SEMI-SUPERVISED SOURCE LOCALIZATION WITH RESIDUAL PHYSICAL LEARNING

Michael J. Bianco and Peter Gerstoft

Marine Physical Laboratory, University of California San Diego

ABSTRACT

Machine learning (ML) approaches to source localization have demonstrated promising results in addressing reverberation. Even with large data volumes, the number of labels available for supervised learning in such environments is usually small. This challenge has recently been addressed using semi-supervised learning (SSL) based on deep generative modeling with variational autoencoders. A problem with ML approaches is they often ignore the intuitions from conventional signal processing approaches. We present a hybrid approach to ML-based source localization, which uses both SSL and conventional, analytic signal processing approaches to obtain source location estimates. An SSL approach is developed which accounts for the residual between analytic source location estimates true locations. Thus, the approach can exploit both labelled and unlabeled data, as well as analytic source location intuition, to provide better localization than either approach in isolation.[1]

Index Terms— Source localization, semi-supervised learning, generative modeling, deep learning

1. INTRODUCTION

Source localization is an important problem in acoustics and many related fields. The performance of localization algorithms is degraded by reverberation, which induces complex temporal arrival structure at sensor arrays. Despite recent advances, e.g., [1, 2, 3], acoustic localization in reverberant environments remains a major challenge [4]. There has been great interest in machine learning (ML)-based techniques in acoustics, including source localization and event detection [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]. A difficulty for ML-based methods in acoustics is the limited amount of labeled data and the complex acoustic propagation in natural environments, despite large volumes of recordings[1, 2]. This limitation has motivated recent approaches for localization based on semi-supervised learning (SSL)[17, 18].

We approach source localization from the perspective of SSL, with the intent of addressing real-world applications of ML. It has been shown that large amounts of synthetic data, resembling real-world sound measurement configurations, can easily be generated. This synthetic data has then been used to train ML-based localization models (e.g.,[8]), with good performance. However, in most real scenarios room geometry includes irregular boundaries, scattering, and diffracting elements (e.g., furniture and uneven surfaces) which may not be convenient to model using acoustic propagation software.

In recent work, it was shown that an SSL localization approach based on deep generative modeling with variational autoencoders (VAEs)[19, 20, 21] generalizes well in the localization task using only sparsely labeled training data [22]. In the proposed approach, VAEs encode and generate the phase of the relative transfer function (RTF) between two microphones [23]. The VAE was trained in parallel with a classifier network to benefit from both labeled and unlabeled examples. In learning to generate RTF-phase, the VAE-SSL system learned the physical model relating the latent model and direction of arrival (DOA) label to the RTF-phase. In addition to improved localization performance, the output of the generative model was physically interpretable.

An important issue with ML approaches to signal processing tasks is that they often do not leverage the power of existing analytical approaches. Such approaches, for example the Steered-Response Power PHAse Transform (SRP-PHAT)[24] and MUltiple SIgnal Classification (MUSIC)[25] in DOA estimation, build on decades of theoretical thought about source localization and separation. Recent work in motion planning and robot control [26, 27, 28] have proposed a residual formulation to ML, in which ML systems are trained to correct the residual error from motion estimates from physical simulator.

Generally, analytic models sacrifice some physical fidelity for analytic tractability. We propose to extend these residual-physical learning concepts to the source localization task, to help compensate the limitations of analytic models with the non-linear modeling capacity of deep neural networks. Similar to [26], our neural networks parameterize stochastic functions, the parameters of which are optimized using variational inference (VI). While such models use an approximate posterior which is often much simpler than the true posterior distribution, the posteriors can provide an estimate of the model uncertainty.

We explore the effect of combining analytic source localization models with semi-supervised and fully-supervised learning approaches. It is found that the hybrid analytic-learned model approach can perform better than either approach in isolation, and that the learned models generalize well. The methods are test using real-world impulse responses from a new dataset [29]. Further comparisons are made with a fully-supervised source localization approach, also with and without residual physical learning.

2. THEORY

We propose a learning-based strategy to localization which utilizes the estimates of conventional localization methods. This allows the system to exploit domain knowledge, while retaining enough capacity to well-model the physics of the reverberant environment.

We modify the architecture developed in [22], a semi-supervised approach to localization based on VAEs deemed VAE-SSL, to use as input estimates from analytic localization models. The VAE-SSL system is formulated using a localization network and a VAE. We incorporate residual physical modeling into the VAE-SSL localization network, to form a hybrid learning system. The VAE provides a deep generative model of the RTF-phase sequences, which helps regularize the learning in the localization network by ensuring the estimates are physically relevant. The analytic localization models use the STFT snapshots from the RTF-phase sequence calculations.

