Semi-supervised learning with deep neural networks for relative transfer function inverse regression

Ziteng Wang, Junfeng Li, Yonghong Yan, Emmanuel Vincent

SEMI-SUPERVISED LEARNING WITH DEEP NEURAL NETWORKS FOR RELATIVE TRANSFER FUNCTION INVERSE REGRESSION

Ziteng Wang, Junfeng Li, Yonghong Yan*

University of Chinese Academy of Sciences
Institute of Acoustics, Beijing, China

Emmanuel Vincent

Inria
F-54600 Villers-lès-Nancy, France

ABSTRACT

Prior knowledge of the relative transfer function (RTF) is useful in many applications but remains little studied. In this paper, we propose a semi-supervised learning algorithm based on deep neural networks (DNNs) for RTF inverse regression, that is to generate the full-band RTF vector directly from the source-receiver pose (position and orientation). Two typical scenarios are discussed: training on labeled RTFs only, or on additional unlabeled RTFs. Both setups utilize the low-dimensional manifold property of RTF in stationary environments. With this property as an additional regularization term, a smooth mapping solution with respect to the manifold is obtained. Experimental simulations show that the proposed method achieves a lower mean prediction error than the free field model with few labeled RTFs, and the unlabeled RTFs are essential in improving the inverse regression performance.

***Index Terms*—** relative transfer function, semi-supervised learning, deep neural network, manifold regularization

1. INTRODUCTION

The relative transfer function (RTF) [1, 2] represents the difference between the signals recorded at two microphones in response to a source signal. The estimation of the RTF is a core task in many applications, such as beamforming and multichannel speech enhancement [3–5], source separation [6–8], and source localization [9]. Generally the estimation is merely based on the microphone observations, and prior knowledge about the RTF given the geometry of the scene remains little studied and exploited. Yet such knowledge could bring additional performance benefits [10, 11].

The RTF is defined as the ratio of two acoustic transfer functions (ATFs) and hence depends on the properties of the acoustic environment and on the *poses* (positions and orientations) of the source and the microphone pair. Conventional room acoustic simulation methods, such as the image-source method [12], rely on explicit physical models to simulate the ATFs. They require precise knowledge of the room geometry and the absorption properties of each material, which is not available in real environments. In such environments, the problem can be reformulated as that of predicting the RTF for a given pose given a set of RTFs or ATFs recorded for other poses. Recent studies have performed ATF interpolation based on room geometry estimation [13], compressed sensing [14, 15] or models derived from the wave propagation equation [16].

*We thank CSC (201604910623) for funding. This work was performed while the author was at Inria. Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`).

In [17], we proposed to predict the full-band RTF vector from a distinctive data-driven perspective, which we coined as *RTF inverse regression*. We trained a deep neural network (DNN) to map the low-dimensional source pose to the high-dimensional RTF. Training was performed in a supervised fashion, based on a set of RTFs and the corresponding poses collected in advance. It turned out that, with dense enough training samples (i.e., when the corresponding poses form a grid with less than 4 cm stepsize), simple linear interpolation [18] achieves a low prediction error. However, the DNN achieves a lower error when the training samples are further apart, which is a more realistic setup.

Unless the training RTFs are collected by fully automated mobile robots [19], labeling them all with the corresponding poses can be cumbersome in a real environment. In this paper, we propose to train the DNN in a semi-supervised way on a set of RTFs of which only few have been labeled with the corresponding pose. This setup is feasible in practice by, e.g., recording from a mobile phone or any other smart device worn by a human freely moving in the environment. As a matter of fact, this setup is the core of a recent series of studies on audio source localization [20]. To train on the unlabeled RTFs, we propose an *encoder-decoder* framework and utilize the low-dimensional manifold property of RTFs [21–23]. Specifically, we introduce a smoothness constraint on the manifold to regularize the encoder network, which provides noisy labels for the unlabeled RTFs that are exploited together with the labeled RTFs to train the decoder network. We investigate the function and the benefit of the unlabeled RTFs experimentally.

