Machine-learning-based estimation and rendering of scattering in virtual reality

Ville Pulkki[1,a] and U. Peter Svensson[2]
[1]*Department of Signal Processing and Acoustics, Acoustics Lab, P.O. Box 13000, Aalto University, FI-00076 Aalto, Finland*
[2]*Department of Electronic Systems, Norwegian University of Science and Technology, 7491 Trondheim, Norway*

(Received 31 July 2018; revised 22 October 2018; accepted 23 October 2018; published online 30 April 2019)

In this work, a technique to render the acoustic effect of scattering from finite objects in virtual reality is proposed, which aims to provide a perceptually plausible response for the listener, rather than a physically accurate response. The effect is implemented using parametric filter structures and the parameters for the filters are estimated using artificial neural networks. The networks may be trained with modeled or measured data. The input data consist of a set of geometric features describing a large quantity of source-object-receiver configurations, and the target data consist of the filter parameters computed using measured or modeled data. A proof-of-concept implementation is presented, where the geometric descriptions and computationally modeled responses of three-dimensional plate objects are used for training. In a dynamic test scenario, with a single source and plate, the approach is shown to provide a similar spectrogram when compared with a reference case, although some spectral differences remain present. Nevertheless, it is shown with a perceptual test that the technique produces only a slightly lower degree of plausibility than the state-of-the-art acoustic scattering model that accounts for diffraction, and also that the proposed technique yields a prominently higher degree of plausibility than a model that omits diffraction.
 https://doi.org/10.1121/1.5095875

[NX] Pages: 2664–2676

I. INTRODUCTION

Scattering occurs when a sound wave reaches a finite object, where the presence of the object causes the sound to be redirected to all directions. The directional radiation of scattering depends heavily on the geometry of the object and on the acoustic characteristics of the surface; typically it follows a complex frequency-dependent pattern that cannot be described in any straightforward manner (Fahy, 2000).

Acoustic virtual reality aims to provide a listener with the same perception of sound as would occur in a corresponding real scenario (Sherman and Craig, 2003). Typical use cases are in acoustic design, computer gaming, and telepresence. In most cases the target is to produce a dynamic rendering of the virtual world, where the sources, receivers, and objects may be moving. To achieve plausible rendering results for such dynamic conditions, the update rate of spatial audio rendering should be relatively high, for example, 5–50 Hz.

The rendering of sound scattered from a finite object is challenging in such virtual realities. If the object has a relatively simple geometry, the scattering phenomenon can be computed accurately; although, even in such scenarios the simulation of scattering may require such high computational resources that is not practical for real use cases. Furthermore, in some cases the geometry is too complex to be simulated or the acoustic parameters of the object are not known, which makes computational modeling impossible in practice.

This article proposes a method where the scattering effect is rendered using a parametric filter structure, and proposes an efficient method to estimate filter-parameters directly from the geometry of a source-object-receiver scenario, using artificial neural networks. The networks are taught using a large number of simulations and/or real measurements where the feature vectors describing the source-object-receiver configurations are associated to parametric descriptions of scattered sound. When rendering a virtual reality, the scattering may then be estimated efficiently while retaining the perception of high plausibility to the listener, even though the physical accuracy of the estimated sound may be relatively low. The possibility to use real measurements in training of the system makes it possible to estimate scattering with acoustically and geometrically complex objects, such as living animals, which opens a large number of potential applications for the approach.

II. BACKGROUND

In order to introduce the reader to the field, this chapter provides the necessary background regarding scattering phenomena, which is followed by a discussion of practical implementations that are utilized in virtual environments. The physically accurate numerical methods utilized in this article to produce the reference data are introduced in

[a]Also at: Department of Electrical Engineering, Hearing Systems, Technical University of Denmark, Kongens Lyngby, Denmark. Electronic mail: Ville.Pulkki@aalto.fi

Sec. II C and finally the earlier applications of artificial neural networks in room acoustics are reviewed in Sec. II D.

A. Acoustic scattering phenomena

The basic acoustic phenomena that occur when an airborne sound wave reaches an object (Savioja and Svensson, 2015; Vorländer, 2013) can be summarized as follows:

(1) The **direct sound** behaves like a sound wave radiating in free-field, with an additional visibility constraint; thus, whenever the line-of-sight from source to receiver is obstructed, the direct sound is set to zero. This can be illustrated as in Figs. 1(a) and 1(b), where no sound reaches receivers that are obstructed, or occluded, by the corner. The sudden disappearance of the direct sound occurs at the shadow zone boundary, in this case with the receiver position at −1 m.
(2) For a flat and smooth surface, a **specularly reflected** sound wave can appear, as if generated by an image source. A visibility constraint also applies to the specular reflection, and a reflection zone boundary appears so that wavefronts are truncated as illustrated in Fig. 1(b) with the receiver position 1 m. In dynamic simulations, these abrupt effects that occur when a receiver passes a zone boundary, can be clearly audible. The amplitude and spectrum of the reflected wave is determined by the (plane-wave) reflection coefficient.
(3) The wavefront truncation can be corrected by the addition of **edge diffraction** waves, which appear as waves from the edges of the finite reflecting surfaces. Figure 1(c) illustrates that adding a diffraction wave to the GA solution in Fig. 1(b), will yield smooth wavefronts. The edge diffraction waves will also cause a substantial change to the spectrum of the reflected wave. In Fig. 1(c) one can observe that the direct sound is weakened also in the visible zone, an effect which is frequency-dependent. Finite objects generate higher orders of diffraction waves, which may be computationally challenging to render. Explicit expressions for edge diffraction waves are available only for rigid surfaces, and a few other special cases (Hewett and Morris, 2015).
(4) **Surface scattering, or diffuse reflection**, results when a reflecting surface is not smooth, but rough. If the roughness has a size which is not negligible compared to the wavelength, then the reflected sound wave will be scattered in the entire hemisphere in front of the diffusely reflecting part of the surface. Various models exist to describe the directivity of this diffuse-reflection phenomenon, including the simple Lambert-model, which has been adapted from optics (Kuttruff, 2017). This wavelength dependence shows that rough surfaces typically reflect specularly for low frequencies and almost completely diffusely at high frequencies. Perceptually, the amplitude and spectrum of the specular reflection might be audibly affected by the diffuse reflection.

The term "scattering" is used here to describe the total reflected sound: the sum of specular and diffuse reflections as well as edge diffraction waves.

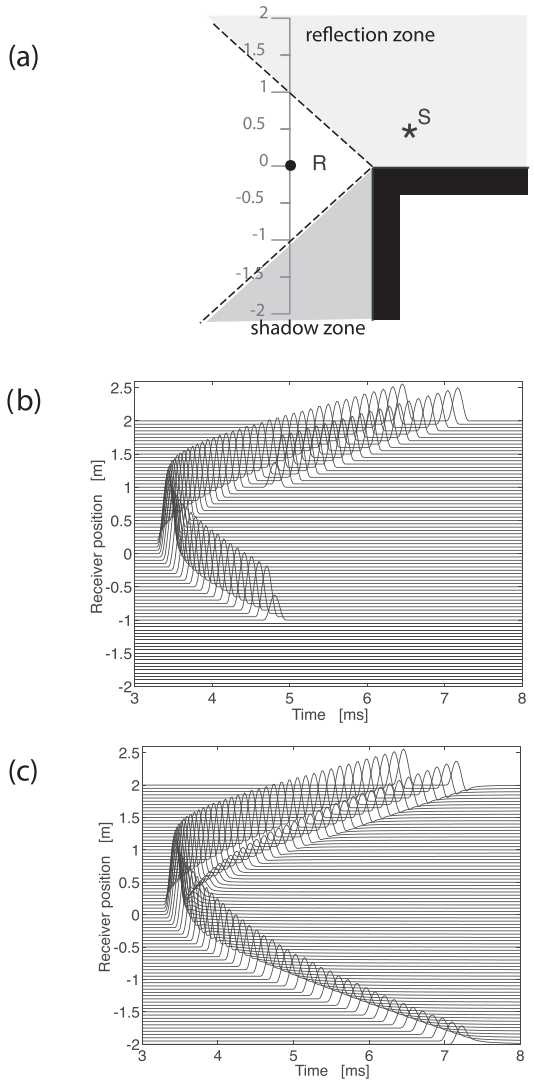

FIG. 1. An example of a first-order diffraction wave for an infinite wedge with closed angle of 90°. (a) The geometry, with a fixed source, S, and a receiver moving across the zone boundaries. (b) The geometrical acoustics (GA) impulse responses, low-pass filtered, for a number of receiver positions, indicating the discontinuities for the direct sound and for the specular reflection. (c) After the addition of diffraction, waves modeled with the edge source (ES) model; note that the wavefronts are perfectly smooth when changing the receiver position.

