Ambisonics Room Impulse Response Estimation From a Single Omnidirectional Measurement Using Deep Neural Networks

Yu, Wangyang; Kleijn, W. Bastiaan

DOI
10.17743/jaes.2022.0181

Publication date
2024

Document Version
Final published version

Published in
AES: Journal of the Audio Engineering Society

Citation (APA)
Yu, W., & Kleijn, W. B. (2024). Ambisonics Room Impulse Response Estimation From a Single Omnidirectional Measurement Using Deep Neural Networks. *AES: Journal of the Audio Engineering Society*, *72*(12), 884-900. https://doi.org/10.17743/jaes.2022.0181

W. Yu and W. B. Kleijn1, "Ambisonics Room Impulse Response Estimation From a Single Omnidirectional Measurement Using Deep Neural Networks,"
J. Audio Eng. Soc., vol. 72, no. 12, pp. 884–900 (2024 Dec.).
https://doi.org/10.17743/jaes.2022.0181.

Ambisonics Room Impulse Response Estimation From a Single Omnidirectional Measurement Using Deep Neural Networks

WANGYANG YU,[1,*] AND W. BASTIAAN KLEIJN[1,2]
(estelle_ywy@outlook.com) (bastiaan.kleijn@vuw.ac.nz)

[1]*Department of Microelectronics, Signal Processing Systems, Delft University of Technology, Delft, The Netherlands*
[2]*School of Engineering and Computer Science, Victoria University of Wellington, Wellington, New Zealand*

Mapping a room impulse response (RIR) to its Ambisonics representation is not always feasible. However, by adding a weak assumption (i.e., the existence of at least two perpendicular walls in the environment), the Ambisonics representation is restricted to be one of a finite set, with known transformations between the set entries. This makes mapping the omnidirectional RIR to the Ambisonics RIR (ARIR) possible. The authors solve the mapping problem with a convolutional neural network and multi-task variational autoencoder. The room is assumed to be rectangular. The proposed method is based on the image source method with frequency-independent reflection coefficients exclusively. The authors focus on the early part of RIRs, where the directional information lies. This method requires only a single RIR. Generalizing to the real world, measurements can obviate the need for specialized hardware for Ambisonics measurement. The proposed method can achieve an SNR of 17.62 dB on estimated first-order ARIRs and 16.15 dB on estimated third-order ARIRs.

0 INTRODUCTION

Augmented reality (AR) is a specific immersive audio-visual environment that provides users with an interactive and enhanced experience in the real world with added artificial objects [1]. It can be used in various applications, such as education and entertainment. Spatial audio, aiming for a 3D audio experience, is a major attribute of a plausible AR system. Consequently, the description of the acoustic environment is of great importance. Different from the commonly used omnidirectional (pressure) room impulse responses (RIRs) that describe the room acoustical environment, Ambisonics RIRs (ARIRs) can provide spatial information based on an orthonormal decomposition of the sound field. Hence, AR commonly makes use of the ARIR. Different types of microphone arrays can measure capsule signals (A-format), which are then encoded in a subsequent step into the Ambisonics B-format. The most commonly used microphone arrays are (coincident) tetrahedral microphone arrays, which consist of four capsules. Using a simple encoding equation, first-order Ambisonics signals can be calculated. In order to measure an Nth-order ARIR, the microphone has to consist of at least $(N + 1)^2$ capsules.

An RIR signal contains information about the configuration of the room acoustical environment implicitly [2]. Estimating the Ambisonics representation from an impulse response is generally infeasible since the authors need to obtain multiple signals from just one input signal. For example, in a free space without a floor, with the ARIR coordinates centered at the receiver, the impulse response is invariant with the movement of the source on a sphere. Hence, the impulse response provides no directional information that can be used to map it to an ARIR. Perhaps surprisingly, with very weak prior information about the environment, i.e., the existence of at least two perpendicular walls, the problem becomes solvable. The authors' method requires an understanding of degeneracy. They select one particular mode to perform the estimation of the ARIR directly from the omnidirectional RIR and then transform among different modes based on the available side information, for example, an image.

The commonly used B-format microphone [3] can only capture first-order signals and the spatial resolution of first-order Ambisonics is low. There exist many works (e.g., [4–8]) that upscale the Ambisonics from the first-order for improved sound quality. RIRs can be considered to be zeroth-order ARIRs and can be measured with an omnidirectional microphone, which has a low cost compared with a microphone array. The estimation of ARIRs from RIRs

*To whom correspondence should be addressed, email: estelle_ywy@outlook.com.

can be interpreted as upscaling the Ambisonics representation from zeroth-order. In this paper, the authors show that deep learning allows them to estimate the ARIR of any order directly from the omnidirectional RIR, thus obviating the need for specialized hardware.

The main contribution of this paper is the ARIR estimation from RIRs using deep neural networks. As mentioned, generating an Ambisonics representation from an omnidirectional signal is not always feasible. This mapping is shown to be possible in some rooms. Specifically, the authors consider rectangular rooms and focus on the specular reflections in the early part of the response. The feasibility relies on the degeneracy of RIRs in a room. This novel method requires only a single RIR without additional information if all that is wanted is to estimate the ARIR and reproduce the immersive environment. If the estimated ARIR is going to be applied in an audiovisual environment, such as AR, the authors need additional information, for example, an image, to determine which mode it belongs to and the alignment between the coordinates of the image and the ARIR. This method is based on the image source method (ISM) [9], aiming at the accurate description of specular reflections of ARIRs.

The paper is organized as follows. The relevant background is reviewed in SEC. 1. In SEC. 2, the ARIR estimation problem is formulated. The ARIR estimation with convolutional neural networks (CNNs) and multi-task variational autoencoders (VAEs) are described in SEC. 3. In SEC. 4, the experimental results are discussed and analyzed in detail. Finally, we conclude the paper in SEC. 5.

1 BACKGROUND

In this section, the authors discuss the relevant background for their work. The RIRs and modeling techniques are described first. SEC. 1.2 introduces Ambisonics. Some algorithms for multi-channel RIRs generation from omnidirectional RIRs are introduced in SEC. 1.3. At the end of this section, CNNs and VAEs are discussed.

1.1 RIRs

ARIRs are estimated based on the RIR. In this subsection, RIRs and the simulation methods are introduced. In the context of this paper, only omnidirectional RIRs are considered.

An RIR is a transfer function between a sound source and receiver in a room, which describes the acoustic environment. An RIR is composed of a direct signal, early reflections and diffuse reverberation. The first high peak is identified as the direct signal. Early reflections refer to the sparsely distributed discrete reflections, which control the spatial impression. The late reverberations are densely distributed reflections and are hard to distinguish, which controls the enveloping. In late reverberation, the energy is statistically equally distributed in the space [10]. The transition from early reflections to diffuse reverberation is characterized by a mixing time [11, 12] and can be determined by statistical measurements [13]. RIRs are widely studied in a variety of work, such as speech dereverberation [14, 15] and room acoustical parameter estimation [16–18].

Many methods exist to simulate RIRs, such as the ray tracing method [19–21] and finite element method [22–24], where the authors highlight the ISM [25, 26, 9]. The ISM was first proposed by Allen and Berkley [9] in 1979. RIRs simulated by the ISM differ from real measured RIRs in several aspects. Firstly, the ISM cannot model frequency-dependent components, such as, frequency-dependent reflection coefficients. Secondly, the ISM cannot be used for curved and nonsmooth reflective surfaces and cannot model diffraction or scattering. Lastly, empty rectangular rooms are always assumed. These assumptions make the simulated RIRs different from real-world RIRs. However, ISM works well for specular reflections.

In this paper, the ISM is used for the shoebox shaped rooms with uniform material properties along each surface because of its computational efficiency, making it suitable for generating a large scale database. An empty rectangular room is assumed and nonspecular reflections are not considered. The method assumes that sound propagates along straight lines. Each reflection is modeled as a pressure wave emitted from an image source in free space. The authors use $\mathbf{p}, \mathbf{m}$ to label each reflection where each element of $\mathbf{p} = (q, j, l)$ can take a value of 0 or 1, indicating the direction of the reflection, and each element of $\mathbf{m} = (m_x, m_y, m_z)$ can take an integer value, indicating the position of the virtual room where image sources locate. The reflection order $O_{\mathbf{p},\,\mathbf{m}}$ can be computed as

$$O_{\mathbf{p},\mathbf{m}} = |2m_x - q| + |2m_y - j| + |2m_z - l|. \tag{1}$$

Let $d_{\mathbf{p},\,\mathbf{m}}$ denote the path length, then one has the time delay $\tau_{\mathbf{p},\,\mathbf{m}} = d_{\mathbf{p},\,\mathbf{m}}/c$. The amplitude of each reflection is determined by the reflection coefficients $\beta_{x_1}, \beta_{x_2}, \beta_{y_1}, \beta_{y_2}, \beta_{z_1}, \beta_{z_2}$, reflection order $O_{\mathbf{p},\,\mathbf{m}}$, and image source position. Assuming the finite and constant reflection coefficients over each wall, then the RIR can be written as [9]

$$h(t) = \sum_{\mathbf{p},\mathbf{m}} \beta_{x_1}^{|m_x - q|} \beta_{x_2}^{|m_x|} \beta_{y_1}^{|m_y - j|} \beta_{y_2}^{|m_y|} \beta_{z_1}^{|m_z - l|} \beta_{z_2}^{|m_z|} \frac{\delta(t - \tau_{\mathbf{p},\mathbf{m}})}{4\pi d_{\mathbf{p},\mathbf{m}}}, \tag{2}$$

which will be used for ARIR computation in SEC. 2.

1.2 Ambisonics and ARIR

Ambisonics [27–29] describes the 3D sound field at a receiver's position instead of depending on the description of specific sound sources. It is suitable for AR systems because head rotations are easily modeled as the rotation of sound fields in the spherical harmonics domain. It describes the sound field by means of a small set of temporal signals. Recent work on Ambisonics often uses higher-order Ambisonics, an extension of the original first-order Ambisonics system developed by Gerzon [28]. Ambisonics is used for spatial audio encoding and transmission and as a basis for rendering. With an Ambisonics representation of sufficient order, a high-quality audio rendering system can give listeners a plausible spatial audio experience. Ambisonics repre-

sents the sound field for the so-called interior case, where all sources lie outside the region of interest. Thus, Ambisonics is a particular representation of the interior-case solution to the acoustic wave equation or, equivalently, the Helmholtz equation. In this subsection, the authors first show how the Ambisonics coefficients are derived. They then define the Ambisonics room response.

