To appear in the proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR) 2025

Sonify Anything: Towards Context-Aware Sonic Interactions in AR

Laura Schütz* Sasan Matinfar Ulrich Eck Daniel Roth Nassir Navab

Technical University of Munich

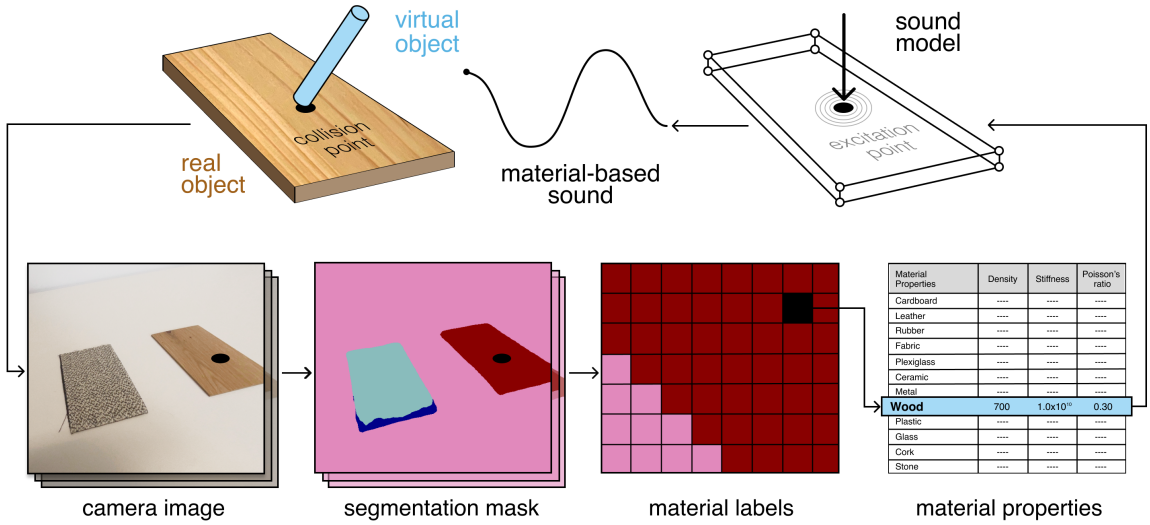

Figure 1: Material-based sonification approach for real-virtual object interactions in Augmented Reality

ABSTRACT

In Augmented Reality (AR), virtual objects interact with real objects. However, the lack of physicality of virtual objects leads to the absence of natural sonic interactions. When virtual and real objects collide, either no sound or a generic sound is played. Both lead to an incongruent multisensory experience, reducing interaction and object realism. Unlike in Virtual Reality (VR) and games, where predefined scenes and interactions allow for the playback of pre-recorded sound samples, AR requires real-time sound synthesis that dynamically adapts to novel contexts and objects to provide audio-visual congruence during interaction. To enhance real-virtual object interactions in AR, we propose a framework for context-aware sounds using methods from computer vision to recognize and segment the materials of real objects. The material's physical properties and the impact dynamics of the interaction are used to generate material-based sounds in real-time using physical modelling synthesis. In a user study with 24 participants, we compared our congruent material-based sounds to a generic sound effect, mirroring the current standard of non-context-aware sounds in AR applications. The results showed that material-based sounds led to significantly more realistic sonic interactions. Material-based sounds also enabled participants to distinguish visually similar materials with significantly greater accuracy and confidence. These findings show that context-aware, material-based sonic interactions in AR foster a stronger sense of realism and enhance our perception of real-world surroundings.

*e-mail: laura.schuetz@tum.de

Index Terms: Audio feedback, Sonic interaction, Audiovisual interaction, Augmented reality, AR, XR, Multisensory perception, Multisensory congruence, Semantic congruence, Material sounds, Context-aware, User interface, Human-computer interaction

1 INTRODUCTION

Our interactions with physical objects naturally create sound. Sound occurs when vibrations set a medium like air into motion. The physical properties of a medium directly shape the sound it produces. For instance, placing a cup on a table, typing on a keyboard, or knocking on a wooden door all make distinct sounds. Therefore, sounds contain information about the source and the event that produced it, enabling people to identify materials from sounds. Audition helps us pick up these sounds and understand what's happening in our environment. The idea of ecological acoustics explains how we use sound to make sense of the world [29]. It explores how sounds carry information about objects, spaces, and events, and how our auditory system interprets them to navigate the world.

Materials are mainly identified in impact sounds via frequency and damping parameters of the sound [21, 27]. However, since frequency changes can also be attributed to changes in object geometry, damping cues have been shown to be more reliable for material identification from hitting actions [28]. It has furthermore been shown that recovering materials from action sounds is harder than identifying the sound producing action, like scraping or rolling [24].

Augmented Reality (AR) applications aim to seamlessly blend virtual and physical elements. However, when interacting with virtual objects, we are robbed of these ecological sounds. The relationship between the sound and the sound source is lost when we use no sound or incongruent sound effects to sonify virtual object interactions, resulting in incongruent multisensory perception. This discrepancy reduces the interaction realism and weakens the user's

sense of presence.

Studies have demonstrated that audiovisual congruence increases presence and realism in Virtual Reality (VR) experiences [19]. Semantic congruence specifically has been shown to improve recognition speed and attentional control, making it a key factor in enhancing interaction realism [23, 6]. Unlike in VR and games, where objects and environments are predefined, and sound samples can be retrieved from a sound library when a sound-producing event occurs, AR operates in constantly changing real-world environments. As a result, pre-recorded audio samples are insufficient to create realistic sonic interactions. There is a need for real-time audio synthesis that dynamically considers the materiality of real-world objects during interactions in AR.

This work aims to bring audiovisual congruence to AR interactions and thereby enhance the perceptual link between users and their physical environment. We introduce a material-based sonification technique that generates dynamic, congruent impact sounds for real-virtual object interactions in AR, using object material properties and impact dynamics to produce accurate audio feedback. Leveraging existing machine learning methods for real-time material segmentation, we extract material information from unseen environments. By combining these techniques with physics-based sound synthesis, we enable real-time, material-driven audio feedback without relying on pre-recorded audio clips or large training datasets. What we do not contribute is a technique for generating highly realistic material sounds, nor a comprehensive study on the perceptual effects of audiovisual congruence in AR. Instead, our work makes the following contributions:

1. We propose a framework for context-aware sounds in AR using material segmentation and physical modelling sound synthesis.

2. We demonstrate a material-based sonification approach that creates physics-based sounds in real-time.

3. We report results from a user study showing that the proposed material-based sounds improve perceived interaction realism over standard generic sounds.

2 Related Work

2.1 Material Segmentation

Material segmentation is a field of study in computer vision that has seen significant advancements in recent years. It is concerned with the task of assigning material labels (e.g., wood, metal) to each pixel in an image. Material segmentation is, for example, used in robotics, for effective decision-making and object interaction [36], or in autonomous driving, where material cues help improve terrain understanding and safety decisions [3]. The Materials in Context Database (MINC) is a large dataset that facilitates deep learning approaches for material recognition [2]. Another recent contribution is the Dense Material Segmentation dataset, which provides 3.2 million dense material annotations for a diverse set of indoor and outdoor scenes, objects, viewpoints, and materials [48]. This makes it especially suitable for augmented reality use cases, where detailed material recognition from diverse viewing angles is required for realistic interactions.

2.2 Sound Synthesis Techniques

To recreate realistic action sounds, we can make use of a variety of sound synthesis methods. Three prominent sound synthesis methods for action sounds are:

Sampling-Based Sound Synthesis: Widely used in games and virtual reality, this method is an easy way to create action sounds from pre-recorded samples [26]. The samples are associated with events or locations in the scene and played back when the interaction occurs. To simulate continuous sounds, periodic elements of the waveform are often looped. Filters and envelopes can be applied to diversify the sound output from a given set of samples [9]. Although sample-based synthesis is easy to implement and computationally efficient, it is limited by the prerecorded sound clips available in a database. As a result, it lacks the flexibility to respond to unexpected changes in the environment, making it less suitable for context-sensitive applications such as Augmented Reality.