In this work, the target is to auralize the effect of scattering using simple parametric filters, whose parameters are estimated using machine learning. To test this approach, the conditions were simplified to shoebox-shaped filled boxes with rigid and smooth surfaces, with object geometries commonly found in domestic environments, such as computer displays and flat-panel TVs. When doing so, the surface scattering effects can be neglected. This was a practical choice, as the simulation of such objects is computationally relatively efficient, taking approximately in the order of 5–30 s per source-object-receiver setup in a modern laptop computer, using the simulation methods described further in Sec. II C. The model has been found to correspond well to reference calculations for the scattering from a rigid cube (Martin *et al.*, 2018). The simulation of arbitrary surfaces would have required more computational resources to model. However, it is still assumed that the results will show the

general trend also with objects with arbitrary surface structures.

B. Simulation of acoustics in virtual reality

The simulation of sound reaching the avatar, a virtual representative for the user, is typically conducted using geometrical acoustics (GA) modeling as described above. In practice this is done by ray-based rendering of direct and reflected sounds and by adding some diffuse reverberation to the scene (Savioja *et al.*, 1999; Vorländer, 2007). Ray-based acoustics can be implemented with relatively low computational resources. In practice each ray is implemented using the following technique, where the sound $s_n(t)$ arriving at the listener, corresponding to the nth ray, is computed by delaying and attenuating the sound emitted by the source according to its propagation path as

$$s_n(t) = s_0(t - \tau) r_0 / r_n, \tag{1}$$

where $s_0(t)$ is the signal captured at the distance of r_0 from the source. τ is additional propagation delay $(r - r_0)/c$, where c is the speed of sound in air $c = 344$ m/s, and r is the length of the shortest travel path between the source and the object.

Additionally, the acoustic properties of reflecting surfaces along the propagation path, in the form of frequency-dependent absorption, may be implemented with digital filtering. Each ray is then made audible to the listener typically using either head-related transfer function (HRTF) based filtering for headphones or loudspeaker-based techniques (Pulkki and Karjalainen, 2015). However, the downside of such methods is that surface scattering and diffraction phenomena are neglected.

Surface scattering is straightforward to implement in ray-based calculations (Schissler *et al.*, 2014). Accurate modeling of surface scattering is usually required in non-real-time ray-based simulations, but the effects may be minor in dynamic situations, so that simplified reverberation will represent scattering effects as well as higher-order reflections.

Edge diffraction effects have been implemented as complements to ray-based models. A simulation method is utilized to estimate an finite impulse response (FIR) or infinite impulse response (IIR) for each diffracted path (Lokki *et al.*, 2002; Pulkki *et al.*, 2002), and they are treated as secondary sources at positions of edges, which scatter sound to all directions in the virtual room. The method described below in Sec. II C can be used in modeling, although it may be computationally too complicated for interactive virtual acoustics applications. A number of simplified and computationally less demanding approaches to modeling the scattering from rigid objects have been suggested, which can be implemented in real-time dynamic scenarios (Schissler *et al.*, 2014; Tsingos *et al.*, 2001; Tsingos and Gascuel, 1998; Tsingos *et al.*, 2011). However, the methods still require relatively simple geometries with hard and flat surfaces, and the application to objects that include curved geometries or absorptive surface material is complicated. The approach suggested here could handle any scattering object as long as one has a parametric filter model that can be trained by data obtained from numerical simulations or measurements.

Alternatively, instead of ray-based methods, wave-based methods can be utilized. In Murphy and Beeson (2003) a digital waveguide mesh was used. An approach has also been proposed, where the rendering is based on pre-computed wave-domain modeling of the virtual world (Raghuvanshi and Snyder, 2014), however, the method requires a relatively large allocation of memory for the purpose, and also that the geometry of the world does not change significantly during run-time. These wave-based methods for estimating the scattering effect are computationally demanding, and it might be challenging to model geometrically complex objects, or objects that have absorptive surfaces.

C. Physical modeling of edge diffraction from plate objects

The tools for edge diffraction modeling that were used to provide the reference cases for training the machine learning systems and for evaluation of them will now be described. As the types of objects for this work have been restricted to plate objects, or more specifically rigid polyhedra, the GA model has to be extended with contributions originating from the edges; note that the edge diffraction waves are described in more detail in point 3 in Sec. II. In terms of rendering virtual reality audio, they improve a GA model notably as they "remedy" the discontinuous sound field caused by the zone boundaries, and they also determine the spectral shape of the reflected sound. In addition, the edge diffraction waves allow sound to propagate around objects to reach occluded receiver positions.

An edge source (ES)-based model to compute diffraction waves was presented in Svensson *et al.* (1999) and Asheim and Svensson (2013) and is implemented in a freely available MATLAB toolbox, "EDtoolbox" (Svensson, 2000). Results computed by this method are denoted the "ES model" from henceforth.

As a general view, the impulse response from the source via a single edge to the receiver is a decaying function that has a sharp onset, possibly a knee point where the level drops to half, and a sharp offset. The sharp onset corresponds to the shortest travel path from the source via the edge to the receiver, and the point on the edge from where the shortest route is traversed is called the apex point of the edge. The temporal response may have a positive or negative onset and it decays asymptotically towards zero, and the decay may include one zero crossing. The shape of the decaying time response is not exactly an exponential function, and its closed-form expression is rather complex (Svensson *et al.*, 1999). Higher-order diffraction denotes the process where the diffracted sound is diffracted again with single or multiple edges. In practice, the response is similar to first-order edge diffraction, with sharp onset and offset points. When all diffraction contributions are summed at the receiver, typically several sharp onsets are seen, since the lengths of travel paths for diffracted components vary.

D. Applications of artificial neural networks in acoustics

Artificial neural networks are an efficient tool for different engineering tasks, such as for classification or regression in the context of complex and non-linear phenomena (Goodfellow *et al.*, 2016; Haykin, 1994). In the field of acoustics, they have advantages in the following tasks: speech recognition (Hinton *et al.*, 2012; Kohonen, 1988); classification of sounds, for example, in everyday sound environments (Cakir *et al.*, 2017); or for the estimation of qualities of physical objects based on sound emitted from them, for example the analysis of roughness of roads from wheel sounds (Ambrosini *et al.*, 2018). In audio signal processing, neural networks can be used to aid in the processing of complex recorded sounds, for example, in source separation (Grais and Plumbley, 2018).

The applications of neural networks in virtual reality audio, room acoustics, or in scattering modeling are sparse. In the work of Schissler *et al.* (2018), the network estimates the acoustic parameters of surfaces from visual images, and in the work of Kon and Koike (2018), networks have been trained to estimate the reverberation time of spaces from visual images. The phenomenon of acoustic scattering has been touched in studies, where in underwater acoustics the back-scatter from a seabed is used to classify it using neural networks (Marsh and Brown, 2009), and in the field of ultrasonics the back-scattered sound is used to recognize the geometry of an object (Watanabe and Yoneyama, 1992).

III. DYNAMIC TEST SCENARIO

A test scenario was designed to illustrate the methods and to compare the proposed methods with reference simulations, and it is also used later in the subjective experiments in this work. The scenario includes a sound source, a diffracting plate, and a receiver, as depicted in Fig. 2. The scenario is dynamic in the sense that the receiver *moves* in relation to the rest of the environment. The geometry of the scenario is designed to ensure relatively high audibility of the scattering effect from the plate.

A. Geometry of the test scenario

A 62 cm × 47 cm × 3 cm plate, corresponding to the size of a normal computer screen, is located with the center of the front plate at the origin. The source is at 250 cm distance in the horizontal plane, in the direction of −45° azimuth. The receiver, which is either a pressure sensor or a binaural human listener, rotates around the plate from 1° to 181° with 5° steps. The offset of 1° from five-based values was introduced since the physical modeling tool used in the study did not produce a result with the direction of 90° of azimuth.

The source is visualized as a loudspeaker, however, the frequency response and directivity of it is assumed to be flat, which makes the effects caused by scattering from the plate more prominent. The sampling frequency was 48 kHz. The source was positioned at a relatively long distance to emphasize the effect of the scattered sound component in the presence of direct sound; when the source is further away, the propagation attenuation has a similar range for both direct and scattered sound components. This is opposed to the case where the source is close to the receiver, where the propagation attenuation is much more prominent for scattered sound than for direct sound.