[1] Codes available at: https://github.com/mikebianco/vaessl-doa

2.1. Measurements

RTFs[23], specifically the instantaneous RTF-phase, calculated using a single STFT frame (i.e., no averaging), are used as the input acoustic feature for our VAE-SSL approach. Since the RTF is independent of source waveform, this feature helps to focus ML on physically relevant features, and thereby reduces the sample complexity of the model.

We use short-time Fourier transform (STFT) domain acoustic recordings of the form

$$d_i(k) = a_i(k)s(k) + u_i(k), \tag{1}$$

with s the source signal a_i the acoustic transfer function relating the source and each of the microphones ($i = \{1, 2\}$ the microphone index and k the frequency index), and u_i sensor noise (spatially white). The relative transfer function (RTF) is defined as[17, 23] $h(k) = a_2(k)/a_1(k)$, with k the frequency index. With d_1 as reference, the instantaneous RTF $\widehat{h}(k)$ is calculated using a single STFT frame (also referred to as a snapshot),

$$\widehat{h}(k) = \frac{d_2(k)}{d_1(k)}. \tag{2}$$

For each STFT frame, a vector of RTFs $\widehat{\mathbf{h}} = [\widehat{h}(1) \dots \widehat{h}(K)]^{\mathrm{T}} \in \mathbb{C}^K$ is obtained with K the number of frequency bins used.

The input to the VAE-SSL and supervised CNN is a temporally ordered RTF-phase sequence. The nth RTF-phase sequence is

$$\mathbf{x}_n = \mathrm{vec}(\mathrm{phase}(\widehat{\mathbf{H}}_n)) \in \mathbb{R}^{KP}, \tag{3}$$

with $\widehat{\mathbf{H}}_n = [\widehat{\mathbf{h}}_n \dots \widehat{\mathbf{h}}_{n+P-1}] \in \mathbb{C}^{K\times P}$, $K = N_{\mathrm{STFT}}/2 - 1$, and P the number of RTF-phase frames in the sequence. We use the wrapped RTF-phase, which is in the interval $[-\pi, \pi]$ radians.

The true DOA labels in the interval $[-90, 90]^\circ$ azimuth, corresponding to each RTF-phase sequence $\mathbf{x}_n$, are represented by $\mathbf{y}_n \in \{0, 1\}^T$, a one-hot encoding, with T the number of DOA classes. The labels estimated using the conventional signal processing approaches are denoted $\widehat{y}_{c,n}$. These estimates are computed using the P STFT frames to estimate the RTFs, with $K = N_{\mathrm{FFT}}/2 + 1$.

We thus have labeled and unlabeled sets, defined by $\{\mathbf{x}_j, \mathbf{y}_j, \widehat{y}_{c,j}\} \in \mathcal{D}_l$ and $\{\mathbf{x}_u, \widehat{y}_{c,u}\} \in \mathcal{D}_u$. Labels for the unlabeled sequences $\mathbf{y}_u$ are reserved to test the performance of the system only after training and validation. The sizes of the sets are $|\mathcal{D}_l| = J$ and $|\mathcal{D}_u| = N - J$. Thus, there are J labeled RTF sequences and $N - J$ unlabeled sequences.

2.2. Semi-supervised source localization with VAEs

We assume each RTF-phase sequence $\mathbf{x}$ is generated by a random process involving the latent random variable $\mathbf{z} \in \mathbb{R}^M$, with M the dimension of the latent space, and source location label $\mathbf{y}$. Thus, the true RTF-phase distribution $p^*(\mathbf{x})$ is approximated with the conditional distribution $p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z})$, where θ are the parameters of the NN used to define the distribution. We use $p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z}) = \mathcal{N}(\cdot|\cdot)$, the truncated normal distribution with mean and variance defined by a NN with parameters (weights and biases) θ. In the optimization stage, the parameters θ are adjusted to fit the data. We discard the variable subscripts to simplify the notation.

The DOA label and latent variable, $\mathbf{y}$ and $\mathbf{z}$, are assumed independent, with their marginal densities $p(\mathbf{y})$ and $p(\mathbf{z})$ the categorical and normal distributions. Thus, the generative model is $p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z})p(\mathbf{y})p(\mathbf{z})$.