Spherical harmonics [30] are a complete set of orthogonal basis functions defined on the surface of a sphere. 3D full normalization is adopted since it is widely used in Ambisonics software packages and is characterized in [31] as the most natural normalization scheme for a physically plausible sound field. The spherical harmonics are

$$Y_n^m(\theta, \phi) = N_n^{|m|} P_n^{|m|}(\sin(\theta)) \begin{cases} \sin(|m|\phi), & \text{for } m < 0 \\ \cos(|m|\phi), & \text{for } m \geq 0 \end{cases}, \quad (3)$$

where $Y_n^m(\theta, \phi)$ is the spherical harmonic of order n and degree m with $-n \leq m \leq n$, ϕ is the azimuth, θ is the elevation, $P_n^{|m|}$ is the associated Legendre function, and $N_n^{|m|}$ is the normalization term. With 3D full normalization, there is

$$N_n^m = \sqrt{2n+1}\sqrt{\frac{2-\delta_m}{4\pi}\frac{(n-|m|)!}{(n+|m|)!}},$$
$$\delta_m = \begin{cases} 1, & \text{if } m = 0 \\ 0, & \text{if } m \neq 0. \end{cases} \quad (4)$$

A complete solution to the Helmholtz equation is now discussed. Let the temporal frequency be denoted by $k = \frac{\omega}{c}$, where ω is the frequency in rad/s and c is the speed of the sound. Then the sound signal p measured at the spherical coordinates $\mathbf{r} = (r, \theta, \phi)$ can be represented as [32]

$$p(\mathbf{r}, k) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} i^n j_n(kr) Y_n^m(\theta, \phi) B_n^m(k), \quad (5)$$

where $j_n(kr)$ is the spherical Bessel function of the first kind and the $B_n^m(k)$ are the *Ambisonics coefficients*.

When Eq. (5) is truncated to a particular N, one can represent the sound field with $(N+1)^2$ temporal signals. Then, the sound field will be accurate within a spherical region near the origin, which is commonly called *sweet area*. The sweet zone increases in size with N. Its size is inversely proportional to frequency. The authors denote by $\mathcal{D}_R^{3D}$ the dimensionality of 3D Ambisonics signals after truncation. The dimensionality is related to N as $\mathcal{D}_R^{3D} = (N+1)^2$.

The authors consider far-field sound sources and model the sound pressure $S_q(k)$ from sound source q as a plane wave arriving from an incidence angle (θ_q, ϕ_q). Let $\mathbf{k}_q$ denote the wavenumber vector, which points in the propagation direction of the plane wave. The spherical harmonic expansion of the plane wave transfer function is described as [29]

$$e^{i<\mathbf{k}_q, \mathbf{r}>} = 4\pi \sum_{n=0}^{\infty} \sum_{m=-n}^{n} i^n j_n(kr) Y_n^m(\theta_q, \phi_q) Y_n^m(\theta, \phi). \quad (6)$$

Using Eqs. (5) and (6), the *Ambisonics coefficients* are found as

$$B_n^m(k) = \sum_q 4\pi Y_n^m(\theta_q, \phi_q) S_q(k). \quad (7)$$

Temporal Ambisonics signals can be obtained by taking the inverse temporal Fourier transform of each $B_n^m(k)$ signal. The first-order Ambisonics signals are the B-format signals [28].

The ARIR (or spatial RIR) refers to the transfer function between a source and receiver in a room that is measured by spherical microphone arrays. ARIRs can be convolved with signals to generate Ambisonics signals and rendered with various approaches [33–41].

Definition 1: When the source signal is an excitation signal, i.e., delta function, the set of $B_n^m(k)$ becomes the Ambisonics room response in the frequency domain. Multiplying a frequency-domain source signal with the $B_n^m(k)$ results in the Ambisonics representation of the sound field around the receiver.

1.3 Multichannel RIR Estimation

Because the authors are not aware of the existing work on ARIR estimation, they reviewed the algorithms that estimate multichannel RIRs from an omnidirectional RIR. These algorithms are similar to the ARIR estimation since the underlying spatial information of the input RIR is used.

An algorithm to estimate an arbitrary number of RIRs from one RIR is proposed in [42]. Similar to the authors' method, it is based on the ISM, the first peak is identified as the direct path, and the direct sound always comes from a positive x direction. The first step is the estimation of the source-receiver distance and room volume as in [43]. Then, the room geometry is determined using a predefined fixed ratio. Secondly, up to four strong peaks are identified as first-order reflections, which are used to determine the source and receiver positions using predetermined rules [42] for the correct direct path distance. The ISM is then applied to calculate the image source positions and reflection coefficients. The diffuse reflections are divided into time sections to model RIR as scattered point sources. RIRs can then be calculated using the ISM and the scattered point sources [42]. The generated RIRs resemble the desired ones experimentally. It performs well in the early reflections but is less efficient or accurate for the complete RIRs. Several approximations exist in the proposed method, for example, the ratio of the room edges.

Binaural RIRs (BRIRs) estimation from an omnidirectional RIR is a problem similar to estimating ARIRs from RIRs since both problems need knowledge of reflections. The difference between these two problems is that BRIR estimation also requires head-related transfer functions (HRTFs) of each reflection direction. The algorithm in [44] assumes knowledge of geometric information of the room volume, the direction of the direct path, and a preprocessed binaural noise. The RIR is divided into three segments by preassigned time slot (i.e., direct sound, early reflections, and diffuse reverberation). The direct sound is filtered by

the HRTF in that direction. The early reflections are filtered with HRTFs of the predefined reflection pattern. Binaural diffuse reverberation is estimated in each frequency bin by shaping the envelope of binaural noise. This method is mathematically and conceptually simplified. It contains several approximations, such as the predefined early reflection pattern. [45–47] improve the method in [44] to allow changes in different aspects, for example, the distance between the source and the receiver, but approximations still exist.

Spatial RIRs can be estimated from one monaural RIR using the parametric method in [48], which extends and improves the method in [44]. Firstly, the proposed method detects the amplitude and the time of arrival of the direct path and early reflections. The direction of arrivals (DOAs) of six to ten selected early reflections can be determined by a pseudorandomized directional distribution or a predetermined DOA pattern or by using the ISM with approximated room geometry. The standard room acoustic parameters are calculated for parametric rendering. The reflection filters are derived to adjust the magnitude spectra of early reflections. Next, the reverberation level, describing the level of the diffuse field in the early reflection part of RIR, is estimated to ensure the preservation of RIR energy when synthesizing BRIRs. Finally, combining the detected early reflections and the ISM, the parameters of an arbitrary position in the room can be calculated to calculate the spatial RIRs or BRIRs. Similar to the previously mentioned methods, the proposed methods consist of several approximations, such as the DOAs of early reflections.

1.4 Deep Neural Networks

Deep learning shows good modeling properties for many applications. In general, it requires high computational capacity and the availability of large databases. Different from conventional modeling methods, deep learning uses neural networks to learn from a large amount of data. Each layer of a neural network can be viewed as a simple function with unknown parameters. Combining multiple layers forms a nonlinear modeling function whose parameters are learned by training with the available dataset. The authors aim to use CNNs and VAEs to estimate ARIRs.

CNNs are commonly used for deterministic maps. [49] first proposed CNNs in the context of visual pattern recognition. CNNs are widely used in many areas, such as image recognition [50–52] and audio classification [53–55]. By parameter sharing and sparse connection, CNNs capture the spatial relationships within the input using kernels. Each convolution layer has a set of feature maps (or the actual input signal) as input and produces as output another set of feature maps. Each feature map corresponds to a channel with the number of output channels set by the network designer. Upsampling is achieved with so-called transposed convolutions [56], which effectively insert zeros in the feature maps prior to a convolution operation. In the context of the current problem, CNNs are used for deterministic mapping from RIRs to ARIRs. The usage of CNNs is consistent with the assumption that RIRs and ARIRs have time-invariant relations between their samples. The accuracy of this assumption is reflected in the experimental outcomes of [18].

VAEs [57–60] can be used as generative models or as methods to remove redundancy from an input representation. VAEs can be used for speech enhancement [61, 62], image classification [63, 64], and so on. Generative models aim to create new data with a probability distribution similar to example data. An autoencoder is a neural network that consists of an encoder that maps the input to a latent representation and a decoder that maps the latent information to an approximation of the input data. It is assumed that the high-dimensional data actually lie in a low-dimensional manifold. Ideally, the bottleneck layer (latent space) of an autoencoder describes the data within the manifold and corresponds to an abstract description of the input data without redundancy. Using the decoder only, a VAE can be used to generate new data by sampling from the latent distribution. Thus, VAEs can be used to remove redundancy or generate new data.

There exist several varieties of VAEs, such as β-VAEs [65] and Bounded Information Rate Variational Autoencoders [66]. Although VAEs are commonly used in a generative setting, the authors use VAEs because of the availability of a latent representation. Variance-constrained autoencoders (VCAEs) [67] are used because they only constrain the variance of the latent layer, which allows a more natural representation of the data. Although sampling from the latent layer is difficult with VCAEs, the authors aim to analyze RIRs rather than generate new data from the latent space.

X, an $\mathbb{R}^d$-valued random variable, is used to represent the signal where d denotes the length of each signal and $X \sim P_D(x)$, whose distribution is determined by the data. A VCAE [67] is composed of an encoder $Q_{Z|X;\psi}$ and a decoder $P_{X|Z;\eta}$ that are implemented by neural networks with parameters ψ and η, respectively. Let Z, an $\mathbb{R}^{d_z}$-valued random variable, represent latent space of dimensionality d_z. The distribution of Z is unknown. VCAEs do not constrain the distribution of z but instead constrain the variance of z. The latent space z follows that $z = \mu_\psi(x) + \epsilon$, where $\epsilon \sim P_\epsilon$ is defined by the system designers. The loss function can be written as [67]

$$\max_{\eta,\psi} E_{X\sim P_D} E_{Z\sim Q_{Z|X;\psi}}[\log p_\eta(X|Z)]$$

$$-\lambda|E_{Z\sim Q_{Z;\psi}}[\|Z - E_{Z\sim Q_{Z;\psi}}[Z]\|_2^2] - v|, \tag{8}$$

where v denotes the target total variance and λ controls the trade-off between the reconstruction performance and the variance of the latent space.

A VCAE is similar to a regular autoencoder but has the ability to control the information rate traveling through each neuron in the latent layer. A VCAE can be viewed as a communication channel [68] where the code is given by $\mu_\psi(x)$, the channel is defined by $p_\psi(\epsilon)$, and the output is given by $z = \mu_\psi(x) + \epsilon$. Choosing $p_\psi(\epsilon) = \mathcal{N}(0, \sigma_\epsilon^2 \cdot I_{d_z})$, the upper bound of the information rate can be computed as

$I_{\text{bits}} = \frac{d_z}{2}\log_2(\frac{1}{\sigma_\epsilon^2})$ [68]. The authors are interested in these information rates in SEC. 3.2.

2 PROBLEM DEFINITION

In this section, the authors formulate the problem they aim to solve, i.e., ARIR estimation from an omnidirectional RIR. As noted in SEC. 1, an ARIR is defined as an Ambisonics representation of the corresponding RIR. The outcome of this work is an estimated ARIR with a simple measurement. The degeneracy of an RIR is analyzed in the first subsection. In the second subsection, the authors discuss their motivation for using deep learning to solve the problem, describe how they compute the ARIRs, and discuss how to estimate the signals.

2.1 Degeneracy

Computing an Ambisonics representation of an omnidirectional signal only is not always feasible. Constraints have to be added to make the computation possible. Degeneracy is defined first. Definition 2: An RIR is M-fold degenerate if, given a set of coordinates, there exist M distinct ARIRs that correspond to the RIR.

