-called non-uniqueness problem. A recent approach uses an extended Kalman filter to identify acoustic systems on nonlinear lower-dimensional manifolds. We extend this approach to MIMO systems. Instead of simply increasing the size of the neural networks, we propose architectural variants to control the number of parameters. We show that restricting the size of the network in exchange for its flexibility is beneficial for online system identification.**

***Index Terms*—echo cancellation, crosstalk cancellation, model learning**

I. Introduction and Relation to Prior Work

Acoustic System Identification (ASI) with multiple inputs and outputs, or MIMO ASI for short, is a common problem in audio signal processing. A prominent application is Acoustic Echo Cancellation (AEC) between multiple loudspeakers and microphones, e.g. in cars or conference rooms [1–3]. A visualization of an AEC scenario with two loudspeakers and two microphones is shown in Fig. 1. The sound emitted by the loudspeakers is also picked up by the microphones. To avoid an echo at the far end, an estimate of that echo must be subtracted from both microphone signals. A more recent application is adaptive binaural Crosstalk Cancellation (CTC), where the acoustic paths between multiple loudspeakers and two microphones at a listener's ears must be tracked in real time to design cancellation filters [4, 5]. In both applications, the acoustic paths are modeled as Finite Impulse Response (FIR) filters, as shown in Fig. 1. The goal of ASI is to track the coefficients of these filters in real time. Fig. 1 also shows why estimation is more difficult for MIMO systems. For example, the filter $\mathbf{h}_{11}(k)$ is estimated using the measurement $y_1(k)$. The desired filter output $d_{11}(k)$ is superimposed on the filter output $d_{12}(k)$, which behaves like measurement noise from this perspective. Consequently, $\mathbf{h}_{12}(k)$ must also be known in order not to interfere with the identification of $\mathbf{h}_{11}(k)$, and vice versa. Compared to the single channel case, more parameters have to be identified, which slows down the identification further. This is detrimental when the acoustic paths are time-varying. Aside from slow convergence, another common problem with MIMO ASI is correlation of the excitation signals. In the example in Fig. 1, this happens because the far-end signals are picked up by a microphone array. In this case, there are multiple filter estimates that all suppress the echo signals equally well, but do not necessarily correspond to the actual acoustic paths. This phenomenon is commonly known as Non-Uniqueness Problem (NUP) [1]. At first glance, this does not appear to be a serious problem, since the wrong solutions suppress the echoes equally well. However, if the individual paths at near or far end change in a time-varying scenario, re-convergence from an incorrect estimate may take longer than from the optimal estimate. The CTC application suffers from the NUP because the binaural excitation signals are inherently correlated. Furthermore, CTC filter design aims to invert the actual acoustic paths, so that the effect of the NUP can be more severe [4].

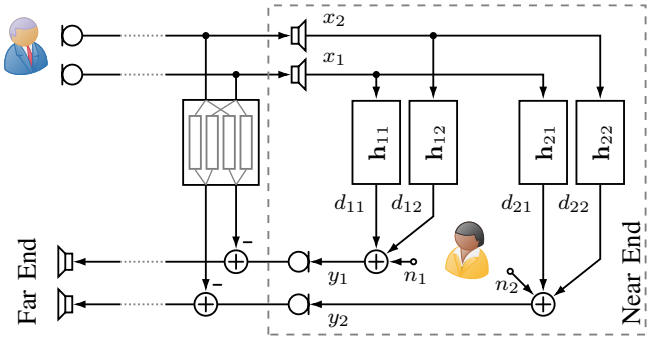

Fig. 1: Model of a full-duplex AEC setup

For ASI a variety of solutions have been presented in the last decades, usually exploiting adaptive filters [6]. Many state-of-the-art approaches rely on optimal step size control of the Kalman Filter (KF) [7, 8] and its formulation in the frequency domain [9, 10]. Since the latter is computationally very efficient, it is an appealing choice for MIMO ASI [3]. Early approaches to overcome the NUP rely on removing the correlation between the excitation signals through distortion [11, 12]. Another way to counter the NUP is to restrict the space of possible estimates [13]. The assumption that acoustic paths in the same enclosure lie on a nonlinear manifold [14] has motivated many approaches to incorporate manifolds in ASI [15–19]. These approaches either project a filter estimate onto the manifold [15, 16] or restrict the solution space to the manifold [17–19].