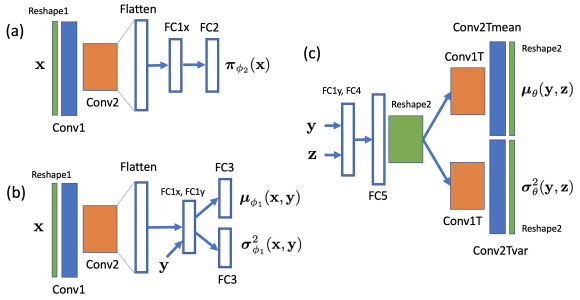

Fig. 1: Neural network configurations. Encoders for (a) label inference (classifier) and (b) latent model. (c) Decoder for generative model.

Now we are presented with the challenge of inferring $\mathbf{y}$ (when no label is provided) and the latent variable $\mathbf{z}$. We illustrate this with Bayes's rule for labeled data $\{\mathbf{x}_j, \mathbf{y}_j\} \sim \mathcal{D}_l$. In the derivation, only the label $\mathbf{y}$ and RTF-phase $\mathbf{x}$ are considered. The posterior of the latent variable $\mathbf{z}$ is

$$p(\mathbf{z}|\mathbf{x}, \mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x}, \mathbf{y})}. \tag{4}$$

Direct estimation of the posterior, e.g., from (4), $p(\mathbf{z}|\mathbf{x}, \mathbf{y})$ is intractable due to $p(\mathbf{x}, \mathbf{y}) = \int p(\mathbf{x}, \mathbf{y}, \mathbf{z})d\mathbf{z}$. Thus, the posteriors are approximated using VI. A variational distribution is defined, $q_{\phi_1}(\mathbf{z}|\mathbf{x}, \mathbf{y})$, which approximates the intractable posterior $p(\mathbf{z}|\mathbf{x}, \mathbf{y})$. $q_{\phi_1}(\mathbf{z}|\mathbf{x}, \mathbf{y})$ is a family of distributions parameterized by the latent inference encoder, with parameters ϕ_1 (see Fig. 1(b)). For $q_{\phi_1}(\mathbf{z}|\mathbf{x}, \mathbf{y})$, a normal distribution is used. The inference model for the DOA label $q_{\phi_2}(\mathbf{y}|\mathbf{x})$, parameterized by label inference encoder with parameters ϕ_2 is obtained starting in (7). For $q_{\phi_2}(\mathbf{y}|\mathbf{x})$, the categorical distribution is used.

Starting with the model for the labeled data, see (4), per VI we seek $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})$ which minimizes the KL-divergence

$$\{\phi_1, \theta\} = \arg\min_{\phi_1,\theta} \mathrm{KL}(q_{\phi_1}(\mathbf{z}|\mathbf{x}, \mathbf{y})||p(\mathbf{z}|\mathbf{x}, \mathbf{y})). \tag{5}$$

via the evidence lower-bound (ELBO) approximation. The objective for labeled data is thus formulated, with $\mathbf{y}$ and $\mathbf{z}$ independent

$$\begin{aligned} -C(\theta, \phi_1; \mathbf{x}, \mathbf{y}) &= \mathbb{E}\big[\log p(\mathbf{x}, \mathbf{y}|\mathbf{z})p(\mathbf{z}) - \log q_{\phi_1}(\mathbf{z}|\mathbf{x}, \mathbf{y})\big] \\ &= \mathbb{E}\big[\log p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z}) + \log p(\mathbf{y}) + \log p(\mathbf{z}) - \log q_{\phi_1}(\mathbf{z}|\mathbf{x}, \mathbf{y})\big]. \end{aligned} \tag{6}$$

This follows [Ref. [20], Eq.(6)].

The objective for unlabeled data $\mathcal{D}_u$ is derived, using $q_\Phi(\mathbf{y}, \mathbf{z}|\mathbf{x}) \approx p(\mathbf{y}, \mathbf{z}|\mathbf{x})$. More details are in [22]. Following [Ref. [20], Eq.(7)]

$$\begin{aligned} -D(\theta, \Phi; \mathbf{x}) &= \mathbb{E}_{q_{\phi_2}(\mathbf{y}|\mathbf{x})}\big[\mathbb{E}_{q_{\phi_1}(\mathbf{z}|\mathbf{x},\mathbf{y})}\big[\log p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z}) \\ &\quad + \log p(\mathbf{y}) + \log p(\mathbf{z}) - \log q_{\phi_1}(\mathbf{z}|\mathbf{x}, \mathbf{y}) - \log q_{\phi_2}(\mathbf{y}|\mathbf{x})\big]\big] \\ &= \sum_{\mathbf{y}} q_{\phi_2}(\mathbf{y}|\mathbf{x})\big[-C(\theta, \phi; \mathbf{x}, \mathbf{y}) - \log q_{\phi_2}(\mathbf{y}|\mathbf{x})\big] \end{aligned} \tag{7}$$

