Semi-Supervised Source Localization in Reverberant Environments With Deep Generative Modeling

MICHAEL J. BIANCO[1], (Member, IEEE), **SHARON GANNOT**[2], (Fellow, IEEE),
EFREN FERNANDEZ-GRANDE[3], (Member, IEEE), **AND**
PETER GERSTOFT[1], (Senior Member, IEEE)

[1]Marine Physical Laboratory, University of California San Diego, San Diego, CA 92093, USA
[2]Faculty of Engineering, Bar-Ilan University, Ramat-Gan 5290002, Israel
[3]Department of Electrical Engineering, Technical University of Denmark, 2800 Kongens Lyngby, Denmark

Corresponding author: Michael J. Bianco (mbianco@ucsd.edu)

ABSTRACT Localization in reverberant environments remains an open challenge. Recently, supervised learning approaches have demonstrated very promising results in addressing reverberation. However, even with large data volumes, the number of labels available for supervised learning in such environments is usually small. We propose to address this issue with a semi-supervised learning (SSL) approach, based on deep generative modeling. Our chosen deep generative model, the variational autoencoder (VAE), is trained to generate the phase of relative transfer functions (RTFs) between microphones. In parallel, a direction of arrival (DOA) classifier network based on RTF-phase is also trained. The joint generative and discriminative model, deemed VAE-SSL, is trained using labeled and unlabeled RTF-phase sequences. In learning to generate and classify the sequences, the VAE-SSL extracts the physical causes of the RTF-phase (i.e., source location) from distracting signal characteristics such as noise and speech activity. This facilitates effective *end-to-end* operation of the VAE-SSL, which requires minimal preprocessing of RTF-phase. VAE-SSL is compared with two signal processing-based approaches, steered response power with phase transform (SRP-PHAT) and MUltiple SIgnal Classification (MUSIC), as well as fully supervised CNNs. The approaches are compared using data from two real acoustic environments - one of which was recently obtained at Technical University of Denmark specifically for our study. We find that VAE-SSL can outperform the conventional approaches and the CNN in label-limited scenarios. Further, the trained VAE-SSL system can generate new RTF-phase samples which capture the physics of the acoustic environment. Thus, the generative modeling in VAE-SSL provides a means of interpreting the learned representations. To the best of our knowledge, this paper presents the first approach to modeling the physics of acoustic propagation using deep generative modeling.

INDEX TERMS Source localization, semi-supervised learning, generative modeling, deep learning.

I. 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]–[3], acoustic localization in reverberant environments remains a major challenge [4]. There has been

The associate editor coordinating the review of this manuscript and approving it for publication was Liangtian Wan.

great interest in machine learning (ML)-based techniques in acoustics, including source localization and event detection [5]–[14]. One 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) [15]–[17].

We approach source localization from the perspective of SSL, with the intent of addressing real-world applications

of ML. It has been shown that it is relatively easy to generate large amounts of synthetic data resembling real-world sound measurement configurations. This synthetic data has then been used to train ML-based localization models (e.g., [9]), with very 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 this paper we contend, along the SSL paradigm, that if there is a microphone system with fixed geometry, recording in a room environment, that reverberant signals convey physical characteristics of the room – that given sufficient time and variety of source locations, the physics of the room may be well modeled using unsupervised ML [15], [18], [19]. It is further observed that if few labels are available, obtained for instance using cell-phone data, such unsupervised representations can be leveraged to localized sources in the room.

An SSL localization approach based on deep generative modeling with variational autoencoders (VAEs) [20]–[22] is proposed. Deep generative models [23], e.g., generative adversarial networks (GANs) [24], have received much attention for their ability to learn high-dimensional data distributions, including those of natural images [25]. GANs in acoustics have had success in generating raw audio [26] and speech enhancement [27]. VAEs have also had much success in speech enhancement [28]. VAEs are inspiring examples of representation learning [29] which use both encoder and decoder neural networks (NNs) to parameterize probability distributions. In this way explicit latent codes are obtained for their generative models.

VAEs are used along with discriminative models to perform SSL. Our work is based on seminal work in SSL with deep generative models [21]. The approach outlined in [21] has become a well-established method in literature for SSL with high-dimensional data. We extend the theory from [21], which focused on image processing, to the fields of acoustics and array processing. The VAE system is used to model physical data in real world acoustic environments. To the best of our knowledge, our paper presents the first approach to modeling the physics of acoustic propagation using deep generative modeling.

In our proposed approach, VAEs are used to encode and generate the phase of the relative transfer function (RTF) between two microphones [18]. The VAE is trained in parallel with a classifier network to benefit from both labeled and unlabeled examples. The resulting model estimates DOA and generates RTF-phase sequences. By learning to generate RTF-phase, the VAE-SSL system learns the physical model relating the latent model and DOA label to the RTF-phase.

This approach is a form of manifold learning [23], [30], [31]. Manifold learning has recently been proposed for semi-supervised source localization in room acoustics [15]. In this work, diffusion mapping is used to obtain a graphical model relating source location to RTF features. It was shown that a lower-dimensional latent representation of the RTFs, extracted by diffusion mapping, correlated well with source positions.

The VAE-SSL method is presented as an alternative to this manifold learning approach. Recent work has indicated the capabilities of deep learning in manifold learning [30]. VAE-SSL uses the non-linear modeling capabilities of deep generative modeling to obtain an SSL localization approach which is less reliant on preprocessing and hand-engineering of latent representations. Thus, the approach can be regarded as nearly *end-to-end*. The VAE-SSL system is designed to not rely on significant preprocessing of the RTFs. Through gradient-based learning it automatically determines the best latent and discriminative representations for the given task. Instead of spectral averaging, we input to the system a sequence of instantaneous RTFs and allow the statistical model to best utilize the patterns in the data.

The generative modeling approach has in general three benefits. First, the system learns a latent embedding of the RTF-phase sequences from the unlabeled data. This representation is leveraged by training with labeled data, which give the latent embedding semantic meaning — in our case, the relationship between the learned embedding and physical source locations in a room. Second, the system allows disentangling of the causes of RTF-phase signal (i.e., source DOA) from other signal factors including source frequency variation [22], [32]. The disentanglement enables the system to learn a task-specific representation of the RTF-phase patterns in input features that are most relevant to DOA estimation task. Finally, the system can be used to conditionally generate new RTF-phase sequences, based on sampling over the latent representation and DOA label. This output can be interpreted physically, and thereby allows us to verify that the system is obtaining a physically meaningful representation.

We build upon our previous work in deep generative modeling for source localization [33]. In this work, the VAE-SSL approach could well learn to localize acoustic sources with a few labeled RTF-phase sequences and extensive unlabeled data. The system was trained and tested on noise signals in simulated environments.

In our current study, the VAE-SSL concepts are extended to more realistic acoustic scenarios and the inference and generative architectures are refined. This includes consideration for the appropriate conditional distributions for the generative model. The learning-based approaches are trained and validated on speech data in two reverberant environments. As part of our study, we have obtained a new acoustic dataset at the Technical University of Denmark (DTU), which is availble online at IEEE Dataport [34]. This acoustic dataset consists of reverberant acoustic impulse responses (IRs) recorded in a classroom at DTU from several source locations. In this dataset, off-grid and off-range measurements are also obtained to test the generalization of the learning-based methods. The implicit acoustic model learned by the generative model is demonstrated and characterized. This is accomplished using the trained VAE-SSL system to conditionally generate RTF-phase sequences.

Reverberant speech is obtained by convolving dry speech with estimated IRs from two real-world datasets: the Multi-channel Impulse Response (MIR) database [35] and the DTU room acoustics dataset. Speech data was obtained from the LibriSpeech corpus [36].

The VAE-SSL method is implemented using convolutional neural networks (CNNs). The performance of the convolutional VAE-SSL in reverberant environments is assessed against two signal processing-based approaches: the steered response power with phase transform (SRP-PHAT) [37] and MUltiple SIgnal Classification (MUSIC) [38]. We further consider the performance of a fully supervised CNN. Comparison of MUSIC and SRP-PHAT in ML DOA estimation literature, as well as DOA estimation techniques more generally, is typical, e.g., [9]. Moreover, the MUSIC method was selected as the baseline method in the recent IEEE Signal Processing Society acoustic source LOCalization And TrAcking (LOCATA) challenge [39].

It is shown that VAE-SSL can outperform conventional source localization approaches, as well as fully supervised approaches for label-limited scenarios. This includes scenarios where the source may be off-range or off-grid from the design case, as well as variations in reverberation time and speech signals. Further, the implicit physical model obtained with VAE-SSL can be used to generate RTF-phase features. The physical characteristics of the generated RTF-phase is assessed by analyzing the phase wrap of the generated phase and the corresponding phase-delay for the generated RTF time domain representation.

A. SUMMARY OF OUR CONTRIBUTIONS

The main contributions of our paper are summarized below.

- A sound source localization approach is developed based on semi-supervised learning (SSL) with deep generative modeling. Variational autoencoders (VAEs) are used as the deep generative model. The VAE is paired with a classifier network and the systems are trained in parallel on labeled and unlabeled data. The approach is deemed VAE-SSL.
- The phase of the instantaneous relative transfer function (RTF) between two microphones is used as the input features to VAE-SSL.
- The inputs are only lightly processed. The gradient-based learning in the VAE-SSL system selects the physically relevant features in a nearly *end-to-end* fashion.
- We obtained an impulse-response (IR) dataset at the Technical University of Denmark (DTU) specifically for our study. The acoustic data was recorded in a reverberant classroom at DTU and included both on- and off-grid source locations. The dataset is available online at IEEE Dataport [34].
- Using IRs from real acoustic environments, the VAE-SSL approach was evaluated against the benchmark SRP-PHAT and MUSIC DOA estimation methods, as well as a fully supervised CNN. In addition to the

DTU dataset, IRs from the multichannel IR (MIR) database were also used for evaluating the methods.
- The generative modeling capabilities of VAE-SSL are demonstrated. The output of the generative model is interpreted physically and is found to correlate well with the expected acoustic characteristics.

II. THEORY

RTFs [18], specifically the RTF-phase, are used as the acoustic feature for our VAE-SSL approach. Since the RTF can be assumed independent of the source, this feature helps to focus ML on physically relevant features, and thereby reduces the sample complexity of the model. A temporal sequence of instantaneous RTF-phases is encoded as a function of source azimuth (direction of arrival, DOA). We choose the instantaneous RTF-phase, calculated using a single STFT frame (i.e., no averaging applied), to minimize the intervention of feature preprocessing. This allows the VAE-SSL method to extract end-to-end the important features for source localization and RTF-phase generation.