In this paper, we extend and evaluate the approach of [19] for the MIMO case. It tracks the filter coefficients in the manifold

Simulations were performed with computing resources granted by RWTH Aachen University under project rwth1260.

coordinate system learned by a neural autoencoder. It uses the step size control of an Extended Kalman Filter (EKF) [20]. In Section II-A we derive the weight update for the MIMO adaptive filter, and in Section II-B we propose several design variants for the network architecture to keep the number of trainable parameters within limits. In Section III we evaluate these variants in a computer simulation.

Throughout this paper, bold lowercase letters denote vectors and bold uppercase letters are matrices. $\mathbb{E}$ denotes an expected value and $\operatorname{diag}\operatorname{square}\mathbf{a}$ is a matrix that contains the squared elements of $\mathbf{a}$ on its diagonal. The identity matrix is given by $\mathbf{I}$ and $\mathbf{0}$ is a matrix with only zeros.

II. Proposed Concept

A. MIMO Adaptive Filtering on Manifolds

Our algorithm uses block adaption, meaning that the recorded signals $y_i(k)$ are buffered to vectors $\mathbf{y}_{i,m}$ of length r. Here $i \in \{1 \dots N_{\text{mic}}\}$ is the microphone index, N_{mic} the number of microphones and m is the frame index. The same notation holds for noise $n_i(k)$ and echo $d_i(k)$ contained in $y_i(k)$. The vector $\mathbf{h}_{ij,m}$ contains l filter taps of the path between microphone i and source $j \in \{1 \dots N_{\text{src}}\}$, during the frame m. The recent $r+l-1$ samples of the excitation signals $x_j(k)$ are rearranged to convolution matrices $\mathbf{X}_{j,m}$ with shape $l \times r$. The current block of the recorded signal is modeled as

$$\mathbf{y}_{i,m} = \sum_{j=1}^{N_{\text{src}}} \mathbf{X}_{j,m}\mathbf{h}_{ij,m} \; + \mathbf{n}_{i,m} \; . \tag{1}$$

In traditional MIMO ASI, there is one adaptive filter per receiver [3]. Using a KF, (1) can be used as observation equation. However, we assume that there is a strong dependency between all paths in the same enclosure. Hence, we aim at a joint adaptation and give the observation equation without loss of generality for $N_{\text{src}} = N_{\text{mic}} = 2$ as

$$\underbrace{\begin{bmatrix}\mathbf{y}_{1,m}\\ \mathbf{y}_{2,m}\end{bmatrix}}_{\mathbf{y}_m} = \underbrace{\begin{bmatrix}\mathbf{X}_{1,m} & \mathbf{X}_{2,m} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{X}_{1,m} & \mathbf{X}_{2,m}\end{bmatrix}}_{\mathbf{X}_m} \underbrace{\begin{bmatrix}\mathbf{h}_{11,m}\\ \mathbf{h}_{12,m}\\ \mathbf{h}_{21,m}\\ \mathbf{h}_{22,m}\end{bmatrix}}_{\mathbf{h}_m} + \underbrace{\begin{bmatrix}\mathbf{n}_{1,m}\\ \mathbf{n}_{2,m}\end{bmatrix}}_{\mathbf{n}_m} \tag{2}$$

