Superhuman Hearing - Virtual Prototyping of Artificial Hearing

A Case Study on Interactions and Acoustic Beamforming

Geronazzo, Michele; Vieira, Luis S.; Nilsson, Niels Christian; Udesen, Jesper; Serafin, Stefania

Published in:
IEEE Transactions on Visualization and Computer Graphics

DOI (link to publication from Publisher):
10.1109/TVCG.2020.2973059

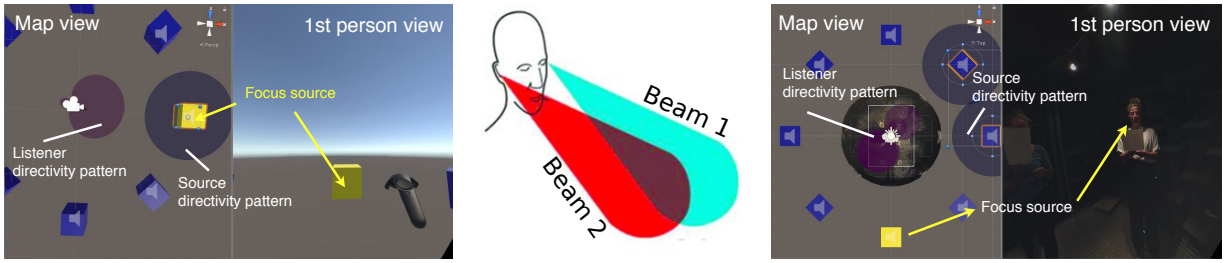
Publication date:
2020

Document Version
Accepted author manuscript, peer reviewed version

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2020.2973059, IEEE Transactions on Visualization and Computer Graphics

Superhuman Hearing - Virtual Prototyping of Artificial Hearing: a Case Study on Interactions and Acoustic Beamforming

Michele Geronazzo, *Senior Member, IEEE*, Luis S. Vieira, Niels Christian Nilsson, Jesper Udesen, and Stefania Serafin



(a) Exp. 1: abstract scenario (b) Ideal beamforming (c) Exp.2: realistic scenario

Fig. 1: Screenshots of the two virtual listening scenarios with map (left) and listener (right) view. Beamforming directivity patterns are also visualized together with an iconic representation of an ideal beamformer (b). Since the abstract scenario (a) was a virtual environment with minimal visual feedback, the realistic scenario (c) was rendered though immersive audio-visual 360° recordings.

Abstract— Directivity and gain in microphone array systems for hearing aids or hearable devices allow users to acoustically enhance the information of a source of interest. This source is usually positioned directly in front. This feature is called acoustic beamforming. The current study aimed to improve users' interactions with beamforming via a virtual prototyping approach in immersive virtual environments (VEs). Eighteen participants took part in experimental sessions composed of a calibration procedure and a selective auditory attention voice-pairing task. Eight concurrent speakers were placed in an anechoic environment in two virtual reality (VR) scenarios. The scenarios were a purely virtual scenario and a realistic 360° audio-visual recording. Participants were asked to find an individual optimal parameterization for three different virtual beamformers: (i) head-guided, (ii) eye gaze-guided, and (iii) a novel interaction technique called dual beamformer, where head-guided is combined with an additional hand-guided beamformer. None of the participants were able to complete the task without a virtual beamformer (i.e., in normal hearing condition) due to the high complexity introduced by the experimental design. However, participants were able to correctly pair all speakers using all three proposed interaction metaphors. Providing superhuman hearing abilities in the form of a dual acoustic beamformer guided by head and hand movements resulted in statistically significant improvements in terms of pairing time, suggesting the task-relevance of interacting with multiple points of interests.

Index Terms—Virtual prototyping, Sonic interactions, Acoustic beamforming, Artificial hearing, Virtual reality, Multi-speaker scenario

1 Introduction

A central feature of human hearing is the listener's ability to focus attention to a certain direction, for example towards a specific sound source [40]. This selective auditory attention relies on a set of mechanisms that work together to produce an understanding of the direction of an incoming sound, as well as what physical features characterize its source, ultimately creating an auditory representation of the source. In this way, the auditory system analyzes the sound field to determine what is relevant content and what is noise. However, listening is a multi-modal experience, where the auditory representation interacts with other modalities as well as with bodily and cognitive mechanisms. It is important to understand to what degree these mechanisms—motion, visual feedback, spatial directivity—influence the listening process during a specific task [62], how they support the auditory behavior, and to what extent they can be used to support artificial hearing devices such as hearing aids and smart headphones, called *hearables* [61]. The ultimate goal of this research is to control these interactions in order to create a set of tools that provide *superhuman hearing*.

People without hearing impairments are usually able to distinguish between meaningful and non-meaningful auditory information; thus, solving the cocktail party problem [11]. On the other hand, this is one of the main challenges for people with hearing impairments who require artificial hearing and hearing aids [61] equipped with digital signal processing algorithms such as *beamforming* [49,70]. The current work is based on the following assumption and long-term vision: even though the currently available technologies do not allow a perfect separation of relevant signals and background noise, one can realistically assume that future developments will raise the bar to make this feature available. This can for example materialize by taking advantage of artificial intelligence to analyze the sound field and separate individual sound sources in adverse listening situations [76]. In the future, when hearing aids can separate individual sound streams in a cocktail party environment, it will be particularly relevant to know how these sound streams should be processed before they are presented to the hearing aid user. One obvious presentation mode would be to preserve

- *Michele Geronazzo, Niels Christian Nilsson, and Stefania Serafin are with Aalborg University Copenhagen, Department of Architecture, Design, and Media Technology, Copenhagen, Denmark, E-mail: {mge,ncn,sts}@create.aau.dk.*
- *Luis S. Vieira is with Khora Virtual Reality, Copenhagen, Denmark, E-mail: luis@khoravr.com.*
- *Jesper Udesen is with GN Audio A/S, Ballerup, Denmark, E-mail: judesen@jabra.com*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

the natural spatial characteristics of the individual sound streams by filtering them with the listener-specific acoustic information. But other more artificial presentation modes could also benefit the hearing aid user, like having a narrow acoustic beam following the head orientation where only sound sources within the beam are presented to the user. Identifying the optimal presentation mode is therefore relevant from a hearing aid technology perspective.

Accordingly, our main research focus relates to optimal interaction between audio streams, which is based on *a priori* knowledge of the separated sound sources and the specific task performed by the listener. From a methodological point of view, our experimental sessions focused on novel forms of interaction aiming to reinforce the listener's selective auditory attention in different situations: from nearly impossible tasks without technological support to aiding solutions for hearing impaired users in practical real-life scenarios. Our case study considered an extremely challenging scenario with multiple concurrent speakers where spatial filtering via beamforming control [70] allowed normal hearing users to acoustically amplify a specific speaker while attenuating unwanted sound sources.

Virtual Reality (VR) was chosen as a prototyping platform because it provided flexibility and faster prototyping of sonic interactions [65] as well as control parameters over different simulations. This paper proposes a platform for virtual prototyping of interactions with hearing aids, where multiple artificial hearing models can be tested. We included two different virtual environments (VEs) providing different levels of realism: (i) one involving minimal visual feedback with eight virtual cubes displayed in a perfect circle around the listener (see Fig. 1a), and (ii) another showing eight real human speakers in the same positions recorded in a 360° video (see Fig. 1c). The participants could individually calibrate a *virtual beamforming* (for an iconic representation see Fig. 1b) with three degrees of freedom: gain, width, and shape. Moreover, the participants were able to control the beam's direction using three different interaction techniques: (i) a standard head pointing, (ii) a more natural eye-gaze pointing recently proposed in [7, 23], and (iii) a novel sophisticated dual beamforming control based on head and hand orientations. In this case study, we investigated the efficacy of the proposed interactions and parameters in terms of completion time, correctness, and perceived task load when trying to find multiple pairs of the same speech segments in highly chaotic situations where the natural listening condition is insufficient to accomplish the task.

The paper is organized as follows. Sec. 2 surveys the theoretical background of the cognitive listening process, hearing aids technologies, and virtual prototyping of optimal hearing support in VR. In Sec. 3, we focus on beamforming and directivity patterns as a tool for artificial hearing. Moreover, this section describes the binaural spatial audio rendering with higher-order ambisonics (HOA) [29] and the technical implementations of the proposed VEs. In Sec. 4, the experimental protocol is described, and Sec. 5 presents the data collected, providing statistical comparisons among interactions and scenarios. Finally, Sec. 6 and 7 discuss the outcomes of the study, concluding with a summary of meaningful contributions and future work.

2 Background

Listening has been an important part of human selective attention research since 1953, starting with Cherry's research on the "cocktail party problem" [14], and the use of dichotic stimuli to test speech intelligibility. Different levels of perception and cognition contribute to human's ability to segregate signals—also referred to as auditory signal analysis [10, 11]. When confronted with multiple simultaneous stimuli (speech or non-linguistic stimuli), it is necessary to segregate relevant auditory information from concurrent background sounds and to focus attention on the source of interest [11, 14]. This action is related to the principles of auditory scene analysis that require a stream of auditory information filtered and grouped into a number of perceptually distinct and coherent auditory objects. Studies on spoken language processing suggest that in multi-talker situations, auditory object formation and selection together with attentional allocation contribute to define a model of cocktail-party listening [32, 66]. Accordingly, Ahrens et al. [2] recently conducted a pilot study with six participants who were able to accurately analyze virtual audio-visual scenes containing up to six concurrent talkers.

