

ONLINE SYSTEM IDENTIFICATION ON LEARNED ACOUSTIC MANIFOLDS USING AN EXTENDED KALMAN FILTER

Till Hardenbicker, Peter Jax

Institute of Communication Systems (IKS), RWTH Aachen University, Germany
{hardenbicker, jax}@iks.rwth-aachen.de

ABSTRACT

Many communication devices require real-time tracking of acoustic paths based on noisy measurements. For this task, solutions such as the Kalman filter are known, which require little or no information about the system to be identified. When the acoustic path is time-varying and the measurements are noisy, these algorithms reach their limits. In this paper, we combine the Kalman filter with neural autoencoders to overcome these limitations. Our approach restricts the set of possible systems to a lower-dimensional manifold, while still exploiting the sophisticated step-size control of the Kalman filter. In simulations, the algorithm shows state-of-the-art performance with reduced computational complexity.

Index Terms—System Identification, Echo Cancellation, Manifold Learning

1. INTRODUCTION AND PRIOR WORK

Acoustic System Identification (ASI) is needed in many signal processing tasks: Adaptive crosstalk cancellation [1], de-reverberation [2], noise cancellation in headphones [3, 4], and Acoustic Echo Cancellation (AEC) [5–7]. Among the most important approaches are the (Normalized) Least Mean Squares (NLMS) algorithm [8] and the Kalman Filter (KF) [9]. These algorithms can track any linear acoustic system with little information about it, but are challenged in noisy and time-varying scenarios. With the increasing availability of measurement data and computing power, data-driven algorithms are becoming more popular in ASI. These algorithms may perform better at ASI because they can better utilize prior information about the system to be identified.

Early approaches include prior knowledge about the acoustic path in ASI by choosing the step size of an adaptive filter based on the exponential decay of known Room Impulse Responses (RIRs) [10, 11]. Approaches from the last decade exploit the assumption that impulse responses of acoustic paths in an enclosure lie on or close to a lower-dimensional manifold [12–14]. This assumption offers great potential for ASI, as fewer parameters are easier to track in noisy and time-varying

scenarios. In multi-channel system identification, the non-uniqueness problem poses a challenge when the excitation signals of the channels are correlated [5]. This as well can be mitigated by restricting the solution space to a manifold [15]. More recently, neural autoencoders, especially Variational Autoencoders (VAEs), have been used to learn such manifolds in different ways: They can learn a manifold of the measurement signals [16, 17] or of the impulse responses themselves [18, 19]. In contrast, there are nonparametric approaches that use subspaces based on nearest neighbors in a training set [20–22]. These subspaces act as tangent planes of the manifolds.

The aim of this paper is to bridge the gap between recent ASI algorithms that use manifolds. The authors of [18] train a VAE on acoustic impulse responses. During ASI, they identify the VAE’s latent representation, rather than identifying the impulse response itself. The authors of [19] track the latent representation in an online algorithm with a scalar step size. In [22], we have plugged the assumption of local linear subspaces into the KF equations, yielding optimal step size control within these subspaces [23]. We now propose to apply the KF in a latent space learned by a VAE, yielding an Extended Kalman Filter (EKF). Compared to [22], this eliminates the computational burden of nearest neighbor search and Principal Component Analysis (PCA) at each time step. Furthermore, this offers an online procedure with step size control, as opposed to [18, 19].

This paper is organized as follows: In Section 2, we introduce the proposed approach. In Section 3, we perform an experimental validation of the proposed algorithm and report the training and model parameters used. Throughout this paper, bold letters denote column vectors and $\hat{\mathbf{a}}$ denotes an estimate of $\mathbf{a}$. $\mathbf{I}_a$ is an identity matrix of size $a \times a$. The 2-norm of $\mathbf{a}$ is written as $\|\mathbf{a}\|$ and its Hadamard square is written as $\mathbf{a}^{\circ 2}$. The operator $\text{diag}(\mathbf{a})$ transforms a column vector into a diagonal matrix and vice versa.