A. RELATIVE TRANSFER FUNCTION (RTF)

Consider 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). Then, the relative transfer function (RTF) is defined as [15], [18]

$$h(k) = \frac{a_2(k)}{a_1(k)}, \tag{2}$$

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),

$$\widehat{h}(k) = \frac{d_2(k)}{d_1(k)}. \tag{3}$$

This estimator is biased since we neglect the noise spectra [Ref. [15], Eq. (5)]. While an unbiased estimator can be obtained, see [40], [41], we observe that the biased estimate $\widehat{h}(k)$ works well for our purposes. For each STFT frame, a vector of RTFs is obtained $\widehat{\mathbf{h}} = [\widehat{h}(1) \dots \widehat{h}(K)]^{\mathrm{T}} \in \mathbb{C}^K$ with K the number of frequency bins used.

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

$$\mathbf{x}_n = \mathrm{vec}(\mathrm{phase}(\widehat{\mathbf{H}}_n)) \in \mathbb{R}^{KP}, \tag{4}$$

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.

B. SEMI-SUPERVISED DOA ESTIMATION

In acoustics we are often faced with scenarios where there are large volumes of acoustic recordings from arrays but potentially only few labels. The recordings themselves contain physical information, but the paucity of labels limit the task-specific value of this physical information. This issue is addressed by formulating an SSL-based approach to source localization. A VAE model is used to obtain latent distributions of the RTF-phase physics for a particular room and corresponding DOAs from labeled and unlabeled examples.

In our SSL formulation to DOA estimation, only a subset of the full dataset containing N RTF sequences have corresponding DOA labels. Here the DOA labels are represented by $\mathbf{y} \in \{0, 1\}^T$, a one-hot encoding, with T the number of DOA classes. We thus have labeled and unlabeled sets, defined by $\{\mathbf{x}_j, \mathbf{y}_j\} \in \mathcal{D}_l$ and $\{\mathbf{x}_u\} \in \mathcal{D}_u$. 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.

The $N - J$ unlabeled RTFs contain physical information, which is extracted by unsupervised learning. The J labeled RTFs have both physical information and corresponding DOA labels, which help give the model task-specific value for DOA estimation by training the classifier. The generative model, which is trained on all samples helps guide the classifier training when labels are not available. Our goal is thus to well-infer the labels corresponding to labeled and unlabeled RTFs based on the trained VAE-SSL model.

C. SEMI-SUPERVISED LEARNING WITH VAEs

VAEs [22] combine the rich, high-dimensional modeling capacity of NNs with the tools of variational inference (VI) [42] to learn generative models of high-dimensional distributions. In VAEs, the distributions are parameterized by NNs, and the NN parameters are updated to best fit the assumed distributions to the data.

We formulate a principled SSL framework based on VAEs [21]. A classifier NN and VAE are trained jointly, using both labeled and unlabeled data. The approach treats the label $\mathbf{y}$ as either latent and observed, depending on whether the data is labeled ($\{\mathbf{x}_j, \mathbf{y}_j\} \sim \mathcal{D}_l$) or unlabeled ($\{\mathbf{x}_u\} \sim \mathcal{D}_u$). This corresponds to the 'M2' model in [21]. To simplify notation, subscripts are disregarded for the theory derivation.

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 data 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. As will be discussed in Sec. II-D, we use $p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z}) = \overline{\mathcal{N}}(\cdot|\cdot)$, 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 $p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z})$ to the data. $\mathbf{y}$ and $\mathbf{z}$ are assumed independent,
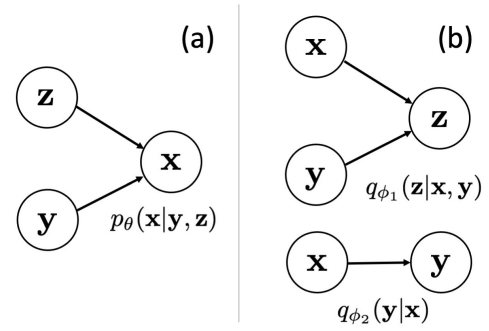
FIGURE 1. Graphical models for VAE-SSL (a) generative and (b) inference models. The generative model conditional distribution is parameterized by a decoder, see Fig. 2(c), and the inference model conditional distributions are parameterized by encoders, see Fig. 2(a,b).

with their marginal densities $p(\mathbf{y})$ and $p(\mathbf{z})$ the categorical and normal distributions (see Sec. II-D). Thus, the generative model is $p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z})p(\mathbf{y})p(\mathbf{z})$, giving the graphical model Fig. 1(a).

Now we are presented with the challenge of inferring $\mathbf{y}$ (when it is not specified) and the latent variable $\mathbf{z}$. For labeled data $\{\mathbf{x}_j, \mathbf{y}_j\} \sim \mathcal{D}_l$, the posterior of the latent variable $\mathbf{z}$ is

$$p(\mathbf{z}|\mathbf{x}, \mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x}, \mathbf{y})} \tag{5}$$

and for unlabeled data $\{\mathbf{x}_u\} \sim \mathcal{D}_u$, the joint posterior of the latent $\mathbf{z}$ and the label $\mathbf{y}$ is

$$p(\mathbf{y}, \mathbf{z}|\mathbf{x}) = \frac{p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z})p(\mathbf{y}, \mathbf{z})}{p(\mathbf{x})}. \tag{6}$$

Direct estimation of the posterior, e.g., from (5), $p(\mathbf{z}|\mathbf{x}, \mathbf{y})$ 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. We define a variational distribution $q_{\phi_1}(\mathbf{z}|\mathbf{x}, \mathbf{y})$ 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. 2(b)). For $q_{\phi_1}(\mathbf{z}|\mathbf{x}, \mathbf{y})$, a normal distribution is used. The graphical model for the corresponding inference model is shown in Fig. 1(b). The inference model for the DOA label $q_{\phi_2}(\mathbf{y}|\mathbf{x})$, parameterized by label inference encoder with parameters ϕ_2 (see Fig. 2(a)) is obtained starting in (10). For $q_{\phi_2}(\mathbf{y}|\mathbf{x})$, the categorical distribution is used. Like the generative model the inference model NN parameters $\Phi = \{\phi_1, \phi_2\}$ are adjusted so the assumed distributions best model the data. The distributions are defined in Sec. II-D.
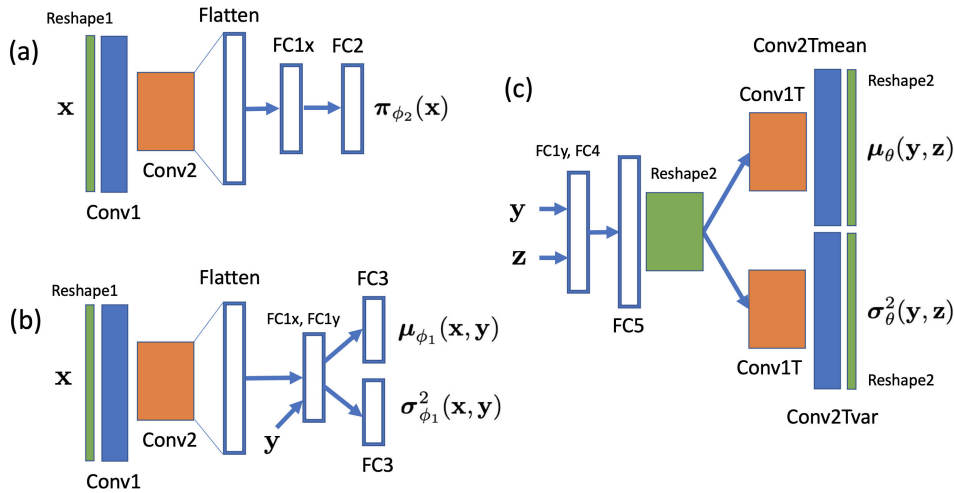
Starting with the model for the labeled data, see (5), per VI we seek $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})$ which minimizes the KL-divergence

$$\{\phi_1, \theta\} = \underset{\phi_1, \theta}{\arg\min} \, \mathrm{KL}(q_{\phi_1}(\mathbf{z}|\mathbf{x}, \mathbf{y})||p(\mathbf{z}|\mathbf{x}, \mathbf{y})). \tag{7}$$

Considering first the labeled data $\mathcal{D}_l$, assessing the KL-divergence, from (7)

$\mathrm{KL}(q_{\phi_1}(\mathbf{z}|\mathbf{x}, \mathbf{y})||p(\mathbf{z}|\mathbf{x}, \mathbf{y}))$
$= \mathbb{E}\big[\log q_{\phi_1}(\mathbf{z}|\mathbf{x}, \mathbf{y})\big] - \mathbb{E}\big[\log p(\mathbf{x}, \mathbf{y}|\mathbf{z})p(\mathbf{z})\big] + \log p(\mathbf{x}, \mathbf{y})$
$= -\mathrm{ELBO} + \log p(\mathbf{x}, \mathbf{y}) \geq 0, \tag{8}$

FIGURE 2. Neural network configurations. Encoders for (a) label inference (classifier) and (b) latent model. (c) Decoder for generative model.

with the expectation $\mathbb{E}$ relative to $q_{\phi_1}(\mathbf{z}|\mathbf{x}, \mathbf{y})$. This reveals the dependence of the KL divergence on evidence $p(\mathbf{x}, \mathbf{y})$, which is intractable. The other two terms in (8) form the evidence lower bound (ELBO). Since the KL divergence is non-negative, the ELBO 'lower bounds' the evidence: ELBO $\leq \log p(\mathbf{x}, \mathbf{y})$. We thus minimize $-$ELBO.

Considering the ELBO terms from (8), the objective for labeled data is formulated, with $\mathbf{y}$ and $\mathbf{z}$ independent (in the generative model, see Fig. 1(a))

$$
\begin{aligned}
-C(\theta, \phi_1; \mathbf{x}, \mathbf{y}) &= \mathbb{E}\big[\log p(\mathbf{x}, \mathbf{y}|\mathbf{z})p(\mathbf{z}) - \log q_{\phi_1}(\mathbf{z}|\mathbf{x}, \mathbf{y}) \big] \\
&= \mathbb{E}\big[\log p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z}) + \log p(\mathbf{y}) + \log p(\mathbf{z}) \\
&\quad - \log q_{\phi_1}(\mathbf{z}|\mathbf{x}, \mathbf{y}) \big].
\end{aligned}
\tag{9}
$$

This follows [Ref. [21], Eq.(6)].