2.1 Dichotic listening and masking effect

Dichotic listening is a psychological test used to investigate selective auditory attention and shows the brain's ability of hemispheric lateralization for speech perception—a feature of importance when listening to different acoustic events presented to each ear simultaneously [33,50]. The right-ear advantage is an interesting finding [39] revealing the direct anatomic connection of the right ear to the left hemisphere, which in most people is specialized in language processing.

Moreover, another relevant aspect of dichotic listening is the effect of interfering speech or other concurrent non-linguistic signals due to their frequency spectrum characteristics and spatial information. Of particular interest here is the anatomy of the human body, head, and ears that introduces a listener-specific acoustic characterization of the stimuli through the so-called head-related transfer function (HRTF), helping the brain to localize the sound in space [74]. The interaural time differences (ITD) and interaural level differences (ILD) allow listeners to locate sound sources on the horizontal plane, and they play an important role in generating high levels of speech recognition in complex listening environments [14, 45]. ITD and ILD combined improve robustness from the masking effect—when the signal of interest shares information in the same frequency bands and/or the same sound pressure level with interfering signals—and increase the decorrelation between left and right ear and thus the separation between background noise and meaningful sounds [75]. The brain also correlates the signals that arrive at both ears through the so-called interaural cross-correlation coefficient (IACC), which is a measure associated with the feeling of spaciousness and envelopment in room acoustics: the higher the value, the more spacious and comfortable the space feels to the listener [8].

2.2 Hearing impairment, hearing aids and artificial hearing

Understanding speech in noisy situations becomes a very difficult task for people with hearing impairments when both speech and noise coexist above their hearing threshold. In such individuals, the ability to focus attention only on the important stimuli benefits from an increase of signal-to-noise ratio (SNR) with respect to masking sources for optimal intelligibility. This is particularly pertinent to multi-speaker scenarios (for a recent review see the work of Falk et al. [22]). Typical hearing losses are located in the cochlea where damage to hair cells can be observed. This damage is often provoked by exposure to loud sound [57]. The hearing threshold, also known as speech reception threshold, defines the lowest level at which a person can separate meaningful signals from noise. This value ranges from a few dB to more than 10 dB, causing severe problems of communication. In many cases, the resonance effect and corresponding frequency perception are deteriorated due to the damage of the outer hair cells. Consequently, the brain is no longer able to benefit from the long-term spectrum fluctuations, where speech is recognized to have larger variance [11] and from the spatial cues which are able to reduce masker interference often referred to as spatial release from masking [26, 57].

Hearing aids equipped with microphone arrays in behind-the-ear (BTE) and in-the-ear (ITE) configurations aim at compensating for these hearing impairments. More recently, alternative designs have come to the market. These are smaller in size featuring a thinner sound tube that connects the hearing device behind the ear to the ear canal, called receiver-in-canal (RIC). Signal processing requirements of for hearing aids are very restricted due to the physical size of the device and optimized due to energy consumption. In general, the signal flow starts by capturing the acoustic input with a microphone array, typically composed of three microphones, which is processed into a single signal within the directional microphone unit. The main frequency-band-dependent processing steps are noise reduction and signal amplification combined with dynamic compression. To address the problem of strong masking and to increase the SNR of the signal output, beamforming or other noise reduction approaches are usually developed [26].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2020.2973059, IEEE Transactions on Visualization and Computer Graphics

2.2.1 Directional microphones and beamformers

To improve the SNR, estimated to be around 4-10 dB for hearing impaired [20, 61], and to help the natural directivity of the outer ear, directional microphones have been used. Such microphones have proven to increase speech intelligibility and the speech reception threshold in the range from 2 to 4 dB [69].

Spatial separation can be exploited to isolate the signal from interferences using a spatial filter at the receiver with beamforming [61]. Such algorithms may be categorized into fixed and adaptive beamforming [70]. Fixed beamformers have a fixed spatial directivity (not dependent on the acoustical environment), and focus on a desired sound source, thereby reducing the influence of background noise, more precisely to attenuate signals outside the line of sight. Examples of fixed beamforming are delay-and-sum beamforming [15, 35], weighted-sum beamforming [24], superdirective beamforming [36], and frequency-invariant beamforming [72]. In the case of adaptive beamforming, directivity is dependent on the acoustical environment. In hearing aids, the directivity is normally adaptive in order to achieve a higher noise suppression effect with coherent noise, i.e., one dominant noise source [56, 61]. The direction from which the noise arrives is continuously estimated and the beamforming directivity pattern is automatically adjusted so that the directivity notch matches the main direction of noise arrival. This process has to be free from artifacts or robust to perceivable changes in the frequency response for the frontal target direction. The adaptation process must be fast enough ($<$ 100 milliseconds) to compensate for head movements and to track moving sources in multi source listening situations

In order to compare different beamforming solutions, the directivity Index (DI) is one of the basic performance measurements [61]. Its definition involves the power ratio of the output signal (in dB) between sound incidence from the front and the diffuse case—from sound coming equally from all directions. A correct DI is of high interest for the improvement of the effective SNR that can be achieved for frontal target sources in a diffuse noise field.

2.3 Virtual prototyping

VR can be considered as an extension of 3D computer graphics with advanced input and output paradigms [34]. In a simulated environment, the user can look, move around, and experience other sensory stimuli, in a natural or artificial way [13, 55]. Computer-aided design for virtual prototyping can be applied in many different manufacturing settings from machining, assembly, inspection to more complex processes, like education [19] and training [4, 5]. Moreover, because of the ability to create the experience of being in an environment without actually be physically there, it might be used in designing rehabilitation, therapy and psychotherapy actions [16].

The use of VR technologies for prototyping implies the development of virtual experiences that should follow specific models of human-computer interaction in an attempt to formalize users' understanding and interactions within specific VEs. The conceptual VR model by Latta and Ober [42] emphasizes the user-centered approach, where the experience is analyzed from the point of view of the user's perception of digital stimuli and how the cognitive abilities/actions offered may influence the task. Moreover, considering a more holistic vision of the user experience could favor the potential user's abilities in triggering effective actions and reactions in VR [54].

In fields such as automotive, architecture, engineering and construction (AEC), VR has been rapidly adopted as part of both development and for showcase products and ideas [46]. Manufacturers use VR headsets to allow engineers and designers to share iterative prototyping activities on the same model remotely and in real-time without the need for sensors or special facilities. In this way, companies reduce costs, simplify collaboration within key functional areas, and showcase their results to stakeholders. The same VR approach significantly facilitates the learning process by enabling trainees to safely work with robots in VR [47]. Similarly, recent work regarding virtual prototypes for smart home systems [3] considered the interaction design (selection and control model) of virtual Internet of Things (IoT) devices.

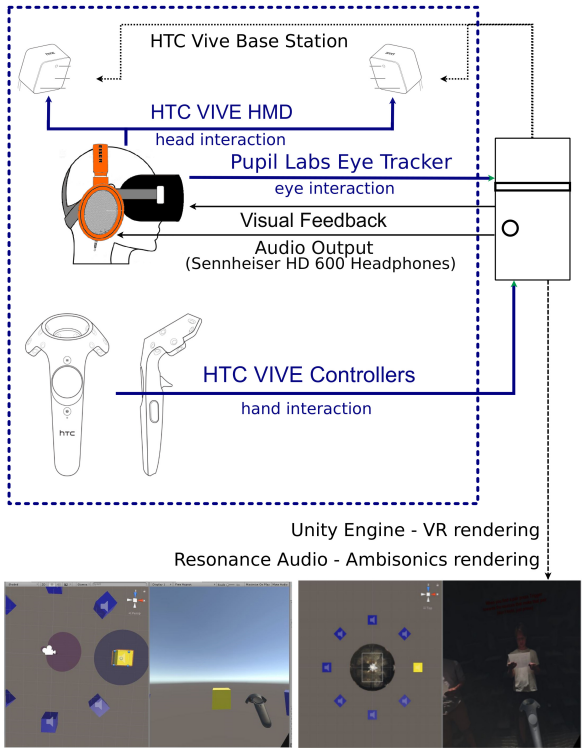


Fig. 2: Hardware and software scheme.

3 Case Study Materials: Virtual Beamforming

The general term "directivity" is commonly used to describe the electroacoustic evaluation of directional properties. In order to investigate user interactions with this aspect in beamforming technologies, the term *directional benefit* can be used to describe situations in which a directional model performs better than an omnidirectional model. Accordingly, research on directional hearing aids sometimes reveals little correlation between listeners' performance and directional benefit [58–60]. This is due to the mixed effect of other hearing aid properties such as signal processing algorithms, frequency shaping characteristics, technical specifications, and performance of the equipment. In contrast, it is assumed that directional benefits reflect the impact of the directivity in the microphone array on the hearing aid processing system, thus relating to the quality of the directional microphone behavior [58, 60].

While directional patterns can provide detailed information relative to directional attenuation provided by a hearing aid across angles, it is sometimes difficult to visualize the total impact of this attenuation in complex listening environments. The additional benefit directional hearing aids will provide is a key behavioral aspect of the evaluation. It can be quantified using objective measures of speech recognition as well as subjective measures of perception of sound quality, benefit, performance, and satisfaction. By far, the most common method for assessing the impact of hearing aids is the quantification of changes in speech recognition in noisy environments [61]. Accordingly, a systematic approach in such evaluation is crucial and an expendable evaluation framework is a perfect application scenario for VR prototyping.