2. PROPOSED ALGORITHM

In this section, we explain how KF and VAE can be combined to obtain adaptive step size control for ASI on manifolds.

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

2.1. Extended Kalman Filter

For ASI with KF a state space model is needed. We start by identifying the observation equation for the general case, i.e., no manifold assumption. Depending on this model, the KF can identify single or multi-channel systems [24] in time or frequency domain [25]. In this contribution we focus on single-channel AEC in time domain. We model the acoustic echo path by a time-varying Finite Impulse Response (FIR) filter. Subsequent additive White Gaussian Noise (WGN) reflects measurement noise. Furthermore, we express the model in vector notation and assume block-wise processing, yielding

$$\mathbf{y}_m = \mathbf{X}_m \mathbf{h}_m + \mathbf{n}_m. \quad (1)$$

Here, $\mathbf{y}_m$ denotes the m^{th} frame of the measurement consisting of r samples, and $\mathbf{n}_m$ represents the noise contained within it. The vector $\mathbf{h}_m$ contains the l FIR filter coefficients, and $\mathbf{X}_m$ is an $r \times l$ convolution matrix of the excitation signal. Now, we modify this equation to confine the filter coefficients $\mathbf{h}_m$ to a manifold learned by a VAE. The VAE consists of an encoder $f_{\text{enc}} : \mathbf{h} \mapsto \mathbf{z}$ and a decoder $f_{\text{dec}} : \mathbf{z} \mapsto \hat{\mathbf{h}}$. We call $\mathbf{z}$ the latent representation of $\mathbf{h}$ with dimension $l' \leq l$. When the filter coefficients are located on the manifold, they can be expressed as $\hat{\mathbf{h}}_m = f_{\text{dec}}(\mathbf{z}_m)$. To track less than l coefficients, we propose that the KF tracks $\mathbf{z}_m$ instead of $\mathbf{h}_m$. Then, the state space model reads

$$\mathbf{y}_m = \mathbf{X}_m f_{\text{dec}}(\mathbf{z}_m) + \mathbf{n}_m \quad (2)$$

$$\mathbf{z}_m = \gamma \mathbf{z}_{m-1} + \delta_m. \quad (3)$$

Here, we assume a first order Markov model in the latent space. The fading factor γ is scalar with $0 \ll \gamma \leq 1$ and δ_m is the process noise in the latent space. The measurement noise $\mathbf{n}_m$ and the process noise δ_m are zero mean random processes whose covariance matrices $\mathbf{Q}_{\mathbf{n},m} = \mathbb{E}\{\mathbf{n}_m \mathbf{n}_m^T\}$ and $\mathbf{Q}_{\delta,m} = \mathbb{E}\{\delta_m \delta_m^T\}$ are estimated at runtime.

Since the observation equation is nonlinear, the EKF is needed [26]. With $\mathbf{z}$ as state of the EKF, the time update reads

$$\hat{\mathbf{z}}_m^- = \gamma \hat{\mathbf{z}}_{m-1}^+ \quad (4a)$$

$$\mathbf{P}_m^- = \gamma^2 \mathbf{P}_{m-1}^+ + \mathbf{Q}_{\delta,m} \quad (4b)$$

where $\mathbf{P}$ is the state error covariance matrix. The superscripts $-$ and $+$ denote prior and posterior estimates, respectively. At $m = 0$, $\hat{\mathbf{z}}_m^-$ and $\mathbf{P}_m^-$ are set to $\mathbf{z}_{\text{init}}$ and $\mathbf{P}_{\text{init}}$.