Data-Driven Sound Synthesis: This approach to sound synthesis uses various data analysis techniques, including statistical methods and machine learning, to infer action sounds. Physics-driven machine learning approaches, such as physics-informed diffusion models, have been developed to synthesize impact sounds from videos [46]. Identifying visual representations of sound-producing actions can be learned from egocentric videos [4] and used to generate action-matching sounds [5]. Large language models (LLMs) have been employed to query for foley sound effects matching the content of a video clip. These sound samples are later adjusted to match the motion dynamics in the video clip [25, 7]. While data-driven approaches can create realistic action sounds from video data, they are constrained by dataset limitations and inference latency.

Physical Modelling Sound Synthesis: Given that material sounds are closely tied to the physicality of the sound-producing objects and actions, model-based synthesis is a promising approach to creating material-based action sounds. This method can create highly realistic impact sounds in real-time, but is more computationally intensive than sample-based sonification. However, several techniques for accelerating physically based sound simulation have been shown to reduce the computational cost, enabling simultaneous simulation of numerous sound models [35], without relying on prior training, as is necessary for data-driven techniques.

The construction of sound models is commonly achieved through physical modelling synthesis [10]. This technique aims to mimic the physical characteristics of real-world instruments, effectively emulating the behavior of actual objects. Several simulation software have been proposed for modeling sounds dynamically based on object properties. Early work in modal synthesis, such as Mosaic [31] and Modalys [11], demonstrated its potential for realistic audio generation. Modalys, in particular, uses the finite element method to perform physical modelling synthesis. By solving differential equations associated with vibrating systems, it can represent key dynamic characteristics such as natural frequencies, damping behavior, and mode shapes, relevant to the creation of realistic material sounds. By precomputing the object's modes of frequencies, this approach enables efficient synthesis, ideal for realtime interactions.

Van den Doel et al. [49] introduced a method for real-time synthesis of contact sounds, such as impact, rolling, and friction, for solid materials using modal synthesis. Later work on modal synthesis for interactive sounds investigated the inclusion of surface information at three levels of resolution (object shape, visible surface bumpiness, microscopic roughness) for synthesizing complex contact sounds in virtual environments [37]. More recently, learning-based methods have been proposed for real-time modal impact sound synthesis [17].

However, a key challenge in using modal techniques is the lack of automatic determination of satisfactory material parameters that recreate realistic audio of sound-producing materials [43]. In AR, this problem is exacerbated. Unlike virtual environments where predefined 3D objects are assigned parameters that will lead to satisfactory audio output, we deal with unknown physical objects, interactions, and collision points in AR. Therefore, we believe that a real-time understanding of the context, objects, and interactions is needed to obtain the required information to create context-aware sonic interactions.

While the works referenced in this section from the fields of

computer vision [46, 5, 4, 7], computer graphics [9, 35, 49], and computer music [10, 31, 11] focus on methods for generating highly-realistic synthesized action sounds, our work, proposing a novel material-based sonic interaction technique, leverages an established sound synthesis technique that has been proven to create high-quality impact sounds [11].

2.3 Audio Interactions in XR

Virtual Reality: Numerous studies have explored the simulation of realistic sonic interactions in virtual environments. Serafin et al. [43] stated that an immersive sonic experience relies on action sounds, binaural rendering, environmental sounds, and sound propagation. Since this study focuses on action sounds - sounds produced by the listener and changing with movement [13]) - we will highlight a few works on action sounds in this section.

Besides object-related contact sounds, the simulation of footstep sounds in VR has also been studied. Nordahl et al. [33] presented an algorithm for simulating walking sounds on solid and aggregated surfaces. In addition, the use of velocity-based variations in walking sounds has been proposed for simulated sneaking in VR [8]. Schütz et al. [42] introduced a multisensory interaction framework for audiovisual interaction with anatomical structures using a physically based sonification approach. Their framework, along with the other physics-based approaches for impact sounds outlined in Sec. 2.2 [49, 37, 17] are effective for fully virtual environments. However, they would require high-resolution 3D reconstructions of physical objects in the scene and real-time computation of the objects' natural frequencies to be feasible for application in AR, limiting their practicality. All the above studies were conducted in entirely simulated or virtual environments. In contrast, only a few works have investigated context-based sounds in AR.

Augmented Reality: To create context-aware sounds in AR, we require knowledge about the real-world environment. This can be achieved by using sensing technology and analyzing the sensor or camera data. Wilson et al. [52] were among the first to propose a system that provides relevant audio feedback about the environment. Using GPS and head orientation to estimate the user's pose, the system dynamically generates spatialized, non-speech audio cues that provide blind or visually impaired users with navigation and inform them about nearby features (e.g., benches or stairs). Medical augmented reality systems using electromagnetic or visual tracking demonstrated precise localization of points of interest within the patient body using parameter mapping sonification [40, 41]. A study by Su et al. [47] introduced a system that generates context-aware sound effects for AR by analyzing the semantics of the virtual augmentations and real-world context. An LLM processes this information to acquire suitable audio through sample retrieval, text-to-sound generation, or text-based sound style transfer. While their approach targets the curation of a set of sound effects that people can choose from to sonify animated AR content, not requiring real-time sound synthesis, our work focuses on generating real-time impact sounds resulting from human-object interactions.

To the best of our knowledge, this is the first work to introduce a context-aware framework for material-based impact sounds in AR. We sonify real-virtual object interactions using real-time material segmentation of physical objects, enabling physically-based sound synthesis in AR. Using this system, semantically and temporally congruent interaction sounds can be generated for unseen environments in real-time.

2.4 Multisensory Congruence in XR

Laurienti et al. [23] report that semantically congruent audiovisual stimuli enhance recognition speed and accuracy in perception tasks. Chen and Spence [6] showed that semantically congruent audiovisual stimuli enhance, whereas semantically incongruent audiovisual stimuli impair, object identification performance. They further highlight the role of temporal congruence in audiovisual perception, showing that users can tolerate slight delays of up to 300 ms between audio and visual stimuli, but that excessive desynchronization disrupts the formation of a coherent multisensory percept. Besides identification accuracy and speed, matching crossmodal stimuli have been shown to enhance the ability to select and hold attention on an object when multiple sensory stimuli compete for attention [51]. Additionally, a study by Fujisaki et al. [12] on audiovisual integration in material perception revealed that sound accuracy significantly affects the perceived material properties of objects.

In VR environments specifically, establishing multisensory congruence has been shown to improve the user's attention and sense of presence [43]. A study by Kim et al. [20] demonstrated that audiovisual congruence significantly improved users' sense of presence, realism, and emotional engagement, while incongruence reduced immersion and increased perceived effort during interaction. Two studies on gait-aware auditory feedback showed that congruent audio feedback enhances presence and immersion in virtual environments [15, 16]. These findings underscore the importance of ensuring multisensory congruence in XR environments, as congruent stimuli enhance realism and presence.

Many of the works from cognitive psychology and neuroscience cited in this section [23, 6, 12, 51] focus on evaluating the effects of varying degrees of multisensory congruence on user perception in controlled psychological experiments. In contrast, our work compares two audiovisual interaction techniques in AR to demonstrate that congruent, material-based sounds can enhance the perceived realism of AR interactions compared to generic sounds.

3 Methods

We propose a material-based approach to sonic interactions in AR using material segmentation to inform the physical modelling sound synthesis. Fig. 1 depicts an overview of the system components. Each component is described in more detail below.

3.1 Scene Understanding

To obtain material information about the objects in the environment, we stream camera images from the left camera of the VisionPro via a WebSocket to a Python script running on a MacBookPro (M1 Max). Camera frames were sent every 200 milliseconds. The Python script runs a pre-trained material segmentation model[1], presented in a paper by Upchurch and Niu [48] on the RGB camera images. The frames are originally captured at 1920×1080 resolution and downscaled to 960×540 to enable faster inference. The resulting segmentation mask - an image with color-coded material labels - is then sent back to the Swift script running on the Vision Pro. Lower image resolutions such as 512×512 or 256×256 significantly degraded classification accuracy. Through empirical testing, 960×540 proved to be the most optimal trade-off between processing speed and segmentation quality for our application.

