Delphine Devallez

Auditory Perspective: perception, rendering, and applications

Ph.D. Thesis

March 23, 2009

Università degli Studi di Verona
Dipartimento di Informatica

Advisor:
prof. Davide Rocchesso

Series N°: **TD-01-09**

Università di Verona
Dipartimento di Informatica
Strada le Grazie 15, 37134 Verona
Italy

Summary

In our appreciation of auditory environments, distance perception is as crucial as lateralization. Although research work has been carried out on distance perception, modern auditory displays do not yet take advantage of it to provide additional information on the spatial layout of sound sources and as a consequence enrich their content and quality. When designing a spatial auditory display, one must take into account the goal of the given application and the resources available in order to choose the optimal approach. In particular, rendering auditory perspective provides a hierarchical ordering of sound sources and allows to focus the user attention on the closest sound source. Besides, when visual data are no longer available, either because they are out of the visual field or the user is in the dark, or should be avoided to reduce the load of visual attention, auditory rendering must convey all the spatial information, including distance. The present research work aims at studying auditory depth (i.e. sound sources displayed straight ahead of the listener) in terms of perception, rendering and applications in human computer interaction.

First, an overview is given of the most important aspects of auditory distance perception. Investigations on depth perception are much more advanced in vision since they already found applications in computer graphics. Then it seems natural to give the same information in the auditory domain to increase the degree of realism of the overall display. Depth perception may indeed be facilitated by combining both visual and auditory cues. Relevant results from past literature on audio-visual interaction effects are reported, and two experiments were carried out on the perception of audio-visual depth. In particular, the influence of auditory cues on the perceived visual layering in depth was investigated. Results show that auditory intensity manipulation does not affect the perceived order in depth, which is most probably due to the lack of multisensory integration. Besides, the second experiment, which introduced a delay between the two auditory-visual stimuli, revealed an effect of the temporal order of the two visual stimuli.

Among existing techniques for sound source spatialization along the depth dimension, a previous study proposed the modeling of a virtual pipe, based on the exaggeration of reverberation in such an environment. The design strategy follows a physics-based modeling approach and makes use of a 3D rectangular

Digital Waveguide Mesh (DWM), which had already shown its ability to simulate complex, large-scale acoustical environments. The 3D DWM resulted to be too resource consuming for real-time simulations of 3D environments of decent size. While downsampling may help in reducing the CPU processing load, a more efficient alternative is to use a model in 2D, consequently simulating a membrane. Although sounding less natural than 3D simulations, the resulting bidimensional audio space presents similar properties, especially for depth rendering.

The research work has also shown that virtual acoustics allows to shape depth perception and in particular to compensate for the usual compression of distance estimates. A trapezoidal bidimensional DWM is proposed as a virtual environment able to provide a linear relationship between perceived and physical distance. Three listening tests were conducted to assess the linearity. They also gave rise to a new test procedure deriving from the MUSHRA test and which is suitable for direct comparison of multiple distances. In particular, it reduces the response variability in comparison with the direct magnitude estimation procedure.

Real-time implementations of the rectangular 2D DWM have been realized as Max/MSP external objects. The first external allows to render in depth one or more static sound sources located at different distances from the listener, while the second external simulates one moving sound source along the depth dimension, i.e. an approaching/receding source.

As an application of the first external, an audio-tactile interface for sound navigation has been proposed. The tactile interface includes a linear position sensor made by conductive material. The touch position on the ribbon is mapped onto the listening position on a rectangular virtual membrane, modeled by the 2D DWM and providing depth cues of four equally spaced sound sources. Furthermore the knob of a MIDI controller controls the position of the mesh along the playlist, which allows to browse a whole set of files by moving back and forth the audio window resulting from the virtual membrane. Subjects involved in a user study succeeded in finding all the target files, and found the interface intuitive and entertaining. Furthermore, another demonstration of the audio-tactile interface was realized, using physics-based models of sounds. Everyday sounds of “frying”, “knocking” and “liquid dripping” are used such that both sound creation and depth rendering are physics-based. It is believed that this ecological approach provides an intuitive interaction.

Finally, “DepThrow” is an audio game, based on the use of the 2D DWM to render depth cues of a dynamic sound source. The game consists in throwing a virtual ball (modeled by a physics-based model of rolling sound) inside a virtual tube (modeled by a 2D DWM) which is open-ended and tilted. The goal is to make the ball roll as far as possible in the tube without letting it fall out at the far end. Demonstrated as a game, this prototype is also meant to be a tool for investigations on the perception of dynamic distance. Preliminary results of a listening test on the perception of distance motion in the virtual tube showed that

duration of the ball's movement influences the estimation of the distance reached by the rolling ball.

Contents

1

Introduction

1.1 Spatialized audio in human computer interaction

Using sound to enhance human computer interfaces is nowadays a trend and a need: with the advent of more and more complex operating environments, rendering more and more data, the traditional visual interface is no more able to efficiently provide the user with the required information at a given time. The most obvious illustrative example of this issue is the mobile, wearable device featuring a reduced visual display: the main challenges are to display information to the user whose eyes and hands may be otherwise engaged, and to avoid a cognitive overload. Therefore visual feedback is not always the best channel, and this is even more true for the visually impaired. The value of sound has been recognized by interface designers, especially for its ability to monitor temporal events. Rendering multiple streams simultaneously however requires the spatial organization of the sound scene. Unlike the visual space that is limited by the screen size, the auditory channel theoretically allows to reproduce a 3D environment thanks to various spatial audio rendering techniques. It is then possible to exploit the ability of the auditory system to discriminate sound sources in space, in particular directional localization to enrich the auditory channel and enhance the user experience. As a promising application, interacting with mobile devices through 3D sound and gestures has been studied by the Interactive Systems Group at the University of Glasgow, UK [1, 10, 95]. Through examples of concrete applications such as a calendar [95] or a progress bar [96], and usability studies [1, 10, 95, 96], the authors showed that spatialized auditory displays enable an efficient interaction with mobile devices while freeing the user's visual attention.

Drawbacks of 3D rendering techniques, which were originally dedicated to virtual reality and which aim at recreating a true-like auditory environment, should however not be neglected: 3D audio technology generally requires headphones for a proper reconstruction of the sound field at the listener's ears, via the use of Head Related Transfer Functions (HRTFs) which are complex individual filters. Usually measured for only one distance and at a fixed number of directions, the HRTFs are then interpolated to build the full 3D space, and the effectiveness of non-individual filters is still an open issue.

In our approach to audio spatialization in human computer interaction, we follow some major directions:

- keep in mind that the objective is to facilitate the transfer of information between the user and the computer,
- consider spatialization as application specific: when creating an auditory interface, one should wonder what is essential for the task and how much realism is necessary. Therefore, the concept of spatialization is defined here in a broad sense as a mean to give a sense of space to the listener, by rendering some (or all) spatial attributes of sound sources. It is therefore not limited to binaural rendering techniques.
- make spatial sound interactive: according to the user input, the organization of multiple sound sources in space may change,
- make use of the knowledge in vision to translate techniques from the graphical to the audio domain. A direct mapping is obviously not the best option given the significant differences between the auditory and the visual systems, and care should be taken to meet the properties of sound perception through experimentation.