3.1 The virtual reality system

The proposed evaluation framework was built for a VR setup using Unity 2017[1], a game engine that allowed the integration of Resonance Audio API[2]. Fig. 2 depicts the high-level structure of the hardware and software adopted in our framework. The framework was based

[1] https://unity3d.com/
[2] https://github.com/resonance-audio

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2020.2973059, IEEE Transactions on Visualization and Computer Graphics

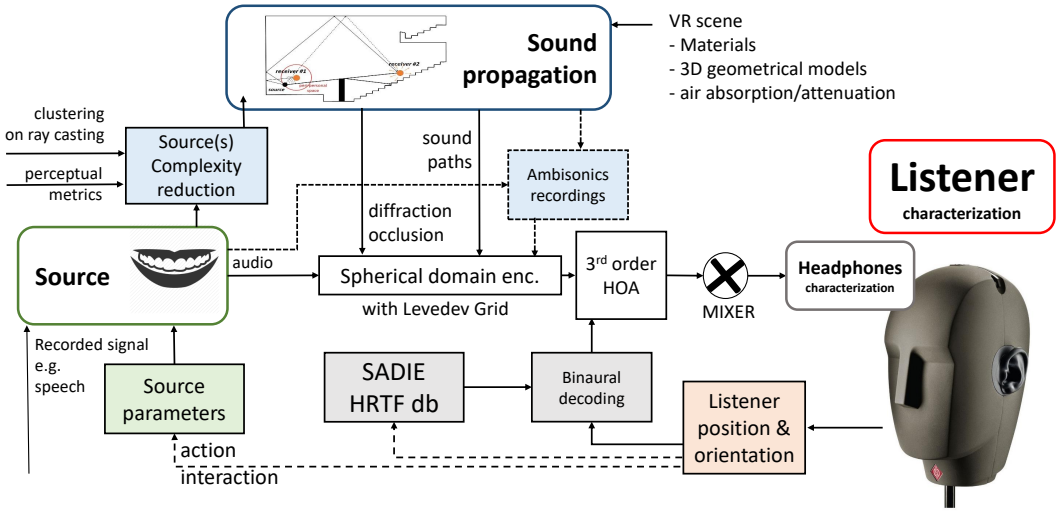


Fig. 3: Simplified block diagram for binaural rendering with Resonance Audio.

on the HTC Vive headset[3] connected to a 64 bit Desktop PC with an i7-4770K CPU, 16 GB of RAM, and a NVIDIA GEFORCE GTX 1070. Moreover, binocular Pupil Lab Eye Tracker[4] allowed the extraction of eye gaze data. For the audio reproduction, a pair of Sennheiser HD600 was combined with a headphone-specific equalization filter, which was convolved with low-latency through *Equalizer APO software*[5]. Such compensation filters were computed averaging the headphone impulse response measurements over more than 100 users from the Acoustic Research Institute of the Austrian Academy of Sciences[6]. The data are available in SOFA format [9]. This equalization process removes the acoustic headphone contribution, and thus reduces spectral coloration.

3.2 Binaural rendering of Ambisonics

For the development of a virtual prototyping framework able to simulate a sound field with multiple sources, and with independent control over their sound propagation, an accurate rendering of the spatial information around the listener is necessary. In this section, the study-specific encoding and binaural decoding of a virtual sound field are explained in order to provide perceptually plausible real-time simulations.

Ambisonics technique is here used for the reproduction of a full 3D virtual acoustical space [17]. A Higher Order Ambisonics (HOA) system provides a scalable solution for wave equation approximation with an increasingly accurate encoding of spatial information for a given three-dimensional volume of space [52]. The rendering process depicted in Fig. 3 consists of a decoder that can weigh the sound pressure in a spherical surface to a finite number of (virtual) loudspeakers array and/or binaural filters over headphones.

3.2.1 Spherical Harmonics

The spherical harmonics represent the sound decomposition into frequency, radial, and angular functions [71], into what leads to the Fourier-Bessel series for a position vector $\vec{r} = (r, \theta, \varphi)$ [17].

$$p(\vec{r}) = \sum_{m=0}^{\infty} i^m j_m(kr) \sum_{0 \le n \le m, \sigma = \pm 1} B_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta, \varphi) \tag{1}$$

with $i = \sqrt{-1}$, $k = 2\pi f/c$, c the speed of sound, and $\sigma \pm 1$ the spin value. For each term of the order m, a radial, spherical Bessel function $j_m(kr)$ is associated with angular functions $Y_m^{\sigma}(\theta, \varphi)$ called spherical harmonics, which define an orthonormal basis within a spherical coordinates system. B_{mn}^{σ} represents the projection of the acoustic pressure on this basis [17, 71].

[3]https://www.vive.com/us/product/vive-virtual-reality-system/
[4]https://pupil-labs.com
[5]https://sourceforge.net/projects/equalizerapo/
[6]http://sofacoustics.org/data/headphones/ari

The aim is the re-synthesis of sound sources from particular spatial directions, either by reproducing dedicated Ambisonics microphone recordings or synthetic signals. Considering an audio signal $f(t)$, which arrives from a certain direction, the representation of the surround audio signal $f(\theta, \varphi, t)$ is constructed using a spherical harmonic expansion up to a truncation order N.

$$f(\theta, \varphi, t) = \sum_{n=0}^{N} \sum_{m=-n}^{n} Y_n^m(\theta, \varphi) \phi_{nm}(t) \tag{2}$$

where Y_n^m represents the spherical harmonics of order n, degree m and $\phi_{nm}(t)$ the expansion coefficients. With increasing order N, the expansion results in a more precise spatial representation. Spherical harmonics are composed of a normalization term $N_n^{|m|}$, the associated Legendre function P_n^m and the trigonometric function [18, 41, 71].

$$Y_n^m(\theta, \varphi) = N_n^{|m|} P_n^m(sin(\varphi)) \begin{cases} sin(|m|\,\theta) & m < 0 \\ cos(|m|\,\theta) & m \ge 0 \end{cases} \tag{3}$$

$P_n^m(x)$ are the associated Legendre functions [73].

3.2.2 Higher Order Ambisonics

Considering our use of Resonance Audio implementation, Ambisonic Channel Numbering (ACN) and SN3D normalization are used in Eq. 2 and 3. ACN defines the ordering sequence for the spherical harmonics channels as $ACN = n^2 + n + m$. The normalization term used is SN3D, often seen in combination with ACN, takes the form of:

$$N_n^{|m|} = \sqrt{(2 - \delta_m)} \frac{(n - |m|)!}{(n + |m|)!} \tag{4}$$

with the Kronecker-delta δ_m is one for $m = 0$ and zero otherwise.

Using this index neatly defines a sequence for the spherical harmonics $Y_n^m(\theta, \varphi) = Y_{ACN}(\theta, \varphi)$ and the ambisonic signals $\phi_{ACN}(t)$ to stack them in the following vector

$$y(\theta, \varphi) = \begin{pmatrix} Y_0(\theta, \varphi) \\ \cdots \\ Y_{(N+1)^2-1}(\theta, \varphi) \end{pmatrix}. \tag{5}$$

The spherical domain components can be considered as the reconstruction of the wave field around the origin using a set number of microphones with multiple directivity patterns that define the magnitude of the signal and the direction of arrival. The higher the order of Ambisonics, the more directivity patterns are assumed with a narrowed region of sensibility and thus a higher spatial resolution could

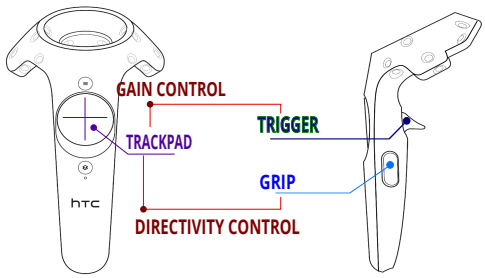


Fig. 4: Parameters control with HTC VIVE hand controller.

be rendered [17, 71]. HOA considers all the spherical domains above the truncation of N = 1. This representation requires $(N+1)^2$ spherical harmonics - HOA signals - and $(2N+1)$ channels for each Ambisonics order.

3.2.3 Binaural decoding

In the field of virtual, augmented, and mixed reality (VR/AR/MR), the binaural rendering of the sound field is in fact the most practical choice of reproduction with headphones. Two different approaches can be used for the binaural decoding [71] over headphones. The first method considers an array of virtual loudspeakers that would form the spherical reproduction as if it was an array of real loudspeakers, and assign two HRTFs to each loudspeaker, for each ear. The output signal for each ear will then be the sum of L loudspeaker signals $\sum_{n=0}^{N}\sum_{m=-n}^{n} B_n^m(\omega)D_{n,l}^m$ convolved with the corresponding HRTFs, $H_{l,left}(\omega)$,$H_{l,right}(\omega)$:

$$S_{ear}(\omega) = \sum_{l=1}^{L} H_l(\omega)\left(\sum_{n=0}^{N}\sum_{m=-n}^{n} B_n^m(\omega)D_{n,l}^m\right) \tag{6}$$

The second approach to binaural HOA reproduction consists of the pre-computation of the spherical harmonics-based HRTFs $H_n^m(\omega)$ by solving the equation:

$$H(\theta,\varphi,\omega) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} H_n^m(\omega)Y_n^m(\theta,\varphi), \tag{7}$$

where the spherical harmonics coefficients can be determined by direct integration or using the least square solution. The decoding process is defined by the channel arrangement only [52].