3.2 Object Interaction

The AR application was developed using Unity[2] (v 6000.0.27f1), with Unity PolySpatial (v 2.1.2) supporting deployment on visionOS (v 2.0). Hand interactions were implemented via the XR Interaction Toolkit (v 3.0.7). To receive segmentation masks within Unity, a dedicated package for Apple Vision Pro camera access[3] was integrated. The package facilitates communication between Swift and Unity through a callback mechanism. The callback function is implemented in C++ (Unity) and invoked in Swift to acquire the segmentation mask images from the Vision Pro. The segmentation masks were received and stored as Texture2D objects. The

[1] https://github.com/apple/ml-dms-dataset

[2] https://unity.com

[3] https://github.com/styly-dev/EnterpriseCameraAccessPlugin

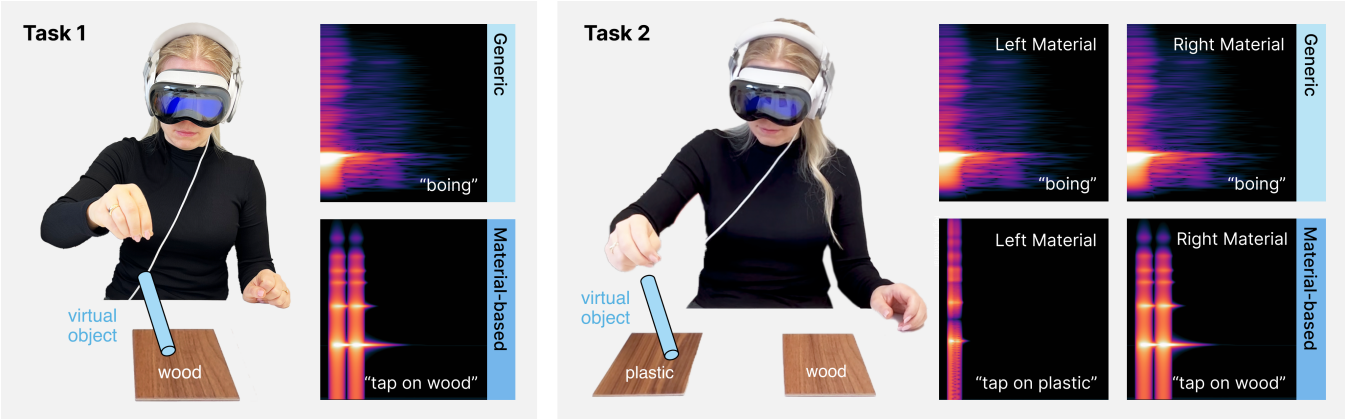

Figure 2: Task 1 (left) and Task 2 (right) of the user study. Task 1 - Participants had to identify one material at a time using the material-based and generic sound condition. Task 2 - Participants had to distinguish two visually similar materials in the material-based and generic sound condition. The same sound effect was used for every material in the generic condition. Individual physical modelling-based sounds were generated for every material in the material-based condition.

ARFoundation (v 6.0.5) AR Plane Manager was used for plane detection. When a virtual object collided with a real-world surface (AR plane), the world-space collision point was transformed into the image plane of the previously recorded segmentation masks using the headset camera intrinsics and the corresponding camera-to-world transformation matrix. For each segmentation mask in the history buffer of up to 5 images, the 3×3 neighborhood around the projected UV coordinate of the collision was sampled to retrieve the most frequent RGB pixel color values. These values were then used to retrieve the associated material class from a dictionary mapping RGB values to material names. A majority vote across all segmentation masks determined the most likely material at the collision point to provide temporal smoothing and robustness against noisy classification results.

3.3 Material-based Sonification

The collision material name is sent from Unity to Max/MSP[4], a software for audio programming, using Open Sound Control [53], a network protocol for interactive computer music. Modalys for Max[5], a physical modelling sound synthesis software, is used to sonify the object interactions. Depending on the object shape, sound models can be created. However, in our study we only focused on evaluating material samples in the shape of plates. Therefore, we used a 3D rectangular plate model in Modalys to represent the geometry of the real-world material samples. The physical material properties of density (kg/m3), stiffness (Young modulus, kPa), and Poisson's ratio were used to parameterize the plate model (see Tab. 1)). Additional parameters included plate thickness and damping coefficients such as constant loss and frequency loss. Modalys numerically solves a set of differential equations that control the dynamics of these modes over time, simulating how the plate responds to the excitation based on its physical parameters. When a collision occurs in Unity, force is applied to excite the virtual plate object in Modalys. All kinds of interactions, like bowing or striking, can be realized within Modalys. To isolate the perception of material properties, we used a one-dimensional force connector, a point-mass excitation at a single location on the rectangular plate. In physical modeling synthesis, complex excitations can introduce nonlinearities that mask the acoustic characteristics of the simulated material. By simplifying the excitation, we reduced its impact on the resulting sound, ensuring that sound differences originated mostly from material properties rather than excitation artifacts. When an excitation occurs, energy is introduced into the model, triggering vibrations across the plate's resonant modes. Modalys picks up the resulting motion at defined listening points, where the sum of the active modes is converted into an audio signal in Max/MSP.

[4]https://cycling74.com/products/max

[5]https://support.ircam.fr/docs/Modalys/current/

4 Study

To determine whether audiovisual congruence established using our material-based sonification approach can enhance interaction realism in AR, we performed a within-subject study comparing our context-ware, material-based sounds to a standard generic sound to investigate the following hypotheses:

- ***H1*** Material-based sounds enhance material identification accuracy (Task 1)
- ***H2*** Material-based sounds enhance material identification confidence (Task 1 & Task 2)
- ***H3*** Material-based sounds enhance sonic interaction realism (Task 1)
- ***H4*** Material-based sounds facilitate distinguishability of visually similar materials (Task 2)

In the AR application, participants were given a virtual stick, which they were instructed to grab with a pinch gesture and use to tap on real, physical material samples. Participants were told that they could imagine the stick as a stiff object similar to a plastic pen. The material of the virtual stick was purposefully left undefined as our approach aimed to exclusively explore the simulation of impact sounds on real-world objects. Participants had to perform two tasks (Fig. 2). Both tasks included two audio conditions: material-based and generic.

4.1 Task 1

In Task 1, each condition included 10 trials, one for every material. Participants were presented with one material at a time and asked to tap on the material sample using the virtual stick to create the impact sounds. Based on the visual appearance of the material and the tapping sound, they were asked to answer three questions related to the material and its properties and one question about the realism of the sonic interaction. After either condition, the participants completed a post-condition questionnaire, which included questions related to their material perception confidence and the helpfulness of the sound (see Sec. 4.1.3 for all questions in Task 1).

Table 1: Physical properties of the materials used in Task 1 and Task 2 based on research in mechanics and material sciences [30, 34, 44, 32, 22]: Density (kg/m3), Stiffness (Young modulus (N/m2)), Poisson's ratio

Material	Density (kg/m3)	Stiffness (N/m2)	Poisson's ratio
Cardboard	689	5.0×10^8	0.33
Ceramic	2600	2.0×10^{11}	0.25
Cork	240	1.0×10^8	0.30
Fabric	1500	1.0×10^6	0.30
Glass	2500	7.2×10^{10}	0.20
Leather	860	1.0×10^8	0.40
Metal	7800	2.0×10^{11}	0.30
Paper	800	5.0×10^8	0.33
Plastic	1100	2.5×10^9	0.35
Rubber	1100	1.0×10^7	0.50
Stone	2700	5.0×10^{10}	0.25
Wood	700	1.0×10^{10}	0.30

4.1.1 Stimulus

Ten realistic indoor surface materials covering a wide range of physical properties, from elastic to stiff, from airy to dense, were included in Task 1: Cardboard, Ceramic, Cork, Fabric, Glass, Leather, Metal, Plastic, Stone, Wood (see Tab. 1 for their physical properties and Fig. 3 for images and sounds). The samples were sized 10x10, 11x15, or 22x22. The thickness of the plates ranged from 0.3 to 1.0 cm.