1.2 On the use of auditory depth cues

Information access in graphical interfaces must address the challenge of bringing the attention of the user on a specific object, i.e. directing the user *focus*. In critical situations, the user's cognitive load may be divided between several applications or information sources, and focus refers to which of the multiple visual objects the user is attending at a given time. A well-known tool in graphical interfaces is the so-called "fisheye views", introduced by Furnas [25], which consists of a nonlinear magnification using distortion to provide a zoom while preserving a view of the global context. This focus-and-context method relies on the ability to focus on an area of interest (which is zoomed in) while still viewing the remaining part of the visual area, namely the context (whose size remains unchanged). In other words, this manipulation acts as a tool for *figure-ground segregation* by bringing the area of focus closer to the user (in the *foreground*) while the context remains more distant (in the *background*). In addition, this magnification is being applied in real time to any region of interest, either by the user (using a mouse or a pointer) to explore the data, or dynamically by the system to highlight the perceptual effect of focus by motion. In short, any visual area of interest may dynamically be attributed to the foreground or the background. Among other techniques, placing important information in the middle of the screen and layering different objects on top of each other are commonly used to help the user focus on a particular source of information or to call his/her attention. As part of perspective effects, layering in addition allows to organize information into a *hierarchy* by ordering the various layers in depth according to their importance.

As a basis for this thesis, we argue that focus and hierarchy may also be achieved in auditory interfaces through sound spatialization, and in particular using auditory distance cues: positioning multiple sound sources at different distances from the user allows to provide a hierarchical ordering of various sources of information, and focus the user attention on the closest sound while preserving the access to the others. The analogy with the visual zooming technique is even stronger by using reverberation as the main dynamic depth cue: if one assumes that space is primarily captured by the visual sense and time by the auditory sense [35], then reverberation destroys the temporal fine structure of sound like spatial details of visual objects are reduced with increasing distance. In this way, changing dynamically the depth of an auditory object allows to modify accordingly its degree of intelligibility and its affordance. Rarely used in practice, the depth dimension of the auditory spatial display may consequently have great potential as a tool for facilitating access and manipulation of large data sets in auditory interfaces. In addition, unlike spatialization based on directional information, the depth dimension allows to reproduce identical signals to both ears since sound sources are always positioned in the median plane.

1.3 Terminology

For a better understanding of what this thesis is about, a few definitions are necessary:

- **Auditory distance perception** is the term used for estimating distances between the listener and sound sources or distances between sound sources, in any and all directions relative to the listener (for instance left-to-right distances between sound sources). It also includes the ability to hear the distance changes of sounds from near to far (and vice versa), and at varying angles.
- If one refers to the definition of depth perception in vision[1], **auditory depth perception** has a more specific and limited meaning, as shown in Figure 1.1. It represents the auditory distance of a sound source straight ahead of the listener. Consequently, auditory depth perception may be defined as looking straight into a tube and estimating forward auditory distances.
- In the same manner, the notion of visual perspective allows to define **auditory perspective** as the technique of reproducing depth relationships between the listener and sound sources at the listeners' ears.
- **Auditory distance cues** enable a person to perceive the distance of sound sources.

This thesis deals with auditory depth and therefore does not consider the potential directional discrimination of sound sources. This restriction allows to reduce the audio space to one dimension, corresponding to a front midline between the left and right ears.

1.4 Objectives of the thesis

This thesis has three main objectives:

[1] `http://www.abledata.com/abledata_docs/Distance_Perception.htm`

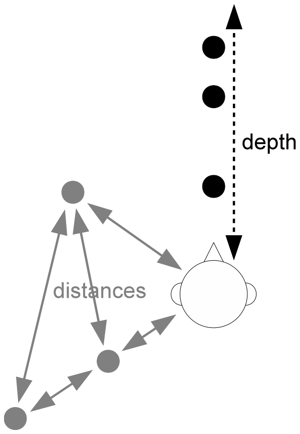


Fig. 1.1. *Definitions of distance (sound sources represented by gray circles) and depth (sound sources represented by black circles).*

1. Contribute to the psychophysical research on auditory depth perception which is lagging behind current knowledges on directional localization. Although the human ability to perceive auditory distance has been receiving more attention for the last decade, some aspects definitely require further investigations.

2. Propose a rendering technique for auditory depth which complies with the aforementioned strategy. In particular, the goal is not to produce an exact copy of a real environment by using complex signal processing, but to make a tool able to provide focus and hierarchy in the auditory domain while preserving the interface usability and quality of interaction.

3. Develop user interface prototypes to demonstrate the potential of the auditory depth dimension through listening tests or user studies.

1.5 Outline

The thesis will start in Chapter 2 with an overview of the aspects of auditory distance perception and studies on the implementation of distance cues in auditory interfaces which are most relevant to the present research work. In particular it will be shown that intensity, and overall direct-to-reverberant energy ratio, are powerful pieces of information for depth perception.
In Chapter 3 an extension to visual depth perception is conducted. The objectives are first to provide a brief summary of existing techniques for producing visual perspective effects, and secondly to experiment the congruence of audio-visual depth cues provided by simple techniques in both modalities.
Based on previous investigations by Fontana et al. [23], a physics-based model, the Digital Waveguide Mesh (DWM), is proposed in Chapter 4 as a rendering technique for auditory depth. It can be considered as a computational acoustic model

for waves traveling in opposite directions, and had already been proposed for use in artificial reverberation systems [76].

As it will be seen in Chapter 2, the human perception of distance is distorted towards a compressed estimation of the physical distance range. Chapter 5 highlights a special case of DWM which allows to linearize the relationship between physical depth and estimated depth. From a theoretical model that predicts the linearity, experimental procedures are presented to validate this finding through listening tests.

The following two chapters present practical applications of the DWM in user interfaces. Chapter 6 proposes an audio-tactile interface that exploits the similarity of the audio and tactile spatial geometries to provide a coherent tool for sound navigation. In Chapter 7, dynamic depth rendering is applied to a virtual rolling ball in an audio game, “DepThrow”.

Finally, general conclusions are drawn in Chapter 8. They include a summary of the achievements as well as implications and open issues.

2

Auditory distance perception

This chapter aims at introducing the human ability to hear auditory distance in order to point out the implications and the potentiality of spatializing sound sources along the depth dimension. Past research on various aspects of auditory distance perception will be reviewed, and a few practical applications in auditory interfaces found in literature will also be presented.

2.1 The human ability to perceive auditory distance

The study of distance perception is not new, even if directional localization has received much more interest. Research studies have indeed allowed to precisely identify the cues that are responsible for directional perception: differences in time and level between the two ears provide robust azimuth information, while monaural spectral cues issued from filtering effects from the head, shoulder, torso and outer ears give elevation information and partially solve front/back confusions (see [61] for a detailed review). As part of sound localization, distance perception was already studied by Blauert [8]. Some earlier studies also reported some phenomena related to distance localization ([15, 26, 27, 82, 94]). However, distance perception is more complex: it is much less accurate than directional localization and may be computed from multiple pieces of information including non-acoustic factors. For the past ten years, some researchers have again shown great interest into distance perception. Several scientists have provided extensive reviews on previous research in distance perception [68, 104], and the aim of this section is not to copy them but rather to give the basis to an easy understanding of what are the relevant properties of the sound reaching the listener's ears to convey distance information, and the implications for rendering distance information in auditory displays.

