MANIFOLD LEARNING-SUPPORTED ESTIMATION OF RELATIVE TRANSFER FUNCTIONS FOR SPATIAL FILTERING

Andreas Brendel, Johannes Zeitler, and Walter Kellermann

Multimedia Communications and Signal Processing, Friedrich-Alexander-Universität Erlangen-Nürnberg,
Cauerstr. 7, D-91058 Erlangen, Germany, e-mail: Andreas.Brendel@FAU.de

ABSTRACT

Many spatial filtering algorithms used for voice capture in, e.g., teleconferencing applications, can benefit from or even rely on knowledge of Relative Transfer Functions (RTFs). Accordingly, many RTF estimators have been proposed which, however, suffer from performance degradation under acoustically adverse conditions or need prior knowledge on the properties of the interfering sources. While state-of-the-art RTF estimators ignore prior knowledge about the acoustic enclosure, audio signal processing algorithms for teleconferencing equipment are often operating in the same or at least a similar acoustic enclosure, e.g., a car or an office, such that training data can be collected. In this contribution, we use such data to train Variational Autoencoders (VAEs) in an unsupervised manner and apply the trained VAEs to enhance imprecise RTF estimates. Furthermore, a hybrid between classic RTF estimation and the trained VAE is investigated. Comprehensive experiments with real-world data confirm the efficacy for the proposed method.

Index Terms— Manifold learning, variational autoencoder, relative transfer function, spatial filtering, unsupervised learning

1. INTRODUCTION AND SIGNAL MODEL

Teleconferencing is an essential technology for modern work places that recently became even more important, e.g., due to the increasing number of people working from home during pandemic situations. An integral aspect of such systems is the enhancement of the voices of the conversation partners under adverse acoustic conditions, i.e., the suppression of undesired sources like people talking in the background. As most of the commonly used teleconferencing devices are equipped with multiple microphones, spatial filtering is an obvious algorithmic choice of maximizing user comfort.

In this paper we consider a source signal observed by a pair of microphones in a reverberant and noisy environment. The Short-Time Fourier Transform (STFT)-domain microphone observation $x_{m,ft}$ at frequency index $f \in \{1,\dots,F\}$, time frame index $t \in \{1,\dots,T\}$ and microphone index $m \in \{1,2\}$ is modeled by

$$x_{m,ft} = a_{m,f} s_{ft} + n_{m,ft} =: c_{m,ft} + n_{m,ft} \tag{1}$$

where $a_{m,f}$, s_{ft}, $c_{m,ft}$ and $n_{m,ft}$ denote the Acoustic Transfer Function (ATF), the anechoic source signal, the reverberant source signal at the microphone (spatial image), and the m-th additive noise signal comprising background noise and interfering signals, respectively. Note that a two-microphone array is chosen for simplicity which does not limit generality of the proposed approach. Knowledge of ATFs enables the construction of several powerful spatial filtering approaches including MPDR, MVDR and LCMV beamformers [1, 2]. However, as the estimation of ATFs is very difficult in practice, Relative Transfer Functions (RTFs) [3, 4] describing the relation between spatial images are employed for spatial filtering instead, as they can directly be estimated from the microphone signals during time intervals when noise and interference are negligibly small. By taking (without loss of generality) the spatial image of the first microphone $c_{1,ft}$ as reference, the spatial image at the second microphone can be expressed via the RTF h_f as $c_{2,ft} = h_f c_{1,ft}$. This allows to describe both microphone signals in vector form by

$$\begin{bmatrix} x_{1,ft} \\ x_{2,ft} \end{bmatrix} = \begin{bmatrix} 1 \\ h_f \end{bmatrix} c_{1,ft} + \begin{bmatrix} n_{1,ft} \\ n_{2,ft} \end{bmatrix} \in \mathbb{C}^2. \tag{2}$$

Simple Least Squares (LS) estimators suffer from being biased when additive noise or interfering sources are present [5]. To address this problem, RTF estimators that rely on knowledge of the statistical properties of additive noise [6] or employ specialized noise estimators [7] have been proposed. Also deep learning-based methods have been applied for selecting useful frequency bins for RTF estimation [8, 9]. When multiple point sources, e.g., speakers, are present, source separation methods like directionally constrained Blind Source Separation (BSS) [10, 11] or simplex analysis [12] have been applied for RTF estimation. However, under acoustically adverse conditions, RTF estimation remains a challenging task and the resulting estimates may suffer from measurement errors which reduce the performance of subsequent spatial filtering tasks.