The structure of the paper is as follows. We define notations in Section 2 and summarize prior work on supervised RTF inverse regression in Section 3. We introduce the proposed semi-supervised encoder-decoder architecture in Section 4 and describe an alternative decoder architecture in Section 5. We evaluate the proposed method experimentally in Section 6. We conclude in Section 7.

2. DEFINITIONS

Let us consider an unknown signal S emitted from a source and captured by a pair of microphones in a fixed environment. For simplicity, we also assume the microphones to be fixed. The source pose is denoted as $\mathbf{p} = [\rho, \theta, \phi, \alpha, \beta, \gamma]$, with ρ, θ, ϕ its distance, azimuth, and elevation with respect to the microphone pair and α, β, γ its yaw, pitch, and roll. In the short-time Fourier transform (STFT) domain, under the narrowband approximation, the signal $A_m(n,k)$ recorded at the m-th microphone can be written as

$$A_m(n,k) = H_m(\mathbf{p},k)S(n,k) + V_m(n,k), \tag{1}$$

with n the time frame index, k the frequency bin index, $H_m(\mathbf{p},k)$ the ATF from the source to the m-th microphone, and $V_m(n,k)$ sen-

sor noise. The RTF $H(\mathbf{p}, k)$ associated with source pose $\mathbf{p}$ is defined as

$$H(\mathbf{p}, k) = \frac{H_2(\mathbf{p}, k)}{H_1(\mathbf{p}, k)}. \tag{2}$$

The RTF is independent of the source signal and, in a stationary environment, it depends only on the source pose. For instance, in a free-field environment, the relationship is given by

$$H^{\text{free-field}}(\mathbf{p}, k) = \exp\left(2\jmath\pi \frac{k}{F} f_{\text{s}} \frac{|r_2(\mathbf{p}) - r_1(\mathbf{p})|}{c}\right), \tag{3}$$

with $\jmath$ the complex unit, F the discrete Fourier transform size, f_{s} the sampling frequency, c the speed of sound, and $r_m(\mathbf{p})$ the Euclidean distance from the source to the m-th microphone. In a reverberant environment, this relationship becomes more complex due to multi-path propagation.

The RTF is often expressed via the interchannel level difference $\text{ILD}(\mathbf{p}, k) = 20\log_{10}|H(\mathbf{p}, k)|$ and the interchannel phase difference $\text{IPD}(\mathbf{p}, k) = \arg(H(\mathbf{p}, k))$. In the following, we consider the $1 \times 3K$ full-band RTF vector $\mathbf{h}$ obtained by concatenating the ILDs and the sines and cosines of the IPDs at all frequencies:

$$\begin{aligned}\mathbf{h} = \big[&\text{ILD}(\mathbf{p}, 0) \quad \cdots \quad \text{ILD}(\mathbf{p}, K-1) \\ &\sin(\text{IPD}(\mathbf{p}, 0)) \quad \cdots \quad \sin(\text{IPD}(\mathbf{p}, K-1)) \\ &\cos(\text{IPD}(\mathbf{p}, 0)) \quad \cdots \quad \cos(\text{IPD}(\mathbf{p}, K-1))\big]\end{aligned} \tag{4}$$

with $K = \frac{F}{2} + 1$ the number of frequency bins.

We define RTF inverse regression as the problem of predicting the full-band RTF vector $\mathbf{h}$ from the source pose $\mathbf{p}$ based on pre-collected training examples, and without making a measurement at that pose. Using machine learning terminology, we refer to $\mathbf{h}$ as a *sample* and to the corresponding pose $\mathbf{p}$ as a *label*.

3. SUPERVISED RTF INVERSE REGRESSION

In [23], probabilistic piecewise affine mapping (PPAM) was used to learn a bijective mapping between $\mathbf{h}$[1] and $\mathbf{p}$ based on a set of L labeled training samples $\{\mathbf{h}_{1:L}, \mathbf{p}_{1:L}\}$. However, this method was only applied to source localization (i.e., mapping $\mathbf{h}$ into $\mathbf{p}$).