3.2.4 Resonance Audio implementation

In our implementation, Ambisonics was used for the rending of complete virtual sound sources with specific spatial information. Resonance Audio performed the encoding of the virtual audio sources, tracking their spatial position in terms of direction and distance and encoding it directly with 3^{rd} order HOA. ACN channel ordering in combination with SN3D normalization were employed. The encoding in the spherical domain is built using pre-computation of HRTFs in the spherical domain, allowing to project multiple sound objects in the Ambisonics sound field. In the case of the 3^{rd} order Ambisonics, a Lebedev grid was used to distribute the matrix of virtual loudspeakers in the most uniform way. From this grid, a matrix of seventeen angles was used in Matlab to pre-compute the spherical harmonics [37, 44] from the HRTF set (1550 measurements) of the Neumann KU 100 dummy head provide by the SADIE database [37].

3.3 Interaction metaphors

This case study considered the simulation of a virtual microphone array equipped with an ideal and versatile virtual beamformer at the user's ears. The system was capable of storing the a priori knowledge for each virtual source, independently. The directivity model of the virtual beamformer took the form of polar pattern coefficients in Ambisonics domain resulted from the following equation:

$$P(\theta) = g\,|(1-\alpha+\alpha cos(\theta))|^{\gamma} \tag{8}$$

where the users can dynamically manipulate the following three parameters, reinforcing the natural/intuitive use of beamforming control for enhancing directional hearing:

- *gain level, g*: gain factor applied to the source signal at the specific θ, in a continuous scale from -20 dB FS and 20 dB FS;
- *directivity alpha, α*: from a full omnidirectional pattern ($\alpha = 0$) to a maximum of a bipolar pattern ($\alpha = 1$), being $\alpha = 0.5$ equivalent to a perfect cardioid sensibility pattern;
- *sharpness, γ*: from a value of $\gamma = 1$ equivalent to the natural sensibility and $\gamma = 10$ the maximum-allowed value for narrowing the beam width.

During a calibration period these values could be changed by using the hand controller and a specific mapping (see Fig. 4 for a graphical representation). The user was able to adjust the directivity parameters and the gain parameters, independently. The grip and up or down in the trackpad would change the directivity pattern, and left and right would affect the width of the pattern. On the other hand, trigger and up or down on the trackpad would increase or decrease the gain value of the associated source. According to Hamacher et al. [26], a linear smooth function was introduced by changing the focus source, enabling a progressive transition between the new selected source and the previous one within approximately half a second.

We developed the beamforming interactions starting from the standard control by head orientation. We extended it with a recently implemented [7, 23] and more natural/embodied control by eye-gaze. Finally, we defined an artificial interaction considering two simultaneous beamformers controlled by head and hand, respectively. Since this choice aimed at exploring superhuman hearing abilities for the specific pairing task, the introduction of a second beamformer is a straightforward path towards an effective extension of available hearing-aid tools.

Head-guided control (H): When using the head interaction, the user was able to choose/focus on a specific target by rotating and positioning the head directly in front of the area of interest. The listener's area is divided into as many areas of interest as the number of stimuli that should be known and tracked by the virtual system. To avoid a sort of selection interference when the participant moves the head, the algorithm waited one second before changing the source in focus. The listener was also able to freely move the head without worrying about constantly selecting random focus sources.

Eye-guided control (HE): The eye control considered eye positions in the two-dimensional projected planes resulting from the intersection of the eye gaze with position of the sources in the space. A collision vector between the eye gazing point and virtual objects in the VE allows the listener to select an audio source by looking directly at it. This interaction enabled the participant to choose between one out of two sources in the same field of view. It is worthwhile to notice that the eye movement is still dependent on the head rotation to be able to select sources from behind. This is the reason why this beamformer interaction was labelled head plus eye (i.e., HE).

Head-plus-controller for a dual beamformer (HC): The HTC Vive controller was used for pointer interaction and the participant could select one source of interest. The controller was tracked in space and its relative position to the focus sources around the listener was computed according to the occupied area by the controller, always connected to a certain source as in the H interaction. The controller had an instantaneous selection timing, so the user could very fast change between focus sources. Since the controller might be considered an extension of the body, it might not be directly affected by the dominance of the head. Accordingly, the participant could be facing one source of interest while pointing the controller at another source. The virtual system managed the information from both the artificially selected source and the head-centric selection, combining parameters from the two beamformers.

4 The User Study

The main aim of the user study was to investigate if the three virtual beamformers could support auditory attention and have an impact on performances in highly challenging listening situations with different levels of visual information.

We conducted a first test where the users were required to complete a small training session (Sec. 4.2.1) and calibration of the directivity parameters for all virtual sound sources (Sec. 4.2.2), followed by a speech pairing task in an abstract VE (Sec. 4.2.3). At this first stage, the goal was to offer the participants a minimal visual environment where the main cues were restrictively auditory by design. The second test was built so that participants performed the task within a 360° video recording where they were surrounded by real speakers in order to recreate a realistic and multimodal (audio-visual) scenario. The main research questions related to this user study are:

- RQ1: What are the main preferences for directivity parameters?
- RQ2: Can users modulate their performance in the presence of realistic visual references?
- RQ3: Will user performance differ between interaction techniques?

We collected data from 17 participants (age 28 $\pm$ 6 years) with self-reported normal hearing. All participants were informed about the experiment and gave consent to data collection.

4.1 Stimuli

The first immersive VE consisted of eight blue virtual cubes defining eight abstract sound sources in a minimalistic open space with light blue sky and gray floor (see Fig. 1a). The cube directly in front of the listener (i.e. the focus source) changed color dynamically based on the user's spatial position when rotating in the scene. Virtual sources were distributed in a perfect circle, separated by a 45° angular distance.

The sound sources were simultaneously played in loops and comprised sets of sentences available from IEEE Recommended Practice for Speech Quality Measurements [1], also known as the Harvard Sentences, used in many different fields of audio engineering (e.g. speech-to-text software and cochlear implants testing). These sentences are considered standard and an optimal research material because all the word lists are phonetically balanced, meaning that the frequency of sounds in these lists corresponds to that of natural language and conversations. The sentences are short and simple; that is, monosyllabic words punctuated by exactly a single two syllable word sentence.

For the 360° video recording setup, 8 concurrent human speakers were placed around the listener's position in an anechoic room every 45° starting from directly in front (see Fig. 1c for a top view map of the recordings imported in unity). The recording contained a 360° video, captured with *Garmin Virb 360* camera[7], of the speakers reading randomly selected Harvard Sentences, and the audio signal from each speaker was recorded with clip microphones, DPA SC4060[8] (see Fig. 5 for two pictures from the recording sessions). The signal cross-feed between microphones was ≈15 dB and it was later reduced with equalization and compression to ≈10 dB difference between the main signal and interfering signal. After the post-editing of both videos and audio signals, the auditory speech stimuli were calibrated at 60 dB SPL, equivalent to a conversational signal level.

For both abstract and realistic scenarios, there were four pairs of two voices: a female and a male speaker, respectively. For each pair, seven sentences (sets hereafter) were selected from Harvard Sentences and synchronously assigned to both sources/speakers. Such a total of 28 sentences were the basis for the creation of seven balanced configurations assigned to participants in our study: sentences order for each set, female-male pairing, and the combination of a set with a pair during the voice synthesis/recording were made following latin squares.

4.2 Protocols

Within-subjects task-based tests were conducted in order to identify meaningful interactions within the two VEs. Participants were asked to find pairs of the same sequences in a group of eight speakers. They were asked to change the shape and the width of a directivity beam for several directions optimizing each interaction technique introduced in Sec.3.3. The three different types of interaction were individually tested by each participant in a within-subject experiment. Moreover, a normal listening condition, (i.e., no hearing aids tool provided) were included to test the level of difficulty of the proposed task. The experiment procedure consisted of two different test blocks with multiple days between the two tests in order to reduce learning effects (see Fig. 6a and 6b for protocol schematics). The first session accounted for training, calibration and pairing task with abstract visuals. The second part evaluated the task difficulty and the realistic VE.

[7]https://buy.garmin.com/da-DK/DK/p/562010

[8]https://www.dpamicrophones.com/dscreet/4060-series-miniature-omnidirectional-microphone

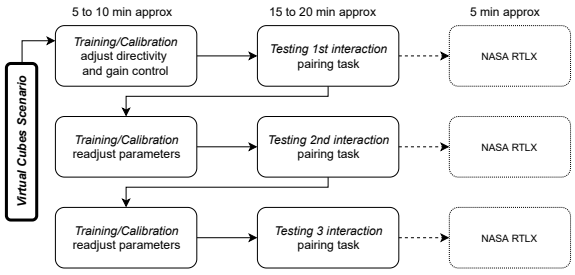
Fig. 5: Recording sessions and positions of the eight speakers.