2.1.1 Auditory distance cues

The main auditory factors that convey distance information are as follows:

1. **Intensity**: this is the most obvious piece of information for distance perception: when the sound moves away from the listener, its intensity decreases.

- For far-field sources in free space, the intensity level varies inversely with the square of the distance [47]. However this law is no longer applicable in reverberant environments where reflections from diverse surfaces reduce the loss of intensity as a function of distance. In a reflective space, although the intensity of the direct sound from a source losses 6 dB per doubling of distance as it would in free field, the reverberant sound (whose level remains roughly constant) gradually takes over as you move away from the source (see Figure 2.1). At some distance, the direct sound and the reverberant sound components have equal levels, this distance is known as the critical distance or reverberation radius. Beyond that distance, the reverberant sound dominates.
- For nearby sources, sound level changes with distance depend on the source direction because of the interaction of the listener's head with the sound waves.

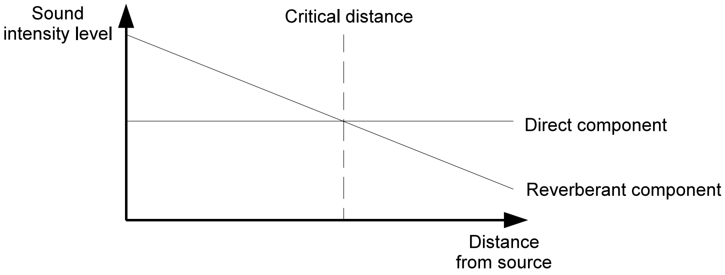


Fig. 2.1. *Definition of the critical distance.*

2. **Direct-to-reverberant energy ratio** (only applicable in reverberant environments):
   - For far sources, the level of reverberation is roughly independent of the location whereas the intensity level of the direct sound varies inversely with the square of the source distance (see Figure 2.1). As a result the direct-to-reverberant energy ratio decreases with increasing distance.
   - For nearby sources, the amount of reverberation varies with the source location because the distance traveled by the reflections can no longer be assumed to be fixed (especially for the first reflections that may come from very nearby objects).
   - Besides the acoustic properties of the environment, the direct-to-reverberant energy ratio also depends on the directivity of the sound source. For the same acoustic power, a directional source will have a greater intensity at a distance r on the acoustic axis than an omnidirectional source at the same distance. Consequently, the direct-to-reverberant energy ratio will be higher on the acoustic axis of the directional source than that produced by the omnidirectional one.
3. **Interaural Level Differences (ILDs)**:

- Level differences at the two ears can arise due to the acoustic interference of the head.
- For nearby sources, the angle from the ear and the center of the head to the source can differ substantially, this effect being known as the "auditory parallax" [13]. As a result of this effect, binaural differences in intensity will change with radial distance and direction of the sound source. They are maximal along the interaural axis and decrease to zero in the median plane.
- For distances higher than 1 m, changes in distance cause no significant change in ILD, whatever the direction. Only changes in direction provide changes in ILDs which occur at high frequencies due to the head shadow effect, and will therefore only provide information about the direction of the source.
- For sound sources in the median plane, ILD is null, independently of the source distance.

4. **Spectrum**: For very large distances (superior to 15 m) sound-absorbing properties of the air slightly attenuate the perceived level of high frequencies (about 3 to 4 dB decrease per 100 meters at 4 kHz [41]) . Furthermore sound-reflective environments may also affect the spectrum of the signal that reaches the ears, and this cue may be seen as a component of the direct-to-reverberant energy cue.

2.1.2 Relative and absolute auditory distance cues

The aforementioned cues do not all provide the same type of information to the listener: either they allow to perceive the absolute distance of the sound source, or they only inform about changes in the relative distance of the sound source.

The intensity cue only provides relative distance information. Indeed, the absolute level of direct sound varies both with distance and with the energy emitted from the source, so that the level at the listener's ears can not judge the absolute distance unless the subject has an a priori knowledge about the sound source level.

Spectral cues are also relative cues since the listener needs a priori information about the spectral characteristics of the source.

The direct-to-reverberant energy ratio is classified as an absolute distance cue, because reverberant rooms have been shown to provide good distance perception independently of the source level [12]. In the free-field condition, Bronkhorst and Houtgast [11] have proved the inability to accurately judge distances, because the most important cue (loudness) is confounded with the level of the sound itself. However the latter result is not easy to achieve in virtual environments where the reverberation is recreated, and an experiment conducted in [103] showed that the threshold in the sensitivity to direct-to-reverberant energy ratio is very high, meaning that this cue provides only a coarse coding of sound source distance.

Binaural cues may be useful to perceive distance only for nearby sound sources, since they no longer vary with distance for far-field sources [84]. The combination of Interaural Intensity Levels (IILs) varying with source distance and constant Interaural Time Differences (ITDs) may provide an absolute judgment of nearby source distance.

2.1.3 Experimental procedures for distance estimation

Estimating the distance of a sound source has been shown to be a difficult task. Therefore, researchers have tried to use other procedures than the typical magnitude estimation based on verbal or written reports of the distance estimates [11, 103]:

- Pairwise comparison scaling method: compared with the magnitude estimation, this method gave rise to similar results [101].
- Perceptually directed walking [53]: this method was developed in reference to a procedure used in vision, the "visually directed action", and consists in walking with closed eyes to the presented location of the sound source. A comparison with the verbal report in feet of the distance estimates did not reveal any significant difference between both procedures.
- "Auditory reachability": instead of asking the subjects directly for the distance of a sound source, Rosenblum et al. [73] got interested in the reachability of the sounding object, which is a more intuitive task. In addition, the experiments dealt with a real sound source (a rattle) in a familiar acoustic environment. When comparing judgments' accuracy to that of studies testing absolute distance perception, the described reaching task provides a much higher performance than experiments dealing with anechoic conditions and/or unnatural stimuli. These results validate the ecological importance of auditory distance perception.
- Magnitude production: in an experiment investigating the perception of distance in a pipe, Fontana and Rocchesso [22] used a procedure where participants had to position a marker along the tube to rate the apparent distance of a moveable loudspeaker. Results showed a larger deviation of distance estimates than that obtained by Zahorik in a study using binaural room impulse responses measured in a small auditorium [103]. However the small amount of data gathered by Fontana and Rocchesso and the difference of conditions and stimuli between both studies make the comparison difficult.

2.1.4 Distance perception accuracy and relative contribution of distance cues

Several studies including [84], [103] and [68] have investigated the accuracy of distance perception and the perceptual relevance of the different distance cues. Here are presented some important results of their experiments:

- Reverberation is an important cue for distance perception. Shinn-Cunningham [85] showed that reverberation improves distance perception, even for nearby sources where the amount of reverberation is low. Besides, in the free-field condition Gardner [27] and Coleman [15] proved the inability to accurately judge distances, because the most important cue (loudness) is confounded with the level of the sound itself. Nielsen [68] also found that reflections are a powerful cue for perceiving distance. Results of the experiments they conducted even showed that in anechoic conditions there was no correspondence between the physical distance of the sound source and the perceived distance, but that judgment was rather based on intensity or timbre changes [15], or influenced by visual sound sources [27]. These results suggest that human ability to perceive distance is inaccurate and gets worse without reverberation.