Often, telecommunication devices are always used within the same acoustic environment, e.g., a hands-free communication system in a car or a teleconferencing toolkit built in a personal computer. Under this assumption, data collected under benign conditions can be used to train a model describing this acoustic environment. This idea has recently been used to learn dictionaries for modeling RTFs [13, 14] or ATFs [15, 16, 17]. While these dictionary-based methods represent linear models, it has been observed that RTFs are well-modeled by a nonlinear manifold [18, 19], i.e., the difference between RTFs is well-described by a nonlinear relation defined by a much lower number of degrees of freedom than the number of RTF filter taps. A manifold learning method for improving RTF estimates based on diffusion maps has been proposed [20, 21] and a semi-supervised deep learning approach to infer RTFs from source positions has been developed in [22]. A Variational Autoencoder (VAE) has been leveraged for source localization from measured RTFs in [23, 24].

In this paper, we propose a method of deep learning-based manifold learning for improving noisy RTF estimates, which is not restricted to be linear such as dictionary-based methods and is based on a highly expressive model represented by a VAE. To this end, a

This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 282835863 – within the Research Unit FOR2457 Acoustic Sensor Networks.

VAE unsupervisedly trained with data collected under benign acoustic conditions to reconstruct RTFs of the considered enclosure is used to enhance inaccurate RTFs measured under acoustically adverse conditions. Furthermore, a hybrid of classic RTF estimation and the trained VAE is proposed. As an additional baseline assuming knowledge about the interfering signals, the trained VAE is fine-tuned to take the noise conditions into account. Comprehensive experiments with real-world data featuring various acoustic conditions confirm the superiority of the proposed method relative to several baselines.

2. MANIFOLD LEARNING-BASED RTF RECONSTRUCTION

In Sec. 2.1, the VAE concept is introduced for the given problem and the training target for the considered VAE structure is introduced, which is used in Sec. 2.2 for RTF reconstruction.

2.1. Variational Autoencoder

Autoencoders (AEs) [25] represent a powerful approach to unsupervised dimensionality reduction, i.e., to learn essential data representations and ignore insignificant data components. To this end, a manifold representation of the input data is generated, accomplished by training a Deep Neural Network (DNN) comprising an encoder and a decoder which are connected by a thin layer (the bottleneck) to reconstruct its input at its output. The output of the encoder, the embedding $\mathbf{z} \in \mathbb{R}^d$, constitutes a low-dimensional representation of the input broadband RTF described by real and imaginary parts

$$\mathbf{h} \coloneqq \big[\underbrace{\mathrm{Re}\{h_1\}, \ldots, \mathrm{Re}\{h_F\}}_{\mathbf{h}_{\mathrm{Re}}}, \underbrace{\mathrm{Im}\{h_1\}, \ldots, \mathrm{Im}\{h_F\}}_{\mathbf{h}_{\mathrm{Im}}}\big]^{\mathsf{T}} \in \mathbb{R}^D, \tag{3}$$

where $2F = D \gg d$. To enforce structured embeddings, i.e., similar input data always correspond to similar embedding representations, VAEs [26] have been proposed that employ stochastic encoders and decoders. Here, instead of trying to directly identify a (generally intractable) generative model $p_{\boldsymbol{\theta}}(\mathbf{h})$ parameterized by the parameter vector $\boldsymbol{\theta}$, the Evidence Lower Bound Objective (ELBO)

$$\mathcal{L}_{\mathrm{VAE}}(\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{h}) \coloneqq \mathcal{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{h})} \log p_{\boldsymbol{\theta}}(\mathbf{h}|\mathbf{z}) - \mathcal{KL}\left\{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{h}) \| p(\mathbf{z})\right\}, \tag{4}$$

which bounds the log-likelihood of $p_{\boldsymbol{\theta}}(\mathbf{h})$ from below, is maximized w.r.t. $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ parameterizing the encoder and decoder, respectively. The second part of (4) represents the Kullback-Leibler Divergence (KLD) between the prior $p(\mathbf{z})$ on the embedding $\mathbf{z}$ and the encoder distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{h})$, which we choose to be normally distributed as

$$p(\mathbf{z}) \coloneqq \mathcal{N}(\mathbf{0}_d, \mathbf{I}_{d\times d}) \quad \text{and} \quad q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{h}) \coloneqq \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{5}$$

respectively. The encoder distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{h})$, which represents a variational approximation of the true posterior of $\mathbf{z}$ is parameterized by a mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and a diagonal covariance matrix $\boldsymbol{\Sigma} \coloneqq \mathrm{diag}\{\boldsymbol{v}\} \in \mathbb{R}^{d\times d}$. Both parameter vectors, $\boldsymbol{\mu}$ and $\boldsymbol{v}$, are estimated from the input $\mathbf{h}$ with an encoder network (the application of the logarithm $\log \boldsymbol{v} \in \mathbb{R}^d$ is meant element-wise)

$$\mathtt{enc}_{\boldsymbol{\phi}} : \mathbb{R}^D \to \mathbb{R}^{2d} \quad \text{with} \quad (\boldsymbol{\mu}, \log \boldsymbol{v}) \coloneqq \mathtt{enc}_{\boldsymbol{\phi}}(\mathbf{h}). \tag{6}$$