4.2.1 Training

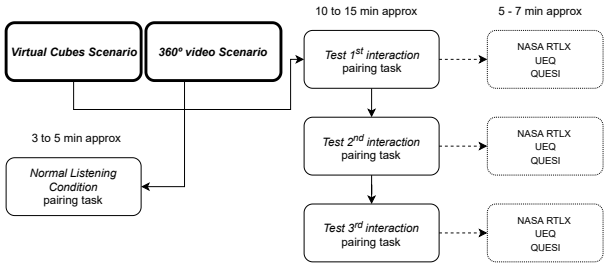
Participants were asked to do a brief tutorial where they became familiar with the HTC Vive platform and the audible effects of changing both directivity and gain values with three different sources in the space. This training procedure helped reduce biases introduced by different confidence levels with VR technologies. We defined four levels:

- (1) an introduction to the hand controller, where the participants were asked to increase the volume for the focus source and shape the directivity of another cube into a very narrow bipolar pattern;
- (2) - (4) the virtual beamformer was already calibrated with arbitrary values for both gain and directivity so that the subject would only need to try to move around using the specific interaction control to select the focus source among the increasing number of concurrent sources.

This training procedure could be repeated as often as necessary.



(a) Part 1



(b) Part 2

Fig. 6: User study, part 1 and 2: protocol schematics.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2020.2973059, IEEE Transactions on Visualization and Computer Graphics

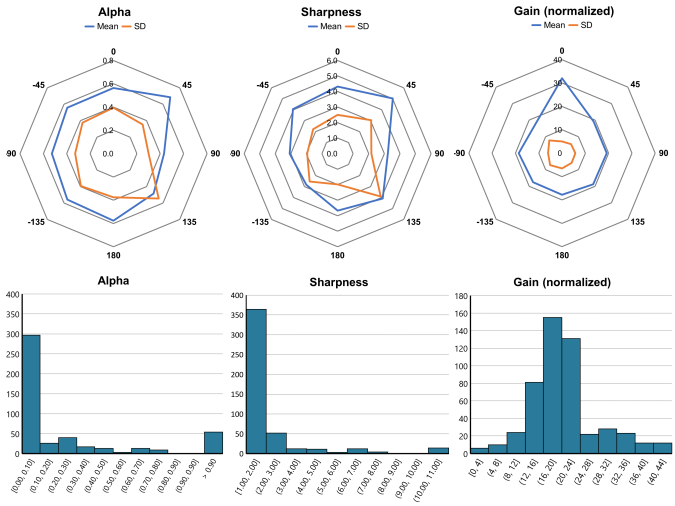


Fig. 7: Top: Radar charts showing the means and standard deviations pertaining to alpha, sharpness and normalized gain. Bottom: histograms showing the distribution of calibration values pertaining to alpha, sharpness and normalized gain. For the sake of visualization, gain values were normalized with an additive constant equals to 20 dB. The 0 db FS corresponds to this value in this chart, and to 60 db SPL.

4.2.2 Calibration

The high variability in auditory spatial perception [25] and cognition [4] among listeners requires to consider biases due to user auditory profiles that might be difficult to model, negatively affecting the final results. Accordingly, participants performed a calibration in VR for each interaction metaphor in the abstract VE surrounded by all eight virtual cubes. The main goal was to identify optimal values of directivity parameters towards each source in order to enhance the natural intelligibility of the frontal sound source (i.e., the focus source) on individual basis. Accordingly, participants were asked to change the values of gain level, directivity alpha, and sharpness of the virtual beamformer related to the focus source, so that they could clearly understand what was being said by the focus voice. Moreover, the calibration procedure also allowed to select direction-dependent parameters for the remaining sources: participants pointed the hand controller towards each masking source and manipulated a sub-beam according to the mapping of Fig. 4 while maintaining the selection on the focus source with head/eye control.[9]

The resulting virtual beamformer could potentially exhibit a multi-lobe radiation pattern. Its parameter values were then used to understand preference distributions, and to practically compensate different user sensitivities during the tests. Since our user population had a limited size, our approach was considered less error prone than considering fixed parameter values that would lead to hardly explainable listener-beamformer mixed effects.

4.2.3 Test

For each beamforming interaction, participants were asked to pair the sources matching the set of sentences. After exposure to both abstract and realistic visuals, the perceived workload was assessed using a modified version of the NASA task load index (TLX) [27] called the NASA Raw TLX (RTLX) [28] in order to identify usability issues on performances due to the beamforming control. The questionnaire included six 5-point rating scales, ranging from 1 to 5 (high ratings indicated high task load). The questionnaire was administered after exposure to each condition and yielded an aggregate score serving as an estimate of the overall perceived workload. Two additional questionnaires were administered after exposure to the realistic scenario. That is, participants' confidence with the three interaction techniques was also assessed by means of the short version of the *user experience questionnaire* (UEQ-S) [64] and the *questionnaire for the subjective consequences of intuitive use* (QUESI) [51]. The UEQ-S includes eight of the 26 semantic difference scales of the original UEQ [43], and the mean of the participants responses to the eight items, the *UX score*, is viewed as a measure of the overall user experience. The QUESI includes 14 items, organized into five subscales pertaining to subjective mental workload, perceived achievement of goals, perceived effort of learning, familiarity, and perceived error rate. The mean of the five subscales, the *QUESI score*, is taken as a measure of the how intuitive participants found the interaction. For the sake of consistency all three questionnaires included 5-point rating scales, where high scores indicated high perceived workload (RTLX), an overall positive user experience (UEQ), and high intuitiveness (QUESI).

The presentation order of the three interaction metaphors was randomized. Participants were always informed about the current interaction metaphor, but no further information regarding the positions of the pairs or which pairs were correct was given.

[9] In HC, the hand controller did not work as an interaction tool for beam control but acted as remote controller for an additional H parameter adjustments.

5 Data Analysis and Results

It is very relevant to note that none of the participants were able to find the first pair (and thus the remaining pairs) in the natural listening condition within the VEs before quitting the experimental session for the high level of difficulty. No statistical analysis was based on this information, even if this result certified the impracticability of such pairing task without artificial hearing support.

The analyzed metrics were:

- **correct pairing**: number of correct pairs found during exposure to each interaction metaphor;
- **pairing time**: time spent exploring the scene while finding the next pair of speakers;
- **pairing action time**: the time required to actually select the two speakers identified as a pair during the pairing time.

All data were treated as interval or ratio data. Significant outliers were identified based on the inspection of boxplots, and Shapiro-Wilk's tests were used to determine if the data were normally distributed. If no outliers were detected and the data was assumed to be normally distributed, one-way repeated-measures ANOVAs were used for statistical comparison. Alternatively, non-parametric Friedman tests were used for statistical analysis, and pairwise comparisons performed using Dunn-Bonferroni tests. All cross-study comparisons were performed using non-parametric methods due to violations in the normality assumption. Specifically, for each measure three Mann-Whitney U tests were used to compare the results obtained during exposure to H, HE and HC in the study involving abstract visuals with the corresponding conditions in the study involving realistic stimuli.

In case of four of the 18 participants exposed to the abstract visuals and four of the 17 participants exposed to realistic visuals, logging errors prevented us from obtaining information about the pairing data. Because these errors did not affect the participants' experiences, the analyses of the questionnaire data involved the full samples. However, the same analyses were also run on the reduced samples for abstract visuals (n = 14) and realistic visuals (n=13). Only discrepancies between the two sets were reported along with the analyses of the full sample.

5.1 Calibration:

In Fig. 7 (bottom), histograms show the data distributions of alpha, sharpness and gain values of the calibration procedure. It is worthwhile to notice that most participants kept the starting values for alpha (0) and sharpness (1) changing only the gain values for each considered direction of the virtual beamformer, shaping their own polar pattern. Accordingly, we conducted the statistical analysis for the gain parameter only. Fig. 7 (right) shows the gain data. They were normally distributed in case of all but one source (i.e., 90° direction). However, significant outliers were found with respect to all sources, except from frontal direction (0°). Thus, the data was analyzed using a non-parametric Friedman test. The test indicated that the median normalized gain differed significantly between sources, $X^2(2) = 47.860, p < .001$. Pairwise comparisons using Dunn-Bonferonni tests indicated that median

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2020.2973059, IEEE Transactions on Visualization and Computer Graphics

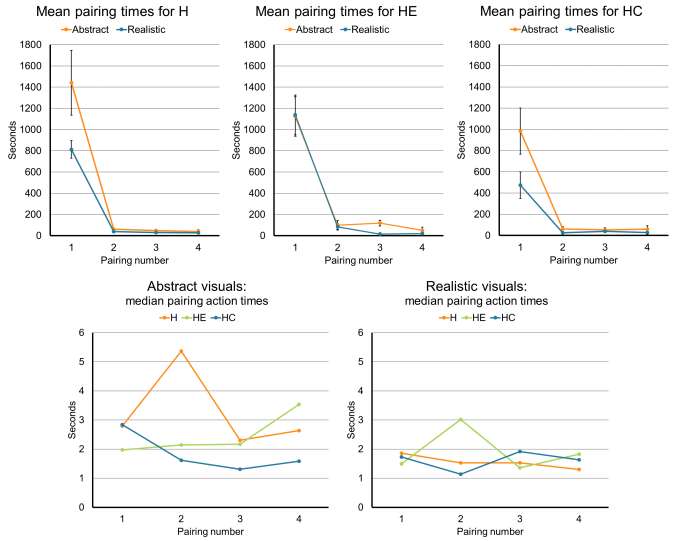


Fig. 8: Top: Mean pairing times across the four pairings for the three conditions HC, H, and HE. Error bars indicate ± 1 SE. Bottom: The median pairing times across the four pairings for abstract visuals (left) and realistic visuals (right).