To obtain the predicted echo signal $\hat{\mathbf{d}}_m$, the latent representation needs to be decoded into filter coefficients, yielding

$$\hat{\mathbf{d}}_m = \mathbf{X}_m f_{\text{dec}}(\hat{\mathbf{z}}_m^-). \quad (5)$$

The EKF also requires the decoder's Jacobian

$$\mathbf{V}_m = \left. \frac{d f_{\text{dec}}}{d \mathbf{z}} \right|_{\mathbf{z}=\hat{\mathbf{z}}_m^-}. \quad (6)$$

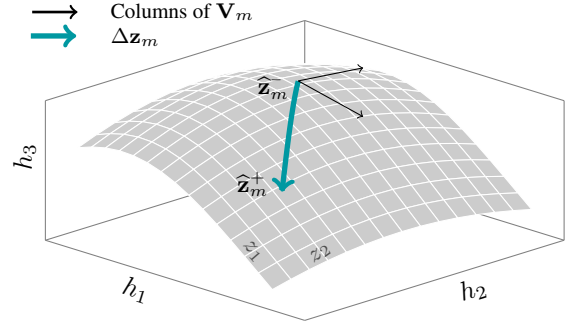


Fig. 1: Visualization of the measurement update. By updating only $\mathbf{z}$, the filter coefficients are confined to the manifold.

Since f_{dec} is implemented as a neural network, $\mathbf{V}_m$ can be computed along with the network output, requiring one additional matrix multiplication per network layer. This offers a significantly reduced computational complexity compared to the nonparametric approaches [21, 22], where a search for nearest neighbors and a PCA are necessary.

The linearization of (2) requires that the observation matrix $\mathbf{X}_m \mathbf{V}_m$ has to be used in the measurement update, yielding

$$\hat{\mathbf{Q}}_{\mathbf{e},m} = \mathbf{X}_m \mathbf{V}_m \mathbf{P}_m^- \mathbf{V}_m^T \mathbf{X}_m^T + \mathbf{Q}_{\mathbf{n},m} \quad (7a)$$

$$\mathbf{K}_m = \mathbf{P}_m^- \mathbf{V}_m^T \mathbf{X}_m^T \hat{\mathbf{Q}}_{\mathbf{e},m}^{-1} \quad (7b)$$

$$\Delta \mathbf{z}_m = \mathbf{K}_m (\mathbf{y}_m - \hat{\mathbf{d}}_m) \quad (7c)$$

$$\hat{\mathbf{z}}_m^+ = \hat{\mathbf{z}}_m^- + \Delta \mathbf{z}_m \quad (7d)$$

$$\mathbf{P}_m^+ = (\mathbf{I}_{l'} - \mathbf{K}_m \mathbf{X}_m \mathbf{V}_m) \mathbf{P}_m^- \quad (7e)$$

The matrix $\hat{\mathbf{Q}}_{\mathbf{e},m}$ can be understood as the covariance of the prediction error $\mathbf{e}_m = \mathbf{y}_m - \hat{\mathbf{d}}_m$, and $\mathbf{K}_m$ is the Kalman gain matrix. Fig. 1 depicts the measurement update.

2.2. Variational Autoencoders

To learn f_{dec} , VAEs can be employed. Although the algorithm is not restricted to a specific choice of f_{dec} , VAEs [27] have already been used successfully for manifold learning of acoustic impulse responses [18, 19] and performed best in preliminary experiments. In general, for neural autoencoders, f_{enc} and f_{dec} are neural networks that are jointly trained to minimize a reconstruction loss $\mathcal{L}(\mathbf{h}, \hat{\mathbf{h}})$. VAEs impose an additional constraint on the encoder: Instead of predicting $\mathbf{z}$, the encoder predicts the mean $\boldsymbol{\mu}$ and the logarithm of the variance of an l' -variate Gaussian distribution with diagonal covariance $\boldsymbol{\Sigma}$. From this distribution, one sample is drawn and fed to the decoder. To avoid that the encoder predicts $\boldsymbol{\Sigma}$ close to $\mathbf{0}$, the Kullback-Leibler Divergence (KLD) between the predicted distribution and a i.i.d. Gaussian distribution is added to the training loss. As a result, points that are close in the full space are also close in the latent space, which is meaningful for the intended application. When the KLD is weighted with a

constant scalar β , the model is called β -VAE [28]. Using a quadratic reconstruction loss, the total loss function reads

$$\mathcal{L}(\mathbf{h}, \hat{\mathbf{h}}) = \left\| \hat{\mathbf{h}} - \mathbf{h} \right\|^2 - \frac{\beta}{2} \left(\log \det \Sigma - \|\boldsymbol{\mu}\|^2 - \text{tr} \Sigma \right), \quad (8)$$

with $\det$ and tr being determinant and trace of Σ , respectively.