It is important for the classifier to learn from the labeled sequences, and we enforce this by adding an auxiliary term $-\log q_{\phi_2}(\mathbf{y}|\mathbf{x})$ to the supervised objective. This is a typical procedure [20].

An overall objective for training the VAE and classifier models using labeled and unlabeled data is derived by combining (6) and (7) with an auxiliary term. It is important for the classifier to learn from

the labeled sequences, and this is enforced by adding an auxiliary term $-\log q_{\phi_2}(\mathbf{y}|\mathbf{x})$. The objective is

$$\mathcal{L} = \sum_{\{\mathbf{x}_j, \mathbf{y}_j, \widehat{y}_{c,j}\} \sim \mathcal{D}_l} C(\theta, \phi_1; \mathbf{x}_j, \mathbf{y}_j) - \alpha \log q_{\phi_2}(\mathbf{y}_j|\mathbf{x}_j) + \sum_{\{\mathbf{x}_u, \widehat{y}_{c,u}\} \sim \mathcal{D}_u} D(\theta, \Phi; \mathbf{x}_u),$$

with α a scaling term. This follows [Ref. [20], Eqs.(8,9)].

2.3. VAE-SSL distributions

For the inference model, the following distributions are used: $q_{\phi_1}(\mathbf{z}|\mathbf{y}, \mathbf{x}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{\phi_1}(\mathbf{x}, \mathbf{y}), \mathrm{diag}(\boldsymbol{\sigma}^2_{\phi_1}(\mathbf{x}, \mathbf{y})))$, with $\mathcal{N}(\cdot|\cdot)$ the normal distribution parameterized by the outputs of the latent inference network $\boldsymbol{\mu}_{\phi_1}(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^M$ and $\boldsymbol{\sigma}^2_{\phi_1}(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^M$ (see Fig. 1(b)); $q_{\phi_2}(\mathbf{y}|\mathbf{x}) = \mathrm{Cat}(\mathbf{y}|\boldsymbol{\pi}_{\phi_2}(\mathbf{x}))$, with $\mathrm{Cat}(\cdot|\cdot)$ the categorical (multinomial) distribution parameterized by the output of the classifier network $\boldsymbol{\pi}_{\phi_2}(\mathbf{x}) \in \mathbb{R}^T$ (see Fig. 1(a)).

For the generative model, $p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z}) = \overline{\mathcal{N}}(\mathbf{x}|\boldsymbol{\mu}_\theta(\mathbf{y}, \mathbf{z}),$ $\mathrm{diag}(\boldsymbol{\sigma}^2_\theta(\mathbf{y}, \mathbf{z})))$, with $\overline{\mathcal{N}}(\cdot|\cdot)$ the truncated normal distribution parameterized by the outputs of the decoder $\boldsymbol{\mu}_\theta(\mathbf{y}, \mathbf{z}) \in \mathbb{R}^{KP}$ and $\boldsymbol{\sigma}^2_\theta(\mathbf{y}, \mathbf{z}) \in \mathbb{R}^{KP}$ (see Fig. 1(c)). The truncated normal distribution is used for the generative conditional distribution $p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z})$ since the wrapped RTF-phase is on the interval $[-\pi, \pi]$.

The marginal densities are $p(\mathbf{y}) = \mathrm{Cat}(\mathbf{y}|\boldsymbol{\pi})$, with $\boldsymbol{\pi} \in \mathbb{R}^T$ the probabilities of the classes, which are assumed equal with $\pi_t = 1/T$; and $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$.

2.4. Residual physical learning

Given a true source location $\mathbf{y}$, the residual physical learning approach augments the outputs of the localization encoder. This formulation allows the same optimization framework used in VAE-SSL [22], with modifications only to the localization encoder.