B. Reference simulations of scattering in the test scenario

Scattering was simulated with the ES model including 15th-order diffraction for 50 frequency points spaced logarithmically between 50 Hz and 12 kHz for each receiver location. The frequency resolution was selected to cover the most important hearing range of humans with a slightly higher resolution than humans possess (Pulkki and Karjalainen, 2015). The resulting location-frequency spectrogram of only the scattered component is shown in Fig. 3(a). It can be noted that the scattered sound has a discontinuity at the shadow zone boundaries, at angles of 120° and 150°, respectively, to compensate for the direct sound's discontinuity. The high-pass-nature of the specular reflection can be seen from receiver azimuth angles 30° to 60°, and the low-pass-nature of occlusion is visible from receiver angles 120° to 150°. The scattering strength is low towards receivers in directions between 80° and 100°. There exists also a comb-filter structure that changes the notch frequencies dynamically as the direction of the receiver changes.

In the dynamic test scenario the sound arriving directly at the receiver is present most of the time, and it is therefore interesting to monitor the interference between the scattered

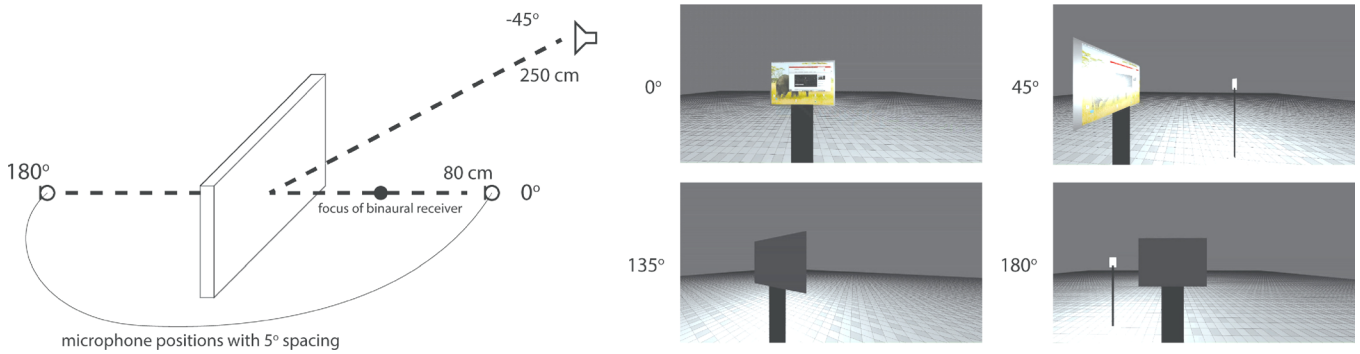

FIG. 2. (Color online) Left: the geometry of the dynamic scene utilized as a test scenario. Both the source and the receiver are located in the horizontal plane, with the source fixed in the direction of −45° and the receiver moving from azimuth angle of 0° to 180°. Right: four snapshots from the corresponding video with annotated viewpoints.

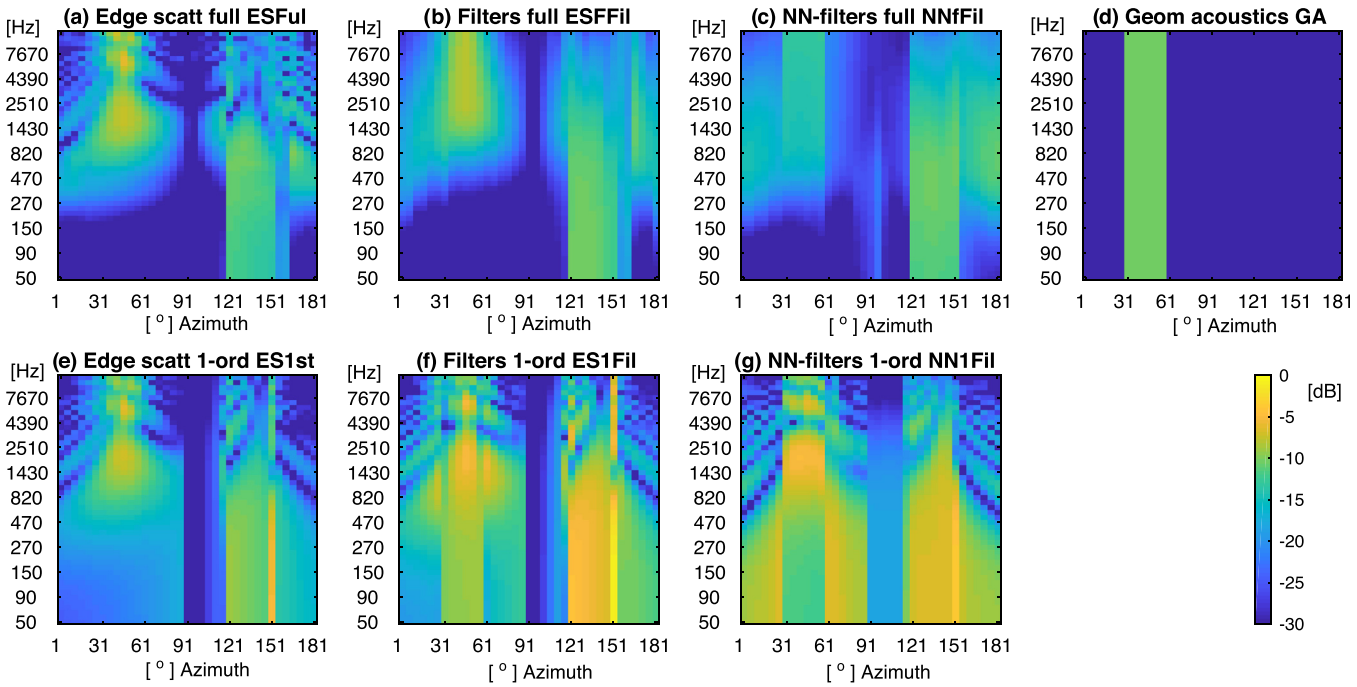

FIG. 3. (Color online) The scattered component arriving from the plate to the receiver in the dynamic scene shown in Fig. 2 rendered with different methods. (a) 15th-order modeling of diffraction, (b) parametric IIR structure directly fitted to the physically modeled spectrum, (c) parameters of the structure estimated from the geometry using neural networks, (d) geometrical acoustic model with direct sound and specular reflection, (e) only first-order edge diffraction simulated physically, (f) each first-order diffraction component rendered with a first-order low-pass filter, (g) parameters estimated for each first-order filter, directly estimated from the geometry using neural networks.

sound and the direct sound, as similar interference occurs also in the ears of the listener. The resulting spectrogram is shown in Fig. 4(a). It can be seen, that the scattered component does have a prominent effect near the regions where the specular reflection occurs and where the occlusion occurs. Consequently, the effect is mild in the region where the level of the scattered sound component is low. When the source becomes audible near the azimuth of 150°, there is a prominent change in the level of sound, which seems larger than one would expect in real cases. It is not known if this effect corresponds to reality, or if the ES model exaggerates the change.

The dynamic situation was also modeled using only first-order diffraction, which is produced by each edge that is