The degeneracy often is finite and can be removed by information from other modalities such as cameras or radar. It is assumed that the walls are either parallel or perpendicular; one side of a single wall defines the considered space, and parallel walls enclose the considered space. Without loss of generality, the axes are assumed to be parallel to existing walls, and the receiver is assumed to be located at the point of origin.

The authors start with discussing the degeneracy of impulse responses under different acoustic scenarios:

1. Free space without walls, ceilings, or floors: The impulse response is composed of a single delta pulse of the direct path. As long as the distance between the source and receiver is the same, the RIR is the same. So, there exists an uncountably infinite degeneracy for ARIRs in this case.
2. One wall (i.e., free space with a floor) or a pair of parallel walls: A rotation of the source with respect to the receiver along the axis orthogonal to the wall or walls does not affect the RIR and hence corresponds to an infinite-fold degeneracy. Mirroring of the room with respect to the floor or parallel walls introduces another two-fold degeneracy. In addition, if source and receiver are exchanged, the RIR does not change, which introduces another two-fold degeneracy. For clarity, there exists infinite $\times 4$-fold degeneracy for ARIRs in this case. If the direct path is parallel to a wall, the corresponding two-fold degeneracy collapses.
3. Two perpendicular walls or two pairs of parallel walls (i.e., 2D room case): Mirroring of the room gives a four-fold degeneracy. The dimensions of the room can additionally be exchanged, which introduces another two-fold degeneracy. The exchange of source and receiver gives another two-fold degeneracy. There is a 16-fold degeneracy for RIRs in total in this case. A figure illustrating the degeneracy is shown as Fig. 1, where the exchange of source and receiver is not included. If the direct path is parallel to a wall, the corresponding degeneracy collapses. Similarly, if the room is square, the degeneracy of the *x*-*y* axis choice collapses. If the room is symmetric around the source, the degeneracy is identical to the one wall case.
4. Three perpendicular walls or three pairs of parallel walls (i.e., 3D room case): Mirroring of the room gives an eight-fold degeneracy. The permutation of room dimensions introduces another six-fold degeneracy. Considering the exchange of source and receiver gives another two-fold degeneracy. There is 96-fold degeneracy for RIRs in total in this case. A collapse of degeneracy happens when the direct path is parallel to a wall or the length of two perpendicular walls is identical.

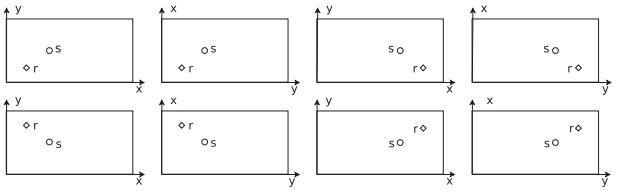

Fig. 1. Degeneracy of 2D room case where *s* and *r* denote source and receiver, respectively.

From the analysis of the degeneracy of RIRs, it can be concluded that by adding at least two perpendicular walls in the acoustic space, the problem is suddenly solvable at the cost of degeneracy. In this paper, the authors assume a rectangular empty room, that three edges of the room are of different lengths, and that the direct path is not parallel to any wall. One mode is defined as one ARIR out of multiple distinct ARIRs corresponding to one RIR. Although there exists a 96-fold degeneracy for an empty rectangular 3D room, the Ambisonics representation can still be made feasible by choosing one default mode out of a 96-fold degeneracy to solve the problem and subsequently mapping from that mode to another based on the requirement or additional information. It is assumed the direct path is always from straight ahead since head rotations are easily modeled as the rotation of sound fields in the spherical harmonics domain. The authors assume they have no knowledge about the DOA of reflections or the environment information, such as room geometry and reflection coefficients.

The degeneracy can also be determined by the first-order reflections as described in [2] with consistent results. The authors derive the condition of one mode as an example and refer for further details to [2]. Three plane coordinates are chosen first. Let (L, W, H) denote the room geometry, (x_s, y_s, z_s) and (x_r, y_r, z_r) denote the coordinates of source and receiver. Assume the first reflection is the pulse that

reflects on $x = 0$, which gives a six-fold degeneracy. This is equivalent to

$$\begin{cases} \|-x_s - x_r\| < \|2L - x_s - x_r\| \\ (-x_s - x_r)^2 + (y_s - y_r)^2 < (x_s - x_r)^2 + (-y_s - y_r)^2 \\ (-x_s - x_r)^2 + (y_s - y_r)^2 < (x_s - x_r)^2 + (2W - y_s - y_r)^2 \\ (-x_s - x_r)^2 + (z_s - z_r)^2 < (x_s - x_r)^2 + (-z_s - z_r)^2 \\ (-x_s - x_r)^2 + (z_s - z_r)^2 < (x_s - x_r)^2 + (2H - z_s - z_r)^2 \end{cases}. \quad (9)$$

Similarly, assuming the next non-x direction first-order reflection reflects on $y = 0$ gives a four-fold degeneracy, which is

$$\begin{cases} \|-y_s - y_r\| < \|2W - y_s - y_r\| \\ (-y_s - y_r)^2 + (z_s - z_r)^2 < (y_s - y_r)^2 + (-z_s - z_r)^2 \\ (-y_s - y_r)^2 + (z_s - z_r)^2 < (y_s - y_r)^2 + (2H - z_s - z_r)^2 \end{cases}. \quad (10)$$

Then, assuming the next non-x and non-y direction first-order reflection reflects on $z = 0$ gives a two-fold degeneracy, which is $\|-z_s - z_r\| < \|2H - z_s - z_r\|$. The exchange of the source and receiver gives another two-fold degeneracy, where one can assume $0 < x_r < x_s$, $0 < y_r < y_s$, and $0 < z_r < z_s$. The conditions for this mode can be summarized as

$$\begin{cases} 0 < x_r < x_s \\ 0 < y_r < y_s \\ 0 < z_r < z_s \\ x_r + x_s < L \\ y_r + y_s < W \\ z_r + z_s < H \\ x_s x_r < y_s y_r < z_s z_r \\ x_s x_r < (W - y_s)(W - y_r) \\ y_s y_r < (H - z_s)(H - z_r) \end{cases}. \quad (11)$$

2.2 ARIR Estimation With Deep Learning

The special and expensive equipment to acquire Ambisonics signals makes the measurement difficult. Limitations on the equipment result in only relatively low-order Ambisonics with a low spatial resolution. Another method to compute Ambisonics signals is based on the estimated room acoustical parameters. However, estimating room acoustical parameters from a single RIR is difficult, especially for real-world measurements, since correct reflections are hard to detect. The existing methods [42, 44–47] either make assumptions on prior knowledge or make approximations on the calculation of room acoustic parameters that introduce errors. These motivate the authors to design a method to compute an Ambisonics representation of the RIR from only an omnidirectional RIR using deep learning. The proposed method does not require special equipment or the approximation or estimation of room acoustical parameters. The first channel of the ARIR signal corresponds to zeroth-order Ambisonics, a scaled version of the omnidirectional RIR that contains no directional information explicitly. Hence, the current problem can also be viewed as an Ambisonics upscaling problem that upscales ARIRs from a zeroth-order to an arbitrary order.

For this ARIR estimation problem, as discussed in SEC. 2.1, the degeneracy of RIR makes it hard to learn with a deep neural network. As a result, the authors first choose one default mode (i.e., define a one-to-one relationship between ARIRs and RIRs). The coordinate system of this default mode is determined based on the first-order reflections as in the authors' previous paper [2]. How to map from one mode to another is discussed in SEC. 3.3.

This computation of ARIRs is based on the ISM [25, 26, 9] since the directions of reflections can be computed with the ISM. Following the ISM, the authors assume rectangular rooms and that each wall has only one reflection coefficient. In addition, since the ISM can only model specular reflections accurately, the focus is on the specular reflections of RIRs that contain the directional information. Using the ISM, an ARIR can be viewed as a composition of multiple Dirac impulses of different amplitudes arriving from the real source and image sources where the amplitudes in channels provide directionality information. Each image source can be viewed as a separate source. An ARIR signal $B_n^m(t)$ can be computed with Eqs. (2) and (7):

$$B_n^m(t) = \sum_q Y_n^m(\theta_q, \phi_q) \times \sum_{\mathbf{p},\mathbf{m}} \beta_{x_1}^{|m_x - q|} \beta_{x_2}^{|m_x|} \beta_{y_1}^{|m_y - j|} \beta_{y_2}^{|m_y|} \beta_{z_1}^{|m_z - l|} \beta_{z_2}^{|m_z|} \frac{\delta(t - \tau_{\mathbf{p},\mathbf{m}})}{d_{\mathbf{p},\mathbf{m}}}. \quad (12)$$

This allows for generating a large-scale database of arbitrary order with RIR-ARIR pairs for deep learning.

3 ARIR ESTIMATION USING DEEP LEARNING

In this section, deep learning–based ARIR estimation is described. First, an ARIR is computed from an omnidirectional RIR under the default mode out of 96-fold degeneracy with CNN and VAE, respectively, in the first two subsections. After that, the transformation matrix of ARIRs among different modes of RIRs is discussed. At the end of this subsection, the authors describe how they can apply their ARIR estimation methods for real-world applications.

3.1 ARIR Estimation With CNN

The ARIR estimation problem can be viewed as a regression problem. Let the pair of random vectors (X, Y) denote the input and output signals of a neural network. These two signals are of the same length. Specifically, in this paper, X is an $\mathbb{R}^d$-valued random variable that represents an RIR where d denotes the length of each RIR signal vector, and Y is an $\mathbb{R}^d$-valued random variable that represents the corresponding ARIR of one channel under default mode. The learned continuous deterministic function h is defined as $\hat{y} = h(x)$ where $\hat{}$ labels an estimate and $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$ is a realization of the random variable pair (X, Y). The loss function $l : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ measures the mapping error of h. The risk R of the model can then be defined as

$$R = \mathbb{E}[l(h(X), Y)], \quad (13)$$

where the expectation $\mathbb{E}$ is calculated with respect to the joint distribution $f_{XY}(X, Y)$. Since the joint distribution

$f_{XY}(X, Y)$ is unknown, the risk R of the model is approximated with the empirical risk R_{emp} on the training set:

$$R_{\text{emp}} = \frac{1}{m}\sum_{i=1}^{m} l(h(x_i), y_i), \tag{14}$$

where m denotes the size of training dataset and each (x_i, y_i) pair is one realization of $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$ in the training dataset. In the context of this problem, the mean square error (MSE) is used as the empirical risk, which measures the squared Euclidean distance between the estimated outputs and corresponding ground truth. The objective function to train the neural network is then defined as

$$l(y, \hat{y}) = \frac{1}{m}\sum_{i=1}^{m} \| y_i - \hat{y}_i \|_2^2, \tag{15}$$

where $\| \cdot \|_2^2$ is the squared l^2-norm, m denotes the size of training dataset, $y \in \mathbb{R}^{m\times d}$ denotes the true ARIR of one channel, and $\hat{y} \in \mathbb{R}^{m\times d}$ denotes the corresponding estimated ARIR of one channel.