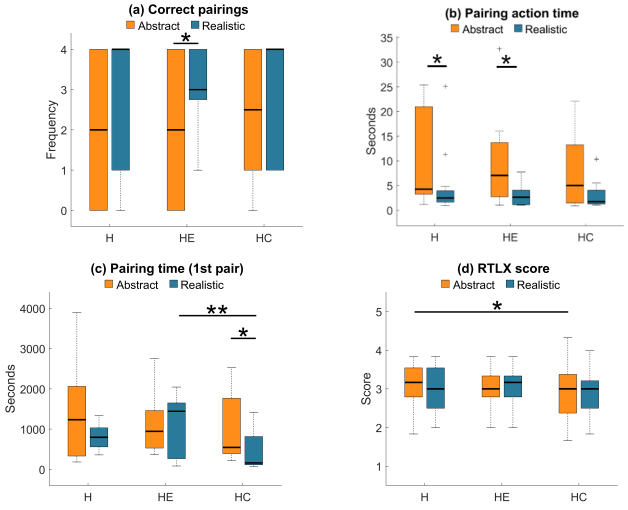


Fig. 9: Boxplots visualizing results for (a) correct pairings, (b) mean pairing action times, (c) first pairing times, and (d) RTLX scores in terms of medians, interquartile ranges, minimum and maximum values, and outliers. Asterisks indicate, where present, a significant difference (*: $p < .05$, **: $p < .01$, ***: $p < .001$ at *post-hoc* test).

normalized gain of direction 0° (Mdn = 13.0 dB) was significantly higher than direction −45° (Mdn = -2,7 dB, $p > .001$), −90° (Mdn = -1.9 dB, $p > .001$), −135° (Mdn = -2 dB, $p > .001$), 180° (-2.7 dB, $p > .001$), 135° (Mdn = -1 dB, $p = .004$), and 45° (Mdn = -1.3 dB, $p = .001$). No significant difference between directions 0° and 90° (Mdn = -1.2 dB) was found, and none of the other sources differed significantly from each other.

5.2 Paring task performances

Fig. 9a shows the results pertaining to the number of *correct pairings* during exposure to abstract and realistic visuals. In regard to abstract visuals, the Friedman test found no significant difference between conditions, $X^2(2) = 1.687, p = .430$. Similarly, the Friedman test used to compare the data from the realistic scenario found no significant difference between conditions ($X^2(2) = 1.226, p = .542$). As apparent, the number of correct pairings was higher when the participants were exposed to realistic visuals across all three conditions. However, the Mann-Whitney U tests only indicated that the median difference was significant when comparing HE across the study realistic visuals (Mdn = 3.0) and the study involving abstract visuals (Mdn = 2.0) , $U = 146, z = 2.309, p = .021$.

In Fig. 8 (top), average pairing time in sequential pairing reveals that the *first pairing* can be clearly distinguished from the subsequent activities. Participants efforts were mainly involved in this first period, which heavily contributes to timing performances in both scenarios. Fig. 9c shows the results pertaining to the first pairing time during exposure to abstract visuals and realistic visuals. The Friedman test revealed no significant difference in pairing time between conditions in the abstract visual scenario, $X^2(2) = 1.000, p = .607$. However, a Friedman test indicated that interactions in realistic condition differed significantly in terms of pairing time, $X^2(2) = 8.769, p = .012$, and the pairwise comparisons indicated that median pairing time for HC (Mdn = 160.8) was significantly lower than the median pairing time of HE (Mdn = 1444.2), $p = .010$. Moreover, the performed Mann-Whitney U tests indicated that pairing times with HC were significantly higher during exposure to abstract visuals (Mdn=546.3) compared to realistic visuals (Mdn=160.9), $U = 47.0, z = -2.135, p = .033$. No significant differences between studies were found with respect to H and HE.

The action of pairing two speakers did not differ in the sequential pairing order due to short action time and high variability among participants. Fig. 8(bottom) displays median values. For this reason, the *pairing action time* was computed without considering pairing order and Fig. 9b shows the results during exposure to abstract visuals ($n = 14$) and realistic visuals ($n = 13$). In regard to the abstract visual scenario, the Friedman test revealed no significant effect of interaction technique on pairing action time, $X^2(2) = 1.857, p = .395$. Moreover, the Friedman test used to compare the results pertaining to the realistic scenario, found no significant difference between conditions, $X^2(2) = 4.154, p = .125$. The Mann-Whitney U tests found no significant difference between HC combined with abstract visuals (Mdn = 5.0) and HC combined with realistic visuals (Mdn = 1.7), $U = 63.0, z = -1.359, p = .185$. However, with respect to H, the medians pairing action time was significantly higher during exposure to abstract visuals (Mdn = 4.2) compared to realistic visuals (Mdn = 2.4), $U = 48.0, z = -2.087, p = .038$. Similarly, with respect to HE, the median pairing action time was significantly higher during exposure to abstract visuals (Mdn = 7.0) compared to realistic visuals (Mdn = 2.6), $U = 42.0, z = -2.378, p = .017$.

5.3 Questionnaires

Fig. 9d shows the results of the NASA RTLX related to the study involving abstract visuals ($n = 18$) and realistic visuals ($n = 17$). A Friedman test indicated that participants answers with abstract visuals were statistically significantly different between conditions, $X^2(2) = 8.842, p = .012$, and pairwise comparisons using Dunn-Bonferroni identified a significant difference between HC (Mdn = 3.00) and H (Mdn = 3.25), $p = .037$, indicating that the perceived task load was slightly higher when the participants relied on head-based only interaction. However, these results were only marginally significant when the same analysis was performed on the reduced sample (n=14), $X^2(2) = 5.915, p = .052$. The questionnaire data related to the realistic scenario were normally distributed, as assessed by Shapiro-Wilk's test ($p > .05$), no significant outliers were identified, and Mauchly's test indicated that the assumption of sphericity was met, $X^2(2) = 1.089, p = .580$. However, the one-way repeated measures ANOVA found no significant difference in RTLX scores between conditions, $F(2, 32) = 1.123, p = .338$. Moreover, the three Mann-Whitney U tests did not reveal any significant differences in regard aggregate RTLX scores.

Fig. 10a shows the results related to aggregate UX scores obtained from the UEQ-S administered after exposure to realistic visuals. The data were normally distributed, as assessed by Shapiro-Wilk's test ($p > .05$), no significant outliers were identified, and Mauchly's test indicated that the assumption of sphericity was met, $X^2(2) = 2.739, p = .254$. However, a one-way repeated measures

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2020.2973059, IEEE Transactions on Visualization and Computer Graphics

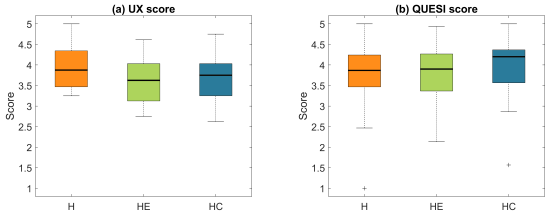


Fig. 10: Boxplots visualizing results related to (a) the UEQ-S questionnaire and (b) the QUESI questionnaire in terms of medians, interquartile ranges, minimum and maximum values, and outliers.

ANOVA found no significant difference in scores between conditions, $F(2,32) = 1.906, p = .165$. Finally, Fig. 10b shows the results related to aggregate QUESI score obtained from the QUESI administered after exposure to realistic visuals. A non-parametric Friedman test was used for analysis. The test did not find a significant difference in scores between the three conditions, $X^2(2) = 1.529, p = .465$.

6 General Discussion

In this study, we decided to let participants decide the shape of their virtual beamformer supported by the great flexibility of a calibration in VR. Answering to RQ1 (directivity parameterization), they mainly adjusted the directional gain (see Fig. 7(bottom)), slightly changing the remaining parameters. No differentiations among interaction techniques were reported. Accordingly, one can assume that all individual parameters referred to the dominance of head-guided interaction metaphors. The typical calibration strategy primarily involved gain tuning for all available directions, followed by changes in alpha and sharpness on an individual basis. Such direction specific modulation resulted in a narrow polar lobe of the virtual beamformer, i.e. $\alpha \approx 0.5$ meaning a cardioid shape in combination with $\gamma > 1$. For many participants, a SNR improvement ranging from 14 to 16 dB resulted from gain adjustments was already satisfactory in their virtual beamformer and in good agreement with the most elevated enhancement reported in the scientific literature [21]. Interestingly, there is also a general trend in parameter tuning for gain and for those participants who also consider α and γ (see radar charts in Fig. 7): right directions exhibited less attenuation, less interference from adjacent directions (i.e., hypercardiod shape), and narrower beams. The average behavior of this virtual beamformer supported the lateralization of auditory attention and the right-ear advantage [30]. However, one can argue that directivity and sharpness parameters were not very useful for the given task, or users did not know how to take advantage of them. A future experimental session might consider an extensive training phase where users might be forced to manipulate one parameter at a time. Preferred and most useful parameter configurations could influence paring strategies.