3. EXPERIMENTAL VALIDATION

3.1. Experimental Design

The aim of the experiment is to compare the proposed algorithm to related state-of-the-art algorithms under different conditions. We assess the performance in an AEC scenario by means of the Echo Return Loss Enhancement (ERLE)

$$\text{ERLE}_m = 10 \log_{10} \left(\frac{\|\mathbf{d}_m\|^2}{\|\mathbf{d}_m - \hat{\mathbf{d}}_m\|^2} \right) \text{ dB}. \quad (9)$$

Here, $\mathbf{d}_m$ is the corresponding frame of the actual echo signal, which is available in simulations. In the experiment we vary the subspace dimension l' in 8 steps between 10 and 2000. For each value of l' we use 20 different speech signals and five WGN signals of 10 s length as excitation. For the speech excitation signals we pick random recordings from the VCTK corpus [29], remove speech pauses with [30] and concatenate the speech parts to the desired length. We simulate two different scenarios: An abrupt change of the source position, and a quasi-continuous movement. We repeat every trial with the Echo-To-Noise Ratios (ENRs) -10 dB, 0 dB and 10 dB. The ENR is defined as $\mathbb{E}_m\{\|\mathbf{d}_m\|^2\} / \mathbb{E}_m\{\|\mathbf{n}_m\|^2\}$. For near-end interference, we simulate 25 WGN signals. This results in 150 trials per algorithm and value of l' . For the proposed approach we train 5 identical VAEs for each value of l' and average over all trials to be independent of the weight initialization.

3.2. RIR Dataset

A useful RIR database for the intended purpose is given by [31]. The authors have measured RIRs between a loudspeaker and multiple microphone arrays. They varied the loudspeaker position on a dense Cartesian grid, yielding $24 \cdot 19 \cdot 9 = 4104$ RIRs per microphone. For our experiments, we use the RIRs associated to the adjacent microphones 7, 8 and 9, and we choose $T_{60} = 0.1$ s. From the RIRs of microphone 8 we leave out one horizontal row consisting of the positions from (1, 3, 5) to (24, 3, 5) for testing. In our test case, the loudspeaker moves along this trajectory. More precisely, we simulate a 0.38 mm jump of the source after every frame (4 ms). We first interpolated the impulse responses linearly and sample-wise in time domain, yielding RIRs between 2500 virtual loudspeakers and microphone 8. The echo signal of the moving loudspeaker is finally obtained as proposed in [32].

3.3. Network Architecture and Training

Prior to the first network layer of f_{enc} , we subtract the global mean vector of the training set from each instance. We transform the result into frequency domain and split it into single-sided real and imaginary part. Although the network could learn this itself, we observed that this transformation favors the convergence during training. After this transformation we put four fully connected layers with weights exponentially decreasing between l and $2 \cdot l'$. The last layer predicts both $\boldsymbol{\mu}$ and $\log \text{diag} \Sigma$ as discussed in Section 2.2 and has a linear activation. For the first three layers we found that the swish activation yields the best results, which is in line with [18]. Statistical sampling was performed with the well known reparametrization trick [27]. The decoder consists of four layers, with the number of neurons exponentially increasing from l' to l . Again, the first three layers use the swish activation and the last layer uses linear activation. Finally, we transform the output of the last layer back into time domain and re-add the mean vector of the dataset. For network training we use the Nadam Optimizer [33] with an initial learning rate of $5 \cdot 10^{-4}$. We leave out 5% of the training data as validation set. When the relative system distance over this set does not decrease by at least 0.01 dB for 7 epochs we decrease the learning rate by a factor of $\sqrt{0.1}$ but not below 10^{-7} . Training terminates when the relative system distance does not decrease by at least 0.01 dB for 20 epochs. In (8) we chose $\beta = 5 \cdot 10^{-6}$.