Hence, by maximizing (4), the KLD term enforces the variational approximation $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{h})$ to stay close to the prior $p(\mathbf{z})$. The first term of (4) is an expectation of the decoder distribution, which is chosen to be Gaussian

$$p_{\boldsymbol{\theta}}(\mathbf{h}|\mathbf{z}) \coloneqq \mathcal{N}\Big(\tilde{\mathbf{h}}, \sigma^2\, \mathbf{I}_{D\times D}\Big), \tag{7}$$

where the mean vector $\tilde{\mathbf{h}} \in \mathbb{R}^D$ is estimated from the embedding $\mathbf{z}$ by the decoder network

$$\mathtt{dec}_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}^D \quad \text{with} \quad \tilde{\mathbf{h}} \coloneqq \mathtt{dec}_{\boldsymbol{\theta}}(\mathbf{z}) \tag{8}$$

and σ^2 is a hyperparameter (we use $\sigma^2 = 0.5$ in the following). When maximizing the ELBO, the first term in (4) enforces accurate reconstruction of the input at the output of the VAE.

With the choices made for the encoder and decoder distribution, we obtain the following cost function for training the VAE

$$\begin{aligned}\mathcal{J}_{\mathrm{VAE}}(\boldsymbol{\phi}, \boldsymbol{\theta}) \coloneqq \gamma\, \frac{\hat{\mathcal{E}}_b\big\{\|\mathbf{h}_b - \tilde{\mathbf{h}}_b\|_2^2\big\}}{\hat{\mathcal{E}}_b\left\{\|\mathbf{h}_b\|_2^2\right\}} \cdots \\ \cdots - \frac{1-\gamma}{2d}\, \hat{\mathcal{E}}_b\left\{\log\det \boldsymbol{\Sigma}_b - \|\boldsymbol{\mu}_b\|_2^2 - \mathrm{tr}\,\boldsymbol{\Sigma}_b\right\}.\end{aligned} \tag{9}$$

Here, weights for the two terms of $\mathcal{J}_{\mathrm{VAE}}$ resulting from the corresponding terms in (4) are added: A convex weight $\gamma \in [0, 1]$ trades off reconstruction performance against the KLD term (cf. [27]). Note that $\gamma = 1$ will reduce the model trained according to $\mathcal{J}_{\mathrm{VAE}}$ to a deterministic AE. To be independent of the input data's energy and vector lengths, we normalize the reconstruction term by the squared norms of the input vectors $\|\mathbf{h}_b\|_2^2$ and the KLD term by the dimension of the embedding vectors d. For robust gradient-descent-based optimization, averages $\hat{\mathcal{E}}_b\{\cdot\} \coloneqq \frac{1}{B}\sum_{b=1}^{B}(\cdot)$ over batches with elements indexed by $b \in \{1, \ldots, B\}$ are incorporated into $\mathcal{J}_{\mathrm{VAE}}$.

2.2. RTF Reconstruction

For RTF estimation, we choose the well-known estimator [5, 3, 7]

$$\hat{h}_f \coloneqq \frac{\hat{\mathcal{E}}_t\{S_{11,ft}S_{12,ft}\} - \hat{\mathcal{E}}_t\{S_{11,ft}\}\hat{\mathcal{E}}_t\{S_{12,ft}\}}{\hat{\mathcal{E}}_t\{S_{11,ft}^2\} - \hat{\mathcal{E}}_t\{S_{11,ft}\}^2}, \tag{10}$$

which showed superior performance in our experiments w.r.t. simple LS-based methods. Here, $S_{11,ft} \coloneqq |x_{1,ft}|^2$ and $S_{12,ft} \coloneqq x_{1,ft}x_{2,ft}^*$ denote instantaneous estimates of the auto and cross Power Spectral Density (PSD) of the observed signals, respectively, and $\hat{\mathcal{E}}_t$ a time average. Under acoustically adverse conditions, the estimated RTFs will be noisy and their usefulness for spatial filtering degrades. To enhance these RTF estimates, we propose two approaches based on the trained VAE model in the following.

Denoising VAE (DN): As the encoder network $\mathtt{enc}_{\boldsymbol{\phi}}$ was trained to represent the most-essential components of an RTF, measurement noise that cannot be considered as a typical RTF ingredient will not be encoded by $\mathtt{enc}_{\boldsymbol{\phi}}$. Hence, applying encoder and decoder successively to an estimated RTF $\hat{\mathbf{h}}$ (defined analogously to (3))

$$\tilde{\mathbf{h}} = \mathtt{dec}_{\boldsymbol{\theta}}(\boldsymbol{\mu}) \quad \text{with} \quad (\boldsymbol{\mu}, \cdot) = \mathtt{enc}_{\boldsymbol{\phi}}\big(\hat{\mathbf{h}}\big) \tag{11}$$

will enhance the estimate $\hat{\mathbf{h}}$ by removing the measurement noise.