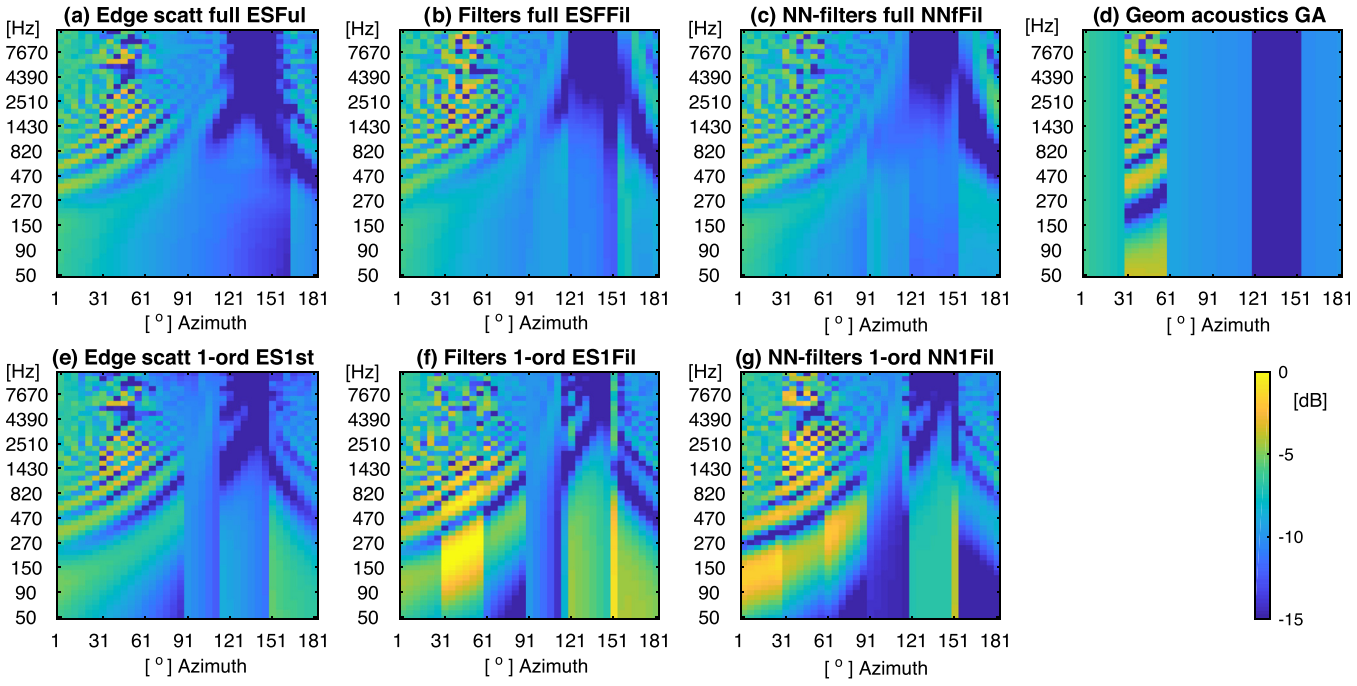

FIG. 4. (Color online) Same as Fig. 3, but with the inclusion of the direct sound contribution.

directly visible from the source and from the receiver. The resulting scattering spectrogram is shown in Fig. 3(e). Although the spectrogram is relatively similar with corresponding 15th-order model spectrogram in Fig. 3(a), there are some distinct differences, such as the level of low-frequency sound in the presence of a specular reflection. The spectrogram computed for the case where the direct sound is present is shown in Fig. 4(e). When it is compared with the 15th-order result in Fig. 4(a), spectral differences are visible. In addition, there exist also prominent discontinuities in the spectrogram with changing azimuth angle of the receiver. For example, the change in spectrum in the direction of 90° can be attributed to the fact that one of the edges becomes invisible to the receiver when the receiver passes 90°, and the contribution of it vanishes abruptly. With 15th-order diffraction, the higher orders of diffraction prevent such discontinuities, as seen in Fig. 4(a).

For the sake of comparison, the output of the GA model described in Sec. II B is shown for only the reflected component in Fig. 3(d), and in Fig. 4(d) the reflected component is summed with direct sound. Prominent differences can be seen when compared with the ground truth cases shown in Figs. 3(a) and 4(a), respectively. It can be seen that the image source model does not render the high-pass effect of the reflected component, it renders occlusion as silence, and also it does not produce smooth spectral transitions when reflection and occlusion effects occur.

IV. PROPOSED APPROACH

A method to extend the GA model with object scattering is proposed in the following. The spectral effect of scattering is rendered using a parametric filter structure, which in practice modifies the spectrum with low-pass or high-pass effects, or effects where frequencies above or below a certain frequency are attenuated or amplified by a certain factor, namely, shelving filtering effects. The implementation with parametric filters differs from earlier approaches, where generic structures such as warped IIRs were used (Lokki *et al.*, 2002), where the filters were directly designed to fit the frequency response of diffracted sound, instead of utilizing parametric filters. The motivation to use parametric filters is that they can be controlled by meaningful measures, such as cut-off frequencies in Hz and frequency-specific output amplitude in dB. This is assumed to provide an intuitive and a computationally efficient means to deliver the effect of scattering to the listener, even in those cases where the scattering has very complex geometry.

Furthermore, it is assumed that an artificial neural network can learn to associate the geometry of the source-object-receiver setup with filter parameters. For example, in reality, when a receiver approaches an occluding object in the shadow, the effect of occlusion is a low-pass effect where the cut-off frequency decreases with decreasing distance between the listener and the object. The machine learning should then associate the angular size of the occluding object from the view point of the listener to the cut-off frequency of the corresponding low-pass filter.

In the proposed approach a large set of real measurements or computer simulations of scattering are utilized to train artificial neural networks. Each network is trained to associate the geometric description of the source-object-receiver to a specific filter parameter value set in the training stage. The target filter parameters values in training are in turn obtained by fitting the parametric filter response to the corresponding measured or simulated data. During the run-time of the virtual reality audio engine, the description of the current geometric configuration is then used as an input to the trained networks, which compute the parameters for the filter structure, used to render the scattering effect.

The motivations for this approach are:

(1) Computational requirements are low. Artificial neural networks are computationally efficient during run-time, and it is assumed that the plausible effect of scattering can be obtained using relatively simple filter structures.
(2) The accuracy of estimation of scattering does not have to be high. It is assumed that relatively high errors can be made in the estimation of the spectrum of scattered sound since the scattered component is often not present alone, but such contributions as direct sound, reverberant field or reflections from other objects are also reaching the listener, in which case, some of spectral errors may be masked by these other sound components.
(3) The goal of the proposed rendering approach is plausibility, rather than authenticity. Since the user does not have a direct reference of scattering available, a *plausible* rendering is targeted (Pellegrini, 2001); where plausibility is defined to be "a suitable reproduction of all required quality features for a given specific application."

V. REPRESENTATION OF SCATTERED COMPONENTS USING LOW-ORDER FILTER STRUCTURE

Two approaches are proposed to render scattered sound using filter structures.

(1) The first approach is to implement the sound arriving from each edge using a separate filter. It can be assumed that the spatial effect of diffraction components arriving from different directions can be reproduced with high fidelity. However, a challenge may be to design the filters accurately enough, as errors in the design may lead into the accumulation of spectral errors due to interference.
(2) The second approach is to implement the total scattered sound component emanating from an object with a single filter, whose output is then spatialized to the direction of the apex point that corresponds to the shortest sound route via the scattering object. Here, the interference problems do not exist, however, the spatial accuracy is lower, although it is not known if the effect is perceptually prominent.

The approaches are described in more detail in the following.

A. Filter-representation of first-order diffraction from individual edges

In this approach, the sound path from a source to an edge and further to the receiver is modeled with a propagation

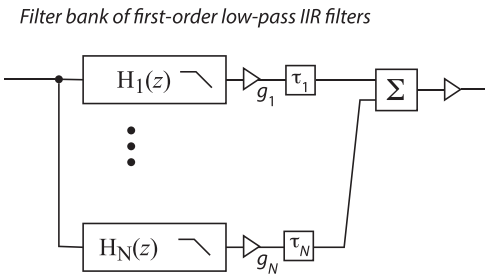

FIG. 5. Filter bank used to render scattering each edge separately.

delay τ_n, propagation attenuation g_n, and a low-order low-pass filter with response $H_n(z)$; where n denotes the index of each individual edge filter. The resulting filter bank is shown in Fig. 5. In rendering, the resulting sound is positioned to the direction from where the sound ray arrives at the listener.

In this study it was chosen to estimate the first-order diffraction effect simply using a first-order low-pass filter with an exponentially decaying impulse response, and a −6 dB/oct slope in the corresponding magnitude response. It is acknowledged that such an approximation is not necessarily accurate, but the implementation represents the simplest computational implementation imaginable. The parameters for such a filter are the level of passband L_{LP}, the cut-off frequency f_{LP} of the filter, and the polarity of the time-domain response p. The response of the filter is denoted as $H_n(z)$.

To illustrate the method, the dynamic test scenario described in Sec. III was rendered by it. For each receiver position, the parameters for each filter in the bank were fitted to the 1st-order diffraction responses computed using EDToolbox. Figures 3(f) and 4(f) show the results without and with direct-sound-cases, respectively. It can be seen that the estimation produces relatively large discrepancies from the physically modeled counterparts and some notable discontinuities exist in the spectrogram with the changing of the angle of the receiver. The perceptual effects of these issues are investigated later in this work.

B. Filter representation of total scattered sound from an object

In this section the scattering effect caused by an object is implemented using a single filter, without considering the edges separately as was described in Sec. V A. The produced spectrum can be assumed to be more complex than the first-order diffraction spectrum of a single edge, as the first- and higher-order diffraction contributions from each edge accumulate at the listener position with arbitrary phase relationships, causing interference effects.