- Reverberation is also a major factor for auditory spaciousness and externalization, which give a sense of open space and where the sound appears to be outside the head [17]. This is particularly true for headphone reproduction, since no natural reverberation is given by the environment.
- While studying the human sensitivity to changes in the direct-to-reverberant energy ratio, Zahorik [105] found that this cue may be relatively imprecise. Indeed, the evaluation of direct-to-reverberant energy discrimination for 4 different types of stimuli (a 50 ms noise burst with abrupt onset/offset, a 300 ms duration noise burst with gradual onset/offset, a speech syllable and an impulse) resulted in a threshold of about 5–6 dB for all stimuli and for the whole range of direct-to-reverberant energy ratio values investigated. This threshold is much more than known just-noticeable difference (JND) for intensity ($\sim$ 1 dB) [43], and this difference consequently suggests that the intensity cue may provide much finer resolution of source distance in relative distance judgment tasks. However in a recent study by Larsen et al. [50] similar to that conducted by Zahorik [105], it was established that JNDs were dependent on the value of the direct-to-reverberant energy ratio, and the minimum JNDs were found to be as small as 2–3 dB.
- For nearby sound sources, Shinn-Cunningham [84] and Zahorik [103] looked into Human Head-Related Transfer Functions (HRTFs) and Binaural Room Impulse Responses (BRIRs) to show that ILDs are not critical for simulating source distance. Zahorik showed that the only change that can be observed is a slight decrease of ILD in high-frequency with increasing distance, corresponding to about 1.6 dB per distance doubling. Shinn-Cunningham studied the variation of the ILDs as a function of distance for nearby sound sources, at various directions in the horizontal plane, and she showed that the effect of this cue is direction-dependent and is the lowest for sound sources directly in front of the subject. Moreover the author suggested that rather the overall level (for a familiar source) and most of all the reverberation were critical for nearby distance judgments.
- Zahorik [103] showed that the manner in which listeners process, or perceptually weight two principal distance cues which are the intensity and the direct-to-reverberant energy ratio, varies substantially across the same stimulus conditions for each subject. Thus he proposes an acoustic cue combination and weighting method which result in a distance estimation. Shinn-Cunningham [84] also showed that the relative utility of overall level, reverberation and ILDs depends on the type(s) of sound source(s) and the kind of acoustic environment to be simulated as well as the range of distances to be encoded.
- "Compressed distance perception": past experiments have allowed to explore the relationship between the perceived distance to the physical distance, and all of them have shown that distance perception is biased. A comparison of a multitude of these psychophysical functions derived from different experimental methodologies has revealed that most of them are well approximated by

a power function of the form $r' = kr^a$, with r' the estimated distance and r the physical source distance. Furthermore it was shown that the fitted exponent a greatly varies between listeners and experimental conditions, and lies within approximately 0.15–0.70 [103]. These results highlight the high variability of distance perception across individuals and experimental procedures, but also the common compression of the psychophysical functions that relate the estimated distance to the physical one: people tend to overestimate close physical distances and underestimate far distances. Bronkhorst [11] proposed another model that relates the effect of the direct-to-reverberant energy ratio with the estimated distance, that also accounts for the "horizon effect" (i.e. the maximum perceived distance). The underlying idea is to take into account the duration of the reverberation (i.e. the number of reflections) in the direct-to-reverberant energy ratio, to cope with his observations of increased perceived distance with the number of reflections. The model proposed uses an integration window of 6 milliseconds in the calculation of the energy of the direct sound. This time constant is also related to the precedence effect, a group of auditory phenomena whereby an acoustic signal arriving first at the ears suppresses the ability to hear any other signal, including reflections, arriving a short time after (up to 30–40 ms) and not significantly louder than the initial signal [8]. This model will be further described in Chapter 5.

- "Specific distance tendency": when subjects are asked to make judgments of sound sources with no distance cues (e.g. unfamiliar sounds in anechoic conditions), the distances reported differ from the corresponding physical distances, and the sound sources are typically perceived to be at a default distance [59]. This phenomenon, called "specific distance tendency", was first attributed to the perception of the visual space [31].

2.2 Non-acoustic cues

- Several studies showed that source familiarity can actually help the listener to estimate the distance from relative cues such as intensity. This is particularly true for speech generated by live talkers [60]. Zahorik [103] also showed that the intensity cue is weighted more than the direct-to-reverberant energy ratio for speech signal, whereas for unfamiliar signals it may become less reliable than the direct-to-reverberant energy ratio (noise-burst signal for instance). In 1962, Coleman [15] had already proved the inability to accurately judge distances of unfamiliar sounds (1-second bursts of wideband noise) in free-field conditions, which was later supported in 1975 by Mershon and King [59].
- "The learning effect": while some researchers have simply suggested that distance perception may become more accurate with experience (i.e. increasing the number of trials), Shinn-Cunningham has formerly shown the effect of practice on localization performance in a reverberant environment [85]. While conducting an experiment on localization performance (for both direction and distance) under weak reverberant conditions, a statistically significant decrease in error magnitude with practice was observed, even if the experiment was performed across several days. In comparison, this effect was not observed in

anechoic conditions. This suggests that people use their past experience to make a localization judgment and are able to calibrate their perception to new reverberant environments.

- vision: in distance perception, vision has been found to increase the accuracy of distance judgments and lower their variability [102]. For this study, an array of five loudspeakers was used in a semireverberant room, and the stimulus, a female voice with a constant level, was presented from one of the speakers. Comparisons were made between blindfolded subjects and others who were allowed to see. Although blindfolded participants were relatively good at estimating the apparent auditory distance, a greater accuracy, a lower judgment variability and a fast learning effect were found under the vision condition. Furthermore, the effect of "visual capture", already known in directional localization as the "ventriloquism effect", has also been demonstrated in distance perception by Gardner [26]: the sound source is perceived at the distance where a visual object (louspeaker) is located. In his study, Gardner further reported the "proximity-image effect": without any available auditory distance cue, people tend to choose the nearest visible loudspeaker as the location of the sound source.

2.3 Distance discrimination

Many studies have been carried out on the perception of concurrently playing sound sources when they are spatially separated in azimuth, and often dealt with the "cocktail party effect", i.e. studying the segregation of speech from a competing speech signal. However very little is known about the effect of spatial separation in distance. On the subject, only the study of Brungart et al. [14] has been found. They also investigated the "cocktail party effect", and in particular the intelligibility of nearby speech when it is spatially separated from an interfering sound source (either noise or another speech signal) in distance. Although distance perception was not directly under study, Brungart examined the effects of distance on the segregation of nearby sources. Under the specific experimental conditions of nearby sound sources located along the listener's interaural axis, he found noticeable differences in intelligibility improvements between a speech masker and a noise masker. While binaural cues facilitate the intelligiblity of speech under the speech-masker condition, the better performance observed under distance separation of the speech with a noise masker is explained by spectral differences between the two signals. This study shows that the impact of spatial separation of sound sources differs according to the nature of the sound sources. However the important limitations of the investigation (nearby sources, types of sound sources, location along the interaural axis) makes it difficult to generalize.