3.4. Reference Algorithms

Linear: This is a simplification of the proposed approach. Instead of a nonlinear manifold, one linear subspace is learned by PCA. Equations in Section 2.1 still hold with f_{dec} being linear. The columns of $\mathbf{V}$ are the principal vectors of the training set and are not updated during runtime.

NN-PCA: This is the nonparametric variant of the proposed approach. Here, $\mathbf{V}$ is not the Jacobian of a function, but the principal vectors of the current filter state's nearest neighbors. For a complete description see [22].

Projection: This algorithm also computes subspaces based on nearest neighbors. A KF operates in full space and soft-projects the posterior estimate onto the subspace. A detailed description is given in [21]. A variant with reduced complexity using precomputed clusters is presented in [20].

TDKF: The block-time domain KF [25] arises when $\mathbf{V} = \mathbf{I}_l$.

FDKF: This is the renowned diagonalized KF in frequency domain [7] without postfilter.

3.5. Parameters and Implementation

We use a sampling rate of 16 kHz, $r = 64$ and $l = 2000$. The constant noise covariance $\mathbf{Q}_n$ is given as oracle knowledge. For all algorithms in time domain we set $\gamma = 1$ and estimate $\mathbf{Q}_{\delta,m} = \alpha \mathbf{Q}_{\delta,m-1} + (1 - \alpha) \text{diag}(\Delta \mathbf{z}_{m-1}^2)$ [34]. For the FDKF, we estimate $\mathbf{Q}_{\delta}$ as proposed in [7] and using $\gamma =$

0.9997. For recursive averaging we use $\alpha = 0.973$ which corresponds to a time constant of 150 ms. We initialize $\mathbf{P}_{\text{init}}$ and $\mathbf{z}_{\text{init}}$ (or $\mathbf{h}_{\text{init}}$) using covariance and mean of the training data. For the proposed variant, we use μ from the encoder f_{enc} to obtain a covariance estimate in latent space. For the FDKF, we set $\mathbf{P}_{\text{init}} = 0.1 \cdot \mathbf{I}$ in frequency domain and $\mathbf{h}_{\text{init}} = \mathbf{0}$, which is reasonable for the data at hand. For [21] and [22] we use the Euclidean distance metric and $2 \cdot l'$ nearest neighbors.

3.6. Results and Discussion

The results are depicted in Fig. 2. The left column shows the results for speech excitation and the right column shows the results for WGN. Each line in a diagram represents another algorithm in dependency of the subspace dimension. Each point represents the median ERLE for all excitation signals and for both moving source and abrupt change.

The results provide various insights: Overall, the ERLE is higher for speech excitation than for excitation with WGN. The reason is that for broadband WGN the entire impulse response needs to be identified. For speech excitation, the identification of low frequency components is sufficient for echo suppression. At low ENR the TDKF performs significantly better than the FDKF. This can be explained by the data-driven initialization of the TDKF.

It is surprising that the KF on a global subspace achieves a higher ERLE than the TDKF, when $l' = l$. This is unexpected because in this case, the global subspace approach operates in the very same space as the TDKF, only in a rotated coordinate system. A consecutive experiment confirmed that the reason for this difference is the estimation of the process noise $\mathbf{Q}_\delta$. We use the estimator from [34] which assumes that $\mathbf{Q}_\delta$ is diagonal. With a global subspace found by PCA, $\mathbf{P}_{\text{init}}$ is diagonal and the assumption is better justified.

For the methods based on nearest neighbors, a strong dependency of the subspace dimension is visible. This might be due to our choice of using $2 \cdot l'$ nearest neighbors for the computation of $\mathbf{V}$. Apparently, for speech excitation, a high number of nearest numbers yields better results. In this case, the set of nearest neighbors does not change a lot between time steps and neither does the subspace. In contrast, it can be observed that with WGN excitation, a small subspace dimension, and so a small number of nearest neighbors yields better results. Thus it can be said that for nonparametric ASI-approaches, the subspace dimension and the number of nearest neighbors is crucial.