In [17], we proposed to train a feedforward DNN $\mathcal{D}$ to directly map $\mathbf{p}$ into $\mathbf{h}$ by minimizing the mean square error (MSE) on the training set

$$\text{MSE} = \sum_{i=1}^{L} \|\mathcal{D}(\mathbf{p}_i) - \mathbf{h}_i\|^2 \tag{5}$$

where $\mathcal{D}(\mathbf{p})$ denotes the DNN output RTF vector from the input pose $\mathbf{p}$. We used a specific output activation function that normalizes the output nodes (o_{K+k}, o_{2K+k}) corresponding to each sine and cosine pair as

$$o_{K+k} \leftarrow \frac{o_{K+k}}{\sqrt{o_{K+k}^2 + o_{2K+k}^2}} \quad \text{and} \quad o_{2K+k} \leftarrow \frac{o_{2K+k}}{\sqrt{o_{K+k}^2 + o_{2K+k}^2}} \tag{6}$$

such that their squared sum is always equal to 1 and they can indeed be interpreted as the sine and cosine of the predicted IPD. In the test phase, the trained DNN was used to generate the RTF of a new pose $\mathbf{p}_t$ as $\widehat{\mathbf{h}}_t = \mathcal{D}(\mathbf{p}_t)$. This model is illustrated in Fig. 1(c).

In [17], we evaluated PPAM in terms of RTF inverse regression performance and compared it with linear interpolation and the above DNN. These three methods are based on supervised training.

[1]The authors used a slightly different representation of $\mathbf{h}$ by combining each $\sin(\text{IPD})$ and $\cos(\text{IPD})$ pair into a unit-norm complex number.

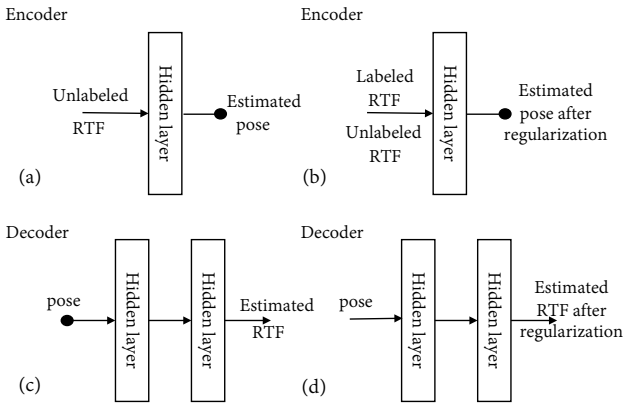

Fig. 1. Proposed encoder and decoder networks for semi-supervised RTF inverse regression. The black dots indicate the connection points of the networks.

4. SEMI-SUPERVISED RTF INVERSE REGRESSION

In the rest of this paper, we consider a semi-supervised setup: there are only L labeled training samples $\{\mathbf{h}_{1:L}, \mathbf{p}_{1:L}\}$ and the remaining U training samples $\{\mathbf{h}_{L+1:L+U}\}$ are unlabeled. To utilize the unlabeled samples in the training process, we propose an intuitive encoder-decoder architecture in Section 4.1. We refine this idea in Section 4.2 by optimizing the encoder network with manifold regularization. The proposed architectures are illustrated in Fig. 1. The network training details are discussed in Section 4.3.

4.1. Encoder-decoder architecture

We refer to the DNN in Section 3 and Fig. 1(c) as the *decoder* network. We introduce an auxiliary *encoder* network $\mathcal{E}$ to help optimize the decoder network under the new loss function:

$$\sum_{i=1}^{L} \|\mathcal{D}(\mathbf{p}_i) - \mathbf{h}_i\|^2 + \lambda \sum_{j=L+1}^{L+U} \|\mathcal{D}(\mathcal{E}(\mathbf{h}_j)) - \mathbf{h}_j\|^2 \tag{7}$$

where λ is a constant scaling factor. The first term is the MSE (5) on the labeled training samples and the second term is the reconstruction error on all training samples. This network architecture is given in Fig. 1(a)+(c). The encoder can be interpreted as a localization network, that maps the RTF space to the pose space and is expected to learn the pose labels implicitly. In the training phase, the encoder and decoder parameters are jointly updated to minimize (7).