2.4 Dynamic cues

Static cues (i.e. when both the sound source and the listener are static) have been studied in priority, and very little literature may be found about dynamic cues, i.e. when the sound source and/or the listener is moving, although this type of information is used to guide locomotion and to anticipate the arrival of a sound source. All the static cues described in Section 2.1.1 may provide auditory motion information given the relative distance change between the listener and the sound source:

- **Intensity changes**: approaching sound sources produce increases in intensity, and receding sound sources produce corresponding decreases. As for a static situation, this cue was found to be dominant for judging the distance travelled by a sound source in a study by Rosenblum et al. [72]. It was found that listeners typically overestimated increasing compared to equivalent decreasing sound intensity [66], and underestimated the starting and stopping distance of approaching sound sources [67]. From an ecological point of view, this behavior could provide an increased margin of safety in response to looming sounding objects. Furthermore, the intensity cue gives rise to another variable, called *acoustic tau*, that may also contribute to auditory motion perception. The acoustic tau has been defined as the time to collision with the source [81], in analogy with the optical τ, and is induced by the rate of change in intensity with distance. According to [81], it could be used to control action, such as reaching tasks. This approach agrees with the possible existence of an auditory system specialized in motion perception as it exists in vision [56]. However, this piece of information was found to hardly improve distance perception [87]. This result was supported by Guski in [33] who suggested that it is not always possible to make simple analogies between visual and auditory variables used in perception, given that vision and hearing have different functions.
- **Binaural cues**: In 1973 Simpson showed that head rotations did not help in perceiving distance more accurately [86]. In a more recent article Speigle and Loomis studied the effect of subject's translation on distance perception [87]. When the subject translates through space and the sound source is stationary, the direction of the source changes (this effect is called *motion parallax*) and therefore may provide additional information about the source distance. However the study concluded that this binaural cue created by the listener's motion only slightly helped in perceiving distance. More precisely, its effect depends on the range of distances considered: at very close distances a small shift of the sound source causes a large change in angular direction but this is no longer true for far distances.
- **Direct-to-reverberant energy ratio changes**: as mentioned in Section 2.1.4, some studies showed that humans are not very sensitive to variations of the direct-to-reverberant energy ratio, which implies that this cue might be relatively unreliable to judge dynamic distance, in comparison to the intensity cue. It is however difficult to separate the direct-to-reverberant energy ratio cue from the intensity cue for dynamic distance perception. In a real environment, both cues are indeed intimately related: when the listener is approaching the sound source (or the sound source is approaching the listener), the level of the direct signal increases, which consequently raises the intensity level as well as the direct-to-reverberant energy ratio.

The **Doppler effect** is also a cue for auditory motion perception, and is named for the 19^{th} century physicist Christian Johann Doppler. A common example of the Doppler effect is the pitch change heard as an emergency vehicle passes by. In theory, when the sound source and the listener are approaching each other (because of the motion of either or both), the received frequency is increased (compared to the emitted frequency), then it is identical at the instant of passing by, and it is decreased when the sound source and the listener are moving apart. However, people tend to perceive rising pitch during the approach, and this illusion was explained by McBeath and Neuhoff as the result of the rise in intensity, and therefore the interaction between frequency and intensity perception [57].

Lufti and Wang studied the relative contributions of three main cues for auditory motion, namely intensity, binaural cues and Doppler effect, for a sound source passing in front of the listener on a straight line [56]. By measuring discrimination thresholds for changes in displacement, velocity and acceleration, the authors found differences according to the range of source velocities. For moderate velocities (below 10 m/s), intensity and binaural cues were the predominant cues, while the Doppler effect rather informed on the velocity and acceleration of the sound source. On the contrary its contribution to the sound source displacement was predominant for higher velocities (50 m/s).

Early reflections might also give some information about dynamic distance. When the direct sound is followed by a strong reflection (with a delay smaller than about 30 ms), the resulting sound is colored because of a superimposed pitch which corresponds to the inverse of the time delay [7]. This phenomenon, called **repetition pitch** [7], may be easily perceived when the original sound is wideband and pitchless itself, like white noise. When one approaches a static sound source, the direct signal arrives earlier at the ears, but the relative delay of the reflected signal increases, and consequently the perceived pitch decreases. Repetition pitch may therefore provide reliable information for dynamic distance, as suggested in [52].

2.5 Depth rendering in auditory interfaces

Recent auditory interfaces have rather taken benefit of research on auditory directional perception to increasingly provide users with spatialized auditory displays, with applications ranging from scientific simulations for research purposes to entertainment and infotainment.

In contrast with directional localization, relatively little attention has been given to auditory depth. However, in 1990 Ludwig et al. proposed to translate the notion of windows in vision to auditory interfaces [55]. In particular, they foresaw the use of “audio windows”, for instance to change the level of detail, and therefore the possibility to render hierarchical auditory information.

Similarly, Schmandt designed a tool called *acoustic zooming* - by analogy with the visual ability of focusing on a specific area of a display - applied to an auditory browsing environment of audio files [77]. The interface is built like a hallway in which the listener moves up and down by head tilting, passing virtual rooms on his left and right sides. Each room holds a cluster of audio files that is represented by an “audio collage” of the various sound files. Inside a room, the sound files are

placed around the listener's head, and in order to be able to discriminate them, only three of them are played simultaneously through an "audio lens" which magnifies what is underneath it. Based on the "Fisheye Lens" proposed by Furnas [25], it allows to smoothly increase the spatial separation of the three sources, and bring the source at the center of the lens into focus by increasing its sound level.
In a very similar manner, Fernström and McNamara created a function called "aura" which restricts the user's spatial range of hearing [19]. This time, the application for browsing and managing sound data sets is both visual and auditory: sound files are represented by visual icons and those inside the aura are played back using stereo panning to separate them in space. A focus effect is also provided inside the aura by applying an inverse square law to the sound level of each file as a function of its distance to the center of the aura. A user study showed that the presence of the aura made the browsing task more efficient.

2.6 Conclusions

1. The main outcome of research on distance perception is that people are generally not able to accurately estimate distance, and this result must be taken into account when developing auditory displays rendering distance information: perceived distance is typically compressed compared with physical distance.

2. Reverberation is a powerful cue for distance perception: it does not need a priori information about the sound sources, it is a monaural cue, and even if the sensitivity to the direct-to-reverberant energy ratio may be slightly higher than that to the intensity, humans may increase their ability to judge distance with practice.

3. While individual HRTFs may be required for directional localization in virtual environments, they do not help in perceiving distance, as shown by Zahorik [104]. To go further, one may wonder why to use HRTFs at all to display depth (i.e. distance directly in front of the listener), at least for far-field sound sources. While binaural cues are ineffective in the median plane, the scattering by the head and torso provide spectral information that might be used for estimating near-field distances [107]. In the far-field, intensity and direct-to-reverberant energy ratio may be sufficient to render depth. Yet, research studies in virtual environments typically use HRTFs as a spatialization tool, motivated by their ability to render directional information as well. It would be interesting to compare the ability to judge auditory depth in the far-field with HRTFs and with a spatial processing algorithm rendering only the overall intensity and the direct-to-reverberant energy ratio cues.