A straightforward solution to the ARIR estimation problem uses a feedforward neural network. The authors hypothesize that the ARIR can be estimated from an omnidirectional RIR without additional information. They make this hypothesis because an RIR signal contains the room acoustical parameters [2], which are sufficient to estimate corresponding ARIRs. Here, a CNN with omnidirectional RIRs is used as input, and the estimated ARIR is used as output. Since the signals are in the time domain, all layers are 1D. This CNN is composed of convolutional layers and transposed convolutional layers, each followed by a batch normalization layer and an activation function, except the last layer. The number of channels increases with depth in the convolutional layers and decreases with depth in the transposed convolutional layers. Instead of learning all channels with a single neural network, one ARIR channel is learned each time.

3.2 ARIR Estimation With a VAE

CNNs learn a deterministic mapping from omnidirectional RIRs to ARIRs without caring about the intrinsic structure of the signals. The authors hypothesize the focus on the features can help to construct ARIRs. This motivates them to use VAEs. An implicit assumption for ARIR estimation they made is that they are able to extract useful information from RIRs to estimate ARIRs. Based on the ISM [9], as discussed in SECS. 1 and 2, the RIR and ARIR signal can be represented by a 15-dim feature vector that contains four features, i.e., 3-dim room geometry, 3-dim source position, 3-dim receiver position, and 6-dim reflection coefficient vector. It is expected that the autoencoder can implicitly perform the ISM to estimate RIR and ARIR and the inverse process to extract room acoustical parameters. However, this turns out to be a difficult task for a neural network, which will be shown in SEC. 4.

In the preliminary test, it was found that if a normal VCAE with a single decoder is used, it focuses on only part of the features and loses some important information required for estimating ARIRs. A multi-task autoencoder can help the latent layer form a good representation [69, 70] and result in a more robust representation of the estimated ARIRs. This motivates the use of a multi-task VCAE to analyze RIRs, estimate ARIRs, and extract these features. Similarly to CNNs, the authors have only omnidirectional RIRs as input. However, they force the latent space of VCAEs to focus on the room acoustic parameters. In addition, the authors are interested to see if the dimensionality of the latent layer corresponds to the known dimensionality.

An important question is how the dimensionality of the latent layer affects the performance of a VAE. In the state-of-the-art VAEs, there is no agreement on the optimal dimensionality of the latent space. The intrinsic dimensionality [71] of a signal refers to the minimum number of parameters necessary for generating the signal. Intrinsic dimensionality can help with the redundancy estimation in the embedded space [72]. In this case, the intrinsic dimensionality of the RIR is 15 by definition. Experiments will be used to analyze how the dimensionality of the latent layer affects the performance of different decoders under a fixed information rate. U-net [73] can outperform the earlier models with connected bypass information. Similarly, a latent layer of this VAE, which is wider than 15 dimensions, can also provide some bypass information. Inspired by U-net, the authors hypothesize that a VAE with a wider latent layer improves performance.

One encoder that takes RIRs as input is used. Multiple decoders are used to perform different tasks. There are four decoders for estimating the four room acoustical parameters respectively. As discussed, the dimensionality of these four features is 15 in total. These four decoders are connected with the first 15 neurons of $\mu_\psi(x)$ to ensure all the information is available. Empirically, it was found that it is difficult to extract RIRs and ARIRs with high accuracy from the first 15 latent neurons alone. Hence, the decoders for the RIRs and ARIRs use additional latent neurons that encode information that the RIR and ARIR decoders find difficult to extract from the first 15 latent neurons alone. This is consistent with the notion that the decoders find it difficult to mimic the ISM and need additional redundancy in the latent layer to perform well.

3.3 Transformations Among Modes of RIRs

The degeneracy of RIRs implies that a different mode results in a different ARIR. Hence, ARIRs must be able to be transformed from one mode to another. The transformations between the modes are linear transforms. From a transformation point of view, as discussed in SEC. 2.1, the relationship among different modes can be classified into mirroring, rotation (i.e., the permutation of room dimensions), and exchange of source and receiver. Each case is dealt with separately.

The mirroring refers to mirroring with respect to $x = 0$, $y = 0$, and $z = 0$. To facilitate the mirroring transformation, the spherical harmonics are first written as direction cosines [74]

$$Y_l = k_l \cdot f_l(u_x, u_y, u_z) \cdot g_l(u_x^2, u_y^2, u_z^2), \tag{16}$$

where l is the Ambisonics channel number of spherical harmonics and can link to (n, m) in Eq. (3) as $l = n(n + 1) + m$,

$$\begin{cases} u_x = \cos(\theta)\cos(\phi) \\ u_y = \sin(\theta)\cos(\phi) \\ u_z = \sin(\phi) \end{cases}, \tag{16}$$

k_l is a scalar, f_l takes the form of $u_x^a \cdot u_y^b \cdot u_z^c$ where a, b, c is either 0 or 1, and g_l is a polynomial of u_x^2, u_y^2, u_z^2. The mirroring can be realized as below [74]. If the sound field is mirrored with respect to $x = 0$, then all terms with u_x are negated. Similarly, If the sound field is mirrored with respect to $y = 0$, then all terms with u_y are negated, and if the sound field is mirrored with respect to $z = 0$, then all terms with u_z are negated.

Rotation is implemented by multiplying the ARIRs of all channels with a rotation matrix $\boldsymbol{Q}$. For simplicity, here, only the rotation matrix for a first-order ARIR rotation around the z axis is shown. Rotation matrices for higher-order Ambisonics for rotation around the x and y axis can be found in [75]. Each element of the matrix $\boldsymbol{Q}$ is denoted as $Q_{n',n}^{m',m}$, and the first-order rotation matrix takes on the form [75]

$$\boldsymbol{Q} = \begin{bmatrix} Q_{0,0}^{0,0} & Q_{0,1}^{0,-1} & Q_{0,1}^{0,0} & Q_{0,1}^{0,1} \\ Q_{1,0}^{-1,0} & Q_{1,1}^{-1,-1} & Q_{1,1}^{-1,0} & Q_{1,1}^{-1,1} \\ Q_{1,0}^{0,0} & Q_{1,1}^{0,-1} & Q_{1,1}^{0,0} & Q_{1,1}^{0,1} \\ Q_{1,0}^{1,0} & Q_{1,1}^{1,-1} & Q_{1,1}^{1,0} & Q_{1,1}^{1,1} \end{bmatrix}. \tag{17}$$

Rotating around the z axis by an angle α corresponds to rotation matrix $\boldsymbol{Q}_Y(\alpha)$ and the element can be calculated as [75]

$$Q_{n',n}^{m',m}(\alpha) = \begin{cases} \cos(m\alpha) & \text{if } n = n' \text{ and } m = m', \\ \sin(m\alpha) & \text{if } n = n' \text{ and } m = -m', \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

The exchange of the source and receiver positions cannot be achieved by a transformation. Consequently, given no prior information, two separate neural networks are trained with different source-receiver position layouts. That is, given an arbitrary input, each ARIR channel is computed with two neural networks. The above transformations can then be applied to both ARIRs to get all Ambisonics representations under different modes of RIRs. If there exists prior information, this additional information can be used to decide which mode is the target one.

3.4 Practical Application

ARIR estimation has many important applications, such as AR, 3D naturalizations, and the transmission of spatial characteristics of room acoustics. Due to the existence of degeneracy of the RIR in a rectangular empty room, in an audio-only environment, one RIR corresponds to the multiple ARIRs given the alignment of all coordinates with a wall orientation. In an AR environment, if one wants to add a virtual object at a position whose RIR is given, it is important to determine the one correct ARIR that gives the user an immersive experience. Different methods can be used to determine the correct ARIR, which will be discussed below. Different methods can combine to increase accuracy.

One method is to use sensors to estimate distances. One can choose from different kinds of sensors based on the resolution requirement and cost, such as radar sensors, light detection and ranging (LiDAR) sensors, ultrasonic sensors, and Bluetooth sensors. The basic underlying principle is similar (i.e., estimating the distance between the user and each wall using return time). Knowing the distance to each wall or one wall from each pair of walls and the relative position of the source, the authors can choose the correct mode and compute the correct ARIR.

Image analysis can also be used to determine the degeneracy. Image analysis can be used to determine the relative positions of the walls, source, and image. Visual simultaneous localization and mapping [76] is one set of methods to locate the user with images only. It includes feature-based, direct, and RGB-D camera-based approaches. [77] proposed a pseudo-LiDAR representation that mimics the LiDAR signal but is converted from image-based depth maps. This method avoids the usage of expensive LiDAR sensors and significantly improves the state-of-the-art image-based method. [78] trained a machine learning model that takes the captured images as input and outputs the distance between the objects and vehicle. After localization from the images, one mode can be determined using the method with sensors.

4 EXPERIMENTS

The experiments are presented in this section. In the first subsection, the setup of the experiments is described. The authors then discuss the experiments on ARIR estimation from RIRs with CNNs and VCAEs in the second and third subsection. Finally, different methods to estimate ARIRs from RIRs are discussed and compared.

4.1 Experimental Setup

In the following, the database used to train and test the model is discussed first. After that, the authors describe the configuration of the neural networks and how they were trained and tested.

4.1.1 Database

To build a clean RIR-ARIR dataset of empty rooms, the ISM was used to simulate RIRs [79], and the methods described in SEC. 2 were used to compute the corresponding ARIRs. The rooms are rectangular and empty. It is assumed the maximum order of reflections can be 100. The speed of sound was set to $c = 340$ m/s. The sampling frequency was set to 8,000 Hz. The length of each RIR was truncated at 1,024 to contain the direct path signal, early reflections, and some of the late reverberation. Each dimension of the room geometry (i.e., length L × width W × height H) was assumed to be independent and identically distributed between 6 × 4 × 2 m and 8 × 6 × 4 m, which covers moderate and small rooms. The reflection coefficient of

Table 1. CNN architecture of ARIR estimation from RIRs.

Operation	Kernel Size	Stride	No. Channels	Output Size
Input				(b, 1,024)
Reshape				(b, 1, 1,024)
Conv1D	16	2	32	(b, 32, 503)
Conv1D	4	1	128	(b, 128, 500)
Conv1D	6	2	512	(b, 512, 248)
Conv1D	8	3	512	(b, 512, 81)
Conv1D	6	1	1, 024	(b, 1,024, 76)
Conv1D	6	2	4, 096	(b, 4,096, 36)
Conv1D	1	1	4, 096	(b, 4,096, 36)
ConvTranspose1d	5	2	1, 024	(b, 1,024, 75)
ConvTranspose1d	4	1	512	(b, 512, 78)
ConvTranspose1d	6	2	128	(b, 128, 160)
ConvTranspose1d	7	1	64	(b, 64, 166)
ConvTranspose1d	3	3	16	(b, 16, 498)
ConvTranspose1d	16	2	4	(b, 4, 1,010)
ConvTranspose1d	15	1	1	(b, 1, 1,024)
Reshape				(b, 1,024)

each wall was simulated as independent and identically distributed between 0 and 1. One source and one receiver were randomly placed in each room under the constraint in Eq. (11) to guarantee a one-to-one mapping function between RIRs and a default ARIR. This prevents the possibility of a one-to-multi relationship that can not be learned by a neural network. The number of simulated RIR-ARIR pairs was 400,000, which was divided into a training dataset, validation dataset, and test dataset with the ratio 7:2:1.