Next, an objective for unlabeled data $\mathcal{D}_u$ is derived. The intractable posterior from (6) is approximated by $q_\Phi(\mathbf{y}, \mathbf{z}|\mathbf{x}) \approx p(\mathbf{y}, \mathbf{z}|\mathbf{x})$. From the KL-divergence, we find the objective (negative ELBO), using terms from (6), as

$$
\begin{aligned}
-D(\theta, \Phi; \mathbf{x}) &= \mathbb{E}_{q_{\phi_2}(\mathbf{y}|\mathbf{x})}\big[\mathbb{E}_{q_{\phi_1}(\mathbf{z}|\mathbf{x},\mathbf{y})}\big[\log p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z}) \\
&\quad + \log p(\mathbf{y}) + \log p(\mathbf{z}) - \log q_{\phi_1}(\mathbf{z}|\mathbf{x}, \mathbf{y}) \\
&\quad - \log q_{\phi_2}(\mathbf{y}|\mathbf{x}) \big]\big] \\
&= \sum_{\mathbf{y}} q_{\phi_2}(\mathbf{y}|\mathbf{x})\big[-C(\theta, \phi; \mathbf{x}, \mathbf{y}) - \log q_{\phi_2}(\mathbf{y}|\mathbf{x}) \big]
\end{aligned}
\tag{10}
$$

This follows [Ref. [21], Eq.(7)]. More details of the derivation are given in Appendix A.

Assessing the terms in the supervised learning objective $C(\theta, \phi_1; \mathbf{x}, \mathbf{y})$ (9), it does not condition the label $\mathbf{y}$ on the sequence $\mathbf{x}$, since for the supervised case the label is assumed known per (5). The term $\log q_{\phi_2}(\mathbf{y}|\mathbf{x})$ is only present in the unsupervised learning objective (10). In this configuration, the classifier network learns only from the unsupervised sequence. 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 [21].

An overall objective for training the VAE and classifier models using labeled and unlabeled data is derived by combining (9) and (10) with the auxiliary term

$$
\mathcal{L} = \sum_{\{\mathbf{x}_j, \mathbf{y}_j\} \sim \mathcal{D}_l} C(\theta, \phi_1; \mathbf{x}_j, \mathbf{y}_j) - \alpha \log q_{\phi_2}(\mathbf{y}_j|\mathbf{x}_j)
$$
$$
+ \sum_{\{\mathbf{x}_u\} \sim \mathcal{D}_u} D(\theta, \Phi; \mathbf{x}_u),
\tag{11}
$$

with α a scaling term, selected by hyperparameter optimization. This follows [Ref. [21], Eqs.(8,9)].

D. DISTRIBUTIONS AND NEURAL NETWORK (NN) PARAMETERS

The VAE-SSL model consists of 3 CNNs which, for training and conditional generation, parameterize the probability distributions in (9) and (10). The networks correspond to the inference model, with parameters Φ, and the generative model, with parameters θ. See Fig. 2 and Table. 1 for NN configurations, and Sec. III-D for further details on implementation.

For the inference model, we specify the following distributions, parameterized by NNs:

- $q_{\phi_1}(\mathbf{z}|\mathbf{y}, \mathbf{x}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{\phi_1}(\mathbf{x}, \mathbf{y}), \mathrm{diag}(\boldsymbol{\sigma}^2_{\phi_1}(\mathbf{x}, \mathbf{y})))$, 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. 2(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. 2(a)).

For the generative model, we specify the distribution as

- $p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z}) = \overline{\mathcal{N}}(\mathbf{x}|\boldsymbol{\mu}_\theta(\mathbf{y}, \mathbf{z}), \mathrm{diag}(\boldsymbol{\sigma}^2_\theta(\mathbf{y}, \mathbf{z})))$, 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. 2(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 defined as $p(\mathbf{y}) = \text{Cat}(\mathbf{y}|\boldsymbol{\pi})$, with $\boldsymbol{\pi} \in \mathbb{R}^T$ the probabilities of the classes, which are assumed equal with $\pi_t = 1/T$; and $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Since the decoder parameterizes a truncated normal density, hyperbolic tangent activation is applied to the outputs of the decoder to constrain the mean, and a scaled sigmoid activation is applied to the decoder variance: $\boldsymbol{\mu}_\theta(\mathbf{y}, \mathbf{z}) \in [-1, 1]$ and variance $\boldsymbol{\sigma}_\theta^2(\mathbf{y}, \mathbf{z}) \in [0, 10]$. The input to VAE-SSL is normalized by π, thus $[-1, 1]$ corresponds to the interval $[-\pi, \pi]$. The variance is limited to improve the training speed, since the truncated normal is implemented using rejection sampling.

The optimization of the VAE-SSL parameters $\{\theta, \Phi\}$ from the objective (11) is performed using stochastic variational inference (SVI) [20]. The neural networks are implemented using Pytorch [43] and SVI is here implemented with the probabilistic programming package Pyro [44], which evaluates the ELBO gradients with Monte Carlo sampling from the ELBO terms directly (instead of reparameterization, e.g., [21]).

During each training epoch, minibatches of RTF-phase sequences are drawn from the labeled $\{\mathbf{x}_j, \mathbf{y}_j\} \sim \mathcal{D}_l$ or unlabeled $\{\mathbf{x}_u\} \sim \mathcal{D}_u$ sets. In the case of $\mathcal{D}_l$, the terms corresponding to $\mathcal{D}_l$ in (11) are evaluated. Similarly, for $\mathcal{D}_u$, only the unlabeled terms are evaluated. Labeled minibatches were used approximately once per N/J unlabeled batches. For each data sample in the minibatch, the ELBO terms are sampled once. This has been found to be sufficient, provided the minibatch size is large enough [21] (we use minibatch size of 256, see Sec. III-D).

E. LABEL ESTIMATION AND CONDITIONAL RTF GENERATION

From the trained inference model, the DOA is estimated by the indicator function

$$y_{\rho n} = \mathbf{1}_{\rho = \hat{t}}, \qquad (12)$$

with

$$\hat{t} = \arg\max_t (\pi_{t, \phi_2}(\mathbf{x}_n)), \qquad (13)$$

and ρ and t the discrete DOA indices, and $\pi_{t, \phi_2}(\mathbf{x}_n))$ the t-th output of $\boldsymbol{\pi}_{\phi_2}(\mathbf{x}_n))$. The DOA angle represented by the one-hot encoding $\mathbf{y}_n$ with active index t is D_{tn} and its estimate $\hat{D}_{tn}$.

RTF-phase sequences can be conditionally generated from the trained generative model for a given label $\mathbf{y}_n$ using the prior $p(\mathbf{z})$ and density $p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z})$

$$\begin{aligned} \mathbf{z}_s &\sim p(\mathbf{z}) \\ \mathbf{x}_s &\sim p_\theta(\mathbf{x}|\mathbf{y}_n, \mathbf{z}_s), \end{aligned} \qquad (14)$$

with subscript s denoting the sampled latent variable and RTF-phase sequence.

RTF-phase sequences can also be reconstructed (approximated) using the trained inference and generative models. For a given RTF-phase sequence $\mathbf{x}_n$ and label $\mathbf{y}_n$, $\mathbf{z}$ is sampled from the inference model

$$\mathbf{z}_s \sim q_{\phi_1}(\mathbf{z}|\mathbf{x}_n, \mathbf{y}_n), \qquad (15)$$

and the reconstruction $\hat{\mathbf{x}}_n$ is obtained by sampling from the generative model by

$$\hat{\mathbf{x}}_n \sim p_\theta(\mathbf{x}|\mathbf{y}_n, \mathbf{z}_s). \qquad (16)$$

III. EXPERIMENTS

We assess the DOA estimation performance of the VAE-SSL approach in moderately reverberant environments against three alternative techniques: SRP-PHAT [45], MUSIC [38], and a supervised CNN baseline. The VAE-SSL and fully supervised CNN are trained and validated using speech to obtain real-world application performance.

The performance of the methods, summarized in Table 2 and 3, is quantified in terms of mean absolute error (MAE) and sequence-level accuracy (Acc.). Further, the generative modeling capabilities of the trained VAE-SSL system are assessed, as shown in Fig. 4–11.

The MAE is defined as

$$\text{MAE} = \frac{1}{N - J} \sum_{\mathcal{D}_u} |D_{tu} - \hat{D}_{tu}|, \qquad (17)$$

with $|\cdot|$ denoting the absolute value. The sequence level accuracy (Acc.) is defined by

$$\text{Acc.} = \frac{100}{N - J} \sum_{\mathcal{D}_u} \mathbf{1}_{D_{tu} = \hat{D}_{tu}}. \qquad (18)$$

A. MEASURED IMPULSE RESPONSES

We use measured IRs from two different real-world datasets: a dataset recently recorded at DTU [34], and the MIR database [35]. We here briefly describe the experimental configuration for the DTU dataset. The data were recorded in a classroom at DTU in June 2020 (Fig. 3). The classroom was approximately rectangular, of dimensions $9 \times 6 \times 3$ m and fully furnished. One of the sidewalls is irregular, with a 100×40 cm extrusion (for heating and ventilation) across the entire wall, in addition to support columns. All walls have scattering elements mounted on them (whiteboards, blackboards, diffusers, and windowpanes), and the sound reflections are not specular. The nominal source-array range was 1.5 m. IRs were obtained from 19 DOAs (10° resolution over the interval $[-90°, 90°]$). The reverberation time in the classroom was estimated to be $\text{RT}_{60} = 500$ ms. There were two microphones, with 8.5 cm spacing. The sampling rate was 48 kHz. The IRs were truncated to 1 s and 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 to test the generalization of learning-based localization methods. The off-grid source DOAs were $[25°, 28°, 45°]$

TABLE 1. Network parameters corresponding to configuration in Fig. 2. Since the kernel size is 3 and stride is 2 in the convolution layers, the output dimension, given input dimension m, is $w(m) = (m - 1)/2$, with m odd.