4.1.2 Independent Variables

Sound The task included two audio conditions: material-based and generic. The material-based condition used the proposed material-based sonification approach to create congruent material interaction sounds for all twelve materials used in the study. In the generic condition, the same audio sample (”Button Pop” from the XR Interaction Toolkit (v 3.0.7)) was played for all materials, emulating the current standard in AR where identical sounds are used for interactions between virtual and physical objects regardless of their materiality.

4.1.3 Dependent Variables

Material & Properties Participants answered the following questions after every trial inside the AR application: ”Which material is it?”. The names of the ten materials included in Task 1 (Sec. 4.1.1) constituted the answer options. In addition, we wanted to assess their perception of the material's physical properties. We asked them to rate the density of the material - ”How dense is the material?” - on a continuous scale from ”As airy as milk foam” (0) to ”As dense as gold” (100). They also rated the stiffness - ”How stiff is the material?” - from ”As elastic as rubber band” (0) to ”As stiff as diamond” (100). These questions inside the AR application used visually uniform sliders without intermediate anchors, allowing participants to select any numeric value along a continuum.

Confidence & Helpfulness In a post-condition, desktop-based questionnaire, participants responded to three 7-point Likert scale questions on a scale from strongly disagree (1) to strongly agree (7) to assess their subjective confidence in their material and material properties answers for all trials in the condition. The items were: ”I was confident in my material assignments.”, ”I was confident in my density estimations.”, and ”I was confident in my stiffness estimations.”. They furthermore answered the 7-point Likert scale question, ”The audio feedback was helpful for classifying the materials.” using the same scale.

Sound Realism To assess the sonic interaction realism, participants responded to the question ”How realistic was the sonic interaction?” on a continuous 0-100 slider from not realistic at all (0) to absolutely realistic (100) after every material inside the AR application.

4.2 Task 2

In Task 2, participants were given two visually similar materials at the same time in each trial. The participants were again tasked to tap on the materials. Based on the visual and auditory cues, they had to identify the materials, rate their confidence in the assignment, and rate the helpfulness of the sound in distinguishing the materials.

4.2.1 Stimulus

We purposefully selected six pairs of visually similar materials, which, however, differed in their actual materiality. In addition, we selected pairs to form two groups:

Ambiguous Material pairs that are visually and audibly similar: wood & wood-printed plastic, glossy paper & glossy plastic, stone & stone-printed plastic.

Unambiguous Material pairs that are visually similar, but audibly different: glass & plexiglass, rubber & milky glass, coated ceramic & coated wood.

4.2.2 Independent Variables

Sound Task 2 again featured the two audio conditions: material-based and generic. In the material-based condition, the material segmentation model was supplemented with marker tracking to ensure congruent sounds, as visually indistinguishable materials cannot be reliably differentiated by current vision-based models. Nevertheless, we included Task 2 to investigate whether material-based sonification could aid users in disambiguating visually similar materials. To ensure consistency, we reused the same material-based sonification model and physical parameters as in Task 1 for generating congruent sounds. In the generic condition, the same audio sample as in Task 1 was applied to all materials.

Audiovisual Ambiguity As described in Sec. 4.2.1, we included two groups of material pairs, a group of visually and auditorily ambiguous materials and a group of visually ambiguous but auditorily distinct materials.

4.2.3 Dependent Variables

Material For each material pair, participants were asked to identify which material corresponded to which sample. They were not allowed to choose the same material for both samples. The items were called: ”Which material is on the left?” and ”Which material is on the right?”. The answer options were, for example, ”Glass” and ”Plexiglass”.

Confidence & Helpfulness Participants rated their confidence in the material assignments and the helpfulness of the sounds on a continuous slider (0-100) from Not confident/helpful at all (0) to Very confident/helpful (100) after every pair inside the AR application. The items were called ”How confident are you in your material assignments?” and ”How helpful was the sound to distinguish the two materials?”.

4.3 Participants

24 participants (12 women, 12 men), with a mean age of 29.54 years (SD = 3.91), took part in the study. Most were PhD or master's students from the fields of biomedical engineering or computer science, with the rest consisting of professionals from very diverse subject areas ranging from business to literature studies and law. The subjects' music experience varied widely. One participant

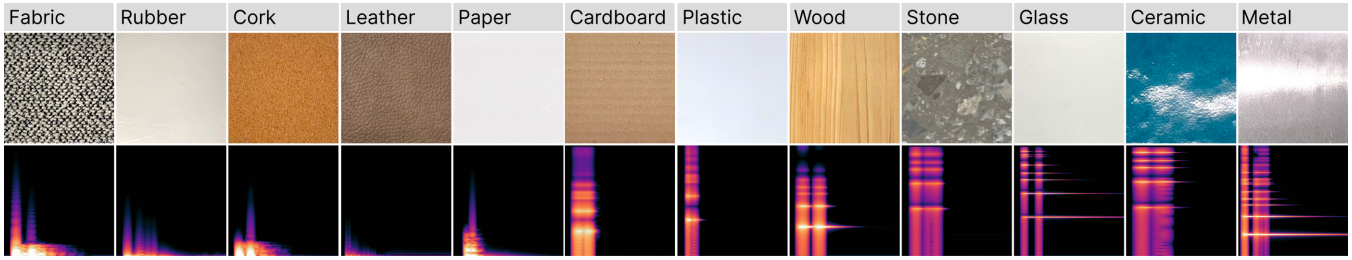

Figure 3: Materials used in Task 1 and Task 2 of the user study along with spectrograms of the material-based sounds generated using physical modelling synthesis. Materials are sorted by increasing stiffness from left to right.

reported being a trained musician, four participants played instruments regularly, twelve of the participants knew how to play one or more instruments but rarely played, and seven participants reported not having any musical training. Experience with using AR or VR was similarly mixed. Fifteen subjects had used it either once or a few times before, eight reported regular or even daily use, and one had never used it before. Gaming habits were diverse as well, with seven people playing daily or weekly, but the rest reported to play very infrequently or never.

4.4 Procedure

Participants were seated in a quiet room at a well-lit table. After providing informed consent and completing a demographics survey, they were equipped with the AR headset, and eye calibration was performed. The material samples were placed out of sight of the participants until retrieved one at a time only for the duration of the trial before being removed from sight again. The audio was delivered to the participants via over-ear headphones. Each task began with a training scene to familiarize participants with the procedure and AR environment. In Task 1, participants completed 10 trials in the first condition followed by a post-condition questionnaire, then proceeded to the second condition and again the questionnaire. In Task 2, they again started with a training scene followed by both conditions of each 3 trials. The order of conditions in both tasks was counterbalanced using Latin-square randomization. The study concluded with a post-study questionnaire, where participants provided qualitative feedback.

5 Results

5.1 Task 1

A Shapiro-Wilk test showed a non-normal distribution of the data for all measures. Outliers in the time data were assessed using the interquartile range method, resulting in the removal of 5% of the data points identified as outliers. No outliers were found for the other measures. Wilcoxon signed-rank tests showed significant differences ($p < 0.001$) between the two sound conditions for all variables except the material recognition accuracy (see Fig. 4 and Tab. 2).

5.1.1 Material-based vs. Generic Sounds

There was a significant main effect of condition ($p<0.05$) on task time. Participants took significantly longer in the material-based condition (46.83 ± 16.02 seconds) than the generic condition (41.44 ± 16.01 seconds). The time per trial did not significantly vary based on which material they were viewing ($p=0.4686$). There was also a significant main effect of condition ($p<0.05$) on the density and stiffness estimations. Materials were rated significantly more dense and more stiff in the material-based than in the generic condition. The sonic interactions were rated significantly more realistic for the material-based condition ($p<0.001$). A per-material comparison of the realism rating for the materials used in the study can be found in Fig. 5.