4.2. Encoder with manifold regularization

Prior work on manifold regularization for audio source localization [20] has shown that the RTFs in a given environment lie on a smooth low-dimensional manifold. A simple validation of this concept is shown in Fig. 2, that plots the Euclidean distance between full-band RTF vectors as a function of the difference between the corresponding source poses. Within a local area (about $\pm 1.6^\circ$ in azimuth and elevation in this example), the distance between the RTFs increases approximately linearly with azimuth and elevation difference and the slope is similar in all directions. In other words, a small shift in the source pose only leads to small changes in the RTF and vice-versa. The detailed setup can be found in Section 6.

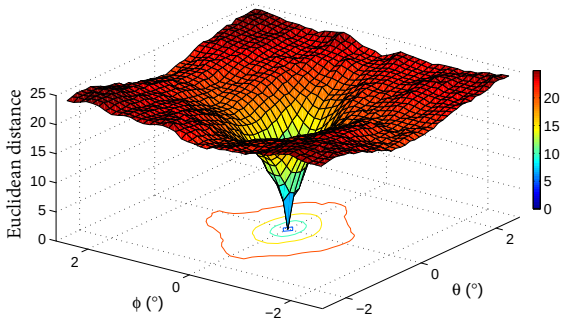

Fig. 2. Euclidean distance between full-band RTF vectors as a function of the azimuth and elevation difference. See Section 6 for the setup. One pose is fixed and the other is varied by up to $\pm 2.5^\circ$.

Inspired by this property, we further refine the encoder network from a localization perspective under the loss function:

$$\sum_{i=1}^{L} \|\mathcal{E}(\mathbf{h}_i) - \mathbf{p}_i\|^2 + \mu \sum_{i=1}^{L+U} \sum_{j=1}^{L+U} W_{ij} \|\mathcal{E}(\mathbf{h}_i) - \mathcal{E}(\mathbf{h}_j)\|^2 \quad (8)$$

that incurs a weighted penalty when similar inputs have different outputs. μ is a scaling factor and the second term is commonly known as graph Laplacian regularization [24], that imposes a smoothness constraint on the final mapping solution. W_{ij} is the weight that reflects the adjacency between encoder inputs $\mathbf{h}_i$ and $\mathbf{h}_j$, and it is close to 0 when the samples are far away. The standard Gaussian kernel function is used for weight calculation as in [20], since it is symmetric positive semi-definite and meets the locality requirements:

$$W_{ij} = \exp\left(-\frac{\|\mathbf{h}_i - \mathbf{h}_j\|^2}{\varsigma^2}\right) \quad (9)$$

where the variance ς^2 controls how fast the weight decays with the distance between RTFs. Accordingly, the new encoder-decoder architecture is given in Fig. 1(b)+(c).

The encoder is first trained to minimize (8) and then kept fixed when the decoder is optimized under loss (7). Further joint tuning didn't bring additional performance gains in our experiments.

4.3. Network training

During training, the networks are initialized following a standard procedure, i.e., the weights are initialized with Gaussian distributed samples and the biases with zeros. ReLU is used as the activation function for all the hidden layers and layer normalization [25] is applied to regularize the parameters. The Adam method [26] is chosen to update the model using an adaptive learning rate.

One issue with the mini-batch based gradient descent is that the regularization in (8) will tend to fail after random shuffling because of the sparse affinity inside the mini-batches, especially with large amount of training data. In [27], a nearest-neighbor graph based solution was discussed to sample the data efficiently. We adopt a simpler technique here. The training samples are first randomly shuffled. We then start from one sample, collect all its neighbors, move on to the next remaining sample and repeat until the mini-batch size has been reached. The regularization is found to gradually take place against all the adjacent samples after several training epochs.

5. ALTERNATIVE DECODER WITH REGULARIZATION

Given the locality property of the RTF manifold as discussed in Section 4.2, we are also motivated to apply regularization directly to the decoder. Similar to (8), the loss function for the new decoder network can be defined by

$$\sum_{i=1}^{L} \|\mathcal{D}(\mathbf{p}_i) - \mathbf{h}_i\|^2 + \nu \sum_{i=1}^{L+U} \sum_{j=1}^{L+U} w_{ij} \|\mathcal{D}(\mathbf{p}_i) - \mathcal{D}(\mathbf{p}_j)\|^2 \quad (10)$$

where ν is a scaling factor, and the weight

$$w_{ij} = \exp\left(-\frac{\|\mathbf{p}_i - \mathbf{p}_j\|^2}{\sigma^2}\right) \quad (11)$$

depends on the distance between poses, with variance σ^2. Note that the loss function (10) doesn't depend on the unlabeled RTF samples anymore. Hence, arbitrarily many pose labels $\{\mathbf{p}_{L+1:L+U}\}$ can be defined without the need to collect unlabeled RTFs. This decoder is illustrated by Fig. 1(d).