Name	Type	Input shape	Output shape	Kernel	Activation
Reshape1	Reshape	$[K \times P]$	$[K, P, 1]$	–	–
Conv1	Convolution	$[K, P, 1]$	$[w(K), w(P), 32]$	$[3, 3]$	ReLU
Conv2	Convolution	$[w(K), w(P), 32]$	$[w(w(K)), w(w(P)), 64]$	$[3, 3]$	ReLU
Flatten	Reshape	$[w(w(K)), w(w(P)), 64]$	$[w(w(K)) \times w(w(P)) \times 64]$	–	–
FC1x	Fully connected	$[w(w(K)) \times w(w(P)) \times 64]$	$[200]$	–	ReLU
FC1y	Fully connected	$[T]$	$[200]$	–	ReLU
FC2	Fully connected	$[200]$	$[T]$	–	Softmax
FC3	Fully connected	$[200]$	$[50]$	–	None
FC4	Fully connected	$[50]$	$[200]$	–	ReLU
FC5	Fully connected	$[200]$	$[w(w(K)) \times w(w(P)) \times 64]$	–	ReLU
Reshape2	Reshape	$[w(w(K)) \times w(w(P)) \times 64]$	$[w(w(K)), w(w(P)), 64]$	–	–
Conv1T	Transpose conv.	$[w(w(K)), w(w(P)), 64]$	$[w(K), w(P), 32]$	$[3, 3]$	ReLU
Conv2Tmean	Transpose conv.	$[w(K), w(P), 32]$	$[K, P, 1]$	$[3, 3]$	Tanh
Conv2Tvar	Transpose conv.	$[w(K), w(P), 32]$	$[K, P, 1]$	$[3, 3]$	Sigmoid
Reshape2	Reshape	$[K, P, 1]$	$[K \times P]$	–	–

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). More details are available online [34].

The MIR database was recorded in a $6 \times 6 \times 2.4$ m room with reverberation time controlled by acoustic panels [35]. The dataset consists of source DOAs on a $15°$ resolution over $[-90°, 90°]$, giving 13 DOAs. Each of the DOAs was obtained at two ranges (1 and 2 m) with three reverberation times (RT_{60} is 160, 360, and 610 ms). The IRs were obtained for 8 microphones located in the center of the room. There were several configurations, with different microphone spacing. We used data with 3 and 8 cm spacing, specifically the two center microphones, with 8 cm spacing. The sampling rate for the MIR database was also 48 kHz. Further, from MIR only the 2 m source range is used, with reverberation times $RT_{60} = 360$ and 610 ms. The MIR IRs were processed in the same manner as the DTU IRs. For more details of the MIR database, see [35].

B. RTF CALCULATIONS

The signal at the microphones is given in (1). RTFs are obtained from the data by (3). 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 $N_{FFT}/2 + 1$, including all frequencies between 0 (DC) up to $N_{FFT}/2$, to support strided transpose convolution without padding). For fair comparison, SRP-PHAT and MUSIC used the $P = 31$ STFT frames to estimate the RTFs, with $K = N_{FFT}/2 + 1$. Thus, the temporal length of the sequences for all methods was 0.26 s.

C. SPEECH DATA PROCESSING

We used speech data from the LibriSpeech corpus [36] development set, which contains 5.4 hours of recorded speech with a 16 kHz sampling rate from public domain audiobooks. The dataset contains equal numbers of male and female speakers (20 each).

The speech segments from LibriSpeech were convolved with the recorded room IRs to obtain reverberant speech. Voice activity detection (VAD) was performed on the dry speech before convolution with the IRs. The WebRTC project VAD system [46] was used. WebRTC is a popular open source VAD system based on pretrained Gaussian mixture models. The VAD settings used were 3 (most aggressive) with a 10 ms analysis window.

VAD was applied to the entire LibriSpeech development corpus, and a total of 40 speech segments 2-3 s in duration were randomly selected for training and validation, 20 segments each. This yielded $\sim 110,000$ RTF sequence for the nominal DTU IRs and $\sim 76,000$ sequences for the MIR IRs (~ 5800 sequences per DOA for both datasets). This further yielded $\sim 17,500$ and $\sim 35,000$ sequences for the DTU off-grid and off-range measurements using the validation speech. Given one active speaker recording segment of time length t, with 50% STFT overlap, the number of RTF-phase sequences in the segment N_{seg} is

$$N_{seg} \approx \frac{2Tf_s t}{N_{\text{FFT}}} - P. \tag{19}$$

For training and validation, J labeled sequences were drawn uniformly from the concatenated reverberant speech sequences (for each DOA). The remaining sequences from the training set were used for unsupervised learning. Sensor noise with 20 dB SNR was added to the microphone signals (see (1)). A range of values for J were considered, for evaluating the effect of the number of labeled sequences on the performance of the learning-based approaches. SRP-PHAT and MUSIC were evaluated using the full dataset of N sequences, for each scenario.

TABLE 2. Localization performance of VAE-SSL, fully supervised CNN, SRP-PHAT, and MUSIC on (a, b) DTU dataset and (c, d) MIR database. Training and validation DOA estimation performance given for unlabeled data $\mathcal{D}_u$. Performance quantified in terms of mean absolute error (MAE) and sequence level accuracy (Acc.).

(a) DTU training data

	VAE-SSL		CNN	
J (# labels)	MAE	Acc.	MAE	Acc.
114	6.34	56.2	31.5	40.3
247	3.38	75.5	25.4	49.9
494	1.85	84.9	20.4	60.1
988	1.36	88.4	17.8	63.4
1995	1.29	89.3	15.0	70.1
3990	1.03	91.6	12.5	74.6
7999	0.73	94.0	10.0	81.2
SRP-PHAT	4.72	66.1		
MUSIC	22.1	25.5		

(b) DTU validation data

	VAE-SSL		CNN	
J (# labels)	MAE	Acc.	MAE	Acc.
114	7.81	52.0	33.1	37.8
247	5.05	68.7	27.6	46.5
494	3.56	76.0	23.2	55.6
988	2.99	79.5	20.3	59.1
1995	3.01	79.9	17.8	64.8
3990	2.92	81.4	16.2	68.0
7999	3.00	80.5	14.3	73.7
SRP-PHAT	5.99	60.9		
MUSIC	22.2	24.9		

(c) MIR training data

	VAE-SSL		CNN	
J (# labels)	MAE	Acc.	MAE	Acc.
117	9.65	82.0	17.0	51.1
247	1.89	92.4	9.32	67.8
494	1.50	92.8	6.11	77.1
988	1.41	93.4	4.69	82.3
1989	1.37	93.8	4.04	84.2
3991	1.38	93.9	3.01	88.5
7995	1.12	95.1	2.09	91.9
SRP-PHAT	6.67	71.5		
MUSIC	43.2	21.7		

(d) MIR validation data

	VAE-SSL		CNN	
J (# labels)	MAE	Acc.	MAE	Acc.
117	12.9	69.3	16.3	47.6
247	3.93	81.4	9.99	60.9
494	3.46	82.1	7.68	67.9
988	3.25	83.3	6.36	72.9
1989	3.32	83.5	5.71	74.4
3991	3.61	82.2	5.31	76.2
7995	3.11	84.4	4.96	77.1
SRP-PHAT	5.34	74.6		
MUSIC	36.3	29.4		

FIGURE 3. Classroom and measurement configuration at Technical University of Denmark (DTU) where room IRs were collected [33].

D. LEARNING-BASED MODEL PARAMETERS

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. 2, and the corresponding parameters are given in Table 1. Each NN used two convolutional layers without pooling, and two fully connected layers. The convolutional layers help process the high-dimensional RTF-sequence data by enabling parameter-sharing between neighboring pixels [23]. Since the one-hot DOA **y** is relatively low-dimensional, it does not necessitate processing by convolutional layers. Thus, it is input directly to the first fully connected layer in the encoder model (Fig. 2(b)) after the convolutional processing of **x**.

Before processing by the CNNs, Fig. 2(a,b), the generated RTF-phase sequence vectors $\widehat{\mathbf{x}}_n \in \mathbb{R}^{KP}$ are reshaped to a matrix with dimensions $P \times K$, so the convolutional filters account for correlations between time and frequency bins. The output of the CNNs (Fig. 2(c)) is reshaped to a vector with length KP.

Several subnetworks reuse the same architecture; hence the names are reused. However, weights for each implementation are trained independently – i.e., no weight sharing is applied. The latent variable **z** dimension for all experiments was $M = 50$, assuming that the hidden representation must account for temporal variation of the RTF-phase sequences, and residual variation of the RTF from speech. Dropout with probability of 0.5 is applied to the large fully connected layers FC1x and FC5 (Fig. 2, Table 1).

For the fully supervised CNN approach, we used the classifier CNN architecture from the VAE-SSL model. The fully supervised CNN was trained using only the labeled data $\mathcal{D}_l$, without integration with the generative model. Like VAE-SSL, the CNN was implemented using Pytorch.

TABLE 3. Localization error (MAE) of VAE-SSL, fully supervised CNN, SRP-PHAT, and MUSIC for off-grid and off-range measurements from DTU dataset.

Method	Off-grid	Off-range
MUSIC	16.5	15.3
SRP-PHAT	6.29	4.16
CNN	8.96	47.7
VAE-SSL	**6.15**	**2.72**
RTF-1NN	7.33	38.0

All NNs were optimized using Adam [47]. For all cases, the learning rate was 5e-5 and the minibatch size was 256. The default momentum and decay values (0.9,0.999) were used from the Pytorch implementation of Adam. The VAE-SSL has the tuning parameter, the auxiliary multiplier α, which weights the supervised objective (11). It was found in the experiments that the performance of VAE-SSL was not very sensitive to α, though the best value per validation accuracy was found by grid search over the interval [500,10000].

E. NON-LEARNING: SRP-PHAT AND MUSIC CONFIGURATION

The same STFT frames used to calculate the RTF-phase sequences for VAE-SSL and CNN were used by the conventional SRP-PHAT [45], [48] and MUSIC [38], [39] approaches. There were P STFT frames in the VAE-SSL and CNN method input sequences $\mathbf{x}$. Thus, the conventional approaches used 13 candidate DOAs over $[-90°, 90°]$ for the MIR IRs and 19 for the DTU IRs. We used the SRP-PHAT and MUSIC implementations from the Pyroomacoustics toolbox [49]. For MUSIC, the number of sources was set to one.