Following the assumption that paths in the same enclosure are located on a manifold with l' dimensions and $l' < N_{\text{src}}N_{\text{mic}}l$, we can express $\mathbf{h}_m = f_{\text{dec}}(\mathbf{z}_m)$. Here $\mathbf{z}_m$ is a coordinate representation within the manifold and f_{dec} is the nonlinear and differentiable decoder function that maps $\mathbf{z}_m \in \mathbb{R}^{l'}$ onto $\mathbf{h}_m \in \mathbb{R}^{N_{\text{src}}N_{\text{mic}}l}$. Following [19] we use an EKF that tracks $\mathbf{z}_m$ instead of $\mathbf{h}_m$. The nonlinear state space system reads

$$\mathbf{y}_m = \mathbf{X}_m f_{\text{dec}}(\mathbf{z}_m) + \mathbf{n}_m \tag{3}$$

$$\mathbf{z}_m = \gamma\mathbf{z}_{m-1} + \boldsymbol{\delta}_m \; . \tag{4}$$

The state $\mathbf{z}_m$ of this system follows a first-order Markov model with fading factor γ and is observed through f_{dec}. The random vectors $\mathbf{n}_m$ and $\boldsymbol{\delta}_m$ are measurement noise and process noise, respectively, and follow multivariate Gaussian distributions with zero mean. Using the Jacobian $\mathbf{V}_m = \frac{\mathrm{d}f_{\text{dec}}}{\mathrm{d}\mathbf{z}_m}$ we can state the EKF equations:

Time Update

$$\widehat{\mathbf{z}}_m^- = \gamma\widehat{\mathbf{z}}_{m-1}^+ \tag{5a}$$

$$\mathbf{P}_m^- = \gamma^2\mathbf{P}_{m-1}^+ + \mathbf{Q}_{\boldsymbol{\delta},m} \tag{5b}$$

Measurement Update

$$\widehat{\mathbf{Q}}_{\mathbf{e},m} = \mathbf{X}_m\mathbf{V}_m\mathbf{P}_m^-\mathbf{V}_m^{\mathrm{T}}\mathbf{X}_m^{\mathrm{T}} + \mathbf{Q}_{\mathbf{n},m} \tag{6a}$$

$$\mathbf{K}_m = \mathbf{P}_m^-\mathbf{V}_m^{\mathrm{T}}\mathbf{X}_m^{\mathrm{T}}\widehat{\mathbf{Q}}_{\mathbf{e},m}^{-1} \tag{6b}$$

$$\widehat{\mathbf{d}}_m = \mathbf{X}_m f_{\text{dec}}\left(\widehat{\mathbf{z}}_m^-\right) \tag{6c}$$

$$\Delta\mathbf{z}_m = \mathbf{K}_m\left(\mathbf{y}_m - \widehat{\mathbf{d}}_m\right) \tag{6d}$$

$$\widehat{\mathbf{z}}_m^+ = \widehat{\mathbf{z}}_m^- + \Delta\mathbf{z}_m \tag{6e}$$

$$\mathbf{P}_m^+ = \left(\mathbf{I}_{l'} - \mathbf{K}_m\mathbf{X}_m\mathbf{V}_m\right)\mathbf{P}_m^- \; . \tag{6f}$$

The superscripts $^-$ and $^+$ denote prior and posterior estimates, respectively. Here, $\mathbf{P}_m^-$ and $\mathbf{P}_m^+$ express the State Error Covariance (SEC) of the states $\widehat{\mathbf{z}}_m^-$ and $\widehat{\mathbf{z}}_m^+$, respectively. The matrices $\mathbf{Q}_{\boldsymbol{\delta},m}$ and $\mathbf{Q}_{\mathbf{n},m}$ express the vector covariance of the unkown noise vectors $\boldsymbol{\delta}_m$ and $\mathbf{n}_m$. The matrix $\mathbf{K}_m$ is called the Kalman gain matrix, and acts as an optimal step size [8].

An import implication from these equations is that the measurements from all N_{mic} microphones influence the estimate $\widehat{\mathbf{h}}_m$ of all employed acoustic paths. So, unlike the traditional approach of N_{mic} independent MISO adaptive filters [3], information is exchanged between microphones. Motivated by efficiency, [3] sets $\mathbf{P}$ as sub diagonal in the frequency domain, meaning that a frequency bin of one acoustic path is coupled to the same frequency bin of all paths related to the same microphone. In contrast, the proposed approach allows for any correlation, depending on the choice of f_{dec}.