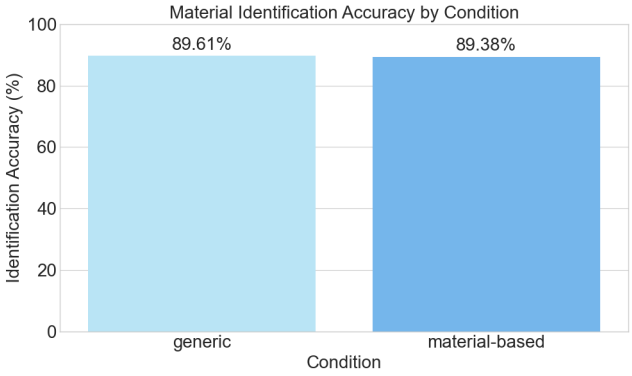

Figure 4: Material identification accuracy by condition in Task 1

Participants furthermore rated the material-based sounds (Mdn=5.5, MAD=0.5) much more helpful than the generic sound (Mdn=1.0, MAD=0.0) for material identification ($p<0.001$). There was no effect of condition on the subjects' material identification confidence. However, there was a significant effect of condition on material properties estimation confidence (density: $p<0.05$, stiffness: $p<0.01$) (see Fig. 6).

Table 2: Results Task 1: material identification accuracy (percent); means, standard deviations and p-values of density (0-100), stiffness (0-100), time per trial (seconds), sound realism (0-100); medians, median absolute deviations, p-values of the 7-point Likert scales on sound helpfulness, material confidence, density confidence, stiffness confidence

Task 1	Generic	Material-based	P-value
Material Accuracy	89.61%	89.38%	> 0.05
Density	54 ± 24	59 ± 23	$<$ **0.001**
Stiffness	52 ± 28	58 ± 27	$<$ **0.001**
Task Time	41.44 ± 16.01	46.83 ± 16.02	$<$ **0.001**
Sound Realism	15 ± 19	66 ± 21	$<$ **0.001**
Sound Helpfulness	1.0 (0.0)	5.5 (0.5)	$<$ **0.001**
Material Confidence	5.0 (1.0)	5.0 (1.0)	> 0.05
Density Confidence	4.0 (1.0)	5.0 (1.0)	$<$ **0.05**
Stiffness Confidence	5.0 (1.0)	5.0 (0.0)	$<$ **0.01**

5.1.2 Hard vs. Soft Materials

We further divided the materials into hard (Ceramic, Glass, Metal, Plastic, Stone, Wood) and soft (Leather, Cardboard, Cork, Fabric). The sonic interaction was perceived to be significantly more realistic for hard materials (73 ± 14) than soft materials (55 ± 17)

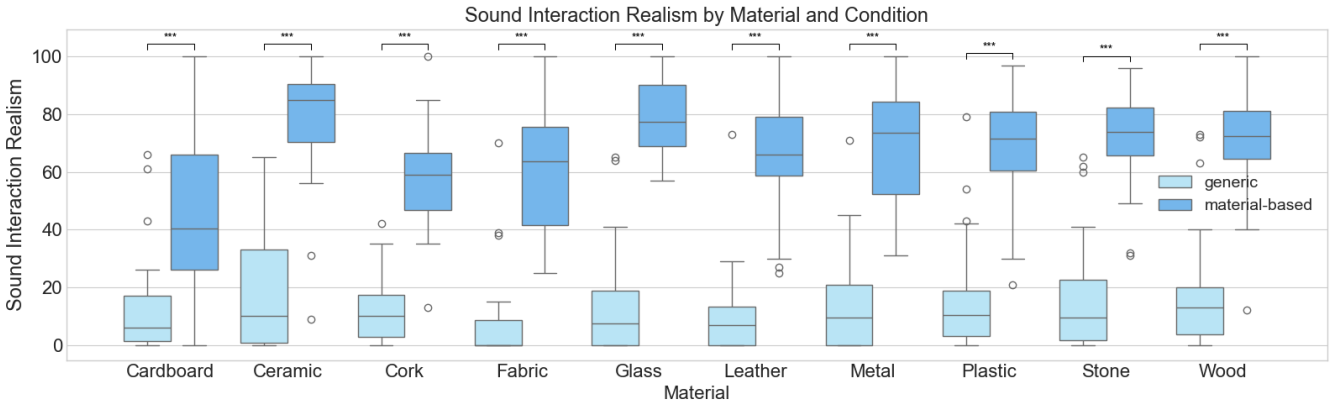

Figure 5: Sonic interaction realism ratings by material for the congruent and incongruent sounds in Task 1, *** = $p < 0.001$

within the generic ($p<0.01$) and material-based ($p<0.001$) condition. Material-based vs. generic sounds led to similar material identification accuracy for both soft and hard materials.

5.2 Task 2

The data showed a non-normal distribution. Outliers were removed from the task time data using the interquartile range method. Three outliers were identified. Please refer to Tab. 3 for a breakdown of all the results in Task 2.

5.2.1 Material-based vs. Generic Sounds

There was a significant main effect of condition ($p<0.05$) on task time. Participants took significantly longer in the material-based (84.31 ± 39.34) than generic (72.70 ± 27.95) condition. The material-based sound feedback significantly improved participants' ability to correctly identify visually similar materials (Chi-square test: $p < 0.001$), showing higher accuracy (92.75%) than the generic condition (61.76%). This was also reflected in the confidence ratings, indicating significantly greater confidence in material assignments when using the congruent, material-based sounds

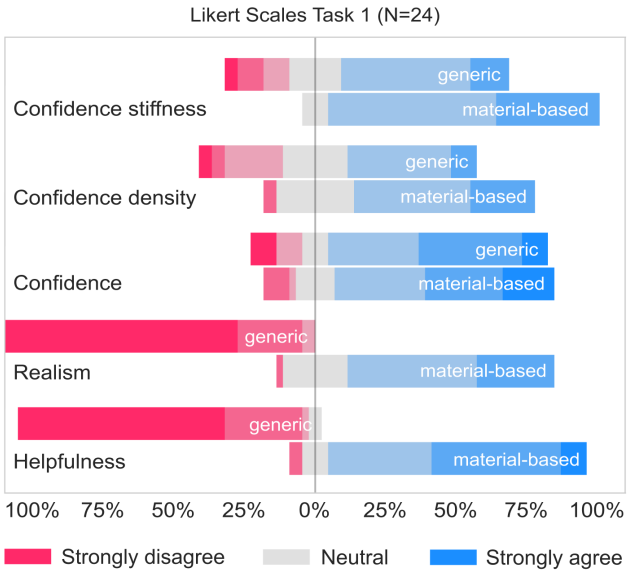

Figure 6: Barplot of Likert Scale responses in Task 1 by question and condition

Table 3: Means, standard deviations, and p-values of the measures in Task 2: material recognition accuracy (percent); material assignment confidence (0: Not confident at all, 100: Very confident), sound helpfulness (0: Not helpful at all, 100: Very helpful), time per trial (seconds)

Task 2	Generic	Material-based	P-value
Material Accuracy	61.76%	92.75%	$<$ **0.001**
Confidence	28.22 ± 25.13	74.14 ± 19.20	$<$ **0.001**
Helpfulness	6.54 ± 12.32	79.16 ± 19.88	$<$ **0.001**
Task Time	72.70 ± 27.95	84.31 ± 39.34	$<$ **0.05**

(Wilcoxon signed-rank test: $p<0.001$) (see Sec. 5.2.1). Participants furthermore perceived the material-based sounds to be significantly more helpful in distinguishing between two materials (Wilcoxon signed-rank test: $p<0.001$) (see Sec. 5.2.1).

5.2.2 Ambiguous vs. Unambiguous Material Sounds

The analysis of differences **between** ambiguous (visually and auditorily similar) and unambiguous (visually similar yet auditorily distinct) material pairs showed no significant effect of auditory ambiguity on task time, sound helpfulness, or material identification accuracy. A significant main effect of auditory ambiguity on material identification confidence was found ($p<0.001$). Participants were significantly more confident in their material assignments for unambiguous (62.20%) than ambiguous (40.21%) pairs.

There was also a significant main effect of condition on material recognition accuracy **within** the ambiguous ($p<0.05$) and unambiguous ($p<0.001$) material groups. The material-based sound feedback resulted in significantly higher material recognition accuracy (84.80%) than the generic sound feedback (60.00%) within the auditorily ambiguous material pairs (see Sec. 5.2.1). The same is true within the unambiguous material pairs, where material-based sounds (100%) also achieved significantly greater material identification accuracy than generic sound feedback (63.60%).