The T DOAs have stochastic one-hot representation $\mathbf{y}$, modeled with the categorical distribution. In training the VAE-SSL, estimates for the labels $\widehat{\mathbf{y}}$ are drawn by $\widehat{\mathbf{y}} \sim q_{\phi_2}(\mathbf{y}|\mathbf{x})$. From the trained inference model, the DOA is estimated by the indicator function

$$\begin{aligned} y_{\rho n} &= \mathbf{1}_{\rho = \widehat{t}}, \\ \widehat{t} &= \arg\max_t (\pi_{t,\phi_2}(\mathbf{x}_n)) \end{aligned} \tag{8}$$

and ρ and t the discrete DOA indices, and $\pi_{t,\phi_2}(\mathbf{x}_n))$.

The conventional source localization approach operates on the same DOA grid as the learned models. In our approach, the one-hot label index i corresponding to $\widehat{y}_c$ with $\widehat{i} = \arg\max_i f_i(\mathbf{x}_n)$ and $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^T$ the beam-power.

2.4.1. Residual physical model formulation

Physical estimates from the conventional model are incorporated into the learned models in the definition of $\boldsymbol{\pi}_{\phi_2}(\mathbf{x})$. We first consider the probabilistic relationships RTF-phase sequence $\mathbf{x}$, and DOA estimates from the localization encoder γ_e and conventional model γ_c. The posterior is assumed

$$\begin{aligned} p(\gamma_e|\gamma_c, \mathbf{x}) &\propto p(\gamma_c|\gamma_e, \mathbf{x}) p(\gamma_e|\mathbf{x}) \\ &\propto p(\gamma_c|\mathbf{x}) p(\gamma_e|\mathbf{x}), \end{aligned} \tag{9}$$

with the conventional estimate γ_c and γ_e independent. We have then $\boldsymbol{\pi}_{\phi_2}(\mathbf{x}, \gamma_c) \propto p(\gamma_e|\gamma_c, \mathbf{x})$, $\boldsymbol{\pi}_{\phi_2,e}(\mathbf{x}) = p(\gamma_e|\mathbf{x})$ the output of the encoder softmax, and $\boldsymbol{\pi}_c(\mathbf{x}) = p(\gamma_e|\mathbf{x})$ a unimodal probability based on the conventional DOA estimate. Thus residual-physical DOA probabilities parameterize the categorical distribution

$$\boldsymbol{\pi}_{\phi_2}(\mathbf{x}, \gamma_c) = C \boldsymbol{\pi}_{\phi_2,e}(\mathbf{x}) \boldsymbol{\pi}_c(\mathbf{x}) \tag{10}$$

with $\boldsymbol{\pi}_c(\mathbf{x})$ defined by

$$\pi_{c,t}(\mathbf{x}) = \begin{cases} p & \text{if } t = \widehat{i} \\ \frac{1-p}{T-1} & \text{otherwise,} \end{cases} \tag{11}$$

and C a normalization factor to ensure $\boldsymbol{\pi}_{\phi_2}(\mathbf{x}, \gamma_c)$ is a proper density. p is the probability assigned to the active DOA index. For this paper, we choose $p = 0.8$ as it works well. Alternatively, p can be chosen via hyperparameter search.

In this residual model configuration, the encoder weights learn to compensate the conventional estimate, thus helping account for the residual between the conventional and true DOAs. The conventional estimates have the added benefit of giving a good initial guess, which improves convergence early in the training, which has been discussed in [26]. The estimates are further beneficial for the results of semi-supervised learning since conventional estimates are available for both labelled and unlabelled RTF-phase sequences.

3. EXPERIMENT

In the following, we assess the performance of the VAE-SSL [22] localization approach, as well as a supervised CNN model with residual physical modeling in a moderately reverberant classroom environment. The learning-based models are compared with the SRP-PHAT [24]. SRP-PHAT also provides conventional estimates for the residual-physical learning approach. The VAE-SSL and CNN are trained and validated using speech to obtain real-world application performance. The performance of the methods, summarized in Figs. 2 and 3, is quantified in terms of mean absolute error (MAE) and sequence-level accuracy (Acc.).

3.1. Measured Impulse Responses

We use measured IRs from a dataset recorded at Technical University of Denmark [29] in 2020. The classroom was approximately rectangular, of dimensions $9 \times 6 \times 3$ m and fully furnished. The nominal source-array range was 1.5 m. IRs were obtained from 19 DOAs (10° resolution over the interval $[-90^\circ, 90^\circ]$). The reverberation time in the classroom was $\mathrm{RT}_{60} = 500$ ms. There were two microphones, with 8.5 cm spacing. The sampling rate was 48 kHz. The IRs were downsampled to 16 kHz for this study.