Evaluating the same task with different levels of visual information has an important methodological validity for multisensory integration of auditory stimuli [31] which drives relevant research questions such as RQ2. In our study, the head-guided beamforming seemed to be highly influenced by the coordination of eye and head-movements, that facilitated human speech processing within a realistic audio-visual scene [68]. The statistical differentiation in pairing action time between abstract and realistic visuals for H and HE confirmed the improved timing performances in space orientation. On the other hand, HC did not follow the same integration mechanism because participants already had the focus on both speakers of a pair thanks to the virtual dual beamformer (i.e. head and hand was spatially anchored to the two identified speakers). Consequently, HC did not require advantages from visual anchors in the pairing action. Moreover, the statistically significant improvement in correct number of pairs in HE (and a reduced variability compared to H and HC) with realistic visual details attested to the increase in reliability while coordinating head and eye.

RQ3 is closely related to RQ2 assuming that it was mainly related to the exploration period in the first pairing which is crucial for finding subsequent pairs and thus for the final performance. The root interaction H showed a positive trend for the first pairing time from abstract to realistic visual references, strengthened by the combination with a hand controller (HC) that exhibited a statistically significant differentiation unlike H alone. However, H and HC did not statistically differ in this metric suggesting future analyses with more participants and alternative performance metrics such as attention switching [53] and head movement quantity. On the other hand, the eye interaction resulted in a tendency of slower exploration. Gaze orientation towards a specific source required more time due to inefficiency in the eye-tracker and unoptimized transitions among focused directions that might require ad-hoc adjustments through gaze-contingent experiments [63]. This latter aspect will be subjected to future investigations.

Summarizing the outcomes regarding the virtual beamforming interactions, HC was the most efficient in the pairing task. The implementation of a dual beamformer could be considered a new artificial feature, perfectly matching task requirements with a superhuman hearing motivation. Moreover, task load decreased in the most challenging situation with minimal visual feedback (Fig. 9d). Since HC resulted in the more versatile technique, H suffered from the lack of visual details in terms of workload and pairing action time. HE closely derived from H, inheriting similar drawbacks at least in this speaker configuration. In particular, the effect of source displacement could be easily changed in future sessions following our methodology of virtual prototyping. High values on UX and QUESI questionnaires supported a future real development of such techniques. For the product design perspective in hearing aid industry, implementing eye tracking would require new technologies such as electrooculogram (EOG) which measures eye movements based on skin mounted electrodes at a high frame-rate [23]. Similarly, supporting hands gesture control would require features comparable to mobile AR/MR devices (see [38] for recent trends in AR research).

Finally, user characterization in terms of listening abilities and pairing strategy will be analyzed within a larger pool of participants, supporting skills for the control of any new virtual beamformers.

7 Conclusion

This study provides a methodological and technological framework for virtual prototyping of the new generation of artificial and augmented hearing devices. Our main motivation is the identification of ideal sonic interactions in VR before investing resources in the actual development of real hardware/software technologies to be included in hearing aids and hearables. This case study disclosed potential superhuman abilities in the form of a virtual acoustic beamformer for the complex cocktail party problem. In particular, a task-specific beamforming control could support an effective experience in challenging (nearly impossible) listening situations. Interactions with a dual virtual beamformer resulted the best solution for a speaker pairing task in multi-source scenario.

The VR framework could be employed in different listening situations considering irregular displacement of listeners and different levels of visual information. Non-anechoic conditions will be considered to evaluate the impact of room acoustics which is a well-known problem for example classroom acoustics [6]. Moreover, dynamic rendering of VEs [48] and the addition of user walking [12] will be extremely relevant in evaluating superhuman hearing tools in a complex system, which will be more and more similar to reality.

It is worthwhile to notice that this virtual prototyping approach could easily integrate non-ideal or realistic virtual beamformers. Ambisonics encoding could incorporate acoustic measurements of directivity patterns from real hearing aids, substituting the pre-computation of spherical-harmonics based HRTFs in Eq. 7. Finally, the assumption of a priori knowledge of all virtual sound sources could be substituted by machine learning algorithms for auditory scene analysis [67]. With the simulated sound field acting as input, the virtual beamformer will be capable of detecting objects, classifying scenes and events.

Acknowledgments

This study was supported by the internationalization grant of the 2016-2021 strategic program "Knowledge for the World" awarded by Aalborg University to Michele Geronazzo.

REFERENCES

[1] Ieee recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17:225–246, 1969.

[2] A. Ahrens, K. D. Lund, and T. Dau. Audio-visual scene analysis in reverberant multi-talker environments. In *In Proc. 23rd Int. Congress on Acoustics*, pages 3890–3896, Aachen, DE, Sept. 2019.

[3] G. Alce, E.-M. Ternblad, and M. Wallergård. Design and Evaluation of Three Interaction Models for Manipulating Internet of Things (IoT) Devices in Virtual Reality. In D. Lamas, F. Loizides, L. Nacke, H. Petrie, M. Winckler, and P. Zaphiris, editors, *Human-Computer Interaction – INTERACT 2019*, Lecture Notes in Computer Science, pages 267–286. Springer International Publishing, 2019.

[4] A. Andreasen, M. Geronazzo, N. C. Nilsson, J. Zovnercuka, K. Konovalov, and S. Serafin. Auditory feedback for navigation with echoes in virtual environments: training procedure and orientation strategies. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):1876–1886, May 2019.

[5] A. Andreasen, J. Zovnercuka, K. Konovalovs, M. Geronazzo, R. Paisa, and S. Serafin. Navigate as a bat. real-time echolocation system in virtual reality. In *Proc. 15th Int. Conf. Sound and Music Computing (SMC 2018)*, pages 198–205, Cyprus, July 2018.

[6] Berg Frederick S., Blair James C., and Benson Peggy V. Classroom Acoustics. *Language, Speech, and Hearing Services in Schools*, 27(1):16–20, Jan. 1996.

[7] V. Best, E. Roverud, T. Streeter, C. R. Mason, and G. Kidd. The Benefit of a Visually Guided Beamformer in a Dynamic Speech Task. *Trends in Hearing*, 21, July 2017.

[8] J. Blauert. Analysis and Synthesis of Auditory Scenes. In J. Blauert, editor, *Communication Acoustics*, pages 1–25. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[9] B. B. Boren, M. Geronazzo, P. Majdak, and E. Choueiri. Phona: a public dataset of measured headphone transfer functions. In *Proc. 137th Conv. Audio Engineering. Society*. Audio Engineering Society, Oct. 2014.

[10] A. S. Bregman. *Auditory scene analysis: the perceptual organization of sound*. MIT Press, Cambridge, Mass., 1990.

[11] A. W. Bronkhorst. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*, 2000.

[12] D. S. Brungart, S. E. Kruger, T. Kwiatkowski, T. Heil, and J. Cohen. The Effect of Walking on Auditory Localization, Visual Discrimination, and Aurally Aided Visual Search. *Human Factors*, page 0018720819831092, Mar. 2019.

[13] S. L. Calvert and S.-L. Tan. Impact of virtual reality on young adults' physiological arousal and aggressive thoughts: Interaction versus observation. *Journal of Applied Developmental Psychology*, 15(1):125–139, Jan. 1994.

[14] E. C. Cherry. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, 1953.

[15] J. Christensen and J. Hald. *Beamforming*. Bruel and Kjaer Technical Review, 2004.

[16] A. Covaci, D. Kramer, J. C. Augusto, S. Rus, and A. Braun. Assessing Real World Imagery in Virtual Environments for People with Cognitive Disabilities. pages 41–48. IEEE, July 2015.

[17] J. Daniel. Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format. *AES 23rd International Conference*, 2003.

[18] J. Daniel, S. Moreau, and R. Nicol. Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging. *AES 114th Convention, Audio Engineering Society*, 2003.

[19] E. Degli Innocenti, M. Geronazzo, D. Vescovi, R. Nordahl, S. Serafin, L. A. Ludovico, and F. Avanzini. Mobile virtual reality for musical genre learning in primary education. *Computers & Education*, 139:102–117, Oct. 2019.

[20] H. Dillon. Hearing aids. *Boomerang Press, Sydney*, 2001.

[21] S. Doclo, S. Gannot, M. Moonen, and A. Spriet. Acoustic Beamforming for Hearing Aid Applications. In S. Haykin and K. J. R. Liu, editors, *Handbook on Array Processing and Sensor Networks*, pages 269–302. John Wiley & Sons, Inc., Hoboken, NJ, USA, Apr. 2010.

[22] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie. Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices: Advantages and limitations of existing tools. *IEEE Signal Processing Magazine*, 32(2):114–124, Mar. 2015.

[23] A. Favre-Félix, C. Graversen, R. K. Hietkamp, T. Dau, and T. Lunner. Improving Speech Intelligibility by Hearing Aid Eye-Gaze Steering: Conditions With Head Fixated in a Multitalker Environment. *Trends in Hearing*, 22, Dec. 2018.

[24] T. Gallaudet and C. de Moustier. On optimal shading for arrays of irregularly-spaced or noncoplanar elements. *IEEE Journal of Oceanic Engineering*, 25(4):553–567, Oct. 2000.

[25] M. Geronazzo, E. Sikström, J. Kleimola, F. Avanzini, A. De Götzen, and S. Serafin. The impact of an accurate vertical localization with HRTFs on short explorations of immersive virtual reality scenarios. In *Proc. 17th IEEE/ACM Int. Symposium on Mixed and Augmented Reality (ISMAR)*, pages 90–97, Munich, Germany, Oct. 2018. IEEE Computer Society.