F. TRAINING AND LOCALIZATION PERFORMANCE

For VAE-SSL and supervised CNN training, J labeled sequences were drawn from the training set and J labeled sequences were drawn from the validation set. The model was chosen based on the validation accuracy for labeled sequences. For VAE-SSL, the additional $N - J$ 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. After training, the performance of the models was evaluated using the unlabeled sequences from the validation dataset. Since for the unlabeled examples, the DOAs are technically unavailable, for unsupervised learning some RTF-phase sequences $\{\mathbf{x}_u\} \in \mathcal{D}_u$ contain frames with different DOA labels. The labeled sequences, and the unlabeled sequences used for performance evaluation, have only one DOA each. All the RTF-phase sequences $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_N]$ were normalized by π for the VAE-SSL and CNN approaches.

The performance of VAE-SSL and the competing approaches for the DTU IR dataset are given in Table 2(a,b). VAE-SSL and fully supervised CNN were trained using 20 speech signals and validated using an additional 20 speech
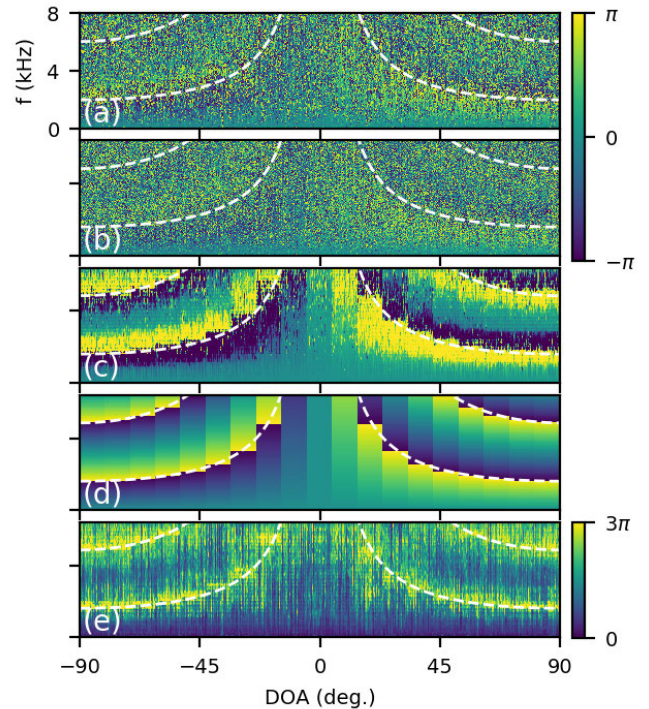
FIGURE 4. Reconstruction of RTF-phase sequences using VAE-SSL (trained using $J = 988$ labels). 10 sequences are reconstructed for each DOA using the DTU IR data convolved with speech. (a) Input RTF-phase sequences $\mathbf{x}_n$. (b) Reconstructed RTF-phase sequences $\hat{\mathbf{x}}_n$ (16). (c) RTF-phase sequence mean $\mu_\theta(\mathbf{y}_n, \mathbf{z}_s)$. (d) Free space RTF (22). (e) RTF-phase sequence standard deviation $\sigma_\theta(\mathbf{y}_n, \mathbf{z}_s)$. Phase-wrap function plotted with white dashed lines. All plot scales in radians, with (e) a larger range than (a–d). Theory, other parameters located in Sec. II D–E, Sec. III D.

signals (see Sec. III-C). We use $J = 247, 494, 988, 1995, 3990, 7999$ supervised sequences, which are multiples of the number of candidate DOAs ($T = 19$) to ensure an equal number of labeled samples for each DOA. It is observed that the VAE-SSL approach generalizes to the validation data, with better performance than SRP-PHAT with as few as $J = 247$ labeled sequences (or 13 per DOA). For both the validation and training data, the performance increases significantly for additional labels. The VAE-SSL outperforms CNN and MUSIC for all the experiments in this paper. It is apparent that in using the minimally processed RTF sequences, that our semi-supervised approach can learn to identify the relevant features from the data.

Similar trends are observed for the MIR database IRs, given in Table 2(d,e). Here VAE-SSL and fully supervised CNN were trained and validated using the same 20 speech signals with two different reverberation times. The training data IR was the $RT_{60} = 610$ ms case, and the validation $RT_{60} = 360$ ms. We use $J = 247, 494, 988, 1989, 3991, 7995$ supervised sequences, which are multiples of the number of candidate DOAs ($T = 13$). It is observed that VAE-SSL significantly outperforms the fully supervised CNN, SRP-PHAT, and MUSIC methods for both the training and validation cases when the number of supervised examples is sufficient. A similar number of labels to the DTU dataset
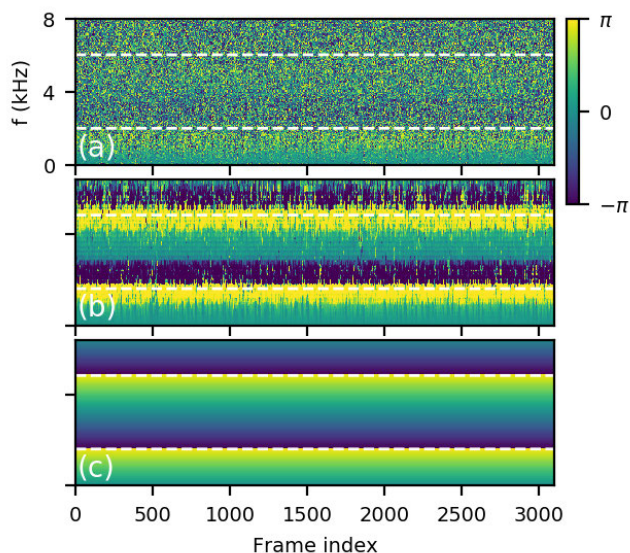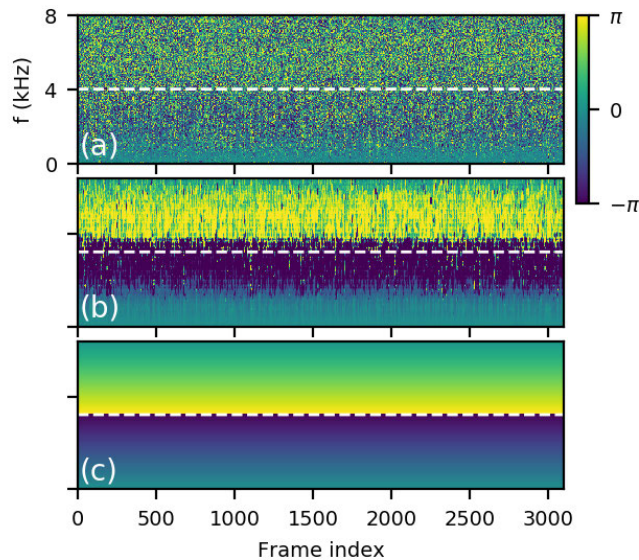
FIGURE 5. Conditionally generated RTF-phase sequences for DOA 90°, using VAE-SSL generative model. VAE-SSL trained with DTU IR data convolved with speech with $J = 988$ labels. (a) Conditionally generated RTF sequences $\mathbf{x}_s$ (14). (b) Generated RTF-phase sequence mean $\mu_\theta(\mathbf{y}_n, \mathbf{z}_s)$. (c) Free space RTF (22). 100 sequences were generated (with 31 frames per sequence, 100*31=3100 frames).

FIGURE 6. Conditionally generated RTF-phase sequences for DOA=−30° (trained using $J = 988$ labels). Same configuration as Fig 5.

were required to outperform SRP-PHAT: $J = 247$ labeled sequences, or in this case 19 labels per DOA. It is shown that the VAE-SSL generalizes well to different reverberation times.

Here it is noted that, in the case of the DTU IR dataset, the CNN performs well in terms of accuracy (which overtakes SRP-PHAT), though the MAE remains large. For the MIR dataset, both CNN and MAE converge to that of SRP-PHAT. The acoustic environment for the DTU dataset, with classroom furniture and irregular wall geometry, is more complex than the MIR database. This can be seen in the reduction of the accuracy for all approaches, despite the reduction in the number of DOAs (13 for DTU vs 19 for MIR). The CNN system cannot generalize well to the unseen data, since not enough labeled frames are available. While the accuracy is good, the mistakes by the CNN are large. As accuracy alone can be misleading, both metrics are considered. For VAE-SSL we find that the system generalizes well to the DTU environment using only few labels, since it extracts physically meaningful features as will be shown in Sec. III-G and III-H.

G. OFF-GRID AND OFF-RANGE GENERALIZATION

We have quantified the performance of the localization methods for different speakers and reverberation times. We now assess the off-grid and off-range localization performance of the methods in a real-room environment with the DTU dataset. This important test examines whether the representations learned by the ML approaches, VAE-SSL and fully supervised CNN, can generalize to source locations that were not present in the training data. In real applications, it can be expected that the source locations will not precisely

correspond to the training locations. The off-grid DOAs are 2 to 5° away from the training DOAs, and the off-range are ±0.5, and + 1 to 2 m away, see Sec. IIIA.

For the learning-based approaches, results are given for systems trained with $J = 988$ labels. For additional comparison, a basic one-nearest-neighbors approach is considered, which uses spectrally averaged RTF features. This manual preprocessing should reduce the noise in the RTF features and give a reasonable baseline for a hand-crafted, but still data-driven, approach.

For the nearest-neighbors approach, here deemed RTF-1NN, spectral averaging is performed over the full RTFs [15] using the STFT frames used to estimate the instantaneous RTF sequence (Sec.II-A). The spectrally averaged RTFs for each DOA are given by

$$\widetilde{h}(k) = \frac{\sum_{s \in S} d_{2s} d_{1s}^*}{\sum_{s \in S} d_{1s} d_{1s}^*}, \tag{20}$$

with $\cdot^*$ the complex conjugate and S the set of STFT frames used. The RTF from the labeled data corresponding to each DOA is $\widetilde{h}_t(k)$, with $|S| = PJ/T$. The RTF from each sequence in the unlabeled data, which is to be classified, is $\widetilde{h}_u(k)$ with $|S| = P$. The DOA for $\widetilde{h}_u(k)$ is estimated by

$$\arg \min_t \|\widetilde{\mathbf{h}}_t - \widetilde{\mathbf{h}}_u\|_2, \tag{21}$$

with $\| \cdot \|_2$ the ℓ_2 norm.