6. EXPERIMENTS AND ANALYSIS

6.1. Setup

The experiments are conducted in a simulated room with $6 \times 6.2 \times 3$ m size and 300 ms reverberation time. Two omnidirectional microphones are positioned at [3 3 1] m and [3.2 3 1] m, respectively. The source is omnidirectional, hence the Euler angles α, β, γ are irrelevant and the notions of source pose and position are used interchangeably. The source position is confined to a spherical cap with radius $\rho = 2$ m, azimuth $\theta \in [10^\circ, 60^\circ]$, and elevation $\phi \in [0^\circ, 30^\circ]$ with respect to the microphone pair, and it is sampled from a grid with 0.125° stepsize in azimuth and elevation on this cap. The setup here is similar to that in [20] and [23], which have been used to evaluate many studies in the field. In total, there are 400×240 poses.

For each pose, a 1 s white noise signal is emitted from the source and captured by the microphones. The sampling rate is 16 kHz. The corresponding acoustic impulses responses are simulated using an efficient implementation of the image-source method [28]. The STFT is applied with frame length 64 ms and 87.5% overlap. The target RTF is then estimated as

$$\widehat{H}(\mathbf{p}, k) = \frac{\sum_n A_2(n,k) A_1^*(n,k)}{\sum_n |A_1(n,k)|^2} \quad (12)$$

with $*$ denoting complex conjugation and the full-band RTF vector is derived as in (4). Note that this estimator is one possible way to measure the RTF in real environments, and a white noise source signal other than a speech signal, provides a reliable estimation at all frequencies [23]. Among the training samples, 24 RTFs, that is only 0.025% of the data, are labeled with their true source poses, creating a grid of 10° stepsize. We use two hidden layers with 1024 nodes each for the decoder and one hidden layer with 1024 nodes for the auxiliary encoder. The scaling factors λ, μ, and ν are heuristically set to 0.01. The variances ς^2 and σ^2 are set such that the weights of samples more than 2° apart are close to 0.

For evaluation, $T = 1000$ extra poses are picked randomly on the same spherical cap. To our knowledge, there is no agreed-upon method to measure the closeness of two high-dimensional RTF vectors. We use the mean absolute error (MAE) to measure the prediction error at each frequency:

$$\text{MAE} = \frac{1}{T} \sum_{t=1}^{T} |\widehat{\mathbf{h}}_t(k) - \mathbf{h}_t(k)|. \quad (13)$$

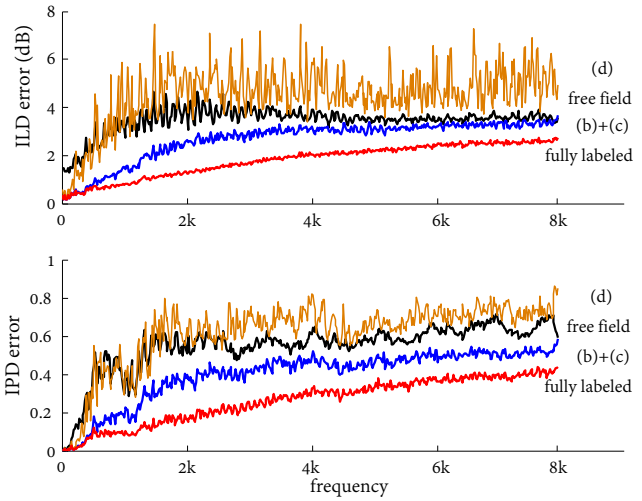

Fig. 3. Mean absolute ILD/IPD prediction errors along frequency for network (d) (orange), the free field model (3) (black), network (b)+(c) (blue), and the decoder trained with fully labeled data (red). The results of network (a)+(c) are close to that of network (d) and are thus not shown for legibility.