Concerning the proposed approach, the results show that the influence of the subspace dimension is small, compared to the approaches based on nearest neighbors. Only for $l' = l$, the performance drops strongly. In this case, the VAE is no longer undercomplete, meaning that it is barely regularized and thus does not learn a meaningful representation. For small dimensions l' , the VAE represents the data well in most scenarios. We conclude that for online ASI on manifolds, it is

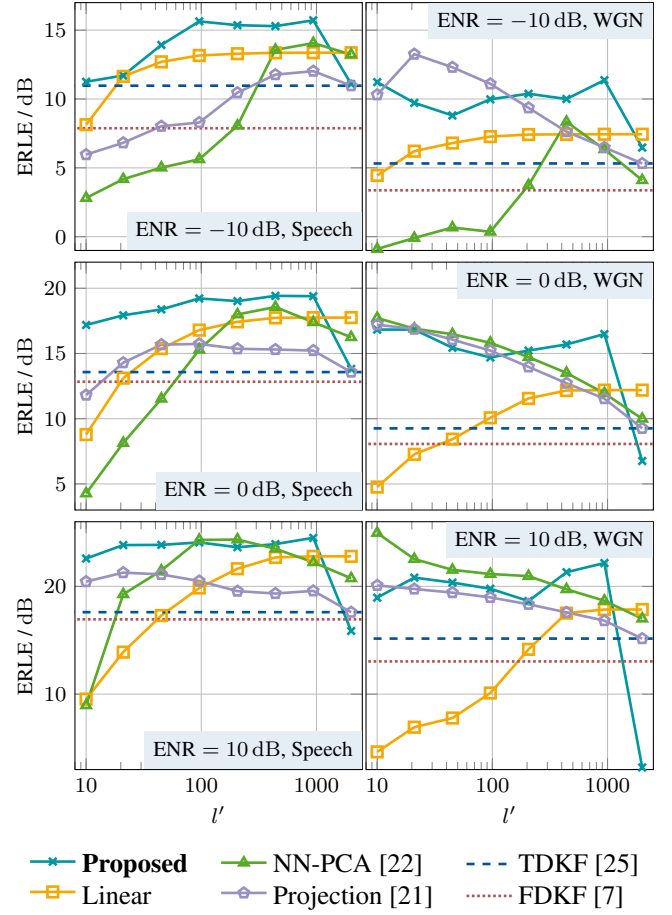


Fig. 2: Median ERLE depending on l'

favorable to compute one function whose tangent planes act as subspaces, rather than computing the subspaces locally during inference. For some configurations, the methods based on nearest neighbors outperform the proposed approach. At this point, we want to reiterate that the proposed algorithm exhibits a reduced computational complexity compared to the methods based on nearest neighbors. In a test run without GPU, using only one computation thread, and setting $l' = 200$, the proposed approach required two thirds less computation time than the NN-PCA approach.

4. SUMMARY AND OUTLOOK

In this paper we presented a novel algorithm for the online identification of time-variant systems. Instead of the impulse response itself, it tracks a lower dimensional representation, learned by a variational autoencoder. Compared to existing subspace approaches based on nearest neighbors, this offers a reduction of computational complexity, since local linearizations of the manifold can be computed more efficiently. In an experimental evaluation, the proposed approach performed as well as or better than state of art approaches.