VAE-based LS estimator (LS): The RTF reconstruction with DN only exploits a previously obtained RTF estimate and neglects the observed signals for its enhancement. In the following, we develop a reconstruction approach that takes both, an RTF estimate and the observed signals, into account. By ignoring the additive noise in (2), the RTF between the microphones could precisely be estimated by an LS approach. To compensate for neglecting the noise, the reconstructed RTF $\tilde{\mathbf{h}}$ is confined to the manifold learned by the VAE

$$\tilde{\mathbf{h}} = \Big[\tilde{\mathbf{h}}_{\mathrm{Re}}^{\mathsf{T}}(\mathbf{z}), \tilde{\mathbf{h}}_{\mathrm{Im}}^{\mathsf{T}}(\mathbf{z})\Big]^{\mathsf{T}} = \mathtt{dec}_{\boldsymbol{\theta}}(\mathbf{z}) \quad \text{with} \quad \tilde{\mathbf{h}}_{\mathrm{Re}}(\mathbf{z}), \tilde{\mathbf{h}}_{\mathrm{Im}}(\mathbf{z}) \in \mathbb{R}^F.$$

The reconstructed RTF $\tilde{\mathbf{h}}$ can now be determined by minimizing the LS cost function

$$\mathcal{J}_{\mathrm{LS}}(\mathbf{z}, \mathbb{X}) \coloneqq \sum_{t=1}^{T} \Big\|\mathbf{x}_{1,t} \odot \Big(\tilde{\mathbf{h}}_{\mathrm{Re}}(\mathbf{z}) + \jmath\tilde{\mathbf{h}}_{\mathrm{Im}}(\mathbf{z})\Big) - \mathbf{x}_{2,t}\Big\|_2^2 \tag{12}$$

	layer name	output shape	activation
$\texttt{enc}_{\boldsymbol{\phi}}$	`input`: $\mathbf{h}_b$	$B \times D$	-
	`encoder1`	$B \times 256$	swish
	`encoder2`	$B \times 128$	swish
	`encoder3`	$B \times 64$	swish
	`postPar`: $\boldsymbol{\mu}_b, \log \boldsymbol{v}_b$	$B \times d \times 2$	linear
	`sampling`: $\mathbf{z}_b$	$B \times d$	-
$\texttt{dec}_{\boldsymbol{\theta}}$	`decoder1`	$B \times 64$	swish
	`decoder2`	$B \times 128$	swish
	`decoder3`	$B \times 256$	swish
	`output`: $\tilde{\mathbf{h}}_b$	$B \times D$	linear

Table 1. Proposed VAE architecture, where B denotes the batch size and $\mathbf{h}_b, \tilde{\mathbf{h}}_b \in \mathbb{R}^D$, $\mathbf{z}_b, \boldsymbol{\mu}_b, \boldsymbol{v}_b \in \mathbb{R}^d$.

subject to $\mathbf{z}$. Here, we introduced the broadband observed signal vector $\mathbf{x}_{m,t} := [x_{m,1t}, \ldots, x_{m,Ft}]^\mathrm{T}$, $m \in \{1,2\}$, the set of observations $\mathbb{X} := \{\mathbf{x}_{m,t} \in \mathbb{C}^F | m \in \{1,2\}, t \in \{1,\ldots,T\}\}$ and the Hadamard product $\odot$. The LS cost function (12) is iteratively minimized by gradient descent ($\texttt{dec}_{\boldsymbol{\theta}}$ is differentiable)

$$\boldsymbol{\mu}_{\mathrm{LS}} \leftarrow \boldsymbol{\mu}_{\mathrm{LS}} - \frac{\alpha}{\sum_{t=1}^{T} \|\mathbf{x}_{1,t}\|_2^2} \frac{\mathrm{d}\mathcal{J}_{\mathrm{LS}}(\mathbf{z}, \mathbb{X})}{\mathrm{d}\mathbf{z}}\bigg|_{\mathbf{z}=\boldsymbol{\mu}_{\mathrm{LS}}}, \qquad (13)$$

where $\alpha \in \mathbb{R}_+$ denotes a step size. For stable convergence the gradient is normalized by the observed signal energy and the iterative process is initialized with the DN solution $(\boldsymbol{\mu}_{\mathrm{LS}}, \cdot) = \texttt{enc}_{\boldsymbol{\phi}}(\hat{\mathbf{h}})$. After termination of the iterative optimization (13), the reconstructed RTF is calculated by $\tilde{\mathbf{h}}_{\mathrm{LS}} = \texttt{dec}_{\boldsymbol{\theta}}(\boldsymbol{\mu}_{\mathrm{LS}})$.