The metric makes sense as we will show in Section 6.3 by applying the predicted RTF to a speech enhancement task. The 95% confidence interval (CI) on the MAE is also calculated to give an idea of the error distribution. The ILD and IPD prediction errors are treated separately. By IPD prediction error, we mean the error on the vector entries (sines and cosines) as in (13), not the angular error in radians.

6.2. RTF inverse regression performance

The prediction error of the three proposed models is shown in Fig. 3 for each frequency bin and the frequency averaged results are summarized in Table 1. For comparison, the free field model (3) and a decoder trained using all 400×240 labels are included as baselines.

The prediction errors clearly increase with frequency and become relatively stable at high frequencies. One possible reason could be that the RTFs vary more rapidly with respect to pose changes at high frequencies as can be concluded from the free field model (3), while the slow-varying RTFs at low frequencies are easier to be captured by the neural networks. The network trained with fully labeled data achieves the lowest mean prediction errors as expected, with 1.80 ± 0.09 dB in ILD and 0.26 ± 0.02 in IPD. It is largely better than other setups where only few training data are labeled. The unlabeled RTFs function differently in different architectures. The (b)+(c) network performs much better than the free field model, while the intuitive encoder-decoder setup (a)+(c), that relies on the same training data, fails to give good predictions.

To see how the unlabeled samples help, we investigate the localization performance achieved by the encoder networks by computing the root mean square error (RMSE) between the encoder outputs and the true source poses. For encoder (a), the localization RMSE is 13.08° in azimuth and 8.74° in elevation. For the encoder (b) with manifold regularization, the results are 2.94° and 2.08°, respectively, which are quite accurate considering that the labeled samples are 10° apart. We can therefore interpret the encoders, and especially encoder (b), as providing the unlabeled RTFs with noisy pose labels. Additional experiments (not shown here) even indicate that encoder (b) slightly outperforms the recently proposed manifold regularization based localization method in [20] on this data.

Table 1. Frequency averaged mean absolute ILD/IPD prediction errors (mean±CI bound) and SBF (dB).

Model	ILD error	IPD error	SBF
(a)+(c)	4.05±0.23	0.64±0.03	-0.60
(d)	3.94±0.20	0.61±0.03	-0.35
free field model	3.47±0.16	0.55±0.03	0.69
(b)+(c)	2.67±0.13	0.39±0.02	2.04
fully labeled	1.80±0.09	0.26±0.02	3.19

The decoder (d) with manifold regularization achieves a similar prediction error at frequencies below 1 kHz compared with the free-field model but its performance falls behind at high frequencies. This is not surprising since it does not make use of unlabeled RTFs.

6.3. Further analysis

It is commonly acknowledged that learning based methods would perform better with more training data and suffer performance loss in mismatched test conditions. These aspects are not further investigated here. Instead we evaluate the generated RTFs in a specific application, the generalized sidelobe canceler (GSC) [1], that requires RTF in the implementation of a blocking matrix to provide the reference noise signal. We define the frequency-domain signal blocking factor (SBF) as

$$\mathrm{SBF} = 10 \log_{10} \frac{\sum_{n,k} \|A_1(n,k)\|^2}{\sum_{n,k} \|A_2(n,k) - H(\mathbf{p},k) A_1(n,k)\|^2} \qquad (14)$$

where the denominator denotes the energy of the leakage signal. The SBF indicates the ability to block the first-channel source image in the second microphone and its value correlates negatively with signal distortion.

The SBF scores of different methods are given in Table 1. The results are consistent with the mean absolute prediction errors, with the fully labeled setup scoring the best (3.19 dB), the (b)+(c) network outperforming the free-field model, and the simple (a)+(c) the scoring the worst (-0.60 dB). Negative scores mean that the generated RTFs are not helpful at all.

The positive score achieved by the (b)+(c) network indicates that the predicted RTF can be considered as reliable prior information derived from the source pose. Further performance benefits are expected if this prior information is incorporated with the observations to achieve maximum a posteriori RTF estimation.