In addition to the nominal source grid for the DTU dataset, several off-range IRs (3 cases) and off-grid IRs (6 cases) were obtained. We use these to test the generalization of learning-based localization methods. The off-grid source DOAs were $[25^\circ, 28^\circ, 45^\circ]$ with 1.5 m range. The off-range source DOAs (ranges) were 0° (1 m), 10° (2 m), 40° (2 m), -30° (2.5 m), -40° (2.5 m), and -30° (3.0 m). For more details, see [29].

3.2. Data processing

The signal at the microphones is given in (1). RTFs are obtained from the data by (2). The RTFs are estimated using single STFT frames with Hamming windowing with 50% overlap and segment length $N_{FFT} = 256$. The VAE and the supervised CNN inputs $\mathbf{x}_n$ use $P = 31$ RTF vectors, giving an input size $K \times P = 127 \times 31$ (neglecting the highest frequencies from full RTF with length

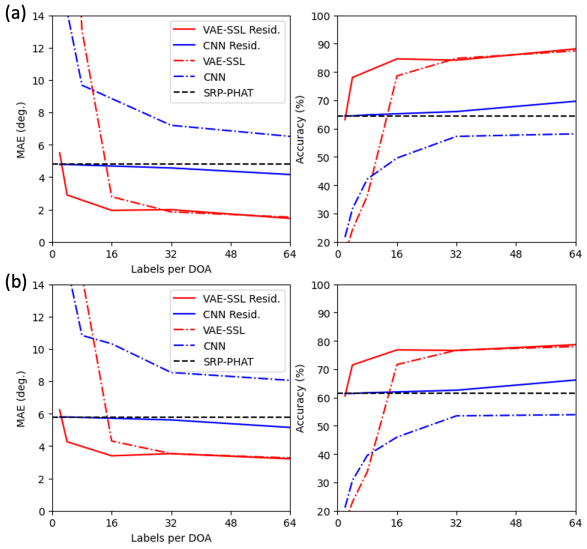

Fig. 2: Localization performance on unlabeled RTF-phase sequences vs. number of labels per DOA for (a) DTU training and (b) validation datasets.

$N_{\rm FFT}/2 + 1$, including all frequencies between 0 (DC) up to $N_{\rm FFT}/2$, to support strided transpose convolution without padding). For fair comparison, SRP-PHAT used the $P = 31$ STFT frames to estimate the RTFs, with $K = N_{\rm FFT}/2 + 1$. The temporal length of the sequences was thus 0.26 s. Sensor noise with 20 dB SNR was added to the microphone signals (see (1)).

40 speech segments 2-3 s in duration were randomly selected from the LibriSpeech development corpus [30]. The segments were isolated using voice activity detection with the WebRTC VAD system [31]. The segments from were convolved with the DTU IRs to obtain reverberant speech. 20 segments each were used for training and validation. This yielded $\sim 110,000$ RTF sequence for the nominal DTU IRs for training and validation. This further yielded $\sim 17,500$ and $\sim 35,000$ sequences for the DTU off-grid and off-range measurements using the validation speech.

3.3. Learning-based model implementation

The VAE-SSL model (classifier, inference, and generative networks) were implemented using strided CNNs, with stride of 2 pixels. The network architectures are given in Fig. 1, and the corresponding parameters are as follows. Each NN used convolutional layers without pooling, with 3x3 kernels. Since the kernel size is 3, and stride is 2 in the convolution layers, this gives transformation of input dimension m as $w(m) = (m-1)/2$, with m odd. The Conv1 layer had input size $[K, P, 1]$ and 32 channels, and Conv2 had input $[w(K), w(P), 32]$ with 64 channels. The fully connected layers (units) are: FC1x ($w(w(K)) \times w(w(P)) \times 64$); FC1y ($T$); FC2, FC3, and FC5 (200); FC4 (50). The transpose convolution layers Conv1T, Conv2Tmean, and Conv2Tvar mirror the shape of the input convolutional layers. Dropout with probability of 0.5 is used in the fully connected layers FC1x and FC5.

Before processing by the CNNs, Fig. 1(a,b), the generated RTF-phase sequence vectors $\widehat{\mathbf{x}}_n \in \mathbb{R}^{KP}$ are reshaped to a matrix with dimensions $P \times K$. The output of the CNNs (Fig. 1(c)) is reshaped to a vector with length KP. All the RTF-phase sequences $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_N]$ were normalized by π for the VAE-SSL and CNN.