6 Discussion

We introduced a framework for material-based sonic interactions in AR and demonstrated improved interaction realism over generic sounds, which are the current standard in AR applications.

More specifically, our results showed that congruent audiovisual feedback in AR generated using our material-based approach leads to significantly more accurate material identification ($p < 0.001$)) of real-world objects when distinguishing between two visually similar materials (H4). However, no significant difference in identifica-

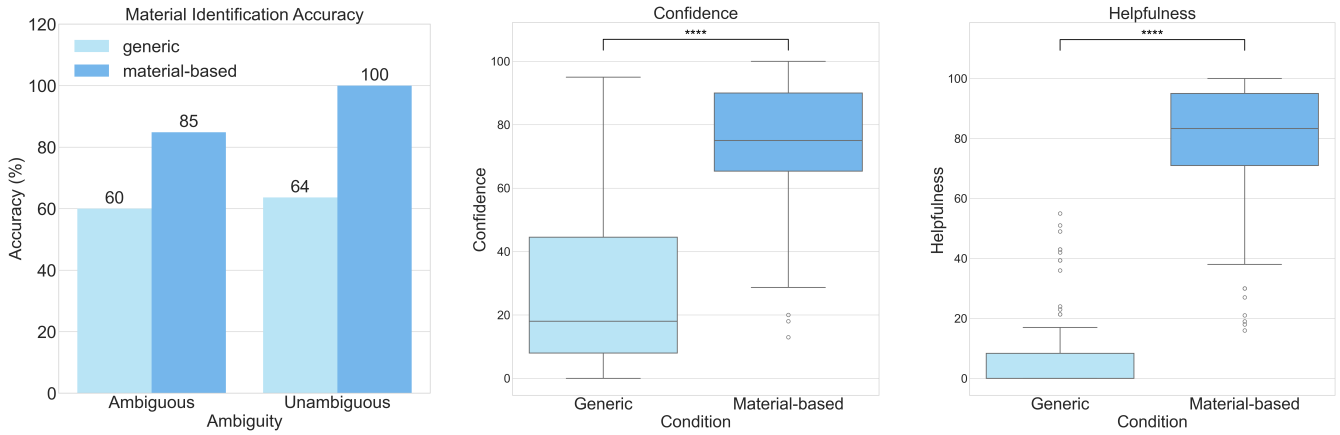

Figure 7: Task 2 results: Material identification accuracy by condition and auditory ambiguity (Ambiguous: Material pairs with visual and auditory similarity, Unambiguous: Material pairs with visual similarity, but audible differences), material assignment confidence by condition, sound helpfulness by condition

tion accuracy was observed in Task 1, where participants identified a single material at a time (H1). This is likely because most of the materials in Task 1 were already visually distinct enough to be reliably identified by sight without the need for matching impact sounds. In everyday contexts, when conflicting visual and auditory cues are present, we often rely on additional haptic feedback to discern materials. In our study, participants were not allowed to touch the materials, as audiovisual stimuli were the subject of examination. It is also noteworthy that in real-world settings, the geometry of an object presents strong cues on object materiality. Through experience, we have learned to associate certain materials with certain object shapes. When visual information on object geometry and surface texture is combined, humans can infer material types more confidently than in our controlled study setting, where shape cues were missing. The use of uniformly shaped material plates might have led to material misclassifications in cases where material-based sounds were ambiguous.

The material identification results were consistent with participants' subjective confidence ratings (H2). In Task 1, where identification accuracy did not differ between conditions, confidence ratings were similarly unaffected. In contrast, in Task 2, where material-based sounds significantly improved identification accuracy, participants also reported significantly higher confidence in their choices ($p < 0.001$) when audiovisual congruence was given. Interestingly, regardless of identification accuracy, participants expressed greater confidence in their estimations of material properties (density, stiffness). This suggests that congruent, material-based sounds carry meaningful information about physical attributes, allowing users to perceive these properties with significantly increased confidence ($p < 0.05$). These findings imply that semantically congruent audio stimuli establish a deeper perceptual link between users and their physical surroundings, even when interacting with the real world through virtual objects.

Furthermore, we were able to show that sounds created using our material-based approach led to significantly greater ($p < 0.001$) sound realism (H3). The average sonic interaction realism was rated 66 out of 100. This result is in line with findings from a study in psychoacoustics that employed the same 0-100 scale for assessing subjective judgment of sound realism. They reported an average score of 68 for recordings of real material interaction sounds, compared to a mean score of 45 for the corresponding sound effects [14]. These findings demonstrate that even real sound recordings are not perceived as fully realistic, highlighting the need to interpret realism ratings relative to other sound conditions. However, there is room for improvement in enhancing the realism of the material-based sounds. The sounds of materials with lower density and stiffness, like fabric, cardboard, cork, and leather, were rated as less realistic than rigid materials. This may be related to the combination of high damping coefficients and low stiffness in soft materials, resulting in lower frequency and shorter temporal evolution of the sounds, thereby offering participants less time to perceive differences in acoustic properties. Alternative sound synthesis techniques, more suitable to soft bodies, e.g., data-driven approaches [46, 45] could be explored to achieve greater sonic interaction realism for soft materials.

Lastly, we saw increased task time during the material-based condition in both tasks. In the generic condition, the same sound effect was played for every material. Therefore, a potential influence on faster trial time may be that the auditory memory only had to refer to one type of sound in every trial, speeding up the decision-making process. Since the material-based sounds carry nuanced information about the materials' properties represented in sound parameters such as frequency and decay [21], corresponding to each material, the sound varied greatly for every trial. This led participants to tap the materials more often in the material-based condition, carefully listening to the complex sound qualities unfold. This was especially true for Task 2, where the material-based trials consisted of two multifaceted sounds, while the generic trials only presented the same sound for both materials. As a result, participants had to integrate more information in the material-based condition to compare and judge the materials, prolonging the trial time.

6.1 Limitations

One limitation of our study is the use of average material property values for the material sounds. However, these values can largely impact the resulting sounds, making them less or more realistic. A potential solution could be two-fold: Firstly, establishing a material segmentation model that is able to provide more granular material classification output and secondly, using material parameter estimation methods, as proposed in Ren et al. [38], to estimate ideal parameters from the vision-based material output to achieve more realistic sounds.

Although both the real and the virtual object are involved in the sound-producing interaction, our study only sonified the tapping actions on the real surfaces in the scene, purposefully excluding the

modeling of the virtual object (stick) involved in the collision to examine the material-based approach in an isolated manner. However, to achieve a comprehensive simulation of the interaction, a bidirectional interaction between both objects, the real and the virtual, should be modeled. On top, adding more diverse action types like sliding, bowing, or scratching would further enhance the interaction capabilities.

In our study, we decided against including a condition without sound, as we were interested in the interplay of multiple senses. However, although the generic audio was incongruent, it still provided an auditory cue, making it difficult to assess its specific influence, e.g., on material identification accuracy, sound helpfulness, and confidence. Generic sounds, beyond being non-helpful, can even be detrimental to the user's ability to correctly identify the material of a real object, especially when the visual appearance of a material is ambiguous. For example, if an object could be either glass or plexiglass, and the generic sound is more similar in pitch and reverb to the impact sound of plexiglass, users may unconsciously rely on this misleading auditory cue, leading to incorrect identification. This can further lead to a misjudgment of decision confidence and sound helpfulness. In contrast, the absence of sound, i.e., in a purely visual condition, might limit confidence but does not introduce conflicting sensory information that biases the user's decision. This perceptual asymmetry is important. While silence preserves ambiguity, generic sounds introduce false specificity. In this sense, generic sounds function not as neutral placeholders but as active confounders, potentially resulting in worse results than a silent condition. Despite the confounders that a generic sound introduces, we purposely chose to use it in our baseline condition to mirror the current industry standard. Our goal was not to compare against silence, but to investigate whether material-based audio offers perceptual benefits over existing, generic audio design. However, we want to emphasize that our findings call attention to the potential negative perceptual consequences of using generic audio in current commercial XR systems and urge headset manufacturers and software developers to consider these modulatory perceptual effects when making audio design choices.