B. Neural Network Architecture

A common choice to obtain f_{dec} is to use a β-VAE [21] [17–19]. Variational Autoencoders (VAEs) are neural autoencoders, where the encoder predicts the parameters of a multivariate Gaussian distribution, and obtains the latent representation $\mathbf{z}$ by sampling from this distribution. During training, the weighted Kullback-Leibler Divergence (KLD) between the predicted distribution and the standard Gaussian distribution is added to the training loss to ensure a contiguous latent space. After the training only the decoder is kept to implement $f_{\text{dec}} : \mathbf{z} \mapsto \mathbf{h}$ for the proposed algorithm.

The decoder in [19] utilizes four fully connected network layers, with an increasing number of neurons, to map the vector $\mathbf{z}$ with the length l' to the vector $\mathbf{h}$ with the length l. However, for MIMO ASI the network is required to output the stacked impulse response vector of length $N_{\text{src}}N_{\text{mic}}l$. This results in an increase in the number of network weights by a factor of approximately $(N_{\text{src}}N_{\text{mic}})^2$. So even for the most compact MIMO configuration, $N_{\text{src}} = N_{\text{mic}} = 2$, the number of required network weights is 16 times greater than for the single channel case. This is detrimental due to a higher memory consumption and increased training time. Most notably, the higher number of trainable parameters necessitates a higher

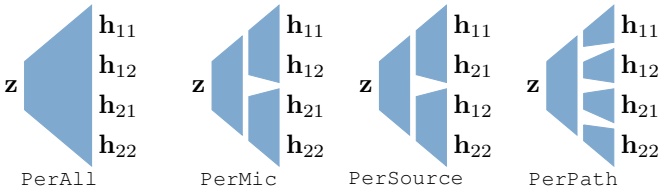

Fig. 2: Design variants

number of training instances, obtained by time-consuming and costly acoustic measurements. Therefore we propose additional network architecture variants that require less parameters. Formally, we divide both the encoder and the decoder into an outer stage and an inner stage. The outer stage refers to first layers of the encoder and the last layers of the decoder. The inner stage refers to the remaining layers. In total, there are seven variants, that can be categorized by two design choices.

The first design decision is on which type of correlation the dimensionality reduction in the outer stage should be based. We distinguish four cases, illustrated for the decoder in Fig. 2: `PerAll`, `PerPath`, `PerMic`, and `PerSource`. The first option `PerAll` corresponds to the generic variant described in the previous paragraph and takes into account any correlation, within and between different acoustic paths. This design has the largest number of parameters. On the other hand, the option `PerPath` has the lowest number of parameters, since it considers only the correlation within each path. This is achieved by using $N_{\text{src}} \cdot N_{\text{mic}}$ smaller sublayers in parallel. The last two options, `PerMic` and `PerSource`, are compromises in terms of the correlation exploited and the number of parameters. In the case of `PerMic`, the outer stage uses N_{mic} sublayers, where a sublayer jointly processes all acoustic paths associated with a microphone. Similarly, the outer stage can use one sublayer for each loudspeaker, resulting in the `PerSource` variant. Regardless of the design, all outputs of the encoders' outer stage are stacked into a single vector so that all paths are processed together in the inner stage. In other contexts, the architectures except for `PerAll` may be called multi headed networks.

The second design choice is only applicable to the multi headed variants. Considering symmetry in many acoustic setups, it is possible to force parallel sublayers to have the exact same weights. This technique is commonly known as parameter sharing. Like the usage of sublayers, parameter sharing exchanges the network's flexibility against a reduced number of parameters. Combining the two presented design choices of sublayers and parameter sharing leads to seven different design variants for the VAE.