4.1.2 Neural Network Description

This part focuses on the configuration of the neural networks for different objectives and the training and testing of the neural networks. An ablation study was performed on network architecture, optimization method, and hyperparameter tuning with a grid search as a preliminary experiment to choose suitable network architectures and hyperparameters. When some database properties changed, an ablation study was performed again on network architecture and hyperparameters with a grid search.

A GPU was used to train the neural network to estimate ARIRs from RIRs. The Adam optimizer [80] was chosen. Its learning rate was set to 0.001, and the coefficients used for computing running averages of the gradient and its square were set to (0.9, 0.999). The maximum iteration epochs were set to 5,000, and early stopping was applied as regularization in the model [81] to prevent overfitting and limit the computational effort. The MSE loss is recorded per epoch on the training set under training mode and the validation set under evaluation mode. Early stopping was performed when the validation error increased in 100 successive epochs. In addition, mini-batch–based training was used to increase computational efficiency [82]. The batch size was set to 100. After training, the model was set to evaluation mode, and the MSE was computed in the test set.

Network architecture of CNN. Table 1 shows the CNN architecture and corresponding parameters for the ARIR estimation from RIRs, where b denotes the batch size. Each (transposed) convolutional layer is always followed by a batch normalization layer and Leaky Rectified Linear Unit layer [83] as the activation function, which is not listed in Table 1 since the output size is not affected.

Network architecture of VCAE. For the multi-task learning with VCAE, the authors used the architecture of the encoder and decoder and the hyperparameters that are presented in Tables 2–4, where b denotes the batch size, v equals to the dimensionality of latent layer, and w de-

Table 2. Network architecture of encoder part of VCAE.

Operation	Kernel Size	Stride	No. Channels	Output Size
Input				(b, 1,024)
Reshape				(b, 1, 1,024)
Conv1D	16	2	32	(b, 32, 503)
Conv1D	4	1	128	(b, 128, 500)
Conv1D	6	2	512	(b, 512, 248)
Conv1D	8	3	512	(b, 512, 81)
Conv1D	6	1	1, 024	(b, 1,024, 76)
Conv1D	6	2	4, 096	(b, 4,096, 36)
Conv1D	1	1	4, 096	(b, 4,096, 36)
Conv1D	1	1	128	(b, 128, 36)
Reshape				(b, 128×36)
Fully connected				(b, v)

Table 3. Network architecture of decoder part of VCAE (RIR reconstruction and ARIR estimation).

Operation	Kernel Size	Stride	No. Channels	Output Size
Input				(b, v)
Fully connected				(b, 128×36)
Reshape				(b, 128, 36)
ConvTranspose1d	1	1	4, 096	(b, 4,096, 36)
ConvTranspose1d	5	2	1, 024	(b, 1,024, 75)
ConvTranspose1d	4	1	512	(b, 512, 78)
ConvTranspose1d	6	2	128	(b, 128, 160)
ConvTranspose1d	7	1	64	(b, 64, 166)
ConvTranspose1d	3	3	16	(b, 16, 498)
ConvTranspose1d	16	2	4	(b, 4, 1,010)
ConvTranspose1d	15	1	1	(b, 1, 1,024)
Reshape				(b, 1,024)

Table 4. Network architecture of decoder part of VCAE (room acoustical parameters).

Operation	Output Size
Input	(b, v)
Fully connected	(b, 40)
Fully connected	(b, w)

notes the length of room acoustical parameters. Similarly to the CNN architecture, after each (transposed) convolutional layer, the authors always applied a batch normalization layer and Leaky Rectified Linear Unit layer [83] as the activation function. For the multi-task learning with VCAE, λ in Eq. (8) was set to be 0.1, and v in Eq. (8) was set to equal to the latent dimensionality.

4.2 Experiments on ARIR Estimation From RIRs With CNN

This subsection presents experiments on first-order and third-order ARIR estimation from RIRs based on a feed-forward neural network.

In the first experiment, the authors predicted first-order and third-order ARIRs from RIRs. The estimation performance was evaluated with SNR, AMBIQUAL [84], and the reflection analysis of first-order ARIRs as in [85]. The SNR is defined as

$$SNR = 10\log_{10}\left(\frac{\|y\|_2^2}{\|y-\hat{y}\|_2^2}\right). \tag{19}$$

The SNR was measured directly on the ARIRs. AMBIQUAL is an objective quality metric (range between 0 and 1 where 1 means a perfect match) proposed for Ambisonic spatial audio, which estimates listening quality and localization quality from Ambisonics. Experiments showed the AMBIQUAL metric to be strongly correlated to the subjective listening tests [84]. In the context of this paper, since B_0^0 is only a scaled version of the omnidirectional RIR, the authors are interested in only the localization quality. To obtain AMBIQUAL scores, the ARIRs were convolved with ten anechoic recordings, which include six speech utterances from the TSP speech database [86] sound and four audio sound signals from the Audio/Video Anechoic Database [87]. The Intensity Binary Mask threshold of AMBIQUAL was set equal to −50 dB.

Table 5. Experimental results of ARIR estimation with CNN.

Signal	Channel	Test SNR (dB)	AMBIQUAL
First-order ARIR	Horizontal	18.48	0.86
First-order ARIR	Vertical	14.95	0.80
Third-order ARIR	Horizontal	14.25	0.76
Third-order ARIR	Vertical	10.79	0.69

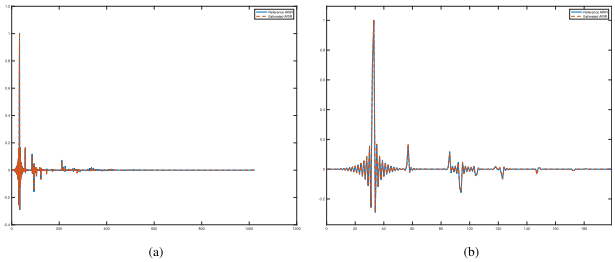

Fig. 2. An estimated first-order ARIR example in the time domain with CNN. (a) Full signal, (b) the first 200 samples.

To evaluate the accuracy of directions of early reflections, the authors adopt the method of [85] where the spherical coordinates of the sound intensity of each time frame are computed and the cross correlation of the spherical coordinates between estimated and reference ARIRs is then used to evaluate the auralization quality. Following the convention of AMBIQUAL, the ARIRs are divided into vertical channels, including B_1^0, B_2^0, and B_3^0, and horizontal channels. The authors average over vertical and horizontal channels, respectively, as the result for the vertical and horizontal channels. The experimental results are shown in Table 5. In addition, Figs. 2−6 show a channel (B_1^1) of estimated first-order ARIRs with average SNR (17.3 dB) and a channel (B_3^1) of estimated third-order ARIRs with average SNR (13.6 dB) as a representative example for a visual impression of the signal quality, where the signal in time domain, the spectrogram, and the spherical coordinates of each time frame of early reflection are included.

The SNR results in Table 5 show that the first-order ARIRs outperform the third-order ARIRs, and horizontal

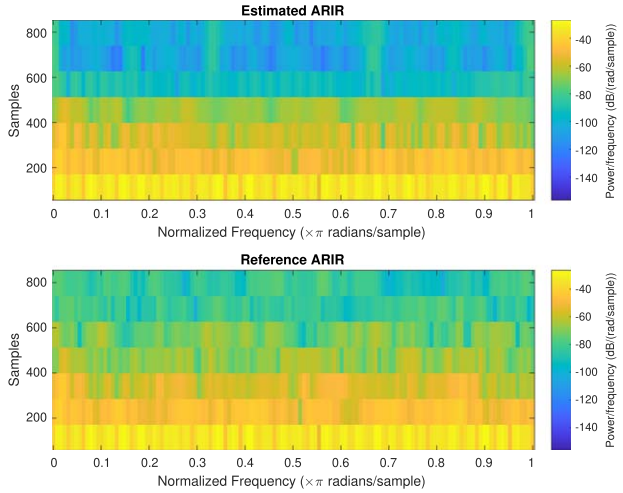

Fig. 3. The spectrogram of an estimated first-order ARIR example with CNN.

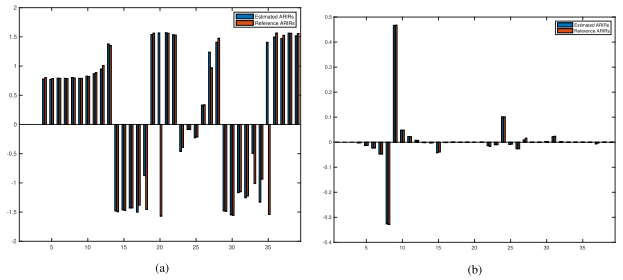

Fig. 4. Reflection analysis of an estimated first-order ARIR example with CNN. (a) Azimuth of the sound intensity of each time frame in early reflections, (b) elevation of the sound intensity of each time frame in early reflections.

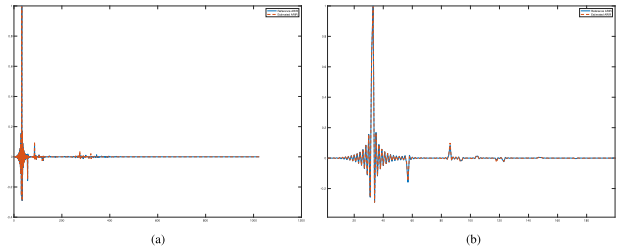

Fig. 5. An estimated third-order ARIR example in the time domain with CNN. (a) Full signal, (b) the first 200 samples.

channels outperform vertical channels. The visual impression is consistent with the SNR results in Table 5. The SNR and the figure show that the estimated ARIRs are reasonable. The AMBIQUAL score confirms the estimated ARIR performs well in terms of localization accuracy for an indoor environment with reverberation. For the auralization quality, the maximum absolute cross correlation for azimuth and elevation are 0.70 and 0.99, respectively, which indicates the estimated directions of early reflections are reasonable and correspond to a reasonable auralization quality.

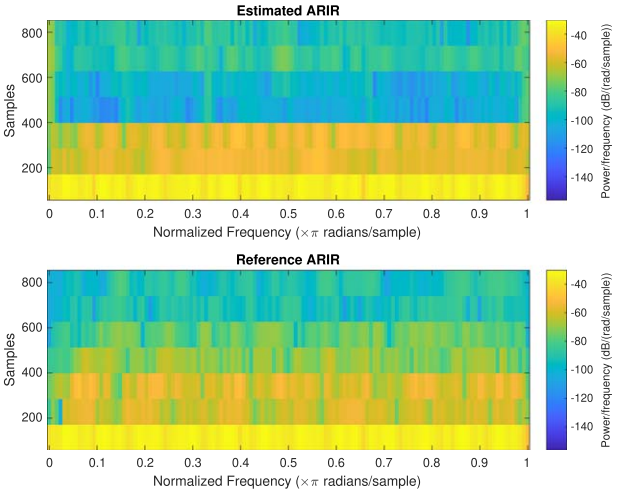

Fig. 6. The spectrogram of an estimated third-order ARIR example with CNN.