3. EXPERIMENTAL SETUP

In the following section, the experimental setup, the proposed VAE architecture, its training as well as the realization of the considered algorithmic variants is discussed.

Datasets: To demonstrate the performance of the proposed method under real-world acoustic conditions, RTFs obtained from the MIRaGe dataset [28] containing measurements from a varechoic lab of dimensions $6\,\mathrm{m} \times 6\,\mathrm{m} \times 2.4\,\mathrm{m}$ at a reverberation time of $T_{60} \in \{0.1\,\mathrm{s}, 0.3\,\mathrm{s}, 0.6\,\mathrm{s}\}$ and speech signals from the ACE dataset [29] are used. To simulate noise and interferers present in real-world acoustic scenes, recorded noise signals including, e.g., meeting, factory, fan or vacuum cleaner noise, from the ACE [29] and NOISEX database [30], as well as own recordings, have been used. The MIRaGe dataset contains recorded White Gaussian Noise (WGN) signals from a loudspeaker placed at grid positions of $2\,\mathrm{cm}$ spacing in x and y direction and $4\,\mathrm{cm}$ in z direction within a cube of dimensions $46\,\mathrm{cm} \times 36\,\mathrm{cm} \times 32\,\mathrm{cm}$ at a height of $1.15\,\mathrm{m}$ of the grid center. The loudspeaker signal has been recorded by several spatially distributed microphone arrays from which we choose a microphone pair at $2\,\mathrm{m}$ distance from the grid center and same height with a spacing of $10\,\mathrm{cm}$. Additionally, several measurements at Out-Of-Grid Positions (OOGPs) at $1\,\mathrm{m}$ distance from the walls are available, which will be used for synthesizing additive noise signals. The dataset for training and evaluation of the proposed methods is created by estimating RTFs of $D = 256$ taps length directly from the recorded WGN signals, i.e., under optimal conditions. From all available 4104 grid positions, we randomly select $N = 200$ for the test set and 100 for the validation set. For data augmentation, the RTFs corresponding to the remaining 3804 grid positions are repeated five times by adding WGN with $1\,\%$ of the average RTF variance yielding the training set.

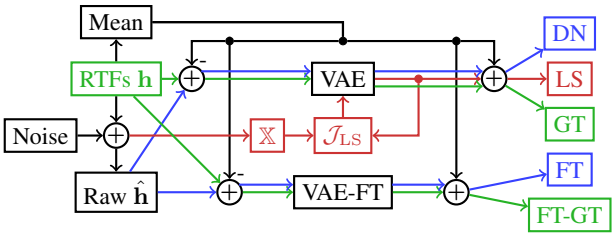

Fig. 1. Overview of the considered algorithmic variants.

Microphone signals: To generate reverberant observations, we use a microphone placed closely to the loudspeaker representing the source during the measurements and estimate the ATFs to the considered microphone pair. The microphone signals are obtained by filtering anechoic speech signals of $10\,\mathrm{s}$ duration from the ACE dataset with the previously estimated ATFs and adding noise signals at a desired Signal-to-Noise Ratio (SNR). Various types of additive noise are considered: WGN, interfering speech or recorded noise signals at a specific OOGP, i.e., a Point Source (PS), and noise signals at multiple available OOGPs, i.e., approximating a diffuse noise source. The interfering speech signals (not used for the desired source) are taken from the ACE database [29].

VAE network and training: The proposed VAE architecture summarized in Tab. 1 takes minibatches $[\mathbf{h}_1, \ldots, \mathbf{h}_B]^\mathrm{T} \in \mathbb{R}^{B \times D}$ of batch size $B = 128$ as input with elements $\mathbf{h}_b \in \mathbb{R}^D$ randomly chosen from the training set. The inputs are processed with three fully-connected layers and swish activation [31] followed by a linear layer `postPar` for the estimation of the parameters of the encoder distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{h})$. To enable backpropagation through the network, `sampling` realizes the 'reparameterization trick' [26]

$$\mathbf{z}_b = \boldsymbol{\mu}_b + \exp\left(\frac{\log \boldsymbol{v}_b}{2}\right) \odot \mathbf{e}_b \quad \text{with} \quad \mathbf{e}_b \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_{d\times d}). \qquad (14)$$

The bottleneck dimension was empirically chosen as $d = 5$. Again, log and exp in (14) denote element-wise operations. The decoder is symmetric to the encoder structure resulting in $\approx 2.15 \cdot 10^5$ trainable parameters in total. To avoid learning of RTF components common to all RTFs in the dataset, their mean is subtracted, the VAE is trained on the residual and the mean RTF is added to the output of the VAE again for reconstruction. The network is trained by minimizing (9) with $\gamma = 0.95$ by ADAM [32] with an initial learning rate of 10^{-3} which is reduced by a factor of five to avoid getting stuck with the training process if the validation loss did not improve by at least 10^{-3} within the last five epochs. To avoid overfitting, early stopping is employed and the network parameters of the epoch with lowest validation loss are restored if the validation loss did not improve by at least 10^{-3} within the last ten epochs.