III. Experimental Validation

A. Experimental Design

For the experimental validation we consider an illustrative mixture of the intended applications AEC and CTC: A listener with microphones at their ears is in a room with two loudspeakers close to a wall. The task is to identify the $N_{\text{src}} \cdot N_{\text{mic}} = 4$ acoustic paths between the loudspeakers and the ear microphones. In the first of two experiments, we investigate the influence of the design variant and the latent space dimension on the ASI performance. In the second experiment we assess the algorithms capability to overcome the NUP. As a metric we consider the Echo Return Loss Enhancement (ERLE)

$$\text{ERLE} = \frac{1}{N_{\text{mic}}} \sum_{i=1}^{N_{\text{mic}}} \frac{\mathbb{E}\left\{d_i^2(k)\right\}}{\mathbb{E}\left\{\left(d_i(k) - \widehat{d}_i(k)\right)^2\right\}}$$

and the relative system distance

$$D = \frac{1}{N_{\text{src}} N_{\text{mic}}} \sum_{i=1}^{N_{\text{mic}}} \sum_{j=1}^{N_{\text{src}}} \frac{\|\mathbf{h}_{ij} - \widehat{\mathbf{h}}_{ij}\|^2}{\|\mathbf{h}_{ij}\|^2} \,.$$

Expected values in the computation of the ERLE are approximated by recursive smoothing with a time constant of $0.15\,\text{s}$.

B. Dataset Design and Testing Scenarios

To train the VAE, data is needed. In this paper we simulated Binaural Room Impulse Responses (BRIRs) using the image source method together with a spherical interpolation of the listeners Head-Related Transfer Function (HRTF) [22]. The head was a FABIAN artificial head measured by [23]. The simulated near end room has a reverberation time of $0.125\,\text{s}$. The sampling rate is $16\,\text{kHz}$ and impulse responses are truncated to $0.125\,\text{s}$ (2000 samples). The simulated sound sources are placed close to a wall and have a distance of $2.5\,\text{m}$ to each other, facing the center of the room with the directional characteristic of a common loudspeaker model. The head was located at 637 random receiver positions, all located in a volume of $0.6 \times 0.6 \times 0.3\text{m}^3$. At each position we simulated the HRTFs for all azimuth angles from $-179°$ to $179°$ in steps of $2°$, resulting in 114.660 BRIRs. We made sure that no receiver position was closer than $5\,\text{cm}$ to the center of the room, since this single position is used for testing.

C. Network Training

For network training we use Nadam [24] with an initial learning rate of 10^{-4}. The loss function is given by the sum of the squared reconstruction loss, the KLD weighted with $\beta = 10^{-7}$, and an additional L_2 weight regularization with a factor of 10^{-8}. All encoders and decoders have two layers each in the outer stages, and two layers in the inner stages. An exception is given by the design variant without sublayers: Due to the high number of parameters it only has three layers each in encoder and decoder. Hidden layers use the swish activation function and output layers use a linear activation. Tab. I shows the number of trainable parameters for the case $l' = 500$.

# Params / 10^6	`PerPath`	`PerSource`	`PerMic`	`PerAll`
Weight sharing	5.0	12.5	12.5	42.5
Individual	18.5	26.5	26.5	

Tab. I: Number of trainable parameters in the decoder

D. State Space Parameters

To initialize $\mathbf{P}_0^-$ and $\mathbf{z}_0^-$ we transform all training instances into the latent space using the trained encoder, without statistical sampling. Then, $\mathbf{P}_0^-$ is the vector covariance and $\mathbf{z}_0^-$ the mean of all encodings. To estimate $\mathbf{Q}_\delta$ we follow [25] and estimate $\mathbf{Q}_{\boldsymbol{\delta},m} = \alpha\mathbf{Q}_{\boldsymbol{\delta},m-1} + (1-\alpha)\,\mathrm{diag\,square}(\Delta\mathbf{z})$ recursively. Similarly and following [26], we estimate the measurement noise covariance $\mathbf{Q}_{\mathbf{n},m} = \alpha\mathbf{Q}_{\mathbf{n},m-1}+(1-\alpha)\,\mathrm{diag\,square}(\mathbf{e}_m)$, where $\mathbf{e}_m$ is the residual error signal $\mathbf{y}_m - \widehat{\mathbf{d}}_m$. We initialize $\mathbf{Q}_{\mathbf{n},m}$ from oracle knowledge. The task of the online estimator is to account for errors due to under-modeling of the decoder and the length of the impulse responses.