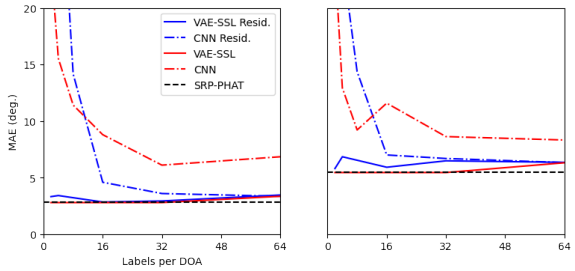

Fig. 3: Localization performance for off-range (left) and off-grid (right) measurements from DTU dataset.

For the fully supervised CNN approach, we used the classifier CNN architecture from the VAE-SSL model without the generative model. All NNs were implemented in Pytorch and optimized using Adam[32] with default settings. Variational inference for VAE-SSL was performed using probabilistic programming with Pyro [33]. For all cases, the learning rate was 5e-5 and the minibatch size was 256. The auxiliary multiplier for VAE-SSL (8), was $\alpha = 5000$.

J labeled sequences were drawn from the training set and J labeled sequences were drawn from the validation set. Models were chosen based on validation accuracy for labeled sequences. For VAE-SSL, $N - J$ unlabeled sequences from the training dataset were used to train the networks with unsupervised learning. Only the labeled sequences were used to train the supervised CNN. We considered a range of $J = [38, 76, 152, 304, 608, 1216]$. With $T = 19$, this corresponded to $[2, 4, 8, 16, 32, 64]$ labels per DOA. SRP-PHAT was evaluated using the full dataset of N sequences.

3.4. Training and localization performance

After training, the performance of the models was evaluated using the unlabeled sequences from the validation dataset. The performance of VAE-SSL and the competing approaches for the DTU IR dataset are given in Fig. 2. It is observed that the VAE-SSL with residual learning generalizes well to the validation data, with better performance than SRP-PHAT with as few as 4 labeled sequences per DOA. The fully-supervised CNN with residual learning also performs well, but less so than the semi-supervised approach. We that VAE-SSL without residual learning converges to the same MAE and Acc. as VAE-SSL with residual learning. The VAE-SSL outperforms fully-supervised CNN for all the experiments in this paper.

The generalization of the methods to off-grid and off-range sources is also examined. Results are given in Fig. 3. It is found that the VAE-SSL and CNN with residual physical modeling generalize about as well as SRP-PHAT for off-grid and off-range. VAE-SSL without residual learning converges to the SRP-PHAT performance for both cases, with sufficient labels.

4. CONCLUSIONS

We have introduced an approach to source localization based on residual physical learning. The approach builds on recent work in semi-supervised learning. The residual modeling approach incorporates physical knowledge into the ML framework and reduces the sample complexity of the learning-based approaches. Improvements in performance of the VAE-SSL supervised CNN models were obtained for sparsely labeled data. It is found that residual physical learning improves the off-design generalization of the methods.

5. REFERENCES

[1] H. Purwins, B. Li, T. Virtanen, J. Schluter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 206–219, May 2019.

[2] M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, and C.-A. Deledalle, "Machine learning in acoustics: Theory and applications," *J. Acoust. Soc. Am.*, vol. 146, no. 5, pp. 3590–3628, 2019.

[3] S. Gannot, M. Haardt, W. Kellermann, and P. Willett, "Introduction to the issue on acoustic source localization and tracking in dynamic real-life scenes," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 3–7, 2019.

[4] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proc. Nat. Acad. Sci.*, vol. 113, no. 48, pp. E7856–E7865, 2016.

[5] H. Nakashima and T. Mukai, "3D sound source localization system based on learning of binaural hearing," in *IEEE Int. Conf. Syst., Man, Cybern.* IEEE, 2005, vol. 4, pp. 3534–3539.

[6] A. Deleforge and R. Horaud, "2D sound-source localization on the binaural manifold," in *IEEE Int. Workshop Mach. Learn. Signal Process.* IEEE, 2012, pp. 1–6.

[7] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement*, John Wiley & Sons, 2018.

[8] S. Chakrabarty and E. A. P. Habets, "Multi-speaker doa estimation using deep convolutional networks trained with noise signals," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 8–21, Mar. 2019.