4. Externalization or "out-of-head" localization is an important issue because it is a necessary condition to perceive distance. In-head localization of sound images is a critical problem in headphone reproduction. In anechoic conditions, it has been shown that HRTFs need to be individualized for proper externalization [36]. However, externalization may be achieved in case of non-individual

HRTFs by adding synthetic reverberation to the stimuli [4]. Then, it would be interesting to evaluate the influence of (non-individual) HRTFs on the externalization of stimuli synthesized or recorded in a reverberant environment in the median plane of the listener.

5. Finally, past studies obviously show that distance perception is facilitated for familiar sound sources in reflective environments. In addition to speech, easily recognized sounds may include a wide range of everyday sounds, that may be used in auditory interfaces to convey information [28].

3

Audio-visual rendering of depth cues

For human-computer interfaces, the combined rendering of both visual and auditory modalities allows to provide much richer information to the user than a one-mode interface such as visual displays. In terms of spatialization of data, on the one hand techniques are widely used to provide visual perspective effects, on the other hand much less attention has been brought to the auditory channel for which spatialization is usually limited to directional discrimination of sound sources. In this context, an experiment on auditory-visual depth perception and its follow-up were carried out in order to investigate the ability of auditory depth cues to compensate for weak visual cues. Before describing these experiments, the chapter will begin with a brief overview of visual depth cues, with a particular focus on partial occlusion. Then, a few past studies on audio-visual interactions will be described to understand the issues raised by multimodal perception, and will introduce the experiments presented in the last section.

3.1 Visual depth rendering techniques

The use of visual perspective effects has a long history and does not necessarily require a high level of data processing. Techniques such as occlusion and perspective on 2D visual displays are largely used nowadays to render layered content and give a sense of depth. In 1993, Bier et al [6] introduced semi-transparent widgets appearing in front of an application and providing the user with tools for operating directly on the application beneath. In first-person engagement in video-games, as in boxing with the Nintendo Wii, transparency is also used to represent the alter ego of the player, and to distinguish him from the contender (see Figure 3.1). With the objective of producing similar results in the auditory domain, it seems natural to get some knowledge about the techniques used in vision.

3.1.1 Visual depth cues

A variety of techniques may be applied to render visual depth on a two-dimensional display. Two main sources of information are available to convey depth: binocular disparity, a depth cue that requires both eyes; and monocular cues, which allow us to perceive depth with just one eye (see [97] for a detailed review of depth cues).

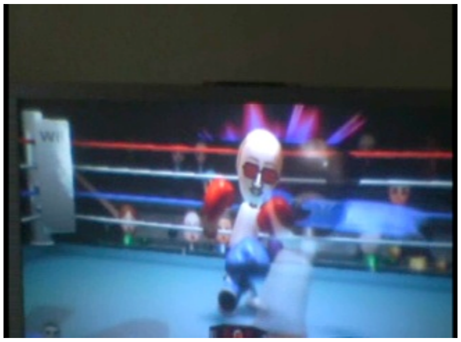


Fig. 3.1. *Semi-transparency is used in Wii Sports boxing to see the contender through the ego of the player.*

- Binocular disparity: because the two eyes are spaced apart, the left and right retinas receive slightly different images. These differences in the left and right images are integrated by the brain into a single three-dimensional image, allowing us to perceive depth and distance. The phenomenon of binocular disparity functions primarily in near space because with objects at considerable distances from the viewer the angular difference between the two retinal images diminishes. Binocular disparity is rendered by presenting two different images for the left and right eyes (called stereo pairs).
- Monocular cues: these cues are effective when viewed with only one eye. The most important ones are:
  - Occlusion (also called interposition or overlapping): when an object is occluding part of another one, the latter is perceived as being further away.
  - Linear perspective: when looking down at two parallel lines, they are perceived to come closer and converge at one point.
  - Relative height and size: the object closer to the horizon is perceived as farther away, and the object further from the horizon is perceived as closer (effect called "Moon illusion").
  - Familiar size: when an object is familiar to us, our brain compares the perceived size of the object to this expected size and thus acquires information about the distance of the object.
  - Texture gradient: the texture of surfaces becomes denser and less detailed as the surface recedes into the background.
  - Aerial perspective: due to the scattering of blue light in the atmosphere, distant objects appear more blue. Contrast of objects also provide clues to their distance. When the scattering of light blurs the outlines of objects, the object is perceived as distant.
  - Shadow: objects dont usually allow light to pass through and therefore cast a shadow, which follows some general rules (for instance, with one source of light, all shadows lie in same direction).
  - Motion parallax: objects at different distances from you appear to move at different rates when you are in motion, the more distant objects appearing to move slower.

3.1.2 Partial occlusion

Among all depth cues, occlusion, motion and binocular disparity have been found to be the strongest cues for computer displays [108]. In particular, occlusion, which is the easiest cue to implement, has been largely used in 3D computer graphics but presents the disadvantage of completely hiding the objects located in the background of another object. Therefore Zhai [108] proposed the use of *partial-occlusion*, which enables to see through the object that overlaps other objects. This cue is produced by *semi-transparency*, which means that the semitransparent surface can still be seen and does not block the view of any object that it occludes. To create the impression that one surface S_1 is in front of another surface S_2 by using semitransparency, the intensity I of the overlapping area is rendered by blending the color intensity of one surface, I_1 with the color intensity of the second surface, I_2, according to:

$$I = \alpha I_1 + (1 - \alpha) I_2 \tag{3.1}$$

where α is the coefficient of transparency, lying between 0 and 1 [20]. If $\alpha = 1$, the surface S_1 is opaque, therefore it appears in front of the surface S_2, and if $\alpha = 0$, the surface S_1 is transparent, therefore the surface S_2 appears in front of S_1. As α varies from 0 to 1, the perceived surface in front will consequently change, from S_1 to S_2. Furthermore $\alpha = 0.5$ refers to the point of equal probability: the probability of seeing S_1 in front equals the probability of seeing S_2 in front. Anderson also called this phenomenon *bistable transparency* [2].

3.2 Review on audio-visual interactions

3.2.1 Introduction: congruence and conflict

Our perception is not infallible: we do not always perceive the world as it is in reality and we sometimes make perceptual errors. This phenomenon is called *illusion* and either involves a single modality or originates from a conflict between the information provided by various modalities. In the second case the congruence of the available information is altered and generates a conflict. In our perception of the real world we generally use simultaneous pieces of information from various senses and their integration allows the unified multimodal percept of an object. Numerous studies have been investigating multimodal illusions, in particular to understand the general mechanisms that explain the interaction between the different sensory modalities in the perception of space and objects located in that space. In particular, research on audio-visual interactions has received increased attention with the development of virtual reality, teleconferencing, gaming and home theater systems. The following literature review is far from being exhaustive on audio-visual interaction. Since the thesis deals with spatial rendering and depth in particular, this review focuses on localization tasks.

3.2.2 Complementary cues

Past studies have demonstrated the improvement of some specific tasks by adding auditory stimuli to the visual ones (see [34] for a review). These include improvement of target detection, decreased reaction times and localization improvement.

In particular, cross-modal benefits are significant when spatial information in one sense is compromised or ambiguous.

In his study, Hairston [34] examined the benefit of acoustical cues under conditions of myopia by presenting light-emitting diodes to the subjects with or without a broadband noise burst coming from the same location. While directional localization accuracy was equivalent for visual and multisensory targets under normal vision, the myopia condition showed a substantial improvement with the addition of auditory cue.