In the remainder of the paper, we discuss and experimentally evaluate the following algorithmic variants illustrated in Fig. 1:

Raw RTF estimation (Raw): All algorithmic variants are based on RTFs estimated by (10). Hence, these raw RTF estimates represent the first baseline for the experimental comparison.

Mean of dataset (Mean): The VAE learns an RTF representation as a refinement of the average of all RTFs in the training set. Hence, this mean RTF represents another natural baseline.

Denoising VAE (DN): The estimated RTFs are denoised by (11).

VAE-based LS estimator (LS): The RTF is estimated by minimizing $\mathcal{J}_{\mathrm{LS}}$ by (13) with $\alpha = 2$. Here, only 20 iterations are executed to avoid overfitting to the interfering signals.

Denoising with fine-tuned VAE (FT): Assuming that pairs of estimated noisy and clean RTFs are available, the VAE can be trained

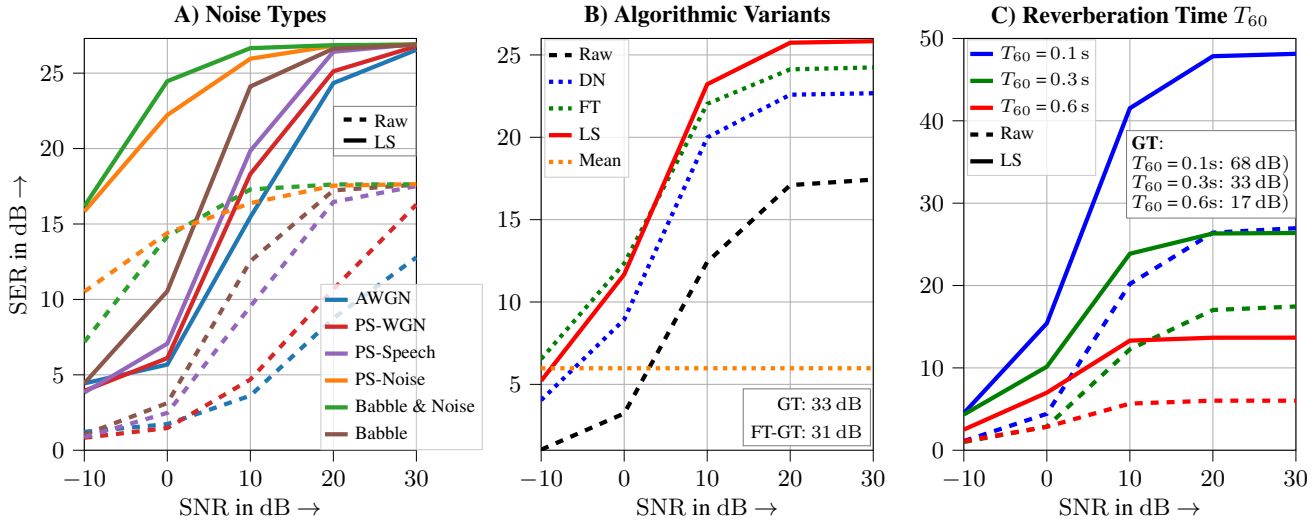

Fig. 2. **A)** Influence of various noise types on RTF estimation and the proposed LS approach. **B)** Comparison of all considered algorithmic variants. **C)** Influence of T_{60} on RTF estimation and the proposed LS approach.

to map noisy RTFs to clean ones instead of reconstructing the input RTF at the output. In this way the typical deterioration of estimated RTFs is taken into account. To this end, we fine-tune the VAE trained on clean RTFs by continuing training with a set of noisy and clean RTFs for additional 15 epochs with the Adam optimizer with a learning rate of 10^{-4}. However, such a training set is rarely given in practice and, hence, FT should be considered as an oracle baseline.

Decoding of ground truth RTF (GT): As an upper bound on the expected performance, the Ground Truth (GT) RTF is reconstructed by the VAE similar to (11): $\tilde{\mathbf{h}} = \mathtt{dec}_{\boldsymbol{\theta}}(\boldsymbol{\mu})$ with $(\boldsymbol{\mu}, \cdot) = \mathtt{enc}_{\boldsymbol{\phi}}(\mathbf{h})$. In this way, the modeling capability of the VAE is evaluated as an upper bound for comparison to the proposed methods. Similarly, we denote the reconstruction of the GT RTF with FT, as **FT-GT**.