The main spectral characteristics of scattering should be implementable with the filter selected for the task. To find such a filter structure, a large quantity of different symmetric and asymmetric boxes with dimensions from 40 cm to 2 m, were simulated with the EDToolbox in a preliminary study and the spectral shapes of scattered components were monitored. Two distinct types of spectra were identified:

(1) If the direct sound is visible, the effect was found to be, in most cases, a high-pass effect with 12 dB/oct stopband response. In some cases, a plateau was also identified at low frequencies. Above the cut-off frequency of the high-pass filter, the spectral shape typically varied wildly.
(2) If the direct sound is not visible, i.e., the source is occluded, a low-pass effect with an approximate −6 dB/oct stopband response was observed, although there could be large frequency-dependent variations in the spectrum.

To implement these effects, a simplified representation of scattered sound is proposed, with the filter structures shown in Fig. 6, one structure for direct-visible cases and the other for the occluded cases. The switching from direct-visible case to occluded case can be determined based on the visibility of direct sound. However, in reality the object scattering produces some low-pass slightly outside the shadow zone boundary, and thus the switching is made where the direct sound is just about visible, and the edge diffraction produces already a low-pass effect. In this case the target filter parameters are computed based on the combined effect of scattered sound and direct sound.

The parameters of the filters are described in the prototype frequency responses shown in Fig. 6. The filter parameters are fitted to the responses with a heuristic approach, where first the high- or low-pass filter is fitted to the spectrum, for direct-sound-visible and -invisible cases, respectively. The fitting is performed by finding the lowest estimation error by sliding the position of the cut-off frequency. After this step, the parameters

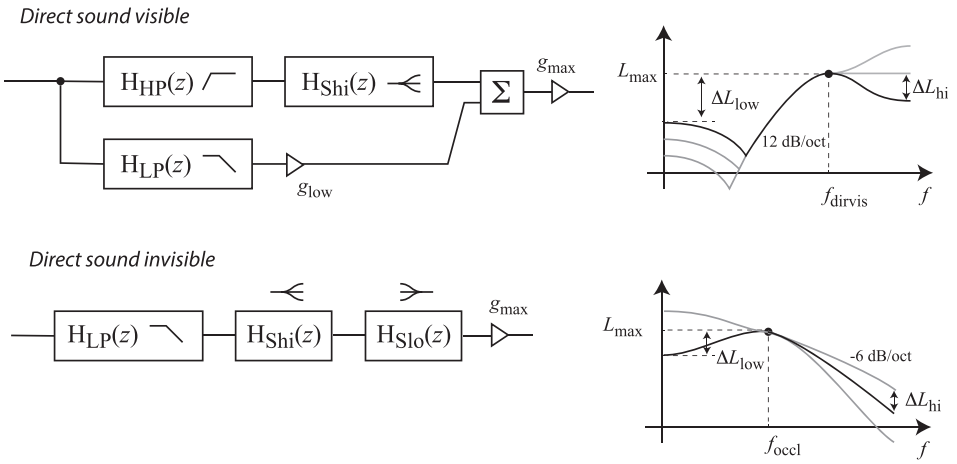

FIG. 6. Left: Filter structures used to render scattering from an object. Right: prototype responses.

for the remaining filter components are computed in a least-squared-error sense, using the methods described in the work of Välimäki and Reiss (2016).

The direct-sound-visible filter consists of parallel first-order low-pass and second-order high-pass filters with the same cut-off frequency. The low-pass contribution is attenuated by a gain factor g_{low}, and the corresponding level parameter is shown in the figure as $\Delta L_{\text{low}} = 20\log(g_{\text{low}}/g_{\text{max}})$. After fitting the high-pass and low-pass filters to the response, a high-frequency first-order shelf filter is subsequently fitted to provide a better approximation of the spectrum. The occlusion filter is formed in a similar fashion, however, the high-pass filter is not included and both high-frequency and low-frequency first-order shelf filters are utilized to approximate the spectral slopes in the modeled spectrum.

To get an indication of the accuracy obtained with such a filter model, the 15th-order edge source responses computed for the dynamic scene shown in Fig. 3(a) were implemented with the proposed filter structure here and the resulting spectra are shown in Fig. 3(b). It can be seen that the filter-structure captures the high-pass and low-pass-effects relatively well, and also the overall level of scattered sound. However, the comb-filter effects are generally lost.

Figure 4(b) shows the resulting spectrogram when the direct sound is included. In this case, the scattered component is implemented as a sound ray that implements the propagation delay and distance attenuation; where the propagation distance is computed using the shortest source-receiver route touching at one position on the surface of the object. It can be seen that the interference patterns in the spectrogram are similar to the ground truth case shown in Fig. 4(a). The introduction of propagation delay to the scattered component produces complex interference patterns in a similar fashion as in the original. The method also provides the smooth fade-in and fade-out behavior of the specular reflection, and low-pass-filtered occlusion effects; although, relatively large differences between figures (a) and (b) are also visible. It may be assumed that the perceptual difference between the physical model result and filter-fitted result is minimal, which is verified in the perceptual testing described in Sec. VII.

VI. ESTIMATION OF FILTER PARAMETERS

In the method proposed in this article the filter parameters that implement diffracted or scattered components are estimated directly from the geometry of the scenario using neural networks. A non-linear regression task is performed with most of the networks, where the input features are geometric descriptors of the source-edge-receiver system, and the output is a parameter for the filter structure. For regression, the classic feed-forward neural network with Bayesian regularization backpropagation training with one or two hidden layers is used with the MATLAB command fitnet. In one case, a network for classification is used using MATLAB command patternnet, where the polarity of the output response is sought. The selection of a linear or logarithmic scale for the target parameter was based on the assumption that the error will be distributed evenly on the selected scale and an even distribution is desired in most cases.

A network capable of estimating all of the parameters at the same time would be desirable, however, in this proof-of-concept implementation it was convenient to train a separate network for the estimation of each individual parameter enabling individual monitoring of the performance. The details of the networks utilized in the evaluation of the system are shown in Table I, which were chosen by testing multiple versions, before selecting the best alternative. The goodness-of-fit values measured with the trained networks, using the test data, are also given in the table as correlation values or, alternatively, as a percentage of correct classification.

A. NN-estimation of parameters for filter modeling of first-order diffraction

The selection of features to be used as input to the neural network is not a straightforward task. However, in this case, the response of first-order diffraction from a single edge is estimated where a closed-form solution exists, and can be computed accurately with a limited set of parameters (Biot and Tolstoy, 1957). It is thus, in principle, clear that the parameters that are fed into the first-order diffraction equation contain all relevant information, although the mapping from the parameters of equations into the parameters of filters is not a trivial endeavor.

A separate network was trained to estimate each value for level L_{LP}, the cut-off frequency f_{LP}, and the polarity p, as shown in Fig. 5. Each network utilizes the same feature vectors, where the values of each component were computed as: (1) dot product between incoming sound direction and the

TABLE I. Neural networks taught to estimate parameters for filter structures used to render diffraction and/or scattering effects. Hid. = hidden.

Single-edge diffraction filter nets					
# input features	Target parameter	Net type	Hid. layer1	Hid. layer2	Regression/classification
13	$\log(f_{\text{LP}})$ [log(Hz)]	fitnet	10	4	$R = 0.84$
13	L_{LP} [dB]	fitnet	10	4	$R = 0.92$
13	p [·]	patternnet	10	5	corr = 72%
Total object scattering nets					
21	$\log(f_{\text{dirvis}})$ [log(Hz)]	fitnet	15	10	$R = 0.85$
21	$\log(f_{\text{occl}})$ [log(Hz)]	fitnet	12	0	$R = 0.68$
21	ΔL_{hi} [dB]	fitnet	22	12	$R = 0.87$
21	ΔL_{lo} [dB]	fitnet	22	12	$R = 0.85$
21	$10^{(L_{\text{max}}/20)}$ [lin]	fitnet	12	0	$R = 0.95$

edge direction [·]; (2) dot product between outgoing sound direction and the edge direction [·]; (3) incoming sound direction in cylindric coordinates of the edge [rad]; (4) outgoing sound direction in cylindric coordinates [rad]; (5) source distance [m]; (6) receiver distance [m]; (7) z cylindric coordinate of source [m]; (8) z cylindric coordinate of receiver [m]; (9) length of the edge [m]; (10) closed angle of the wedge [m]; (11) direct sound visible [boolean]; (12) reflection visible [boolean]; and (13) difference between shortest route from source to receiver and from source to apex to receiver [m].