6.2 Application Areas

Multiple potential application areas for the proposed sonification framework come to mind. For one, the material-based sounds could be used to convey more detailed information about object materiality to blind or visually impaired users. Multiple works have shown that audio augmented reality can support blind or visually impaired users in perceiving and navigating their environment [39, 50, 18]. Our system could, for example, be used to augment a white cane with more granular sensory information about the materiality of the ground to facilitate navigation and support tactile understanding of the environment.

This material-based sonification method could furthermore be applied to medical use cases. Schütz et al. [42] already introduced a physics-based sonification approach for medical applications such as tumor localization. While their work is limited to pre-defined models of the human anatomy, the framework presented in this paper could expand the use of physics-based sounds to medical AR applications. In minimally invasive laparoscopic procedures, for instance, the surgeon inserts instruments into the abdominal region via small incisions in the skin. This rids the surgeon of direct visual, audio, and haptic interaction with the anatomy. In this case, real-time audio feedback based on endoscopic images from inside the body could provide important textural information about the human tissue and thus increase awareness of human tissue properties. This could be helpful in distinguishing between cancerous and healthy tissue, potentially enhancing surgery outcomes. However, as our study showed, sounds can influence or alter our perception of materiality. As sound has the power to influence our decision-making and behavior [1], sonic interactions in surgical applications must be designed with careful consideration of their perceptual impact. To ensure high safety standards in medical applications, robust auditory cues should be targeted.

Besides the outlined use cases in accessibility and medical technology, material-based sonification could enhance AR training applications where users interact with virtual tools on real objects (e.g., assembling parts). Here, material-based audio could improve realism and skill transfer. In addition, telepresence in remote collaboration settings [54] could benefit from realistic interaction sounds to make remote user actions feel more believable, thereby improving presence.

6.3 Future Work

A possible extension of our work could be the inclusion of the real object's shape, size, and context into the sonification pipeline. Achieving this within a modal synthesis framework would require real-time 3D reconstruction to obtain object geometries. With rapid advances in computer vision and the increasing computational capabilities of AR headsets, this may soon become feasible.

Touch interactions with physical objects are accompanied by visual, auditory, and haptic feedback. Touch was not considered in our work, but is instrumental to achieving fully realistic interaction experiences. Future work could investigate the addition of this sensory modality.

It would also be interesting to investigate the impact of congruent and incongruent material sounds on user perception. This would necessitate the inclusion of a condition presenting ever-changing, incongruent material sounds instead of the constant generic sound used in our study. We also did not include a "no sound" or a "real sound" condition. We see benefits in including either or both in future studies where appropriate. A real sound condition could be especially relevant in work primarily focused on achieving high-fidelity sound synthesis. To implement a "real sound" condition, high-quality, controlled audio recordings of all relevant material-object interactions would be required. These recordings would then need to be mapped to specific interactions in the AR system, requiring precise real-time collision detection, timing synchronization, and spatial audio rendering to ensure perceptual alignment. Implementing such a control condition introduces significant technical complexity, but it might be worth incorporating in a study focused on audio fidelity.

7 Conclusion

We introduced a material-based sonification framework that uses scene understanding and physical modeling sound synthesis to generate context-aware sound for AR interactions in real-time. The resulting material-based sounds establish audiovisual congruence during real-virtual object interactions in AR. In a user study comparing our congruent, material-based sounds to incongruent, generic sound effects, participants rated interactions with material-based sounds as significantly more realistic. Furthermore, material-based sounds enabled more accurate and confident differentiation of visually similar materials. These results suggest that context-aware, material-based auditory feedback can strengthen the perceptual connection between users and their physical environments during AR interactions. We hope that future research will build upon our findings and that commercial AR systems will increasingly incorporate realistic, context-aware sound feedback to enhance user experience.

Acknowledgments

The authors wish to thank Prof. Dr. Stephan Krusche at the Technical University of Munich for his support with managing Apple Developer certificates.

REFERENCES

[1] E. H. Anne de Haas and L.-H. Lee. Deceiving audio design in augmented environments : A systematic review of audio effects in augmented reality. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 36–43, 2022. doi: 10.1109/ISMAR-Adjunct57072.2022.00018 9

[2] S. Bell, P. Upchurch, N. Snavely, and K. Bala. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2

[3] S. Cai, R. Wakaki, S. Nobuhara, and K. Nishino. Rgb road scene material segmentation. *Image Vision Comput.*, 145(C), May 2024. doi: 10.1016/j.imavis.2024.104970 2

[4] C. Chen, K. Ashutosh, R. Girdhar, D. Harwath, and K. Grauman. Soundingactions: Learning how actions sound from narrated egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27252–27262, 2024. 2, 3

[5] C. Chen, P. Peng, A. Baid, Z. Xue, W.-N. Hsu, D. Harwath, and K. Grauman. Action2sound: Ambient-aware generation of action sounds from egocentric videos. In *European Conference on Computer Vision*, pp. 277–295. Springer, 2024. 2, 3

[6] Y.-C. Chen and C. Spence. When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. *Cognition*, 114(3):389–404, 2010. 2, 3

[7] Z. Chen, P. Seetharaman, B. Russell, O. Nieto, D. Bourgin, A. Owens, and J. Salamon. Video-guided foley sound generation with multimodal controls. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18770–18781, 2025. 2, 3

[8] S. Cmentowski, A. Krekhov, A. Zenner, D. Kucharski, and J. Krüger. Towards sneaking as a playful input modality for virtual environments. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 473–482, 2021. doi: 10.1109/VR50410.2021.00071 3

[9] P. Cook. Sound production and modeling. *IEEE Computer Graphics and Applications*, 22(4):23–27, 2002. doi: 10.1109/MCG.2002.1016695 2, 3

[10] P. R. Cook. Physically informed sonic modeling (phism): Percussive synthesis. In *Proceedings of the 1996 International Computer Music Conference*, pp. 228–231. The International Computer Music Association, 1996. 2, 3

[11] G. Eckel. Sound synthesis by physical modelling with modalys. *Proc. ISMA'95*, pp. 478–482, 1995. 2, 3

[12] W. Fujisaki, N. Goda, I. Motoyoshi, H. Komatsu, and S. Nishida. Audiovisual integration in the human perception of materials. *Journal of Vision*, 14(4):12–12, 2014. 3

[13] M. Geronazzo and S. Serafin. *Sonic Interactions in Virtual Environments: The Egocentric Audio Perspective of the Digital Twin*, pp. 3–45. Springer International Publishing, Cham, 2023. doi: 10.1007/978-3-031-04021-4_1 3

[14] L. M. Heller and L. Wolf. When hybrid sound effects are better than real recordings. *Proceedings of Meetings on Acoustics*, 46(1):050002, 08 2022. doi: 10.1121/2.0001581 8

[15] M. Hoppe, J. Karolus, F. Dietz, P. W. Woźniak, A. Schmidt, and T.-K. Machulla. Vrsneaky: Increasing presence in vr through gait-aware auditory feedback. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, p. 1–9. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3290605.3300776 3

[16] N. Ibers and M. Wölfel. Effects on presence, agency, and walking in immersive virtual reality of self-induced and estimated footstep-sounds. In *2023 IEEE 2nd International Conference on Cognitive Aspects of Virtual Reality (CVR)*, pp. 000063–000068, 2023. doi: 10.1109/CVR58941.2023.10395789 3

[17] X. Jin, S. Li, T. Qu, D. Manocha, and G. Wang. Deep-modal: Real-time impact sound synthesis for arbitrary shapes. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, p. 1171–1179. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3394171.3413572 2, 3

[18] O. B. Kaul, K. Behrens, and M. Rohs. Mobile recognition and tracking of objects in the environment through augmented reality and 3d audio cues for people with visual impairments. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3411763.3451611 9

[19] A. C. Kern and W. Ellermeier. Audio in vr: Effects of a soundscape and movement-triggered step sounds on presence. *Frontiers in Robotics and AI*, 7, 2020. doi: 10.3389/frobt.2020.00020 2

[20] H. Kim and I.-K. Lee. Studying the effects of congruence of auditory and visual stimuli on virtual reality experiences. *IEEE Transactions on Visualization and Computer Graphics*, 28(5):2080–2090, 2022. doi: 10.1109/TVCG.2022.3150514 3