The performance of all investigated RTF estimation methods is measured by the Signal-to-Error Ratio (SER) (larger values correspond to better performance)

$$\mathrm{SER} \coloneqq \frac{10}{N} \sum_{n=1}^{N} \log_{10} \frac{\|\mathbf{h}_n\|_2^2}{\|\mathbf{h}_n - \tilde{\mathbf{h}}_n\|_2^2}, \tag{15}$$

where the samples of the test set are indexed by $n \in \{1, \ldots, N\}$.

4. RESULTS

The experimental results corresponding to the methodology described in Sec. 3 are shown in Fig. 2 and are discussed below:

A) Noise Types: On the left of Fig. 2, LS and Raw are evaluated for various noise types at $T_{60} = 0.3\,\mathrm{s}$, where it can be seen that LS achieves an improvement over Raw up to about 10 dB for some scenarios. WGN added to the microphone signals (AWGN) caused the worst performance in these experiments. As a second kind of noise, randomly placed PSs for the evaluation of each of the $N = 200$ test RTFs, which emitted either WGN, speech or a recorded noise signal, simulate scenarios with an interfering source, i.e., an undesired speaker or a compact noise source in the background. Here, LS achieves similar results for WGN and speech, which are also close to the results for AWGN. An interfering noise PS was less detrimental in our experiments. To simulate ambient noise, we place in a third set of experiments several PSs at OOGPs emitting speech or recorded noise signals. We denote these experimental conditions by 'Babble' if only speech signals are used and 'Babble & Noise' if a set containing speech and recorded noise signals is distributed over the OOGPs. Here, the 'Babble & Noise' performs similar to a PS noise interferer and 'Babble' shows slightly better but similar performance than a PS speech interferer.

B) Algorithmic Variants: In the middle of Fig. 2, the different algorithmic variants described in Sec. 3 are compared experimentally for $T_{60} = 0.3\,\mathrm{s}$ and additive 'Babble' noise. The SER of the raw RTF estimate is increased by DN, FT and LS, where DN yields the lowest improvements. While slightly worse than FT for very low SNRs, LS yields the best results of the considered enhancement approaches above $\mathrm{SNR} = 10\,\mathrm{dB}$. However, it should be noted that while FT assumes knowledge about estimation errors as outlined in Sec. 3, LS does not need such prior knowledge. DN, FT and LS show better performance than the trivial baseline represented by the mean of the RTF data set (Mean) whereas Raw becomes better than Mean for SNRs above 0 dB. The SER of the reconstructed GT RTFs is slightly lower for FT (31 dB) than without fine tuning (33 dB), which is to be expected as FT is optimized for noisy input and clean output RTFs and not for perfect reconstruction of RTFs.

C) Reverberation Time T_{60}: On the right of Fig. 2 the achieved SER of LS and Raw is shown for additive 'Babble' noise and varying T_{60}. LS significantly improves the SER of Raw for all T_{60}, where the largest improvements are obtained for $T_{60} = 0.1\,\mathrm{s}$. The overall RTF estimation performance as well as the GT decreases with increasing T_{60}, which is to be expected as the more and more complex RTFs are modeled with a filter of same length D.

5. CONCLUSION

In this contribution, we propose a VAE-based manifold model for RTFs and leverage it for enhancing RTF estimates. We show that the proposed LS-based RTF estimator regularized by the trained VAE increases the RTF quality relative to VAE-based denoising of the RTF estimates. All experiments are conducted with measured data which emphasizes the real-world applicability of the proposed method. As next steps, we will combine and evaluate the proposed method with spatial filtering algorithms. A further improvement over the proposed method is expected by employing complex-valued networks and estimators for the statistics of additive noise.

6. REFERENCES

[1] H. L. Van Trees, *Optimum array processing*, Number 4 in Detection, estimation, and modulation theory. Wiley, New York, USA, 2002.

[2] S. Gannot et al., "A Consolidated Perspective on Multi-Microphone Speech Enhancement and Source Separation," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 25, no. 4, pp. 692–730, Jan. 2017.

[3] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. on Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[4] E. A. P. Habets et al., "On the application of the LCMV beamformer to speech enhancement," in *IEEE WASPAA*, New Paltz, NY, USA, Oct. 2009, pp. 141–144.

[5] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE Trans. on Signal Process.*, vol. 44, no. 8, pp. 2055–2063, Aug. 1996.

[6] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *IEEE ICASSP*, South Brisbane, Queensland, Australia, Apr. 2015, pp. 544–548.