4.3 Experiments on Multi-Task VCAE–Based ARIR Estimation

This subsection presents the experiments on VCAE-based ARIR estimation as described in SEC. 3. The authors show the performance of the different decoders under different dimensionality of the latent space. In addition, the performance of VCAE is compared with CNN.

The experiments were performed on different dimensionalities at the same information rate. The reference dimensionality was set to be 15 since it was preassumed the features of an RIR can be described by a 15-dim vector as described in SEC. 3. The dimensionality of the latent space was also set to 10, 30, 60, 80, 100, 200, and 400 for comparison. These experiments indicate that, as long as each neuron of the latent layer can be allocated with more than one bit information rate on average, a higher information rate does not improve the experimental results. As a result, the information rate was set to 600 bits for training to make sure the multi-task VCAE of different dimensionalities had enough information rate. Although a multi-task autoencoder was used, the authors aimed to synthesize the ARIRs. Consequently, the different models were compared based on the performance of the ARIRs. Since the different ARIR channels perform similarly, only channel B_1^0 was used to compare the different latent dimensionalities.

The relationship between latent dimensionality and the performance for the estimated ARIR is shown in Fig. 7. It shows that the model with latent dimensionality 200 performed best on estimated ARIR. It proved that a wider latent layer (before reaching the plateau) improved the results, although the signal can be described by a 15D latent layer. This is consistent with the experiment in [88] where complex-valued as well as the magnitude and instantaneous frequency of the short-time Fourier transform result in a better performance than the time-domain waveform. Both these experiments and the results of [88] show that neural networks have difficulty learning some classes of complex relations.

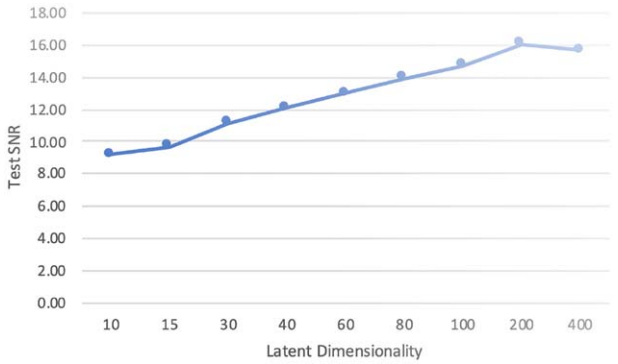

Fig. 7. SNR of estimated ARIR under different dimensionality.

Table 6. Experimental results of ARIR estimation with VCAE.

Signal	Channel	Test SNR (dB)	AMBIQUAL
First-order ARIR	Horizontal	18.40	0.87
First-order ARIR	Vertical	16.05	0.84
Third-order ARIR	Horizontal	16.67	0.83
Third-order ARIR	Vertical	14.11	0.80

Table 7. Experimental results of RIR reconstruction and the estimation of room acoustical parameters with VCAE.

	Test SNR (dB)
RIR	22.5773
Receiver position	37.40
Source position	39.10
Room geometry	43.40
Reflection coefficients	21.36

The previous experiment shows the model with latent dimensionality 200 is the best model. Experimental results of horizontal and vertical ARIR channels with SNR and the AMBIQUAL results were presented. Since the authors reconstructed RIR and estimated room acoustical parameters when they estimated each channel of ARIR, they averaged over these estimates and compared them with the ground truth in terms of SNR. In addition, similarly to the CNN based method, the early reflections of first-order ARIRs were analyzed, and the auralization quality was evaluated.

The experimental results on estimated ARIRs are shown in Table 6. The method performs better on first-order ARIR estimation than on the more difficult third-order ARIR estimation because of the higher resolution and more detailed description of higher-order ARIRs. Horizontal channels outperform vertical channels, which is likely related to the vertical room dimension being smaller. Table 7 presents the experimental results on reconstructed RIRs and estimated room acoustical parameters. The multi-task autoencoder structure also shows reasonable performance on these bypass tasks. In addition, example channels of estimated first-order ARIR (B_1^1) were plotted with average SNR (17.62 dB) and third-order ARIR (B_3^1) with average SNR (16.15 dB) in Figs. 8–12 for a visual impression on the signal quality, where the signal in the time domain, the spectrogram, and the spherical coordinates of each time frame of

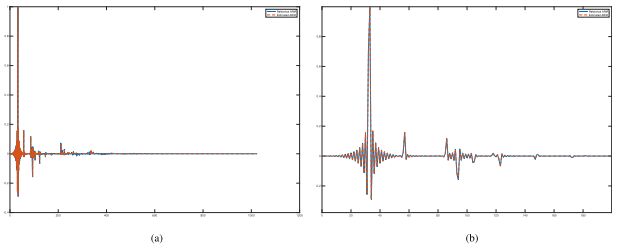

Fig. 8. An estimated first-order ARIR example in the time domain with VCAE. (a) Full signal, (b) the first 200 samples.

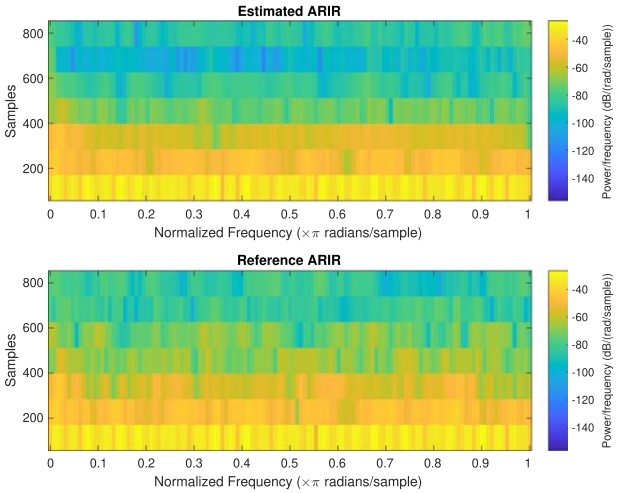

Fig. 9. The spectrogram of an estimated first-order ARIR example with VCAE.

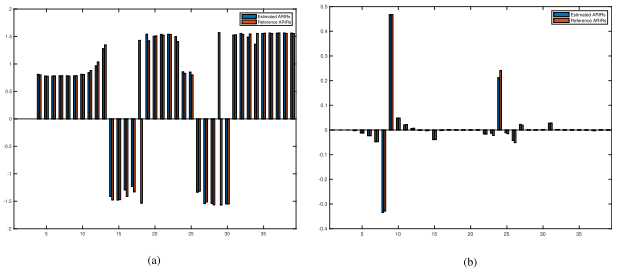

Fig. 10. Reflection analysis of an estimated first-order ARIR example with VCAE. (a) Azimuth of the sound intensity of each time frame in early reflections, (b) elevation of the sound intensity of each time frame in early reflections.

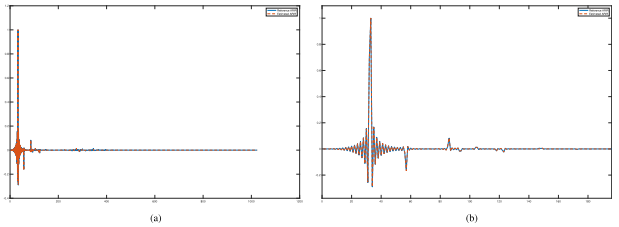

Fig. 11. An estimated third-order ARIR example in the time domain with VCAE. (a) Full signal, (b) the first 200 samples.

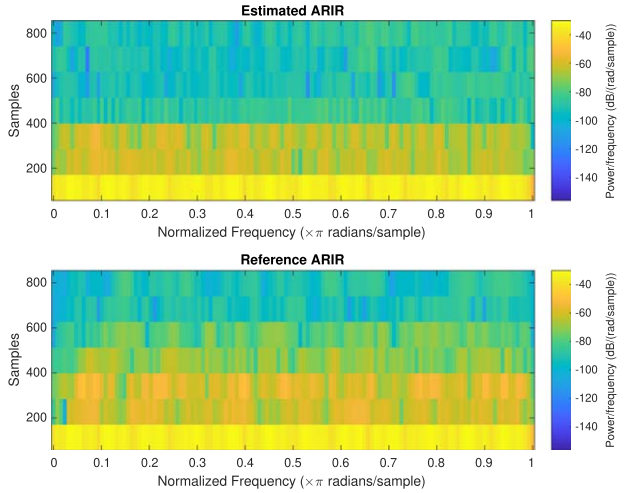

Fig. 12. The spectrogram of an estimated third-order ARIR example with VCAE.

early reflection are included. For the auralization quality, the maximum absolute cross correlation for azimuth and elevation are 0.73 and 0.99, respectively. From the SNR, AMBIQUAL score, auralization quality, and figures, it can be concluded that the performance of VCAE-based estimated ARIRs is good.

At the end of this section, the performance is compared with the CNN in Table 5 and VCAE in Table 6. The SNR, AMBIQUAL, and auralization quality all confirm that the VCAE-based method outperforms the CNN-based method, especially for the third-order ARIRs. As discussed in SEC. 3.2, the VCAE uses features useful for both RIRs and ARIRs, and then all important features are passed to the main decoder for the ARIR estimation task. This helps to formulate a good representation of the latent layer and results in a more robust estimation of ARIRs. Consequently, the VCAE-based method shows a better performance of ARIR estimation than the CNN-based method.

5 CONCLUSION

In this paper, the authors showed it is possible to estimate ARIRs from omnidirectional RIRs under the assumption that there exist at least two perpendicular walls within a finite set of degeneracy. The proposed method only requires a single RIR between a source and receiver as input. The proposed method works in rectangular rooms where each wall has a single reflection coefficient based on the ISM. The generalization of the real measured data obviates the need for special measurement equipment. Two methods were used to achieve this mapping, a feedforward mapping with CNN and a multi-task VCAE. The experiments showed that the multi-task VCAE performs better than the feedforward mapping, especially for higher-order ARIRs, since the structure is more suitable for the estimation of ARIRs. An extension of this work can focus on the generalization of real-world measurements.

The proposed method is just an initial step in ARIR estimation. There exist many directions for developing this work. Firstly, the proposed method is based on the ISM. The ISM makes some assumptions, and some sound properties cannot be modeled, for example, the scattering and frequency-dependent reflection coefficients. Future work can relax these assumptions and model these effects in the Ambisonics representations. Secondly, the room is assumed to be rectangular. Future work can generalize the rooms to include nonrectangular room shapes. The directivity of the loudspeaker and microphone can be taken into account. Finally, listening tests can be conducted to evaluate the performance of the estimated ARIRs.

6 REFERENCES

[1] P. Cipresso, I. A. C. Giglioli, M. A. Raya, and G. Riva, "The Past, Present, and Future of Virtual and Augmented Reality Research: A Network and Cluster Analysis of the Literature," *Front. Psychol.*, vol. 9, paper 2086 (2018 Nov.). https://doi.org/10.3389/fpsyg.2018.02086.

[2] W. Yu and W. B. Kleijn, "Estimation of Source and Receiver Positions, Room Geometry and Reflection Coefficients From a Single Room Impulse Response," *arXiv preprint arXiv:2301.09198* (2023 Jan.). https://doi.org/10.48550/arXiv.2301.09198.