The results of the different methods are shown in Table 3. We find that the VAE-SSL approach generalizes well to off-grid and off-range sources. It is shown that the trained VAE-SSL outperforms the other approaches. Again, this performance is achieved using minimal preprocessing of the input features. Particularly, for the off-range scenario, the CNN and RTF-1NN generalize quite poorly.
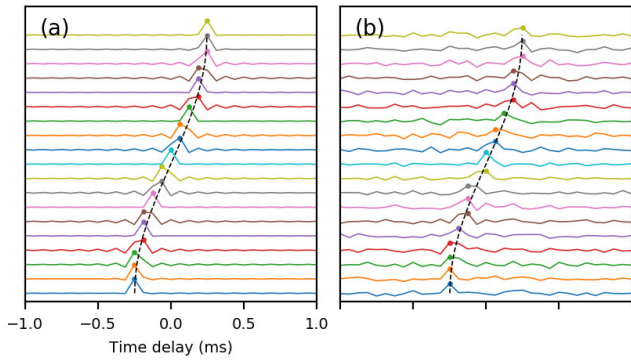
FIGURE 7. Inverse fast Fourier transforms (IFFTs) of RTF (from RTF-phase), see Sec. III-H1. (a) Simulated anechoic time delay based on DTU dataset source-receiver parameters. (b) Time delays from reconstructed RTF-phase mean from reverberant DTU dataset (see Fig. 4(b) for RTF-phase), calculated from one RTF realization per DOA. Hypothesized time delays, corresponding to the 19 DOAs in the dataset, are shown as black dashed line.
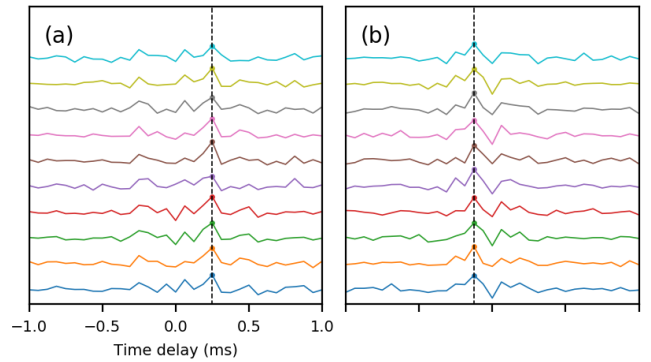

FIGURE 8. IFFTs of VAE-SSL conditionally generated RTFs (time delay), see Sec. III-H1. Time delays from conditionally generated RTF-phase means from reverberant DTU dataset for (a) DOA = 90° (see Fig. 5(b) for RTF-phase) and (b) DOA = −30° (see Fig. 6(b) for RTF-phase). Calculated for 10 RTF realizations. Hypothesized time delay shown as black dashed line.

H. RTF GENERATION WITH VAE-SSL

The trained generative model from VAE-SSL can be used to generate new RTF sequences (see Sec. II-E). RTF-phase generation is demonstrated using the VAE-SSL model trained on the DTU IRs for $J = 988$ labeled sequences. The physical interpretation of the generated phase is discussed. Generated RTF-phase sequences are shown in Fig.4–6. For display, the generated RTF-phase sequence vectors $\widehat{\mathbf{x}}_n \in \mathbb{R}^{KP}$ are reshaped to $P \times K$. We use the phase-wrap of the RTF ($\pm\pi$ rad.) (as a function of sensor separation and DOA θ_n, $f = c/(2r|\sin\theta_n|)$) to help qualify the physics learned by VAE-SSL, with r the microphone spacing. This is plotted along with the RTF-phase frames from the design case room configuration. Further, the generated RTF-phases are compared with the corresponding free-space RTF, which is calculated by

$$\text{phase}(H(f_k, \theta)) = \frac{2\pi f_k r}{c} \sin\theta. \tag{22}$$

In Fig. 4, 10 RTF sequences for each DOA ($T = 19$ DOA, giving 190 sequences) are reconstructed per (16). It is observed that the RTF-phase is well-reconstructed, and that the reconstruction (Fig. 4(b)) and phase mean (Fig. 4(c)) conform to the phase-wrap function. Further, the free-space RTF-phase (22) is shown in Fig. 4(d). The generated mean RTF-phase correlates well with the free-space phase. It is also observed that the standard deviation of the reconstructed RTF-phase is, in general, peaked in the region of the expected phase-wrap. This is expected since the phase in these regions is ambiguous.

Using the trained generative model from VAE-SSL with DTU IRs and speech, we conditionally sample $\mathbf{x}_s$ for fixed label $\mathbf{y}_n$ per (14). In Fig. 5, $\mathbf{y}$ corresponding to DOA 90° is used, and 100 RTF-phase sequences are sampled. Similarly, in Fig. 6, $\mathbf{y}$ corresponding to DOA −30° is used. It is observed that the generated phase and its corresponding mean in each case is well-correlated with the predicted RTF-phase wrap

function. In the case of the conditionally generated RTFs, they also correlate well with their corresponding free-space RTFs. The output sampled from the generative model (see (14)) is multiplied by π, since normalization was applied during training.

1) RTF TIME DELAY

The generated RTF-phases obtained by VAE-SSL are further evaluated by considering their corresponding time delays.

The free-space phase delay for a given sensor is

$$a_i(f_k, \theta) = e^{j\frac{2\pi f_k r_i}{c}\sin\theta} \tag{23}$$

with r_i the sensor spacing. With microphone 1 is used as reference ($r_1 = 0$) and microphone 2 located at r_2, the free-space RTF is

$$\begin{aligned} h(f_k, \theta) &= \frac{a_2(f_k, \theta)}{a_1(f_k, \theta)} \\ &= e^{j\frac{2\pi f_k r_2}{c}\sin\theta}, \end{aligned} \tag{24}$$

and $\tau = (r_2/c)\sin\theta$ is the time delay.

Given the derivation for RTF time delay in free-space, synthetic free-space RTFs are generated per (24) for the same source-microphone geometry as the DTU dataset (Sec. III-A). The inverse fast Fourier transform (IFFT) of the RTFs is obtained, Fig. 7(a). These time delays are compared to those from the IFFT of the VAE-generated RTF mean values from the reverberant DTU dataset (Sec. III-A). In the previous section, the DTU RTF-phase was reconstructed based on input the VAE. The results are shown in Fig. 4. One generated RTF from each of the 19 DOAs from the dataset (corresponding to phases in Fig. 4) were used to obtain time delays, and these delays are plotted in Fig. 7(b). The maximum value of the IFFT is indicated in each plot.

Overall, the peak of the IFFT of the generated reverberant RTF means correspond to the free-space RTF delay. In Fig. 7(b) the peak correlates well with the hypothesized
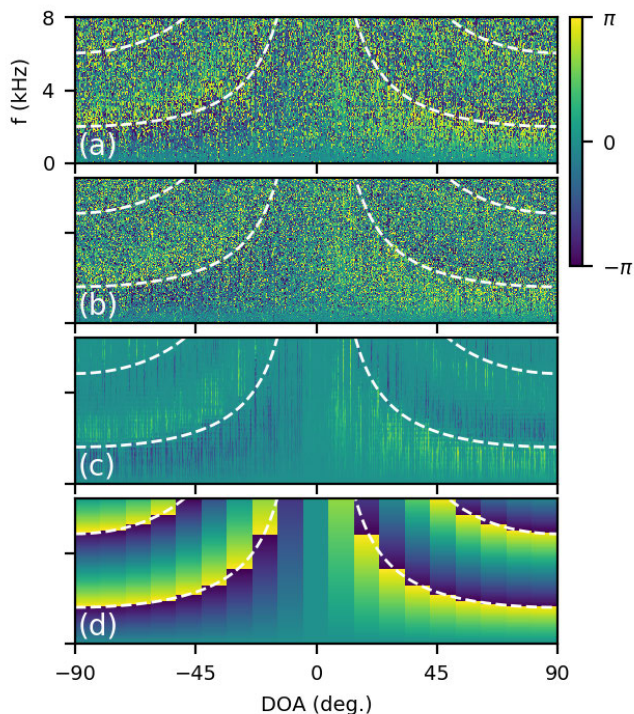
FIGURE 9. Reconstruction of RTF-phase sequences using VAE-SSL (trained using *J* = 988 labels) with normal conditional distribution. Same arrangement as Fig. 4.
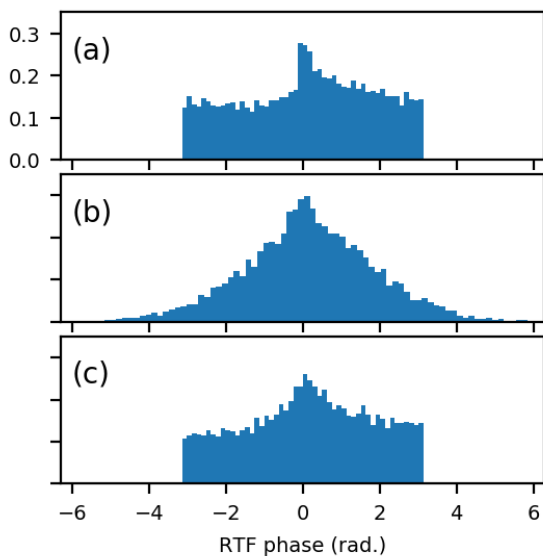
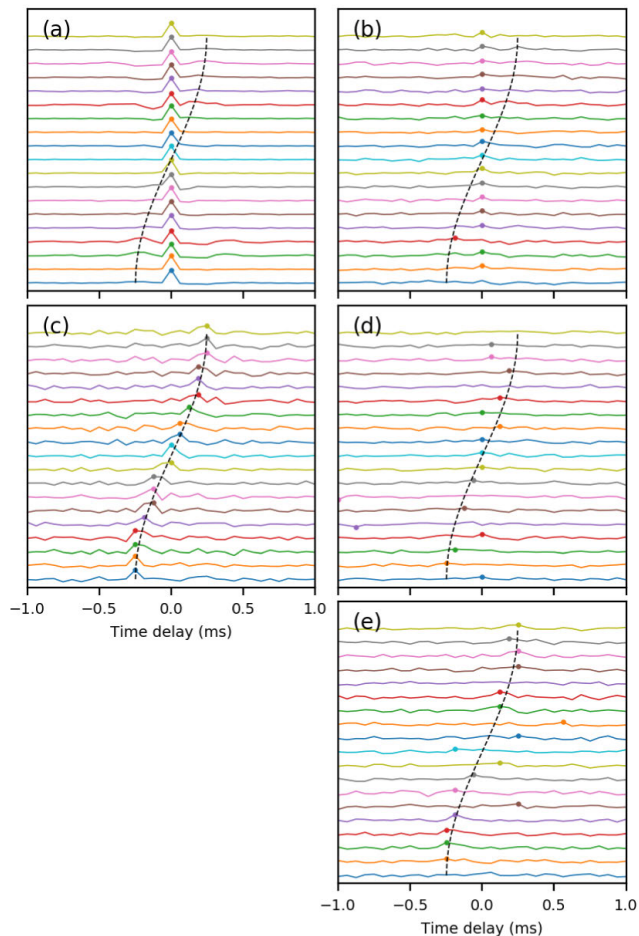**FIGURE 10.** Histograms of RTF-phase at DOA = 20° for 100 RTF frames from (a) input, (b) generated with normal conditional distribution $p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z})$, and (c) generated with truncated normal conditional distribution (Sec. II D–E, Sec. III D). The truncated normal distribution (c) better approximates the input RTF-phase statistics (a). Histograms are normalized.