[7] I. Cohen, "Relative Transfer Function Identification Using Speech Signals," *IEEE Trans. on Speech and Audio Process.*, vol. 12, no. 5, pp. 451–459, Sept. 2004.

[8] S. E. Chazan, J. Goldberger, and S. Gannot, "DNN-Based Concurrent Speakers Detector and its Application to Speaker Extraction with LCMV Beamforming," in *IEEE ICASSP*, Calgary, AB, Canada, Apr. 2018, pp. 6712–6716.

[9] J. Málek, Z. Koldovský, and M. Bohac, "Blockonline multichannel speech enhancement using deep neural networksupported relative transfer function estimates," *IET Signal Process.*, vol. 14, no. 3, pp. 124–133, May 2020.

[10] K. Reindl et al., "Geometrically Constrained TRINICON-based relative transfer function estimation in underdetermined scenarios," in *IEEE WASPAA*, New Paltz, NY, USA, Oct. 2013.

[11] A. Brendel, T. Haubner, and W. Kellermann, "A Unified Probabilistic View on Spatially Informed Source Separation and Extraction Based on Independent Vector Analysis," *IEEE Trans. on Signal Process.*, vol. 68, pp. 3545–3558, 2020.

[12] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Source Counting and Separation Based on Simplex Analysis," *IEEE Trans. on Signal Process.*, vol. 66, no. 24, pp. 6458–6473, Dec. 2018.

[13] Z. Koldovský et al., "Semi-Blind Noise Extraction Using Partially Known Position of the Target Source," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 21, no. 10, pp. 2029–2041, Oct. 2013.

[14] Z. Koldovský and S. Gannot, "Dictionary-Based Sparse Reconstruction of Incomplete Relative Transfer Functions," in *European Signal Process. Conf. (EUSIPCO)*, Dublin, Ireland, Aug. 2021.

[15] T. Koren, R. Talmon, and I. Cohen, "Supervised system identification based on local PCA models," in *IEEE ICASSP*, Kyoto, Japan, Mar. 2012, pp. 541–544.

[16] M. Fozunbal, T. Kalker, and R. W. Schafer, "Multi-channel echo control by model learning," in *Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle, WA, USA, 2008.

[17] T. Haubner, A. Brendel, and W. Kellermann, "Online Supervised Acoustic System Identification Exploiting Prelearned Local Affine Subspace Models," in *IEEE MLSP*, Espoo, Finland, Sept. 2020.

[18] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "A Study on Manifolds of Acoustic Responses," in *Int. Conf. on Latent Variable Analysis and Signal Seperation (LVA/ICA)*, pp. 203–210. Liberec, Czech Republic, Aug. 2015.

[19] A. Deleforge, F. Forbes, and R. Horaud, "Acoustic Space Learning for Sound-Source Separation and Localization on Binaural Manifolds," *Int. J. of Neural Syst.*, vol. 25, no. 01, Feb. 2015.

[20] R. Talmon and S. Gannot, "Relative transfer function identification on manifolds for supervised GSC beamformers," in *EUSIPCO*, Marrakech, Morocco, Sept. 2013.

[21] A. Sofer et al., "Robust Relative Transfer Function Identification on Manifolds for Speech Enhancement," in *European Signal Process. Conf. (EUSIPCO)*, Dublin, Ireland, Aug. 2021.

[22] Z. Wang et al., "Semi-Supervised Learning with Deep Neural Networks for Relative Transfer Function Inverse Regression," in *IEEE ICASSP*, Calgary, AB, Canada, Apr. 2018, pp. 191–195.

[23] M. J. Bianco, S. Gannot, and P. Gerstoft, "Semi-Supervised Source Localization with Deep Generative Modeling," in *IEEE MLSP*, Espoo, Finland, Sept. 2020.

[24] M. J. Bianco et al., "Semi-supervised source localization in reverberant environments with deep generative modeling," *arXiv:2101.10636 [cs, eess]*, Apr. 2021.

[25] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[26] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *Int. Conf. on Learning Representations (ICLR)*, Banff, Canada, Dec. 2013.

[27] I. Higgins et al., "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," in *Int. Conf. on Learning Representations (ICLR)*, Toulon, France, Apr. 2017.

[28] J. Čmejla et al., "MIRaGe: Multichannel Database of Room Impulse Responses Measured on High-Resolution Cube-Shaped Grid," in *European Signal Process. Conf. (EUSIPCO)*, Amsterdam, The Netherlands, Jan. 2021, pp. 56–60.

[29] J. Eaton et al., "Estimation of Room Acoustic Parameters: The ACE Challenge," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 24, no. 10, pp. 1681–1693, Oct. 2016.

[30] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, July 1993.

[31] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for Activation Functions," *arXiv:1710.05941 [cs]*, Oct. 2017.

[32] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Jan. 2017.