References

- [1] T. Kabzinski and P. Jax, "An adaptive crosstalk cancellation system using microphones at the ears," in *Audio Engineering Society Convention*, 2019.
- [2] B. Schwartz, S. Gannot, and E. A. Habets, "An online dereverberation algorithm for hearing aids with binaural cues preservation," in *Workshop on Applications of Signal Processing to Audio and Acoustics*, 2015.
- [3] E. Bjarnason, "Analysis of the filtered-x lms algorithm," *Trans. on Speech and Audio Processing*, 1995.
- [4] J. Fabry, F. Konig, S. Liebich, and P. Jax, "Online secondary path estimation with masked auxiliary noise for active noise control," in *Int. Workshop on Acoustic Echo and Noise Control*, 2018.
- [5] 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," *Trans. on Speech and Audio Processing*, 1998.
- [6] G. Schmidt, "Applications of acoustic echo control - an overview," in *European Signal Processing Conf.*, 2004.
- [7] G. Enzner and P. Vary, "Frequency-domain adaptive kalman filter for acoustic echo control in hands-free telephones," *Signal Processing*, 2006.
- [8] B. Widrow and M. E. Hoff, "Adaptive switching circuits," *Tech. Rep.*, 1960.
- [9] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. of Basic Engineering*, 1960.
- [10] S. Makino, Y. Kaneda, and N. Koizumi, "Exponentially weighted stepsize nlms adaptive filter based on the statistics of a room impulse response," in *Trans. on Speech and Audio Processing*, 1993.
- [11] C. Antweiler, J. Grunwald, and H. Quack, "Approximation of optimal step size control for acoustic echo cancellation," in *Int. Conf. on Acoustics, Speech, and Signal Processing*, 1997.
- [12] H. Buchner, "A systematic approach to incorporate deterministic prior knowledge in broadband adaptive mimo systems," in *Asilomar Conf. on Signals, Systems and Computers*, 2010.
- [13] R. Talmon, I. Cohen, S. Gannot, and R. R. Coifman, "Diffusion maps for signal processing: A deeper look at manifold-learning techniques based on kernels and graphs," *Signal Processing Mag.*, 2013.
- [14] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "A study on manifolds of acoustic responses," in *Lecture Notes in Computer Science*, 2015.
- [15] 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.
- [16] S. S. Rangapuram, M. W. Seeger, J. Gasthaus, and et al., "Deep state space models for time series forecasting," in *Advances in Neural Information Processing Systems*, 2018.
- [17] I. Buchnik, D. Steger, G. Revach, R. J. G. van Sloun, T. Routenberg, and N. Shlezinger, "Learned kalman filtering in latent space with high-dimensional data," in *Int. Conf. on Acoustics, Speech, and Signal Processing*, 2023.
- [18] 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*, 2022.
- [19] 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*, 2023.
- [20] 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)*, 2020.
- [21] —, "Online acoustic system identification exploiting kalman filtering and an adaptive impulse response subspace model," *J. of Signal Processing Systems*, 2022.
- [22] T. Hardenbicker and P. Jax, "Fast tracking of time-variant systems using local affine subspaces," in *ITG Conf. on Speech Communication*, 2023.
- [23] 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.*, 2015.
- [24] S. Malik and G. Enzner, "Recursive bayesian control of multi-channel acoustic echo cancellation," *IEEE Signal Processing Letters*, 2011.
- [25] T. Kabzinski and P. Jax, "A unified perspective on time-domain and frequency-domain kalman filters for acoustic system identification," in *European Signal Processing Conf.*, 2022.
- [26] 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*, 1962.
- [27] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013.
- [28] 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*, 2017.
- [29] C. Veaux, J. Yamagishi, K. MacDonald et al., "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research*, 2017.
- [30] "py-webrtcvad," <https://github.com/wiseman/py-webrtcvad>.
- [31] J. Čmejla, T. Kounovský, S. Gannot, Z. Koldovský, and P. Tandeitnik, "Mirage: Multichannel database of room impulse responses measured on high-resolution cube-shaped grid," in *European Signal Processing Conf.*, 2021.
- [32] C. Urbanietz and G. Enzner, "Binaural rendering of dynamic head and sound source orientation using high-resolution hrtf and retarded time," in *Int. Conf. on Acoustics, Speech, and Signal Processing*, 2018.
- [33] T. Dozat, "Incorporating nesterov momentum into adam," 2016.
- [34] 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*, 2014.