FIGURE 11. Comparison of IFFTs (time delays) of the generated RTFs using (a,b) the normal and (c,d) truncated normal conditional distribution $p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z})$, from the VAE-SSL generative model trained on the DTU dataset. In (e) the time delays of RTFs corresponding to the input are shown. See Fig. 4 for input and generated RTF-phase using the truncated normal conditional distribution, and Fig. 9 for those generated with the normal distribution. The time delays obtained using the normal distribution are biased, but with the truncated normal they are not. These results were calculated for one RTF realization per DOA (*T* = 19 DOAs). The hypothesized time delay shown as black dashed line.

DOA location based on $\tau = (r_2/c)\sin\theta$, shown as a dashed black line in both subfigures.

We also consider the time delays corresponding to the conditionally generated RTF-phase (see Fig. 5 and 6) from VAE-SSL trained on DTU dataset. The IFFT of the RTFs corresponding to the conditionally generated phases are shown

in Fig. 8. Again, the peak correlates well with the hypothesized time delay.

As a final comparison, similar analyses are performed using VAE-SSL trained using the conventional normal distribution for the generative model conditional distribution $p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z})$. Results are given for $J = 988$ labels in Fig. 9–11. In Fig. 9, we reconstruct the input RTF-phase sequences, similar to Fig. 4. In Fig. 10 the RTF-phase distributions generated using the normal and truncated normal distributions are shown. Fig. 11 depicts the RTF-phase and mean RTF-phase generated by VAE-SSL using the normal and truncated normal distributions, similar to Fig. 7. Since the wrapped RTF-phase is on the interval $[-\pi, \pi]$, the truncated distribution to the same range is more physically meaningful. Further, the large peak of the full normal distribution-generated RTF-phase (Fig. 10(b)) is related to the

bias in the IFFT of the RTF corresponding to the generated phase (see Fig. 11) since zero RTF-phase implies broadside source $D_t = 0°$. In all cases, the physical characteristics of the wrapped RTF-phase are better modeled using the truncated normal conditional distribution.

IV. ALGORITHM COMPUTATION TIME

We conclude our discussion by comparing the speed of the localization approaches. The algorithmic complexity of MUSIC is dictated primarily by the construction and eigen-decomposition of the cross-spectral matrix. SRP-PHAT and the learning-based approaches, VAE-SSL, and fully supervised CNN, consist primarily of matrix (tensor) multiplication operations. For localization with the trained VAE-SSL only the classifier CNN (Fig. 2(a)) is used. The fully supervised CNN and VAE-SSL have the same complexity for the localization task. We note that the number of multiplication operations for CNNs can be quite large, though efficient implementations in deep learning toolboxes help with wall-time.

Average CPU wall time of the approaches for one STFT or RTF-phase sequence (31 frames) were obtained using an Intel Xeon server with 16 threads. The times for VAE-SSL and fully supervised CNN were 0.6 ms, for SRP-PHAT 3.1 ms, and for MUSIC 95.1 ms. Thus, the learning-based approaches can be quite fast relative to conventional localization.

V. CONCLUSION

We have proposed a semi-supervised approach to acoustic source localization in reverberant environments based on deep generative modeling with VAEs, deemed VAE-SSL. To the best of our knowledge, this study presents the first application of deep generative modeling in source localization and acoustic propagation modeling. It is found that VAE-SSL can outperform both conventional signal processing-based approaches SRP-PHAT and MUSIC, as well as fully supervised CNNs in label-limited scenarios. Further, VAE-SSL generalizes well to source locations that were not in the training data. This performance is demonstrated using real IRs, obtained from two different environments, and speech signals from the LibriSpeech corpus.

By learning to generate RTF-phase from minimally pre-processed input data, VAE-SSL models, end-to-end, the reverberant acoustic environment. The approach exploits structure in the unlabeled data to improve localization performance relative to using only the labeled sequences. The VAE-SSL system separates the physical causes of the RTF-phase (i.e., source location) from signal characteristics such as noise and speech activity which are not salient to the task. Further, the trained VAE-SSL system can generate new RTF-phase samples. We interpreted the generated RTF-phase and verified the VAE-SSL approach well-learns the physics of the acoustic environment. The generative modeling used in VAE-SSL provides interpretable features.

It is observed that deep generative modeling can improve ML model interpretability in the context of acoustics. Such models are robust to light preprocessing of input features and can automatically obtain the appropriate task-specific representation in an end-to-end fashion. The processing of only short time windows may facilitate the localization of moving sources, namely speaker tracking problems. In our future work, we will extend this approach to multi-source and moving-source scenarios. Moreover, the effect of acoustic feature processing on the VAE-SSL generalization and sample complexity will be considered further.

APPENDIX A
DERIVATION OF UNSUPERVISED OBJECTIVE

We here give additional details of the derivation of the objective for unlabeled data (10). The steps also clarify the derivation of the supervised objective, which follows a very similar development. Starting with Bayes' rule we have for unlabeled data

$$p(\mathbf{y}, \mathbf{z}|\mathbf{x}) = \frac{p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z})p(\mathbf{y}, \mathbf{z})}{p(\mathbf{x})}. \tag{25}$$

Per VI we seek an approximate density $q_\Phi(\mathbf{y}, \mathbf{z}|\mathbf{x}) \approx p(\mathbf{y}, \mathbf{z}|\mathbf{x})$ which minimizes the KL-divergence

$$\{\Phi, \theta\} = \underset{\Phi, \theta}{\arg\min} \, \mathrm{KL}(q_\Phi(\mathbf{y}, \mathbf{z}|\mathbf{x})||p(\mathbf{y}, \mathbf{z}|\mathbf{x})). \tag{26}$$

The KL divergence is for two arbitrary distributions is (and can be factored as)

$$\begin{aligned}
&\mathrm{KL}(Q(x)||P(x)) \\
&= \int_{-\infty}^{\infty} Q(x) \log \frac{Q(x)}{P(x)} dx \\
&= \int_{-\infty}^{\infty} Q(x) \log Q(x) dx - \int_{-\infty}^{\infty} Q(x) \log P(x) dx \\
&= \mathbb{E}_{Q(x)}\big[\log Q(x)\big] - \mathbb{E}_{Q(x)}[\log P(x)],
\end{aligned} \tag{27}$$

with $\mathbb{E}_{Q(x)}$ the expectation with respect to $Q(x)$.

With the definitions in (25) and (27), the KL-divergence (26) is assessed

$$\begin{aligned}
&\mathrm{KL}(q_\Phi(\mathbf{y}, \mathbf{z}|\mathbf{x})||p(\mathbf{y}, \mathbf{z}|\mathbf{x})) \\
&= \mathbb{E}\big[\log q_\Phi(\mathbf{y}, \mathbf{z}|\mathbf{x})\big] - \mathbb{E}\big[\log \frac{p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z})p(\mathbf{y}, \mathbf{z})}{p(\mathbf{x})}\big] \\
&= \mathbb{E}\big[\log q_\Phi(\mathbf{y}, \mathbf{z}|\mathbf{x})\big] - \mathbb{E}\big[\log p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z})p(\mathbf{y}, \mathbf{z})\big] \\
&\quad + \mathbb{E}\big[\log p(\mathbf{x})\big] \\
&= -\mathrm{ELBO} + \log p(\mathbf{x}),
\end{aligned} \tag{28}$$

with the expectation $\mathbb{E}$ relative to $q_\Phi(\mathbf{y}, \mathbf{z}|\mathbf{x})$.

Considering the ELBO terms from (28), we find the objective (negative ELBO) for the unlabeled data as

$$\begin{aligned}
-D(\theta, \phi; \mathbf{x}) &= \mathbb{E}\big[\log p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z})p(\mathbf{y}, \mathbf{z}) - \log q_\Phi(\mathbf{y}, \mathbf{z}|\mathbf{x})\big] \\
&= \mathbb{E}\big[\log p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z}) + \log p(\mathbf{y}) + \log p(\mathbf{z}) \\
&\quad - \log q_{\phi_1}(\mathbf{z}|\mathbf{x}, \mathbf{y}) - \log q_{\phi_2}(\mathbf{y}|\mathbf{x})\big].
\end{aligned} \tag{29}$$

We can factorize the expectation in (29) with $\mathbf{y}$ and $\mathbf{z}$ independent. Since $\mathbf{y}$ has discrete states, we have for an arbitrary

density $P(\mathbf{y}, \mathbf{z})$

$$
\begin{aligned}
\mathbb{E}_{q_\Phi(\mathbf{y}, \mathbf{z}|\mathbf{x})}&\big[P(\mathbf{y}, \mathbf{z})\big] \\
&= \int \int q_\Phi(\mathbf{y}, \mathbf{z}|\mathbf{x})P(\mathbf{y}, \mathbf{z})d\mathbf{y}d\mathbf{z} \\
&= \int q_{\phi_2}(\mathbf{y}|\mathbf{x})\bigg[\int q_{\phi_1}(\mathbf{z}|\mathbf{x}, \mathbf{y})P(\mathbf{y}, \mathbf{z})d\mathbf{z}\bigg]d\mathbf{y} \\
&= \sum_{\mathbf{y}} q_{\phi_2}(\mathbf{y}|\mathbf{x})\bigg[\int q_{\phi_1}(\mathbf{z}|\mathbf{x}, \mathbf{y})P(\mathbf{y}, \mathbf{z})d\mathbf{z}\bigg]. \qquad (30)
\end{aligned}
$$