The training corpus was formed using the EDToolbox. Edges with lengths ranging from 0.3 to 0.95 m with 5 cm increments were formed with three different closed-wedge angles: 90°, 67°, and 45°. A source was simulated at 20 different random directions with the source-distance varying from 60 cm to 2 m. For each source, 30 receivers were also simulated at random directions and distances varying from 60 cm to 2 m. If the acoustic simulation resulted in a valid response, i.e., if the travel path of sound was not obstructed, the computed target filter parameters were used to obtain the response. This resulted in approximately 56 000 data points. 15% of data were used as test data, and the regression value for the networks for L_{LP}, the cut-off frequency f_{LP} were obtained as 0.92 and 0.84, respectively, which may be considered as a relatively good fit to the data. However, the polarity network did not fit to the data well, with only 72% of the cases providing the correct response. It is not known if this is a feature of the data, or the filter-simplification of the rendering, or if it is caused by some other unidentified reason.

Erroneous polarity estimation can produce severe aberrations in the rendering, as the corresponding impulse-response pulse will be phase-inverted and complex interference effects will occur since the pulses arriving from each edge are summed together at the receiver. To illustrate the accuracy of the method, the test scenario described in Sec. III was simulated with the trained networks and the resulting spectrogram is shown in Fig. 3(g). It can be observed that the method simulates the components relatively well, however, some clear differences between original and NN-simulated spectrograms are visible. Furthermore, especially at regions near the zone borders, the level changes with changing azimuth, producing clear discontinuities.

In the simulation case, where the direct sound is present, in Fig. 4(g), it can be seen that although the response is relatively similar in some regions, large discrepancies can be identified when compared with the 1st-order reference case in Fig. 4(e). Furthermore, the discontinuities in the spectrogram corresponding to the changing of receiver azimuth are also more prominent than in the filter-modeled version in Fig. 4(f). Such errors and discontinuities may cause audible artifacts in sound rendering, which is investigated with perceptual/subjective tests later in this work. Although the estimation accuracy can be observed to be limited, based on a visual comparison of the spectrograms, the current implementation is still utilized in the tests to find out how much these aberrations degrade the perceived plausibility of virtual reality rendering.

B. Estimation of parameters for filter modeling of total scattering from a finite object

In this case the total scattered sound from an object is modeled with the filter structure described in Sec. V B. A logarithmic scale is used for the target parameters in training to ensure an even distribution of error. However, in the total object scattering network is trained to estimate L_{max} using a linear scale, since the best accuracy is desired for high values of L_{max}; this is because low-level scattered components will be masked by other sounds in many cases.

There is no self-evident set of feature values that could be utilized to estimate the filter parameters. The geometry of the scenario is presented for the EDToolbox as coordinate values of corners, which were not viewed as a viable approach in this context. A set of features was selected heuristically, which described the main angular and absolute dimensions and orientations of the object, from the viewpoints of the receiver and the source. Additionally, the geometry of the shortest path of sound traveling from the source via the scattering object to the receiver is described.

In this work, the orientation and size of the object is described by analyzing the corner points of the object with principal component analysis (PCA), which results in a three-dimensional vector base $\mathbf{o}_{\mathrm{PCA}}$ revealing the dimensions and the orientation of the object.

Each network was then trained using feature vectors comprised of the following components: (1) distance from source to object-center [m]; (2) distance from object-center to receiver [m]; (3) turning angle of sound path (negative value if source is not visible) [rad]; (4) direct visible [boolean]; (5) reflection visible [boolean]; (6) angular area covered by the object from the viewpoint of the source [rad^2]; (7) angular area covered by the object from the viewpoint of the receiver [rad^2]; (8) incoming angle with respect to the normal of the front plate [rad]; (9) outgoing angle with respect to the normal of the front plate [rad]; (10–12) incoming sound direction dot products with $\mathbf{o}_{\mathrm{PCA}}$ $[\cdot\,\cdot]$; (13–15) outgoing sound direction dot products with $\mathbf{o}_{\mathrm{PCA}}$ $[\cdot\,\cdot]$; (16–18) angular width of each $\mathbf{o}_{\mathrm{PCA}}$ vector from the viewpoint of the receiver; and (19–21) angular width of each $\mathbf{o}_{\mathrm{PCA}}$ vector from the viewpoint of the source.

For training, 78 rectangular plate objects were simulated, where the width and height of the plate varied from 30 to 90 cm, and the depth from 2 to 4 cm. The responses to 50 receiver positions from 20 source positions were computed, where the sources and receivers were in random directions, with distances varying from 40 cm to 3 m. The geometry of the plate used in the dynamic scenario was not included in the simulation. This resulted in 78 000 spectral responses, of which 15% were used as the test data. In training of the network to estimate the cut-off frequency; it was found, that the estimation did not converge for cases where the direct sound was occluded. To overcome this, separate networks were trained for direct-sound-visible and direct-sound-occluded-cases to estimate the cut-off frequency. In 67 000 responses the source was visible, and in the remaining 11 000 responses it was obstructed. Both sets of cases were further divided into training and testing sets with the same

propositions as in other experiments in this work. The trained networks are described in Table I.

The test scenario described in Sec. III was implemented with trained networks and the resulting spectrogram is shown in Fig. 3(c). It can be seen that the networks clearly capture the main features of reflections and occlusion. However, the accuracy is notably lower when compared to the filtering approach. When the direct sound is also taken into account, as shown in Fig. 4(c), the similarity to the ground truth case in Fig. 4(a) is relatively high, although some spectral details are either different or smeared. For example, in the region of occlusion, the interference patterns caused by comb-filtering are different, however, the perceptual prominence of the errors has to be tested with human listeners, as described in Sec. VII.

VII. PERCEPTUAL EVALUATION

The motivation for the experiment was to ascertain the perceptual effects of sound rendering for virtual reality regarding the plausibility when utilizing: (a) simplified rendering of scattered spectra by using the proposed filter structures and (b) neural-network-based estimation of filter parameters from the geometry. This is to be measured for both of the proposed modeling variants: individual-edge-diffraction and total-object-scattering cases.

A. Composition of audio and video

In the test, a video of the situation was shown on a standard computer screen and the audio was played back using headphones (Sennheiser HD-600) in a quiet listening booth. Three different sound samples were used: (1) white noise filtered by a 6 dB/oct lowpass filter with cutoff at 4 kHz; (2) an anechoic recording of a single snare drum shot, repeated; and (3) a male utterance of the English word "base," repeated. The signals, therefore, have a broad spectrum where the spectral effects caused by scattered sound should be audible; however, they still differ in impulsiveness, spectral balance, and tonality.

In the experiment the described test scenario was auralized for the listeners via binaural reproduction, which was segregated into two scenes. In the *Reflection* scene, the receiver moves from azimuth direction 1° to 91°, see Fig. 2, so that the specular reflection fades in and out. In the *Occlusion* scene, the receiver moves from azimuth direction of 91° to 181°, so that the direct sound fades out and then in. In both cases the distance between the listener and the origin was 80 cm, which was also the center point of the frontal plane of the diffracting object. To maximize the audibility of the reflected sound, the head orientation of the listener was kept constant, facing towards the point 40 cm in front of the plane, as shown in Fig. 2. This allowed the direct sound and reflected sound to arrive from different lateral sides of the listener, causing relatively high audibility of the scattered component, since the direct sound is shadowed by the head.

The scattering was simulated with the ES model, using either only first-order diffraction components, or up to 15th-order diffraction components. With first-order components, the sound simulated for each component was auralized to the direction of the apex point of the corresponding edge, resulting in 4–6 virtual sources. In the case of 15th-order diffraction the diffracted spectrum of the scattered sound was not available for individual edges, but only as a spectrum representing all of the scattered sound from the whole plate. For this reason, the sound was rendered to the direction of the apex point of the edge, which corresponds to shortest route of diffracted sound. All the tested methods are presented in Table II.

A set of individual HRTFs that were obtained by scanning a real person and utilizing numeric simulation of the acoustic field around the head (Huttunen *et al.*, 2014) was used, and no head tracking was employed. The dynamic rendition was implemented simply by convolving the source signal with a length of 374 ms with the computed filters, and performing overlap-add synthesis with 16 ms cross-fade between samples computed for successive receiver positions.