[21] R. L. Klatzky, D. K. Pai, and E. P. Krotkov. Perception of material from contact sounds. *Presence*, 9(4):399–410, 2000. doi: 10.1162/105474600566907 1, 8

[22] R. F. Landel and L. E. Nielsen. *Mechanical properties of polymers and composites*. CRC press, 1993. 5

[23] P. J. Laurienti, R. A. Kraft, J. A. Maldjian, J. H. Burdette, and M. T. Wallace. Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental brain research*, 158:405–414, 2004. 2, 3

[24] G. Lemaitre and L. M. Heller. Auditory perception of material is fragile while action is strikingly robust. *The Journal of the Acoustical Society of America*, 131(2):1337–1348, 2012. 1

[25] K. Lin and S. Liu. Foley agent: Automatic sound design and mixing agent for silent videos driven by llms. In N. Magnenat Thalmann, X. Hu, B. Sheng, D. Thalmann, T. Peng, W. Meng, J. Huang, L. Zhu, and X. Wei, eds., *Computer Animation and Social Agents*, pp. 177–192. Springer Nature Singapore, Singapore, 2025. 2

[26] D. B. Lloyd, N. Raghuvanshi, and N. K. Govindaraju. Sound synthesis for impact sounds in video games. In *Symposium on Interactive 3D Graphics and Games*, I3D '11, p. 55–62 PAGE@7. Association for Computing Machinery, New York, NY, USA, 2011. doi: 10.1145/1944745.1944755 2

[27] S. McAdams, A. Chaigne, and V. Roussarie. The psychomechanics of simulated sound sources: Material properties of impacted bars. *The Journal of the Acoustical Society of America*, 115(3):1306–1320, 02 2004. doi: 10.1121/1.1645855 1

[28] S. McAdams, V. Roussarie, A. Chaigne, and B. L. Giordano. The psychomechanics of simulated sound sources: Material properties of impacted thin plates. *The Journal of the Acoustical Society of America*, 128(3):1401–1413, 2010. 1

[29] J. H. McDermott, V. Agarwal, and J. Traer. Physics, ecological acoustics and the auditory system. *Current Biology*, 34(20):R1006–R1013, 2024. doi: 10.1016/j.cub.2024.05.056 1

[30] G. W. Morey. The properties of glass. 1954. 5

[31] J. D. Morrison and J.-M. Adrien. Mosaic: A framework for modal synthesis. *Computer Music Journal*, 17(1):45–56, 1993. 2, 3

[32] P. Niemz, W. Sonderegger, T. Keplinger, J. Jiang, and J. Lu. Physical properties of wood and wood-based materials. In *Springer handbook of wood science and technology*, pp. 281–353. Springer, 2023. 5

[33] R. Nordahl, L. Turchet, and S. Serafin. Sound synthesis and evaluation of interactive footsteps and environmental sounds rendering for virtual reality applications. *IEEE Transactions on Visualization and Computer Graphics*, 17(9):1234–1244, 2011. doi: 10.1109/TVCG.2011.30 3

[34] R. A. Paquin. Properties of metals. *Handbook of optics*, 2:35–49, 1995. 5

[35] N. Raghuvanshi and M. C. Lin. Physically based sound synthesis for large-scale virtual environments. *IEEE Computer Graphics and Applications*, 27(1):14–18, 2007. doi: 10.1109/MCG.2007.16 2, 3

[36] S. K. Ravipati, E. Latif, R. Parasuraman, and S. M. Bhandarkar. Object-oriented material classification and 3d clustering for improved semantic perception and mapping in mobile robots. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9729–9736, 2024. doi: 10.1109/IROS58592.2024.10801936 2

[37] Z. Ren, H. Yeh, and M. C. Lin. Synthesizing contact sounds between textured models. In *2010 IEEE Virtual Reality Conference (VR)*, pp. 139–146, 2010. doi: 10.1109/VR.2010.5444799 2, 3

[38] Z. Ren, H. Yeh, and M. C. Lin. Example-guided physically based modal sound synthesis. *ACM Trans. Graph.*, 32(1), Feb. 2013. doi: 10

.1145/2421636.2421637 8

[39] F. Ribeiro, D. Florencio, P. A. Chou, and Z. Zhang. Auditory augmented reality: Object sonification for the visually impaired. In *2012 IEEE 14th international workshop on multimedia signal processing (MMSP)*, pp. 319–324. IEEE, 2012. 9

[40] L. Schütz, T. El Chemaly, E. Weber, A. T. Doan, J. Tsai, C. Leuze, B. Daniel, and N. Navab. Interactive shape sonification for tumor localization in breast cancer surgery. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3613904.3642257 3

[41] L. Schütz, E. Weber, W. Niu, B. Daniel, J. McNab, N. Navab, and C. Leuze. Audiovisual augmentation for coil positioning in transcranial magnetic stimulation. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 11(4):1158–1165, 2023. doi: 10.1080/21681163.2022.2154277 3

[42] L. Schütz, S. Matinfar, G. Schafroth, N. Navab, M. Fairhurst, A. Wagner, B. Wiestler, U. Eck, and N. Navab. A framework for multimodal medical image interaction. *IEEE Transactions on Visualization and Computer Graphics*, 30(11):7419–7429, 2024. doi: 10.1109/TVCG.2024.3456163 3, 9

[43] S. Serafin, M. Geronazzo, C. Erkut, N. C. Nilsson, and R. Nordahl. Sonic interactions in virtual reality: State of the art, current challenges, and future directions. *IEEE Computer Graphics and Applications*, 38(2):31–43, 2018. doi: 10.1109/MCG.2018.193142628 2, 3

[44] S. Siegesmund and H. Dürrast. Physical and mechanical properties of rocks. In *Stone in architecture: properties, durability*, pp. 97–225. Springer, 2010. 5

[45] F. Su and C. Joslin. Procedurally-generated audio for soft-body animations. In *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion*, AM '18. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3243274.3243285 8

[46] K. Su, K. Qian, E. Shlizerman, A. Torralba, and C. Gan. Physics-driven diffusion models for impact sound synthesis from videos. *CVPR*, 2023. 2, 3, 8

[47] X. Su, J. E. Froehlich, E. Koh, and C. Xiao. Sonifyar: Context-aware sound generation in augmented reality. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST '24. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3654777.3676406 3

[48] P. Upchurch and R. Niu. A dense material segmentation dataset for indoor and outdoor scene parsing. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, eds., *Computer Vision – ECCV 2022*, pp. 450–466. Springer Nature Switzerland, Cham, 2022. 2, 3

[49] K. van den Doel, P. G. Kry, and D. K. Pai. Foleyautomatic: physically-based sound effects for interactive simulation and animation. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, p. 537–544. Association for Computing Machinery, New York, NY, USA, 2001. doi: 10.1145/383259.383322 2, 3

[50] K. van den Doel, D. Smilek, A. Bodnar, C. Chita, R. Corbett, D. Nekrasovski, and J. McGrenere. Geometric shape detection with soundview. In *Proceedings of ICAD 04-Tenth Meeting of the International Conference on Auditory Display*, pp. 1–8. ICAD, Sydney, Australia, 2004. 9

[51] R. Van Ee, J. J. Van Boxtel, A. L. Parker, and D. Alais. Multisensory congruency as a mechanism for attentional control over perceptual selection. *Journal of Neuroscience*, 29(37):11641–11649, 2009. 3

[52] J. Wilson, B. N. Walker, J. Lindsay, C. Cambias, and F. Dellaert. Swan: System for wearable audio navigation. In *2007 11th IEEE International Symposium on Wearable Computers*, pp. 91–98, 2007. doi: 10.1109/ISWC.2007.4373786 3

[53] M. Wright. Open sound control: an enabling technology for musical networking. *Organised Sound*, 10(3):193–200, 2005. 4

[54] K. Yu, G. Gorbachev, U. Eck, F. Pankratz, N. Navab, and D. Roth. Avatars for teleconsultation: Effects of avatar embodiment techniques on user perception in 3d asymmetric telepresence. *IEEE Transactions on Visualization and Computer Graphics*, 27(11):4129–4139, 2021. 9