[3] E. Benjamin and T. Chen, "The Native B-Format Microphone," presented at the *119th Convention of the Audio Engineering Society* (2005 Oct.), paper 6621.

[4] L. Gölles and F. Zotter, "Directional Enhancement of First-Order Ambisonic Room Impulse Responses by the 2+2 Directional Signal Estimator," in *Proceedings of the 15th International Audio Mostly Conference (AM)*, pp. 38–45 (Graz, Austria) (2020 Sep.). https://doi.org/10.1145/3411109.3411131.

[5] E. Hoffbauer and M. Frank, "Four-Directional Ambisonic Spatial Decomposition Method With Reduced Temporal Artifacts," *J. Audio Eng. Soc*, vol. 70, no. 12, pp. 1002–1014 (2022 Dec.). https://doi.org/10.17743/jaes.2022.0039.

[6] A. Wabnitz, N. Epain, and C. T. Jin, "A Frequency-Domain Algorithm to Upscale Ambisonic Sound Scenes," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 385–388 (Kyoto, Japan) (2012 Mar.). https://doi.org/10.1109/ICASSP.2012.6287897.

[7] W. B. Kleijn, A. Allen, J. Skoglund, and F. Lim, "Incoherent Idempotent Ambisonics Rendering," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 209–213 (New Paltz, NY) (2017 Oct.). https://doi.org/10.1109/WASPAA.2017.8170025.

[8] W. B. Kleijn, "Directional Emphasis in Ambisonics," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 1079–1083 (2018 Jul.). https://doi.org/10.1109/LSP.2018.2841652.

[9] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950 (1979 Apr.). https://doi.org/10.1121/1.382599.

[10] A. B. Barry, "An Interdisciplinary Synthesis of Reverberation Viewpoints," *J. Audio Eng. Soc.*, vol. 49, no. 10, pp. 867–903 (2001 Oct.).

[11] J.-D. Polack, "Modifying Chambers to Play Billiards: the Foundations of Reverberation Theory," *Acta Acust. un. Acust.*, vol. 76, no. 6, pp. 256–272 (1992 Jul.).

[12] A. Lindau, L. Kosanke, and S. Weinzierl, "Perceptual Evaluation of Model- and Signal-Based Predictors of the Mixing Time in Binaural Room Impulse Responses," *J. Audio Eng. Soc.*, vol. 60, no. 11, pp. 887–898 (2012 Nov.). https://doi.org/10.14279/depositonce-15236.

[13] R. Stewart and M. B. Sandler, "Statistical Measures of Early Reflections of Room Impulse Responses," in *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx)* (Bordeaux, France), pp. 59–62. (2007 Sep.).

[14] N. Mohanan, R. Velmurugan, and P. Rao, "Speech Dereverberation Using NMF With Regularized Room Impulse Response," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4955–4959 (New Orleans, LA) (2017 Mar.). https://doi.org/10.1109/ICASSP.2017.7953099.

[15] M. Joorabchi, S. Ghorshi, and A. Sarafnia, "Single-Channel Speech Dereverberation in Acoustical Environments," in *Proceedings of the International Symposium on Electronics in Marine (ELMAR)*, pp. 1–4 (Zadar, Croatia) (2014 Sep.). https://doi.org/10.1109/ELMAR.2014.6923353.

[16] A. H. Moore, M. Brookes, and P. A. Naylor, "Room Geometry Estimation From a Single Channel Acoustic Impulse Response," in *Proceedings of the 21st European Signal Processing Conference (EUSIPCO 2013)*, pp. 1–5 (Marrakech, Morocco) (2013 Sep.).

[17] Y. E. Baba, A. Walther, and E. A. P. Habets, "3D Room Geometry Inference Based on Room Impulse Response Stacks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 5, pp. 857–872 (2018 May). https://doi.org/10.1109/TASLP.2017.2784298.

[18] W. Yu and W. B. Kleijn, "Room Acoustical Parameter Estimation From Room Impulse Responses Using Deep Neural Networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 436–447 (2021 Dec.). https://doi.org/10.1109/TASLP.2020.3043115.

[19] J. He and M. Zhu, "Simulation of Combined Head and Room Impulse Response Based on Sound Ray Tracing in Frequency Domain," in *Proceedings of the IET International Conference on Smart and Sustainable City (ICSSC)*, pp. 361–365 (Shanghai, China) (2013 Aug.). https://doi.org/10.1049/cp.2013.1957.

[20] A. Alpkocak and M. K. Sis, "Computing Impulse Response of Room Acoustics Using the Ray-Tracing Method in Time Domain," *Arch. Acoust.*, vol. 35, no. 4, pp. 505–519 (2010 Sep.).

[21] C. Gu, M. Zhu, H. Lu, and B. Beckers, "Room Impulse Response Simulation Based on Equal-Area Ray Tracing," in *Proceedings of the International Conference on Audio, Language and Image Processing*, pp. 832–836 (Shanghai, China) (2014 Jul.). https://doi.org/10.1109/ICALIP.2014.7009911.

[22] F. J. Serón, F. J. Sanz, M. Kindelan, and J. I. Badal, "Finite-Element Method for Elastic Wave Propagation," *Commun. Appl. Numer. Methods*, vol. 6, no. 5, pp. 359–368 (1990 Jul.). https://doi.org/10.1002/cnm.1630060505.

[23] J. D. De Basabe and M. K. Sen, "Grid Dispersion and Stability Criteria of Some Common Finite-Element Methods for Acoustic and Elastic Wave Equations," *Geophysics*, vol. 72, no. 6, pp. T81–T95 (2007 Nov.). https://doi.org/10.1190/1.2785046.

[24] L. L. Thompson, "A Review of Finite-Element Methods for Time-Harmonic Acoustics," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1315–1330 (2006 Mar.). https://doi.org/10.1121/1.2164987.

[25] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia* (Academic Press, Cambridge, MA, 1994).

[26] S. G. McGovern, "Fast Image Method for Impulse Response Calculations of Box-Shaped Rooms," *Appl. Acoust.*, vol. 70, no. 1, pp. 182–189 (2009 Jan.). https://doi.org/10.1016/j.apacoust.2008.02.003.

[27] D. Jerome and M. Sebastien, "Further Study of Sound Field Coding With Higher Order Ambisonics," presented at the *116th Convention of the Audio Engineering Society* (2004 May), paper 6017.

[28] F. Hollerweger, "An Introduction to Higher Order Ambisonic," http://decoy.iki.fi/dsound/ambisonic/motherlode/source/HOA_intro.pdf (2005 Apr.).

[29] M. A. Poletti, "Three-Dimensional Surround Sound Systems Based on Spherical Harmonics," *J. Audio Eng. Soc.*, vol. 53, no. 11, pp. 1004–1025 (2005 Nov.).

[30] T. M. MacRobert, *Spherical Harmonics: An Elementary Treatise on Harmonic Functions, With Applications* (Dover Publications, Mineola, NY, 1948), 2nd ed.

[31] M. Chapman, W. Ritsch, T. Musil, et al., "A Standard for Interchange of Ambisonic Signal Sets. Including a File Standard With Metadata," presented at the *Ambisonics Symposium* (Graz, Austria) (2009 Jun.).

[32] N. Epain and C. T. Jin, "Spherical Harmonic Signal Covariance and Sound Field Diffuseness," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 10, pp. 1796–1807 (2016 Oct.). https://doi.org/10.1109/TASLP.2016.2585862.

[33] M. A. Poletti, "A Unified Theory of Horizontal Holographic Sound Systems," *J. Audio Eng. Soc.*, vol. 48, no. 12, pp. 1155–1182 (2000 Dec.).

[34] F. Zotter and M. Frank, "All-Round Ambisonic Panning and Decoding," *J. Audio Eng. Soc.*, vol. 60, no. 10, pp. 807–820 (2012 Oct.).

[35] L. S. Davis, R. Duraiswami, E. Grassi, et al., "High Order Spatial Audio Capture and Its Binaural Head-Tracked Playback Over Headphones With HRTF Cues," presented at the *119th Convention of the Audio Engineering Society* (2005 Oct.), paper 6540.

[36] G. Enzner, M. Weinert, S. Abeling, J.-M. Batke, and P. Jax, "Advanced System Options for Binaural Rendering of Ambisonic Format," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 251–255 (Vancouver, Canada) (2013 May). https://doi.org/10.1109/ICASSP.2013.6637647.

[37] M. Noisternig, A. Sontacchi, T. Musil, and R. Holdrich, "A 3D Ambisonic Based Binaural Sound Re-

production System," in *Proceedings of the 24th AES International Conference on Multichannel Audio: The New Reality* (2003 Jun.), paper 1.

[38] M. Zaunschirm, M. Frank, and F. Zotter, "Binaural Rendering With Measured Room Responses: First-Order Ambisonic Microphone vs. Dummy Head," *Appl. Sci.*, vol. 10, no. 5, paper 1631 (2020 Feb.). https://doi.org/10.3390/app10051631.

[39] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in *Proceedings of the DAGA*, vol. 44, pp. 339–342 (Munich, Germany) (2018 Mar.).

[40] M. Zaunschirm, C. Schörkhuber, and R. Höldrich, "Binaural Rendering of Ambisonic Signals by Head-Related Impulse Response Time Alignment and a Diffuseness Constraint," *J. Acoust. Soc. Am.*, vol. 143, no. 6, pp. 3616–3627 (2018 Jun.). https://doi.org/10.1121/1.5040489.

[41] T. McKenzie, D. T. Murphy, and G. Kearney, "Diffuse-Field Equalisation of Binaural Ambisonic Rendering," *Appl. Sci.*, vol. 8, no. 10, paper 1956 (2018 Oct.). https://doi.org/10.3390/app8101956.

[42] M. Kuster, "Multichannel Room Impulse Response Rendering on the Basis of Underdetermined Data," *J. Audio Eng. Soc.*, vol. 57, no. 6, pp. 403–412 (2009 Jul.).

[43] M. Kuster, "Reliability of Estimating the Room Volume From a Single Room Impulse Response," *J. Acoust. Soc. Am.*, vol. 124, no. 2, pp. 982–993 (2008 Aug.). https://doi.org/10.1121/1.2940585.

[44] C. Pörschmann and S. Wiefling, "Perceptual Aspects of Dynamic Binaural Synthesis Based on Measured Omnidirectional Room Impulse Responses," in *Proceedings of the 3rd International Conference on Spatial Audio (ICSA)* (Graz, Austria) (2015 Sep.).

[45] C. Pörschmann and P. Stade, "Auralizing Listener Position Shifts of Measured Room Impulse Responses," in *Proceedings of the DAGA*, pp. 1308–1311 (Berlin, Germany) (2016 Mar.).

[46] C. Pörschmann, P. Stade, and J. M. Arend, "Binauralization of Omnidirectional Room Impulse Responses-Algorithm and Technical Evaluation," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pp. 345–352 (Edinburgh, Scotland) (2017 Sep.).