Using (30), we expand (29)

$$
\begin{aligned}
-D(&\theta, \Phi; \mathbf{x}) \\
&= \mathbb{E}_{q_{\phi_2}(\mathbf{y}|\mathbf{x})}\big[\mathbb{E}_{q_{\phi_1}(\mathbf{z}|\mathbf{x},\mathbf{y})}\big[\log p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z}) \\
&\quad + \log p(\mathbf{y}) + \log p(\mathbf{z}) - \log q_{\phi_1}(\mathbf{z}|\mathbf{x}, \mathbf{y}) - \log q_{\phi_2}(\mathbf{y}|\mathbf{x})\big]\big] \\
&= \sum_{\mathbf{y}} q_{\phi_2}(\mathbf{y}|\mathbf{x})\big[-C(\theta, \phi; \mathbf{x}, \mathbf{y}) - \log q_{\phi_2}(\mathbf{y}|\mathbf{x})\big]. \qquad (31)
\end{aligned}
$$

Thus, it follows from the derivation with Bayes' rule that in the unlabeled case, we marginalize over the states of $\mathbf{y}$. This requires sampling the terms in (31) over all states of $\mathbf{y}$ for unlabeled data $\{\mathbf{x}_u\} \sim \mathcal{D}_u$.

REFERENCES

[1] H. Purwins, B. Li, T. Virtanen, J. Schlüter, 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. Amer.*, 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, Mar. 2019.

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

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

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

[7] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "Dcase 2017 challenge setup: Tasks, datasets and baseline system," in *Proc. DCASE*, Nov. 2017, pp. 1–8.

[8] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*. Hoboken, NJ, USA: Wiley, 2018.

[9] 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.

[10] 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.*, vol. 13, no. 1, pp. 34–48, Mar. 2019.

[11] 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. Amer.*, vol. 147, no. 6, pp. 3895–3904, Jun. 2020.

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

[13] X. Zhu, H. Dong, P. S. Rossi, and M. Landrø, "Feature selection based on principal component analysis for underwater source localization by deep learning," 2020, *arXiv:2011.12754*. [Online]. Available: http://arxiv.org/abs/2011.12754

[14] H. Hammer, S. E. Chazan, J. Goldberger, and S. Gannot, "Dynamically localizing multiple speakers based on the time-frequency domain," *EURASIP J. Audio, Speech, Music Process.*, vol. 2021, no. 1, pp. 1–10, Dec. 2021. [Online]. Available: https://rdcu.be/cilAr

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

[16] R. Opochinsky, B. Laufer-Goldshtein, S. Gannot, and G. Chechik, "Deep ranking-based sound source localization," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2019, pp. 283–287.

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

[18] 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.

[19] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[20] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–15.

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

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

[23] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*, vol. 1. Cambridge, MA, USA: MIT Press, 2016.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[25] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," 2019, *arXiv:1912.04958*. [Online]. Available: http://arxiv.org/abs/1912.04958

[26] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*. [Online]. Available: http://arxiv.org/abs/1609.03499

[27] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5024–5028.

[28] S. Leglaive, L. Girin, and R. Horaud, "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2019, pp. 101–105.

[29] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[30] E. Peterfreund, O. Lindenbaum, F. Dietrich, T. Bertalan, M. Gavish, I. G. Kevrekidis, and R. R. Coifman, "Local conformal autoencoder for standardized data coordinates," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 49, pp. 30918–30927, Dec. 2020.

[31] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Data-driven multimicrophone speaker localization on manifolds," *Found. Trends Signal Process.*, vol. 14, nos. 1–2, pp. 1–161, 2020.

[32] N. Siddharth, B. Paige, J.-W. Van de Meent, A. Desmaison, N. Goodman, P. Kohli, F. Wood, and P. Torr, "Learning disentangled representations with semi-supervised deep generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5925–5935.

[33] M. J. Bianco, S. Gannot, and P. Gerstoft, "Semi-supervised source localization with deep generative modeling," in *Proc. IEEE 30th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2020, pp. 1–6.

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

[35] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. 14th Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Sep. 2014, pp. 313–317.

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

[37] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1997, pp. 375–378.

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

[39] C. Evers, H. W. Lollmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge: Acoustic source localization and tracking," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1620–1643, Apr. 2020.

[40] S. Markovich-Golan, S. Gannot, and W. Kellermann, "Performance analysis of the covariance-whitening and the covariance-subtraction methods for estimating the relative transfer function," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 2499–2503.

[41] Z. Koldovsky, J. Malek, and S. Gannot, "Spatial source subtraction based on incomplete measurements of relative transfer function," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 8, pp. 1335–1347, Aug. 2015.

[42] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *J. Amer. Stat. Assoc.*, vol. 112, no. 518, pp. 859–877, Apr. 2017.

[43] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and A. Desmaison, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[44] 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.*, vol. 20, no. 1, pp. 973–978, 2018.

[45] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*. Berlin, Germany: Springer, 2001, pp. 157–180.

[46] Google. (2011). *WebRTC*. [Online]. Available: https://webrtc.org

[47] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2014, pp. 1–15.

[48] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2007, pp. 1–121.

[49] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A Python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 351–355.

He is also the Faculty Vice Dean and the Bar-Ilan University Data Science Institute Deputy Director. He is also a Moderator for arXiv in the field of audio and speech processing. His research interests include statistical signal processing and machine learning algorithms, including manifold learning and deep learning, with applications to single- and multi-microphone speech processing, distributed algorithms for *ad hoc* microphone arrays, speech enhancement, noise reduction and speaker separation, dereverberation, single microphone speech enhancement, speaker localization, and tracking.

Dr. Gannot is an IEEE Fellow for contributions to acoustical modeling and statistical learning in speech enhancement. He has been a member of the Technical and Steering Committee of the International Workshop on Acoustic Signal Enhancement (IWAENC), since 2005; the Audio and Acoustic Signal Processing Technical Committee of the IEEE Signal Processing Society (SPS), since January 2010; the IEEE SPS, Conferences Board, since 2019; the Executive Subcommittee of the IEEE SPS, Conferences Board, since 2020; and the IEEE SPS Education Board, since 2020. He is also a member of the Data Science Initiative of IEEE SPS. He was a recipient of the Bar-Ilan University Outstanding Lecturer Award, in 2010 and 2014, and the Rector Innovation in Research Award, in 2018. He was a co-recipient of twelve best paper awards. He was the Area Chair for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE processing, from 2013 to 2017 and since 2021. He was the General Co-Chair of IWAENC, Tel-Aviv, Israel, in August 2010, and the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, in October 2013. He was the Committee Chair, from 2017 to 2018. He was an Associate Editor of the *EURASIP Journal of Advances in Signal Processing*, from 2003 to 2012, and the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, from 2009 to 2013. He was an Editor of several special issues on multi-microphone speech processing the *EURASIP Journal of Advances in Signal Processing*. He was a Guest Editor of several special issues for *Speech Communication* (Elsevier) and *Signal Processing* (Elsevier) journal. He was a Lead Guest Editor for the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING Special Issue on speaker localization, in 2019. He is also a reviewer for many IEEE journals and conferences. He was selected (with colleagues) to present tutorial sessions with ICASSP 2012, EUSIPCO 2012, ICASSP 2013, EUSIPCO 2013, and EUSIPCO 2019, and was a Keynote Speaker for IWAENC 2012, the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA) 2017, the Informationstechnische Gesellschaft im VDE (ITG) Conference on Speech Communication 2018, and the Audio Analysis Workshop 2018.

MICHAEL J. BIANCO (Member, IEEE) received the B.Sc. degree in aeronautical and astronautical engineering from the College of Engineering, Purdue University, in 2007, and the M.Sc. and Ph.D. degrees in applied science from the University of California at San Diego (UCSD), in 2015 and 2018, respectively. From 2008 to 2012, he was an Engineer with Rocketdyne, Los Angeles, USA. He is currently an Assistant Project Scientist with the Marine Physical Laboratory, UCSD. His research interests include signal processing and machine learning, with application to acoustics and inverse problems.

SHARON GANNOT (Fellow, IEEE) received the B.Sc. degree *(summa cum laude)* from the Technion-Israel Institute of Technology, Haifa, Israel, in 1986, and the M.Sc. *(cum laude)* and Ph.D. degrees from Tel-Aviv University, Tel Aviv, Israel, in 1995 and 2000, respectively, all in electrical engineering.

In 2001, he held a postdoctoral position with the Department of Electrical Engineering, KU Leuven, Leuven, Belgium. From 2002 to 2003, he held a research and teaching position with the Faculty of Electrical Engineering, Technion-Israel Institute of Technology. From 2018 to 2019, he was a part-time Professor with the Technical Faculty of IT and Design, Aalborg University, Denmark. He is currently a Full Professor with the Faculty of Engineering, Bar-Ilan University, Ramat-Gan, Israel, where he is heading the Acoustic Signal Processing Laboratory and the Data Science Program.

EFREN FERNANDEZ-GRANDE (Member, IEEE) received the B.Sc. degree in telecommunication engineering from the Polytechnic University of Madrid (UPM), in 2006, and the M.Sc. and Ph.D. degrees in acoustics from the Technical University of Denmark (DTU), in 2008 and 2012, respectively. He completed a postdoctoral fellowship from the Danish Council for Independent Research, from 2013 to 2015. He is an Associate Professor with the Acoustic Technology Group, Department of Electrical Engineering, DTU. He teaches several courses in physical acoustics and signal processing. Since 2018, he has been the Head of the M.Sc. in engineering acoustics with DTU. His research interests include signal processing for source identification and localization, sound radiation, acoustic holography, and spatial audio. He is a fellow of the Acoustical Society of America and a VILLUM Young Investigator Fellow.

PETER GERSTOFT (Senior Member, IEEE) received the Ph.D. degree from the Technical University of Denmark, Lyngby, Denmark, in 1986. Since 1997, he has been with the University of California at San Diego. His current research interests include signal processing and machine learning applied to acoustic, seismic, and electromagnetic signals.

• • •