Conversely, Shelton and Searle [82] showed that the vision of loudspeakers rendering the sound stimuli had a powerful effect on the accuracy of directional localization judgments in the horizontal plane. Even though less powerful, the effect was still present for sources outside the field of vision. They explained this observation by suggesting that vision gave a visual frame of reference to localization tasks.

When the visual sense gives ambiguous information, it has been shown that auditory cues may be able to resolve the ambiguity. This phenomenon, called "cross-modal compensation" by Begault [5], is defined as the ability to create relative shifts in perception in one modality by changing another. Sekuler et al. [79] conducted an experiment on the perception of motion of two disks. Without any other cue, the visual stimulus may result into two different interpretations: either the disks stream through, or they bounce off each other. However, since collision often produce sounds characteristic of the impact, the absence of sound rather leads to the perception of streaming through. The perception of the scene was changed with the addition of a brief click at or near the point of coincidence, and promoted the perception of bouncing. Besides showing the effect of sound on visual motion, the authors also reported that the auditory stimulus did not need to be in perfect synchrony with the visual one but could be presented up to 150 ms before or after the visual coincidence point.

Another similar experiment was conducted by Ecker and Heller in [18]. This time, the ambiguous visual stimulus consisted of a rolling ball that could either roll back in depth on the floor of a box, or jump in the frontal plane. Moreover, other ball's paths of different types and curvature in between were also presented to the subjects. The moving ball was either shown alone, accompanied with the sound of a ball rolling, or the sound of a ball hitting the ground. Similarly to the results of Sekuler [79], it was found that sound influences the perception of the ball's trajectory, depending on the type of sound.

Frassinetti [24] also reported an improvement of visual tasks under auditory-visual conditions. Her experiment showed that the perceptual sensitivity for luminance detection of a green LED masked with four red LEDs was facilitated when an auditory stimulus (white noise burst) was presented at the same location and simultaneously to the visual stimulus.

Reaction time may also be speeded up by the presence of cues in different sensory modalities. Laurienti et al. [51] studied in particular the effect of semantically congruent auditory-visual stimuli on response time, using circles of red or blue color and the words "red" and "blue". Either unimodal or congruent bimodal stimuli (i.e. a red circle with the word "red") were presented to the subjects. A significant decrease of the response time was found under the auditory-visual conditions in comparison with unimodal auditory or visual conditions.

Another case of auditory-visual interactions that has been reported involves perception of distance when no auditory distance cue is present. In other words, subjects are asked to judge distance in free field. This experiment was conducted by Gardner [26]. In anechoic conditions, people were asked to select among five aligned loudspeakers the one playing a speech signal. Without any reverberation cue and no possibility to use the relative sound level cue since only one level was used during the experiment, they all reported that the sound was coming from the nearest loudspeaker, even if a further loudspeaker was actually activated. This effect, called "proximity-image effect" and already mentioned in Section 2.2, suggested that in the absence of distance cues, the sound source is perceived coming from the closest rational visible location.

Zahorik [102] also studied the effect of vision on the perceived distance and reused the experiment performed by Gardner. However this time, the experiment was performed in a semi-reverberant room, all the five loudspeakers were used to render the stimulus, and half of the listeners were blindfolded. With the presence of the direct-to-reverberant energy ratio and the relative intensity cues, people were able to judge distance accurately, especially in the vision condition. This result provided evidence that in the conditions of the experiment, which are relatively close to normal living conditions, visual information dominates the judgment of distance.

All these experiments show that auditory and visual information are complementary and increase the accuracy of localization or the reaction time, and in case that one sense does not give reliable cues (or any cue at all), the perception relies on a more accurate sense (see the proximity-image effect).

3.2.3 Conflicting cues

What happens if the multisensory cues generate conflicting information? It has been widely suggested that the brain processes all the inputs from diverse senses to give rise to a single percept, but the way this process weights each individual modality is unknown, so that it is unlikely that one may be able to predict the outcome of conflicting cues.

However in the case of auditory-visual interactions, it was well established that vision was the dominant sense, and the well-known "ventriloquism effect" [39] corroborated this idea. This phenomenon involves a conflict between the spatial location of an auditory stimulus (speech sounds) and a spatially disparate visual stimulus. Under specific conditions that enable the formation of a unitary percept, the perceived location of the event is predominantly determined by the location of the visual stimulus.

Another compelling study about auditory-visual interactions related to speech is the "McGurk effect" named after Harry McGurk [58]. While it is well-known that speech intelligibility is enhanced by looking at the speaker's lips [78], the experiment by McGurk demonstrates that visual speech influences what people hear when the visual and auditory speech are different. Subjects are shown a person saying a simple consonant-vowel syllabe and simultaneously hear a recording of the same type of syllable but different from the visual one. The results of the conflict between the acoustic and visual information vary among subjects: either they perceive one of the two syllables that were actually presented, or they perceive a new composite sound based on the two syllables. The most famous example is that when an acoustic [ba] is synchronized with a visual [ga], most people report perceiving [da].

Recently, Shams et al. [80] related an opposite phenomenon, called "vision illusion", where audition influences vision. The outcome of their experiment is that a single flash of light accompanied with multiple beeps is perceived as multiple flashes. It was also found that the induced illusion started to degrade when the time between two consecutive beeps reached about 100 ms, but that any other parameter manipulation in the sound or the light did not influence the outcome. This robust effect against stands in favor of cross-modal interactions and demonstrates that vision may not be always the dominant sense.

Similarly, Morein-Zamir et al. carried out experiments on the perception of two lights in a visual temporal order judgment task [64]. In addition, two sounds (clicks) were presented either simultaneously or before the first light and after the second light. Results reveal a phenomenon called by the authors "temporal ventriloquism": sounds can bias visual perception in the temporal domain. In detail, sounds presented before the first light and after the second light pull the two lights further apart in time while they lead to a decrease of performance when presented between the lights.

In [35] Handel explains the human integration of conflicting cues based on the *unity assumption*. The latter considers temporal and spatial aspects of the auditory and visual inputs: if they are temporally synchronous and appear to come from the same spatial location, then they refer to a single object. In the event that information from the two modalities is too conflicting, humans may decide that auditory and visual information come from two distinct objects. Besides the same author concluded from previous experiments on auditory-visual interactions that in general the auditory sense captures temporal patterns (see the vision illusion effect), and the visual sense captures the spatial location (see the ventriloquism effect) in case of discrepant auditory and visual inputs. The perceived multimodal object seems to result from a process of multisensory weighting, where the weights attributed to the different sensory inputs depend on their "appropriateness" with respect to the context of the task (e.g. predominance of the spatial or temporal dimension) [83].