E. Reference Algorithms

As a baseline we consider the very efficient subdiagonal Frequency Domain Kalman Filter (FDKS) [3]. It consists of two MISO Kalman filters, such that correlation is only considered between speakers and between equal frequency bins. Process noise is estimated as proposed in [3] using $\gamma = 0.995$, measurement noise is estimated as proposed in [26]. On the other hand we consider an (exact) Time Domain Kalman Filter (TDKE) that is able to account for any kind of correlation but has a high complexity due to its large SEC matrix. To analyze the need for a nonlinear manifold, we consider the linear simplification of the proposed approach, where $\mathbf{V}$ is computed only once by Principal Component Analysis (PCA) over the training data. For comparability, the state and the SEC of all reference algorithms are initialized from the training data, i.e. its mean and the covariance in the corresponding space. This results in an optimal step size [8]. For all algorithms, we set $r = 64$ which corresponds to $4\,\mathrm{ms}$.

F. First Experiment: Network Architecture

In this experiment we train networks with all seven design variants and with nine different subspace dimensions, resulting in 63 trained decoders to be tested in the proposed ASI algorithm. As test case, the receiving head performs an azimuthal rotation from $-90°$ to $90°$ within $10\,\mathrm{s}$, to test the algorithm's capability of tracking highly time-variant systems. The impulse responses are truncated only after $0.21\,\mathrm{s}$ to account for under-modeling in the training data. The Echo-To-Noise Ratio $\mathrm{ENR} = \mathbb{E}\,d_i^2(k)\,/\mathbb{E}\,n_i^2(k)$ is set to $20\,\mathrm{dB}$ for both ears.

For the excitation signals $x_j(k)$ we follow Fig. 1. A person is speaking into two microphones that are located $5\,\mathrm{cm}$ from each other, as an example. The person is located randomly in front of the microphones, so that the recorded signals are correlated but not identical. The dry signals are $10\,\mathrm{s}$ of freeform speech each from the first ten subjects in [27].

The results are shown in Fig. 3. Each point represents the median of the metric for all ten test signals, considering only frames where speech is active in the microphone signal. Dotted and solid lines represent the proposed algorithm with and without parameter sharing, respectively. Dashed lines show the baseline algorithms. The results provide a variety of insights:

1 – The number of trainable parameters can be reduced without sacrificing performance, as the overall depence on the network architecture is weak. On average, the variants that use parameter sharing achieve better results. As an additional benefit, these variants scale better with increasing number of loudspeakers and microphones. For larger values of l', decoders without parameter sharing even fail completely. We conclude that too large models are difficult to train.

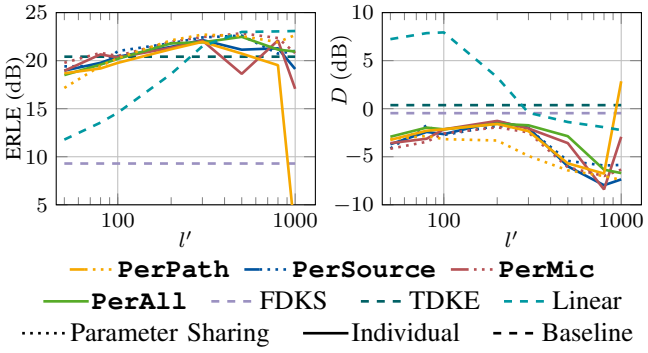

Fig. 3: ERLE and relative system distance depending on the dimension l' of the latent space.