7. CONCLUSIONS

We considered the RTF inverse regression task from a practical perspective where the training data are partially labeled and introduced a semi-supervised learning approach. Several possible neural network architectures were discussed. The proposed encoder with manifold regularization and decoder architecture outperformed the free-field model, in terms of both mean prediction error and signal blocking ability in the GSC application. Additional experiments showed that unlabeled RTFs and manifold regularization are both necessary to achieve good performance. Incorporating and evaluating the generated RTFs as prior information in other applications is worth further investigation in the future.

8. REFERENCES

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

[2] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 451–459, 2004.

[3] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 546–555, 2009.

[4] K. Reindl, Y. Zheng, A. Schwarz, S. Meier, R. Maas, A. Sehr, and W. Kellermann, "A stereophonic acoustic signal extraction scheme for noisy and reverberant environments," *Computer Speech and Language*, vol. 27, no. 3, pp. 726–745, 2013.

[5] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multi-microphone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[6] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, pp. 1833–1847, Aug. 2007.

[7] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components with estimating propagation model parameters in frequency-domain blind source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1592–1604, 2007.

[8] M. Taseska and E. A. P. Habets, "Relative transfer function estimation exploiting instantaneous signals and the signal subspace," in *23rd European Signal Processing Conference (EUSIPCO)*. IEEE, 2015, pp. 404–408.

[9] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of relative transfer function in the presence of stationary noise based on segmental power spectral density matrix subtraction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 320–324.

[10] R. Talmon and S. Gannot, "Relative transfer function identification on manifolds for supervised GSC beamformers," in *Proceedings of the 21st European Signal Processing Conference (EUSIPCO)*. IEEE, 2013, pp. 1–5.

[11] Z. Koldovský, J. Málek, P. Tichavský, and F. Nesta, "Semi-blind noise extraction using partially known position of the target source," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2029–2041, 2013.

[12] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[13] A. Asaei, M. Golbabaee, H. Bourlard, and V. Cevher, "Structured sparsity models for reverberant speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 3, pp. 620–633, 2014.

[14] R. Mignot, L. Daudet, and F. Ollivier, "Room reverberation reconstruction: Interpolation of the early part using compressed sensing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2301–2312, 2013.

[15] R. Mignot, G. Chardon, and L. Daudet, "Low frequency interpolation of room impulse responses using compressed sensing," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 1, pp. 205–216, 2014.

[16] P. Samarasinghe, T. Abhayapala, M. Poletti, and T. Betlehem, "An efficient parameterization of the room transfer function," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 12, pp. 2217–2227, 2015.

[17] Z. Wang, E. Vincent, and Y. Yan, "Relative transfer function inverse regression from low dimensional manifold," *arXiv preprint arXiv:1710.09091*, 2017.

[18] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: Datasets, tasks and baselines," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 126–130.

[19] J. Le Roux, E. Vincent, J. R. Hershey, and D. P. W. Ellis, "MICbots: collecting large realistic datasets for speech and audio research using mobile robots," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5635–5639.

[20] B. Laufer, R. Talmon, and S. Gannot, "Semi-supervised sound source localization based on manifold regularization," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 8, pp. 1393–1407, 2016.

[21] A. Deleforge and R. Horaud, "2D sound-source localization on the binaural manifold," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2012, pp. 1–6.

[22] B. Laufer, R. Talmon, and S. Gannot, "Relative transfer function modeling for supervised source localization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2013, pp. 1–4.

[23] A. Deleforge, F. Forbes, and R. Horaud, "Acoustic space learning for sound-source separation and localization on binaural manifolds," *International Journal of Neural Systems*, vol. 25, no. 01, 2015, 1440003.

[24] Z. Yang, W. W. Cohen, and R. Salakhutdinov, "Revisiting semi-supervised learning with graph embeddings," in *International Conference on Machine Learning*, 2016, pp. 40–48.

[25] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[26] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] S. Thulasidasan and J. Bilmes, "Acoustic classification using semi-supervised deep neural networks and stochastic entropy-regularization over nearest-neighbor graphs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2731–2735.

[28] E. A. P. Habets, "Room impulse response (RIR) generator," https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator, 2014.