[26] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass. Signal processing in high-end hearing aids: State of the art, challenges, and future trends. *EURASIP Journal on Applied Signal Processing*, 2005.

[27] S. G. Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage Publications Sage CA: Los Angeles, CA, 2006.

[28] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[29] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties. MPEG-H 3d Audio – The New Standard for Coding of Immersive Spatial Audio. *IEEE Journal of Selected Topics in Signal Processing*, 9(5):770–779, Aug. 2015.

[30] M. Hiscock and M. Kinsbourne. Attention and the right-ear advantage: What is the connection? *Brain and Cognition*, 76(2):263–275, July 2011.

[31] N. P. Holmes and C. Spence. Multisensory integration: Space, time, & superadditivity. *Curr Biol*, 15(18):R762–R764, Sept. 2005.

[32] A. Ihlefeld and B. Shinn-Cunningham. Disentangling the effects of spatial cues on selection and formation of auditory objects. *The Journal of the Acoustical Society of America*, 124(4):2224–2235, 2008.

[33] J. C. L. Ingram. *Neurolinguistics: an introduction to spoken language processing and its disorders*. Cambridge University Press, 2007.

[34] S. Jayaram, H. I. Connacher, and K. W. Lyons. Virtual assembly using virtual reality techniques. *Computer-Aided Design*, 29(8):575–584, Aug. 1997.

[35] D. H. Johnson and D. E. Dudgeon. *Array Signal Processing: Concepts and Techniques*. Simon and Schuster, Inc, 1992.

[36] J. M. Kates. Superdirective arrays for hearing aids. *The Journal of the Acoustical Society of America*, 94(4):1930–1933, Oct. 1993.

[37] G. Kearney and T. Doyle. An HRTF Database for Virtual Loudspeaker Rendering. In *Proc. of the Audio Engineering Society Convention 139*. Audio Engineering Society, Oct. 2015.

[38] K. Kim, M. Billinghurst, G. Bruder, H. B.-L. Duh, and G. F. Welch. Revisiting Trends in Augmented Reality Research: A Review of the 2nd Decade of ISMAR (2008–2017). *IEEE Transactions on Visualization and Computer Graphics*, 24(11):2947–2962, Nov. 2018.

[39] D. Kimura. Cerebral dominance and the perception of verbal stimuli. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 15(3):166–171, 1961.

[40] B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, and J. Verhey, editors. *Hearing - From Sensory Processing to Perception*. Springer-Verlag, Berlin Heidelberg, 2007.

[41] M. Kronlachner and F. Zotter. Spatial transformations for the enhancement of ambisonic recordings. *2nd Int. Conf. on Spatial Audio (ICSA)*, 2014.

[42] J. N. Latta and D. J. Oberg. A conceptual virtual reality model. *IEEE Computer Graphics and Applications*, 14(1):23–29, Jan. 1994.

[43] B. Laugwitz, T. Held, and M. Schrepp. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and Usability Engineering Group*, pages 63–76. Springer, 2008.

[44] P. Lecomte, P.-A. Gauthier, C. Langrenne, A. Garcia, and A. Berry. On the use of a lebedev grid for ambisonics. *The Journal of the Acoustical Society of America*, 2015.

[45] L. H. Loiselle, M. F. Dorman, W. A. Yost, S. J. Cook, and R. H. Gifford. Using ild or itd cues for sound source localization and speech understanding in a complex listening environment by listeners with bilateral and with hearing-preservation cochlear implants. *Journal of Speech, Language and Hearing Research*, 2016.

[46] D. Ma, J. Gausemeier, X. Fan, and M. Grafe, editors. *Virtual Reality & Augmented Reality in Industry*. Springer-Verlag, Berlin Heidelberg, 2011.

[47] E. Matsas and G.-C. Vosniakos. Design of a virtual reality training system for human–robot collaboration in manufacturing tasks. *International*

Journal on Interactive Design and Manufacturing, 11(2):139–153, May 2017.

[48] R. Mehra, A. Rungta, A. Golas, M. Lin, and D. Manocha. WAVE: Interactive Wave-based Sound Propagation for Virtual Environments. *IEEE Transactions on Visualization and Computer Graphics*, 21(4):434–442, Apr. 2015.

[49] A. H. Moore, J. M. de Haan, M. S. Pedersen, P. A. Naylor, M. Brookes, and J. Jensen. Personalized signal-independent beamforming for binaural hearing aids. *The Journal of the Acoustical Society of America*, 145(5):2971–2981, May 2019.

[50] F. E. Musiek and G. D. Chermak. Chapter 18 - Psychophysical and behavioral peripheral and central auditory tests. In M. J. Aminoff, F. Boller, and D. F. Swaab, editors, *Handbook of Clinical Neurology*, volume 129 of *The Human Auditory System*, pages 313–332. Elsevier, Jan. 2015.

[51] A. Naumann and J. Hurtienne. Benchmarks for intuitive interaction with mobile devices. In *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*, pages 401–402. ACM, 2010.

[52] M. Noisternig, A. Sontacchi, T. Musil, and R. Höldrich. A 3d ambisonic based binaural sound reproduction system. *AES 24th International Conference on Multichannel Audio*, 2012.

[53] J. Oberem, I. Koch, and J. Fels. Intentional switching in auditory selective attention: Exploring age-related effects in a spatial setup requiring speech perception. *Acta Psychologica*, 177:36–43, 2017.

[54] N. Parés and R. Parés. Towards a Model for a Virtual Reality Experience: The Virtual Subjectiveness. *Presence*, 15(5):524–538, Oct. 2006.

[55] K. Pimentel and K. Teixeira. *Virtual Reality: Through the New Looking Glass*. McGraw-Hill, Inc., New York, NY, USA, 1993.

[56] T. Powers and V. Hamacher. Three-microphone instrument is designed to extend benefits of directionality. *The Hearing Journal*, 55(10), Oct. 2002.

[57] H. Puder. Hearing aids: An overview of the state-of-the-art, challenges, and future trends of an interesting audio signal processing application. *6th International Symposium on Image and Signal Processing and Analysis*, 2009.

[58] T. Ricketts. The impact of head angle on monaural and binaural performance with directional and omnidirectional hearing aids. *Ear and Hearing*, 21(4):318–328, Aug. 2000.

[59] T. Ricketts and S. Dhar. Comparison of performance across three directional hearing aids. *Journal of the American Academy of Audiology*, 10(4):180–189, Apr. 1999.

[60] T. Ricketts, G. Lindley, and P. Henry. Impact of compression and hearing aid style on directional hearing aid benefit and performance. *Ear and Hearing*, 22(4):348–361, Aug. 2001.

[61] T. A. Ricketts. *Directional Hearing Aids*. Westminster Publications, 2001.

[62] O. Rummukainen and C. Mendonça. Task-relevant spatialized auditory cues enhance attention orientation and peripheral target detection in natural scenes. *Journal of Eye Movement Research*, 9(1), Jan. 2016.

[63] D. R. Saunders and R. L. Woods. Direct measurement of the system latency of gaze-contingent displays. *Behav Res Methods*, 46(2):439–447, June 2014.

[64] M. Schrepp, A. Hinderks, and J. Thomaschewski. Design and evaluation of a short version of the user experience questionnaire (ueq-s). *IJIMAI*, 4(6):103–108, 2017.

[65] S. Serafin, M. Geronazzo, N. C. Nilsson, C. Erkut, 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.

[66] B. G. Shinn-Cunningham and V. Best. Selective attention in normal and impaired hearing. *The Journal of the Acoustical Society of America*, 2008.

[67] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley. Detection and Classification of Acoustic Scenes and Events. *IEEE Transactions on Multimedia*, 17(10):1733–1746, Oct. 2015.

[68] Tsuhan Chen. Audiovisual speech processing. *IEEE Signal Processing Magazine*, 18(1):9–21, Jan. 2001.

[69] M. Valente. *Use of Microphone Technology to Improve User Performance in Noise, Singular Publishing Group*. The textbook of hearing aid amplification, 2000.

[70] B. Van Veen and K. Buckley. Beamforming: a versatile approach to spatial filtering. *IEEE ASSP Magazine*, 5(2):4–24, Apr. 1988.

[71] J. Vennerød. Binaural reproduction of higher order ambisonics a real-time implementation and perceptual improvements. Master's thesis, NTNU - Trondheim, Norwegian University of Science and Technology, Norway, 2014.

[72] D. B. Ward, R. A. Kennedy, and R. C. William. Fir filter design for frequency invariant beamformers. *IEEE Signal Processing Letters*, 1996.

[73] E. G. Williams. *Fourier acoustics: Sound radiation and nearfield acoustical holography*. Academic Press, 2005.

[74] B. Xie. *Head-Related Transfer Function and Virtual Auditory Display*. J. Ross Publishing, Plantatation, FL, May 2013.

[75] W. A. Yost, R. H. Dye Jr., and S. Sheft. A simulated "coctail party" with up to three sound sources. *Perception & Psychophysics*, 58(7):1026–1036, 1996.

[76] T. Zhang, F. Mustiere, and C. Micheyl. Intelligent hearing aids: The next revolution. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 72–76, Aug. 2016.