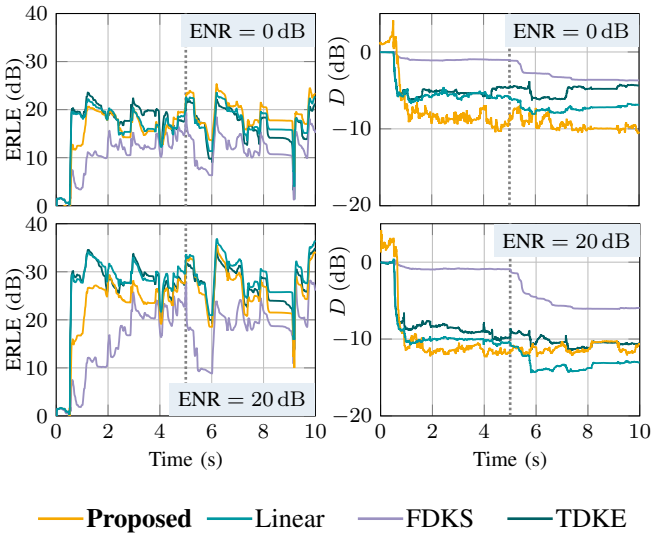

Fig. 4: ERLE and relative system distance during a movement of the far-end speaker.

2 – Nonlinear manifolds outperform linear manifolds in terms of ERLE, when the subspace dimension is small. But it has to be considered that the linear variant is less complex, so the subspace dimension can possibly be increased to achieve comparable performance at comparable complexity.

3 – The proposed algorithm is able to compete with or even surpass the TDKE, which has a much higher complexity. It is worth to mention that an important contributor to the TDKEs performance is the data driven initialization of the SEC.

G. Second Experiment: Non-Uniqueness Problem

In this experiment we investigate the algorithms' capabilities to overcome the NUP. The head with microphones is fixed in the center of the room and does not move. Instead, the speaker in the far end room moves. First he is slightly closer to the

right microphone and after $5\,\mathrm{s}$ he moves so that he is slightly closer to the left microphone. For the proposed algorithm we use the configuration with $l' = 1000$, sublayers `PerPath` and with weight sharing. For the FDKS we set $\gamma = 0.9999$. We investigate two ENRs, i.e. $0\,\mathrm{dB}$ and $20\,\mathrm{dB}$.

Fig. 4 shows the ERLE and the relative system distance over time. The effect of the NUP is clearly visible: For both ENRs the FDKS has a slightly lower ERLE than the other algorithms. However, the relative system distance of the FDKS is much higher. After the positional change of the far end speaker, the relative system distance of the FDKS becomes lower but the filter needs to re-converge, resulting in a lower ERLE. The other algorithms are not affected by the NUP. The proposed algorithm and its linear variant restrict the space of possible solutions, so that the filter cannot converge to a false solution. The TDKE does not make assumptions about a subspace, but the data-driven initialization of its SEC steers the weight updates into the right direction.

IV. Summary

In this paper we extended a single-channel manifold-ASI algorithm to the multichannel case. Restricting the space of solutions to a lower dimensional subspace reduces the computational complexity compared to the TDKE and helps to overcome the NUP. In addition, we proposed variants of the neural network architecture which reduce the number of trainable parameters. According to simulation results, the reduction of parameters improves performance in real-time BRIR identification. Future research should consider the algorithms performance in real-world scenarios and its robustness to changes in the reverberation time.

References

[1] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 156–165, 1998.

[2] H. Buchner, J. Benesty, and W. Kellermann, "Generalized multichannel frequency-domain adaptive filtering: Efficient realization and application to hands-free speech communication," *Signal Processing*, vol. 85, no. 3, pp. 549–570, 2005.

[3] S. Malik and G. Enzner, "Recursive Bayesian control of multichannel acoustic echo cancellation," *IEEE Signal Processing Letters*, vol. 18, no. 11, pp. 619–622, Nov. 2011.

[4] B. Masiero and M. Vorländer, "A framework for the calculation of dynamic crosstalk cancellation filters," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1345–1354, 2014.

[5] T. Kabzinski and P. Jax, "An adaptive crosstalk cancellation system using microphones at the ears," in *Audio Engineering Society Convention*, Oct 2019.

[6] S. Haykin, *Adaptive Filter Theory*, 5th ed. Pearson Education Ltd., 2014.

[7] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. of Basic Engineering*, vol. 82, no. 1, pp. 35–45, Mar. 1960.