B. Perceptual test

The envisioned application of the technique is in interactive virtual reality, where the original sound field is not necessarily defined, and the main target is to provide a believable reproduction to the listener. Also in the case of the test scenario the "original" sound field is not available, and no "original" or "reference" sound item can be delivered to the listener. Instead of this, all of the stimuli were presented in a multiple-stimulus test, where the listener had to grade the *plausibility* of each sound sample presented with the video. The ability to compare between samples enabled a convenient way to report the differences in plausibility. It was assumed that the 15th-order diffraction model without filter fittings would produce the highest plausibility grades and that all other results would be degraded in relation to that. A continuous slider with 100 non-numerally indicated positions was presented to the test subjects. The text "Grade the plausibility of sound" was visible to the listeners, and verbal anchors "Excellent," "Good," "Fair," "Poor," and "Bad" were clearly marked at positions 90, 70, 50, 30, and 10 of the slider, respectively.

TABLE II. Methods utilized in the perception experiment.

Name	Description
Direct	Only direct sound positioned to the direction of the source (DS)
GA	DS + specular broad-band reflection positioned to the direction of image source
ESFul	DS + 15th-order diffraction implemented as a 256-tap FIR filter, positioned towards nearmost apex
ESFFil	as *ESFul*, but implemented with a 6th-order filter structure response fitted to the spectrum
NNfFil	as *ESFFil*, but the parameters of filter structure estimated using neural networks
ES1st	DS + 1st-order diffraction from each visible edge implemented as FIR, positioned towards corresponding the apex points
ES1Fil	as *ES1st*, except the FIRs implemented as 1st-order low-pass IIRs fitted directly to FIRs
NN1Fil	as *ES1Fil*, except the parameters of each IIR estimated using neural networks

The number of subjects was 8, with ages between 20 and 40 years with no reported hearing deficits. The authors did not participate in the test. The listeners were first given written instructions for the test, after which they were presented with a graphical user interface, where they were asked to listen to each sample using a large array of buttons, where each sound sample was randomly assigned to a unique button. After these preparations, the test was conducted, where a multi-stimulus user interface was presented for each combination of sound sample and scenario. The number of multi-stimulus tests was 12, resulting in from three sound samples in two sound scenes with two repetitions. Each multi-stimulus test had eight systems to be tested. The multi-stimulus tasks of all combinations of samples and scenes were represented in a random order to the listeners.

C. Test results

The results were subjected to analysis of variance assuming fixed effects with two-factor interactions over the independent variables *listener*, *method*, *scene*, and *sound*. All the variables and their interactions were found to produce significant effects on the data, although many of them were relatively small. All the interactions were monitored by plotting the data and only the most relevant cases are discussed here.

A two-way analysis of variance yielded a main effect for *method*, $F(7, 767) = 193.35$, $p < 0.0001$. The means and 95% confidence intervals were computed, and are shown in Fig. 7. As was assumed, the *ESFul* method was graded with high plausibility with the value of "good" in the ITU scale. This was expected since it is the most complete and accurate modeling method utilized in this study. An encouraging result is that the method *ESFul* and *ESFFil* obtained very similar scores in the test, with values of 67.4 ± 4.2 and 66.1 ± 4.3, respectively. This supports the assumptions that the deviations in spectrogram of the scattered components observed in Fig. 3(b) relative to Fig. 3(a) do not prominently decrease the degree of plausibility. In turn, the corresponding neural-network-based method *NNfFil* received a somewhat lower score of the value of 56.7 ± 4.5, within the range of the verbal attribute "fair," but still close to the border of "good." This may be interpreted as a satisfying result, as the plausibility is still graded clearly higher than with the physically based methods *Direct*, *GA*, and even *ES1st*.

The 1st-order diffraction rendering method *ES1st* was rated in the middle of the range of "fair" with value of 52.6 ± 4.8, and the corresponding filter-fitted, (*ES1Fil*), and NN-estimated versions, (*NN1Fil*), to notably lower degree of plausibility with the values of 45.6 ± 4.6 and 33.9 ± 3.7, respectively. Based on informal listening and comments of the listeners, it can be assumed that the rapid changes in the rendered spectrum visible in Fig. 4(e) did lower the grade, as in reality all changes in sound in such cases are gradual and smooth.

The second-strongest effect was caused by the factor *scene*, with $F(1, 767) = 75.54$, $p < 0.0001$, with the mean for the *Reflection* scene approximately 10 units higher when compared to the corresponding value for *Occlusion*. The third strongest effect was the factor *listener*, with $F(7, 767) = 69.95$, $p < 0.0001$. When the corresponding mean values of the grades of each listener were monitored, some differences were observed. Evidently the listeners used the scale in individual manner, either slightly emphasizing the positive or the negative end of the scale. The effect caused by *sound* was relatively mild, with $F(2,767) = 10.58$, $p < 0.0001$. The effect was monitored, and it was found that the noise sound was graded about 5 points lower in average than the speech or drum sounds.

There exists a prominent interaction between *method* and *scene*, with $F(7, 767) = 24.19$, $p < 0.0001$, which was an interesting finding in the scope of this article. The interaction is shown in Fig. 7, where it can be seen that many techniques received lower scores for the *Occlusion* scene than for the *Reflection* scene. With the methods *Direct* and *GA*, it is clear that the total vanishing of the direct sound in the *Occlusion* scene degraded the plausibility. In contrary, the ES 1st-order technique, *ES1st*, received higher grades for the *Occlusion* scene than for the *Reflection* scene. It is assumed that the advantage of the method in the *Occlusion* scene is due to the fact that each edge produces an individual edge filter and individual a virtual source, resulting in a smooth spatial transition of sound. The advantage does not exist with filter-versions of 1st-order diffraction methods *ES1Fil* and *NN1Fil*, possibly explained by the modeling issues discussed in Sec. VI A.

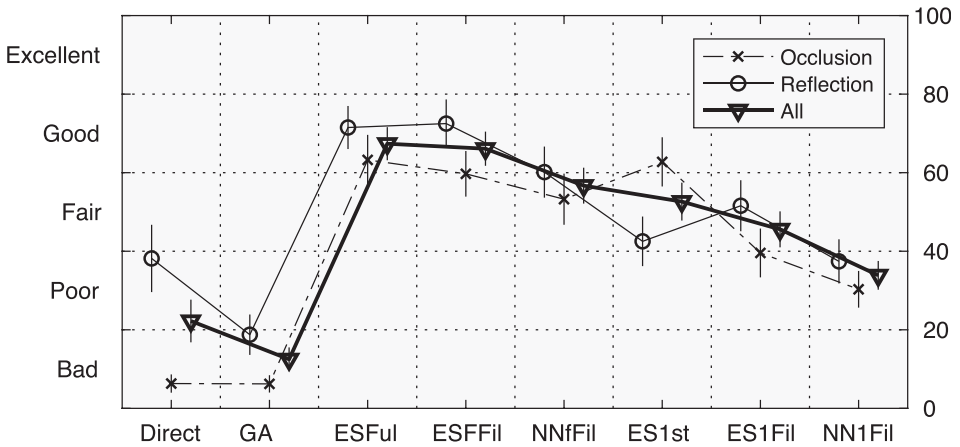

FIG. 7. The plausibility of each system and system*scene interaction evaluated in a multi-stimulus listening test with 8 listeners and 3 sound samples. The plot shows the mean values and the whiskers show the 95% confidence intervals.

A clear feature of the interaction is also the effect of the scene on the *Direct* method, which received grades bordering between "poor" and "fair" for the *Reflection* scene, and it was graded into the "bad" category for the *Occlusion* scene. The difference is largely due to the fact that in the *Occlusion* scene, the sound vanished totally for a period of time, which was clearly a disturbing factor to the listeners. In the *Reflection* scene the rendered audio does not have any detrimental artifacts with the *Direct* method, whereas for the *GA* method, the image source appears and vanishes abruptly. This rapid change resulted in a lower degree of plausibility, when compared with the technique where the reflection was not present at all.

As was anticipated, the plausibility was lower for the individual-edge-filter methods, *ES1Fil* and *NN1Fil*, than for total-object-scattering methods. The *ES1Fil* method received "fair" and *NN1Fil* received "poor" gradings. The informal comments of the listeners revealed that the rapid changes in the spectrogram and unnatural colorations of timbre caused them to grade these methods with lower plausibility values on average.

VIII. DISCUSSION

This article suggests the use of supervised learning for modeling the acoustic scattering from finite objects. The implementation should be viewed as a proof-of-concept, as physical modeling of restricted set of geometries was used as target values in the training of the networks. However, it is the belief of the authors that the method has great potential, as the supervised-learning-based approach enables the ability to render measured acoustic scattering occurring in objects, which cannot be physically modeled in practice.