[9] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE J. Sel. Topics Signal Process.*, 2019.

[10] G. Ping, E. Fernandez-Grande, P. Gerstoft, and Z. Chu, "Three-dimensional source localization using sparse bayesian learning on a spherical microphone array," *J. Acoust. Soc. Am.*, vol. 147, no. 6, pp. 3895–3904, 2020.

[11] E. Ozanich, P. Gerstoft, and H. Niu, "A feedforward neural network for direction-of-arrival estimation," *J. Acoust. Soc. Am.*, vol. 147, no. 3, pp. 2035–2048, 2020.

[12] X. Zhu, H. Dong, P. S.o Rossi, and M. Landrø, "Feature selection based on principal component analysis for underwater source localization by deep learning," *arXiv preprint arXiv:2011.12754*, 2020.

[13] H. Hammer, S. E. Chazan, J. Goldberger, and S. Gannot, "Fcn approach for dynamically locating multiple speakers," *ArXiv e-prints*, 2020, https://arxiv.org/abs/2008.11845.

[14] Y. Wu, R. Ayyalasomayajula, M. J. Bianco, D. Bharadia, and P. Gerstoft, "Sslide: Sound source localization for indoors based on deep learning," *IEEE Int. Conf. on Acoust. Speech and Signal Process. (ICASSP)*, pp. 4680–4684, 2021.

[15] P. Gerstoft, Y. Hu, M. J. Bianco, C. Patil, A. Alegre, Y. Freund, and F. Grondin, "Audio scene monitoring using redundant ad-hoc microphone array networks," *IEEE J. Internet of Things*, 2021, DOI: 10.1109/JIOT.2021.3103523.

[16] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A review of sound source localization with deep learning methods," *ArXiv e-prints*, 2021, https://arxiv.org/abs/2109.03465.

[17] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Semi-supervised sound source localization based on manifold regularization," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 8, pp. 1393–1407, Aug. 2016.

[18] R. Opochinsky, G. Chechik, and S. Gannot, "Deep ranking-based doa tracking algorithm," in *EUSIPCO*, 2021.

[19] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *Proc. Int. Conf. Learn. Represent.*, 2014.

[20] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Adv. Neural Info. Process. Sys.*, 2014, pp. 3581–3589.

[21] D. P Kingma, M. Welling, et al., "An introduction to variational autoencoders," *Found. and Trends in Machine Learn.*, vol. 12, no. 4, pp. 307–392, 2019.

[22] M. J. Bianco, S. Gannot, E. Fernandez-Grande, and P. Gerstoft, "Semi-supervised source localization in reverberant environments with deep generative modeling," *IEEE Access*, 2021, DOI: 10.1109/ACCESS.2021.3087697.

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

[24] J.H DiBiase, H.F Silverman, and M.S Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, pp. 157–180. Springer, 2001.

[25] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[26] A. Ajay, J. Wu, N. Fazeli, M. Bauza, L. P. Kaelbling, J. B. Tenenbaum, and A. Rodriguez, "Augmenting physical simulators with stochastic neural networks: Case study of planar pushing and bouncing," in *IEEE/RSJ Int. Conf. Intel. Robots Sys. (IROS)*. IEEE, 2018, pp. 3066–3073.

[27] T. Johannink, S. Bahl, A. Nair, J. Luo, A. Kumar, M. Loskyll, J.A. Ojea, E. Solowjow, and S. Levine, "Residual reinforcement learning for robot control," in *Int. Conf. Robotics Automat. (ICRA)*. IEEE, 2019, pp. 6023–6029.

[28] A. Zeng, S. Song, J. Lee, A. Rodriguez, and T. Funkhouser, "Tossingbot: Learning to throw arbitrary objects with residual physics," *IEEE Trans. Robotics*, vol. 36, no. 4, pp. 1307–1319, 2020.

[29] E. Fernandez-Grande, M. J. Bianco, S. Gannot, and P. Gerstoft, "Dtu three-channel room impulse response dataset for direction of arrival estimation 2020," *IEEE Dataport*, 2021, DOI: 10.21227/c5cn-jv76.

[30] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *IEEE Int. Conf. Acoust. Speech and Signal Process. (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[31] Google, "WebRTC," 2011.

[32] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, 2014.

[33] E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, . Karaletsos, R. Singh, P. Szerlip, P. Horsfall, and N. D. Goodman, "Pyro: Deep Universal Probabilistic Programming," *J. Mach. Learn. Res.*, 2018.