[47] C. Pörschmann, P. Stade, and J. M. Arend, "Binaural Auralization of Proposed Room Modifications Based on Measured Omnidirectional Room Impulse Responses," *Proc. Mtgs. Acoust.*, vol. 30, no. 1, paper 015012 (2017 Oct.). https://doi.org/10.1121/2.0000622.

[48] J. M. Arend, S. V. A. Garí, C. Schissler, F. Klein, and P. W. Robinson, "Six-Degrees-of-Freedom Parametric Spatial Audio Based on One Monaural Room Impulse Response," *J. Audio Eng. Soc*, vol. 69, no. 7/8, pp. 557–575 (2021 Jul./Aug.). https://doi.org/10.17743/jaes.2021.0009.

[49] K. Fukushima, "Neocognitron: A Hierarchical Neural Network Capable of Visual Pattern Recognition," *Neural Netw.*, vol. 1, no. 2, pp. 119–130 (1988 Sep.). https://doi.org/10.1016/0893-6080(88)90014-7.

[50] Y. LeCun, B. Boser, J. S. Denker, et al., "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551 (1989 Dec.). https://doi.org/10.1162/neco.1989.1.4.541.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916 (2015 Sep.). https://doi.org/10.1109/TPAMI.2015.2389824.

[52] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823 (Boston, MA) (2015 Jun.). https://doi.org/10.1109/CVPR.2015.7298682.

[53] S. Hershey, S. Chaudhuri, D. P. W. Ellis, et al., "CNN Architectures for Large-Scale Audio Classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131–135 (New Orleans, LA) (2017 Mar.). https://doi.org/10.1109/ICASSP.2017.7952132.

[54] S. Jingzhou, W. Yongbin, and C. Xiaosen, "Audio Segmentation and Classification Approach Based on Adaptive CNN in Broadcast Domain," in *Proceedings of the IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*, pp. 1–6 (Beijing, China) (2019 Jun.). https://doi.org/10.1109/ICIS46139.2019.8940257.

[55] T. V. Kumar, R. S. Sundar, T. Purohit, and V. Ramasubramanian, "End-to-End Audio-Scene Classification From Raw Audio: Multi Time-Frequency Resolution CNN Architecture for Efficient Representation Learning," in *Proceedings of the International Conference on Signal Processing and Communications (SPCOM)*, pp. 1–5 (Bangalore, India) (2020 Jul.). https://doi.org/10.1109/SPCOM50965.2020.9179600.

[56] V. Dumoulin and F. Visin, "A Guide to Convolution Arithmetic for Deep Learning," *arXiv preprint arXiv:1603.07285* (2016 Mar.). https://doi.org/10.48550/ARXIV.1603.07285.

[57] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv preprint arXiv:1312.6114* (2013 Dec.). https://doi.org/10.48550/arXiv.1312.6114.

[58] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic Backpropagation and Approximate Inference in Deep Generative Models," in *Proceedings of the 31st International Conference on Machine Learning*, vol. 32, no. 2, pp. 1278–1286 (Beijing, China) (2014 Jun.). https://doi.org/10.48550/arXiv.1401.4082.

[59] C. Doersch, "Tutorial on Variational Autoencoders," *arXiv preprint arXiv:1606.05908* (2016 Jun.). https://doi.org/10.48550/arXiv.1606.05908.

[60] D. P. Kingma and M. Welling, "An Introduction to Variational Autoencoders," *Found. Trends Mach. Learn.*, vol. 12, no. 4, pp. 307–392 (2019 Nov.). https://doi.org/10.1561/2200000056.

[61] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "A Recurrent Variational Autoencoder for Speech Enhancement," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 371–375 (Barcelona, Spain)

(2020 May). https://doi.org/10.1109/ICASSP40776.2020.9053164.

[62] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational Autoencoder for Speech Enhancement With a Noise-Aware Encoder," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 676–680 (Toronto, Canada) (2021 Jun.). https://doi.org/10.1109/icassp39728.2021.9414060.

[63] X. Chen, Y. Sun, M. Zhang, and D. Peng, "Evolving Deep Convolutional Variational Autoencoders for Image Classification," *IEEE Trans. Evol. Comput.*, vol. 25, no. 5, pp. 815–829 (2020 Dec.). https://doi.org/10.1109/TEVC.2020.3047220.

[64] C. Varano, "Disentangling Variational Autoencoders for Image Classification," https://cs231n.stanford.edu/reports/2017/pdfs/3.pdf (2017).

[65] I. Higgins, L. Matthey, A. Pal, et al., "Beta-VAE: Learning Basic Visual Concepts With a Constrained Variational Framework," presented at the *5th International Conference on Learning Representations (ICLR)* (Toulon, France) (2017 Apr.).

[66] D. T. Braithwaite and W. Kleijn, "Bounded Information Rate Variational Autoencoders," *arXiv preprint arXiv:1807.07306* (2018 Jul.). https://doi.org/10.48550/arXiv.1807.07306.

[67] D. Braithwaite and W. B. Kleijn, "Speech Enhancement With Variance Constrained Autoencoders," in *Proceedings of the INTERSPEECH*, pp. 1831–1835 (Graz, Austria) (2019 Sep.). https://doi.org/10.21437/Interspeech.2019-1809.

[68] D. T. Braithwaite, M. O'Connor, and W. B. Kleijn, "Variance Constrained Autoencoding," *arXiv preprint arXiv:2005.03807* (2020 May). https://doi.org/10.48550/arXiv.2005.03807.

[69] M. Crawshaw, "Multi-Task Learning With Deep Neural Networks: A Survey," *arXiv preprint arXiv:2009.09796* (2020 Sep.). https://doi.org/10.48550/arXiv.2009.09796.

[70] C. Fifty, E. Amid, Z. Zhao, et al., "Efficiently Identifying Task Groupings for Multi-Task Learning," in *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS)*, pp. 27503–27516 (Online) (2021 Dec.). https://doi.org/10.48550/ARXIV.2109.04617.

[71] R. S. Bennett, "Representation and Analysis of Signals Part XXI. The Intrinsic Dimensionality of Signal Collections," Tech. Rep. AD0475844 (1965 Apr.).

[72] S. Gong, V. N. Boddeti, and A. K. Jain, "On the Intrinsic Dimensionality of Image Representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3982–3991 (Long Beach, CA) (2019 Jun.). https://doi.org/10.1109/CVPR.2019.00411.

[73] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, vol. 9251, pp. 234–241 (Springer, Cham, Switzerland, 2015).

[74] M. Chapman, "Symmetries of Spherical Harmonics: Applications to Ambisonics," presented at the *Ambisonics Symposium* (Graz, Austria) (2009 Jun.).

[75] J. G. Tylka and E. Y. Choueiri, "Algorithms for Computing Ambisonics Translation Filters," 3D3A Tech. Rep. 2 (2019 Mar.).

[76] T. Taketomi, H. Uchiyama, and S. Ikeda, "Visual SLAM Algorithms: A Survey From 2010 to 2016," *IPSJ Trans. Comput. Vis. Appl.*, vol. 9, no. 1, paper 16 (2017 Jun.). https://doi.org/10.1186/s41074-017-0027-2.

[77] Y. Wang, W.-L. Chao, D. Garg, et al., "Pseudo-LiDAR From Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8437–8445 (Long Beach, CA) (2019 Jun.). https://doi.org/10.1109/CVPR.2019.00864.

[78] J. A. Musk, S. K. Sahai, and A. K. Elluswamy, "Estimating Object Properties Using Visual Image Data," US Patent 10,956,755 B2 (2021 Mar.).

[79] Laboratories Erlangen International Audio, "RIR Generator," https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator (2014).

[80] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," presented at the *International Conference on Learning Representations (ICLR)* (San Diego, CA) (2015 May). https://doi.org/10.48550/arXiv.1412.6980.

[81] L. Prechelt, "Early Stopping — But When?" in G. Montavon, G. B. Orr, K.-R. Müller (Eds.), *Neural Networks: Tricks of the Trade*, pp. 53–67 (Springer, Berlin, Germany, 2012). https://doi.org/10.1007/978-3-642-35289-8_5.

[82] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: Primal Estimated Sub-Gradient Solver for SVM," *Math. Program.*, vol. 127, no. 1, pp. 3–30 (2011 Mar.). https://doi.org/10.1007/s10107-010-0420-4.

[83] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier Nonlinearities Improve Neural Network Acoustic Models," in *Proceedings of the International Conference on Machine Learning (ICML)*, vol. 28, paper 3 (Atlanta, GA) (2013 Jun.).

[84] M. Narbutt, J. Skoglund, A. Allen, et al., "AMBIQUAL: Towards a Quality Metric for Headphone Rendered Compressed Ambisonic Spatial Audio," *Appl. Sci.*, vol. 10, no. 9, paper 3188 (2020 May). https://doi.org/10.3390/app10093188.

[85] P. Małecki, "Spatial Impulse Response Assessment in Room Acoustics Auralization," *Acta Phys. Pol. A*, vol. 128, no. 1–A, pp. 17–21 (2015 Jul.). https://doi.org/10.12693/APhysPolA.128.A-17.

[86] P. Kabal, "TSP Speech Database," https://www.mmsp.ece.mcgill.ca/Documents/Data/ (2002).

[87] D. Thery and B. F. G. Katz, "Anechoic Audio and 3D-Video Content Database of Small Ensemble Perfor-

mances for Virtual Concerts," in *Proceedings of the 23rd International Congress on Acoustics (ICA)*, pp. 739–746 (Aachen, Germany) (2019 Sep.).

[88] J. Nistal, S. Lattner, and G. Richard, "Comparing Representations for Audio Synthesis Using Generative Adversarial Networks," in *Proceedings of the 28th European Signal Processing Conference (EUSIPCO)*, pp. 161–165 (Amsterdam, The Netherlands) (2021 Jan.). https://doi.org/10.23919/Eusipco47968.2020.9287799.

THE AUTHORS

Wangyang Yu

W. Bastiaan Kleijn

Wangyang Yu received a B.Sc. degree in information engineering from the Beijing Institute of Technology, Beijing, China, in 2015, and M.Sc. degree in electrical engineering from the Delft University of Technology, Delft, The Netherlands, in May 2017. She obtained her Ph.D. degree with the Signal Processing Systems Group, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands, in May 2024. Her research interests include room acoustics, Ambisonics, audio signal processing, and deep learning.

•

W. Bastiaan Kleijn is Professor at Victoria University of Wellington in New Zealand (since 2010) and Research Scientist at Google (since 2011). He was Professor and Head of the Sound and Image Processing Laboratory with KTH, Stockholm, 1996–2014; Professor at TU Delft, 2011–2021; and prior to that, Member of the Technical Staff at the Research Division of AT&T Bell Laboratories. He holds Ph.D. degrees in Electrical Engineering from TU Delft and Soil Science from the University of California, Riverside. He was a Founder of Global IP Solutions, which provided the enabling audio technology to Skype and was later acquired by Google. He has served on the editorial Boards of four IEEE journals and was the Technical Chair of ICASSP 1999 and EUSIPCO 2010. He is a Fellow of the IEEE (1999) and of the Royal Society of New Zealand (2021).