[8] D. P. Mandic, S. Kanna, and A. G. Constantinides, "On the intrinsic relationship between the least mean square and Kalman filters [lecture notes]," *Signal Processing Mag.*, vol. 32, no. 6, pp. 117–122, Nov. 2015.

[9] G. Enzner and P. Vary, "Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones," *Signal Processing*, vol. 86, no. 6, pp. 1140–1156, Jun. 2006.

[10] F. Kuech, E. Mabande, and G. Enzner, "State-space architecture of the partitioned-block-based acoustic echo controller," *Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 1295–1299, 2014.

[11] M. M. Sondhi, D. R. Morgan, and J. L. Hall, "Stereophonic acoustic echo cancellation - An overview of the fundamental problem," *IEEE Signal processing letters*, vol. 2, no. 8, pp. 148–151, 1995.

[12] D. R. Morgan, J. L. Hall, and J. Benesty, "Investigation of several types of nonlinearities for use in stereo acoustic echo cancellation," *IEEE transactions on speech and audio processing*, vol. 9, no. 6, pp. 686–696, 2001.

[13] M. Fozunbal, T. Kalker, and R. W. Schafer, "Multi-channel echo control by model learning," in *Int. Workshop on Acoustic Echo and Noise Control*, 2008.

[14] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "A study on manifolds of acoustic responses," in *Lecture Notes in Computer Science*, Springer. Springer International Publishing, 2015, pp. 203–210.

[15] T. Haubner, A. Brendel, and W. Kellermann, "Online supervised acoustic system identification exploiting prelearned local affine subspace models," in *Int. Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2020, pp. 1–6.

[16] ——, "Online acoustic system identification exploiting Kalman filtering and an adaptive impulse response subspace model," *J. of Signal Processing Systems*, vol. 94, no. 2, pp. 147–160, Feb. 2022.

[17] A. Brendel, J. Zeitler, and W. Kellermann, "Manifold learning-supported estimation of relative transfer functions for spatial filtering," in *Int. Conf. on Acoustics, Speech, and Signal Processing*, May 2022.

[18] K. Helwani, P. Smaragdis, and M. M. Goodwin, "Generative modeling based manifold learning for adaptive filtering guidance," in *Int. Conf. on Acoustics, Speech, and Signal Processing*, Jun. 2023.

[19] T. Hardenbicker and P. Jax, "Online system identification on learned acoustic manifolds using an extended Kalman filter," in *Int. Workshop on Acoustic Echo and Noise Control*, 2024, pp. 339–343.

[20] G. L. Smith, S. F. Schmidt, and L. A. McGee, *Application of statistical filter theory to the optimal estimation of position and velocity on board a circumlunar vehicle*. NASA, 1962.

[21] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework." *International Conference on Learning Representations*, vol. 3, 2017.

[22] F. Brinkmann and S. Weinzierl, "Aktools - An open software toolbox for signal acquisition, processing, and inspection in acoustics," in *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

[23] F. Brinkmann, M. Dinakaran, R. Pelzer, J. J. Wohlgemuth, F. Seipl, and S. Weinzierl, "The hutubs HRTF database," *DOI 10.14279/depositonce-8487*, 2019.

[24] T. Dozat, "Incorporating Nesterov momentum into adam," 2015. [Online]. Available: https://cs229.stanford.edu/proj2015/054_report.pdf

[25] C. Paleologu, J. Benesty, S. Ciochina, and S. L. Grant, "A Kalman filter with individual control factors for echo cancellation," in *Int. Conf. on Acoustics, Speech, and Signal Processing*, May 2014, pp. 5974–5978.

[26] J. Franzen and T. Fingscheidt, "Improved measurement noise covariance estimation for n-channel feedback cancellation based on the frequency domain adaptive Kalman filter," in *Int. Conf. on Acoustics, Speech, and Signal Processing*, 2019, pp. 965–969.

[27] J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, and T. Gerkmann, "EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation," in *ISCA Interspeech*, 2024, pp. 4873–4877.