The usage of neural networks in this task also makes it possible to use different networks for different objects, where the networks have been trained using different geometric descriptors. For example, a certain network could be used for cars, another for animals, and one for pieces of furniture; as it may be assumed, for instance, that the scattered responses from different models of cars are comparatively similar, and a relatively small amount of descriptors is enough to describe their differences for machine learning. Similar descriptor sets could then be formed for other subsets of objects, such as vertebrates, fish, and trees.

Furthermore, this approach is considered an appealing solution for such objects that could be modeled in principle, however, are too computationally demanding with conventional methods for the virtual reality use case. Examples of such cases are objects with round and rough surfaces, acoustically soft materials, and porous objects. For instance, it is intuitively clear that the more absorbing an object is, the lower the scattering level. It is assumed that an appropriately selected porousness index could be associated with a neural network to accommodate this lower level of scattered sound, using acoustic measured data as reference.

IX. CONCLUSION

A machine-learning-based method to estimate and render sound scattered from finite objects in virtual reality is proposed in this work. The method creates the spectral effect of scattering by filtering the sound arriving at the object with a parametric filter structure. Two distinct approaches are proposed, either to render the edge diffraction caused by each individual edge with an independent low-pass filter, or to render the total scattering effect caused by an object using a filter structure, as a combination of low-pass, high-pass, and shelving filters.

The parameters for the filters are estimated using artificial neural networks. The input data in training contain information about the configuration formed by the source, scattering object, and the receiver. The target data for the training are obtained by fitting the response of the filter structure to a spectrum obtained from measurements or from acoustic modeling.

A perceptual experiment, where the scattering from a thick-plate object, at a distance of 80 cm from the listener, was rendered for a dynamic scene using different methods. The results show that (1) the highest plausibility was obtained when the total scattering response was modeled, rather than modeling each edge individually; (2) a relatively simple parametric filter delivered similar plausibility as a detailed acoustic model; (3) the total scattering filter parameters estimated using a neural network degraded the plausibility only slightly from the detailed acoustic model; and (4) the increase in plausibility caused by the rendering of the total scattering is prominent, when compared to the plausibility obtained with the image-source method disregarding diffraction effects.

ACKNOWLEDGMENTS

This research was supported by the Academy of Finland.

Ambrosini, L., Gabrielli, L., Vesperini, F., Squartini, S., and Cattani, L. (**2018**). "Deep neural networks for road surface roughness classification from acoustic signals," in *144th AES Convention*.

Asheim, A., and Svensson, U. P. (**2013**). "An integral equation formulation for the diffraction from convex plates and polyhedra," J. Acoust. Soc. Am. **133**(6), 3681–3691.

Biot, M. A., and Tolstoy, I. (**1957**). "Formulation of wave propagation in infinite media by normal coordinates with an application to diffraction," J. Acoust. Soc. Am. **29**(3), 381–391.

Cakir, E., Parascandolo, G., Heittola, T., Huttunen, H., Virtanen, T., Cakir, E., Parascandolo, G., Heittola, T., Huttunen, H., and Virtanen, T. (**2017**). "Convolutional recurrent neural networks for polyphonic sound event detection," IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) **25**(6), 1291–1303.

Fahy, F. J. (**2000**). *Foundations of Engineering Acoustics* (Elsevier, Amsterdam).

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (**2016**). *Deep Learning* (MIT Press, Cambridge), Vol. 1.

Grais, E. M., and Plumbley, M. D. (**2018**). "Combining fully convolutional and recurrent neural networks for single channel audio source separation," in *144th AES Convention*.

Haykin, S. (**1994**). *Neural Networks: A Comprehensive Foundation* (Prentice Hall, Upper Saddle River, NJ, USA).

Hewett, D. P., and Morris, A. (**2015**). "Diffraction by a right-angled impedance wedge: An edge source formulation," J. Acoust. Soc. Am. **137**(2), 633–639.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (**2012**). "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Sign. Process. Mag. **29**(6), 82–97.

Huttunen, T., Vanne, A., Harder, S., Paulsen, R. R., King, S., Perry-Smith, L., and Kärkkäinen, L. (**2014**). "Rapid generation of personalized HRTFs," in *Proceedings of the AES 55th Conference: Spatial Audio*.
Kohonen, T. (**1988**). "The 'neural' phonetic typewriter," Computer **21**(3), 11–22.
Kon, H., and Koike, H. (**2018**). "Deep neural networks for cross-modal estimations of acoustic reverberation characteristics from two-dimensional images," in *144th AES Convention*.
Kuttruff, H. (**2017**). *Room Acoustics*, 6th ed. (Taylor & Francis, London).
Lokki, T., Svensson, P., and Savioja, L. (**2002**). "An efficient auralization of edge diffraction," in *Proceedings of AES 21st Conference: Architectural Acoustics and Sound Reinforcement*.
Marsh, I., and Brown, C. (**2009**). "Neural network classification of multibeam backscatter and bathymetry data from Stanton Bank (Area IV)," Applied Acoustics **70**(10), 1269–1276.
Martin, S. R., Svensson, U. P., Slechta, J., and Smith, J. O. (**2018**). "Modeling sound scattering using a combination of the edge source integral equation and the boundary element method," J. Acoust. Soc. Am. **144**(1), 131–141.
Murphy, D. T., and Beeson, M. J. (**2003**). "Modelling spatial sound occlusion and diffraction effects using the digital waveguide mesh," in *Proceedings of AES 24th Conference: Multichannel Audio, The New Reality*.
Pellegrini, R. S. (**2001**). "Quality assessment of auditory virtual environments," *Proceedings of 2001 International Conference on Auditory Display (ICAD)*.
Pulkki, V., and Karjalainen, M. (**2015**). *Communication Acoustics: An Introduction to Speech, Audio and Psychoacoustics* (Wiley, London).
Pulkki, V., Lokki, T., and Savioja, L. (**2002**). "Implementation and visualization of edge diffraction with image-source method," in *Proceedings of 112th AES Convention*.
Raghuvanshi, N., and Snyder, J. (**2014**). "Parametric wave field coding for precomputed sound propagation," ACM Trans. Graphics **33**(4), 38.
Savioja, L., Huopaniemi, J., Lokki, T., and Väänänen, R. (**1999**). "Creating interactive virtual acoustic environments," J. Audio Eng. Soc. **47**(9), 675–705.
Savioja, L., and Svensson, U. P. (**2015**). "Overview of geometrical room acoustic modeling techniques," J. Acoust. Soc. Am. **138**(2), 708–730.
Schissler, C., Loftin, C., and Manocha, D. (**2018**). "Acoustic classification and optimization for multi-modal rendering of real-world scenes," IEEE Transactions Visual. Comput. Graphics **24**(3), 1246–1259.
Schissler, C., Mehra, R., and Manocha, D. (**2014**). "High-order diffraction and diffuse reflections for interactive sound propagation in large environments," ACM Trans. Graphics **33**(4), 39.
Sherman, W., and Craig, A. (**2003**). *Understanding Virtual Reality: Interface, Application, and Design* (Morgan Kaufmann, San Francisco).
Svensson, P. (**2000**). "Edge diffraction Matlab toolbox (EDtoolbox)," https://github.com/upsvensson/Edge-diffraction-Matlab-toolbox (Last viewed April 20, 2018).
Svensson, U. P., Fred, R. I., and Vanderkooy, J. (**1999**). "An analytic secondary source model of edge diffraction impulse responses," J. Acoust. Soc. Am. **106**(5), 2331–2344.
Tsingos, N., Funkhouser, T., Ngan, A., and Carlbom, I. (**2001**). "Modeling acoustics in virtual environments using the uniform theory of diffraction," in *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, *SIGGRAPH'01* (ACM, New York), pp. 545–552.
Tsingos, N., and Gascuel, J.-D. (**1998**). "Fast rendering of sound occlusion and diffraction effects for virtual acoustic environments," in *Proceedings of 104th AES Convention*.
Tsingos, N., Jiang, W., and Williams, I. (**2011**). "Using programmable graphics hardware for acoustics and audio rendering," J. Audio Eng. Soc. **59**(9), 628–646.
Välimäki, V., and Reiss, J. D. (**2016**). "All about audio equalization: Solutions and frontiers," Appl. Sci. **6**(5), 129.
Vorländer, M. (**2007**). *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality* (Springer, Berlin).
Vorländer, M. (**2013**). "Computer simulations in room acoustics: Concepts and uncertainties," J. Acoust. Soc. Am. **133**(3), 1203–1213.
Watanabe, S., and Yoneyama, M. (**1992**). "An ultrasonic visual sensor for three-dimensional object recognition using neural networks," IEEE Trans. Robotics Automat. **8**(2), 240–249.