3.3 Experiments on auditory-visual depth perception

The interesting result for the present experiment is that if the spatial and temporal rules of multisensory integration are followed, auditory cues may help to resolve ambiguous visual information, especially for localization tasks. We present two experiments investigating the influence of auditory cues on the perceived visual order in depth. The unity assumption plays a major role in auditory-visual displays because it is a necessary condition for multimodal interactions. If the unity is too weak, the interaction between the two senses will be strongly reduced. With this in mind, it was decided to use colors as stimuli for the experiment: visual colors of two squares, e.g. red and blue, were associated with the recorded spoken words "red" and "blue" and the rendered visual orders in depth were consistent with the rendered audio orders in depth, i.e. if the red square appeared in front on the visual display, the audio signal "red" would be louder than (or equal to) the audio signal "blue", and vice versa. Visual and auditory depths were rendered by manipulating respectively the coefficient of transparency of the overlapping surface and the intensity difference between the two audio signals. Semi-transparency was used as the visual depth cue because, as already seen in Section 3.1.2, it is very easy to implement and enables to see through the object that overlaps other objects. For auditory depth perception, the intensity cue is as well the most obvious and easy-to-implement piece of information: when the sound source moves away from the listener, its intensity decreases. Other cues provide information about sound source depth, such as the direct-to-reverberant energy ratio, however in the case of speech signals Zahorik [103] showed that the intensity cue is weighted more than the direct-to-reverberant energy ratio.

3.3.1 Experiment 1

This experiment may also be found in Paper A.

Method

Participants

Sixteen Italian volunteers (6 women and 10 men) participated in the experiment. Their ages ranged from 20 to 44 years and all of them had at least basic knowledge of the English language. All reported to have normal or corrected-to-normal vision and normal hearing. All studied or worked at the University of Verona, Italy. None of them worked in the field of crossmodal interactions and they were all naive as for the purpose of the experiment.

Stimuli and Apparatus

Visual stimuli. Each stimulus appeared in the middle of an Apple MacBook Pro 15-inch Widescreen Display (1440×900 pixels). The viewing distance was about 70 cm from the display. The patterns in Figure 3.2 illustrate the stimulus shape. The stimulus consisted of two overlapping 7.8 cm × 7.8 cm squares in the middle of a permanent white rectangular background corresponding to the display area. One square was red ($C_1 = (1, 0, 0)$ in the RGB color space and the second one was blue ($C_2 = (0, 0, 1)$). To simulate transparency, the color C_3 of the overlapping area was a linear combination of the red and blue colors, such that

$$C_3 = C_1 * (1 - \alpha) + C_2 * \alpha \tag{3.2}$$

where α was the coefficient of transparency which took nine values from 0.3 to 0.7 with a 0.05 increment. The overlapping squares appeared for 1 second, then the subsequent stimulus appeared 3 seconds after the subject answered by pressing a key. Theoretically, the point of bistable transparency arises at $\alpha = 0.5$ (which was verified during a preliminary visual experiment), while α values smaller than 0.5

make the red square appear in front of the blue one, and α values greater than 0.5 make the blue square appear in front of the red one. During the aforementioned preliminary visual experiment, the whole range of α values were explored from 0 to 1, and people were asked to determine under visual conditions only which square appeared to be in front of the other. It was found that no confusion arose for α values smaller than 0.3 or higher than 0.7.

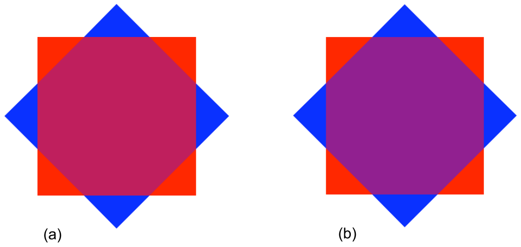


Fig. 3.2. *Visual stimulus used for the experiment. (a) The red square appears in front of the blue square (α = 0.3). (b) Bistable transparency (α = 0.5).*

Auditory stimuli. Each visual stimulus was paired with an auditory stimulus consisting of the words “red” and “blue” presented simultaneously. These two sounds were recorded separately in a quiet room using a Marantz portable audio recorder PMD660 set at the same sound level for both sound signals. The speaker was the author herself and the two words were recorded in stereo with the built-in microphone in the uncompressed wav format, at 44.1 kHz sampling frequency. The two sound files were then time-aligned, shortened to 1 second, normalized to the average sound level, and an equal loudness contour was applied to both signals. Figure 3.3 shows the time responses of the resulting left auditory signals, and their magnitude responses smoothed with 1/3-octave filters. In order to create an effect of auditory depth between the two signals, the sound level of each signal was manipulated digitally, while keeping the total sound level constant. The sound level difference ΔL was either -12, -6, -2, 0, 2, 6 or 12 dB, where negative values indicate that the sound level of “red” is greater than the sound level of “blue”, and positive values indicate that the sound level of “blue” is greater than the sound level of “red”. During the experiment auditory stimuli were presented over a pair of Beyerdynamic DT 770 headphones.

Design

The whole experiment consisted of three sessions per test subject, separated with breaks of about ten minutes. The visual and auditory stimuli were synchronized and had both a duration of 1 second. Besides they were congruent, i.e. both of them theoretically led to the same square being in front: negative ΔL values were combined with α values smaller or equal to 0.5, and positive ΔL values were combined with α values greater or equal to 0.5. The case where there was no audio

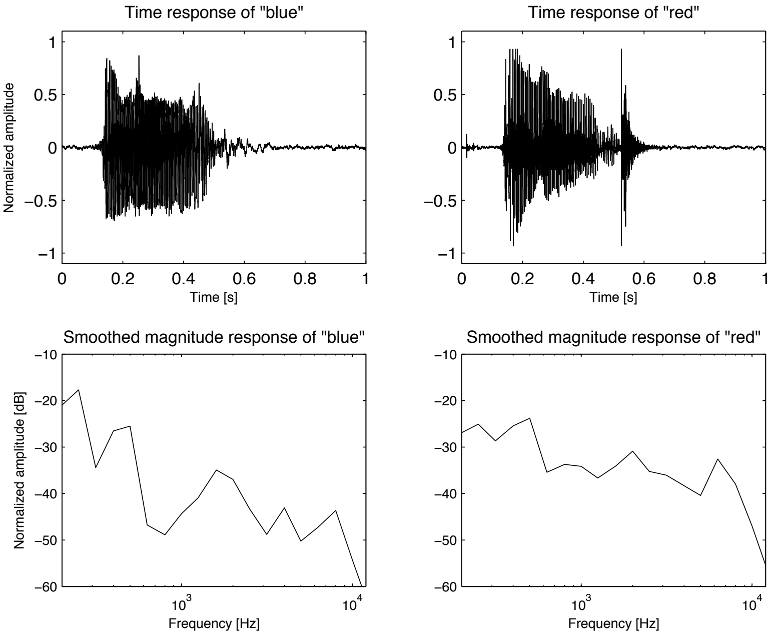


Fig. 3.3. *Time and smoothed magnitude responses of the words "blue" and "red".*

cue, i.e. $\Delta L = 0$, was combined with all the values of α. Therefore the test included 39 different combinations of visual and auditory stimuli, ordered randomly for each session and each subject. Furthermore, between two consecutive trials the blue and red squares were exchanged in order to avoid bias from a specific visual configuration. As a result each pair of α and ΔL was rendered twice in a different visual configuration during each session, giving a total of 78 visual-auditory stimuli per session and per subject.

Procedure

Before the experiment, a written instruction was given to each subject. Participants sat at a viewing distance of about 70 cm from the computer screen and wore headphones which played back the auditory stimuli. Possible auditory-visual interactions were not suggested to the subjects. For each stimulus people were asked to determine which square appeared to be in front of the other and press the key of the corresponding color. To answer, the "V" and "N" keys of the MacBook Pro keyboard were covered respectively with red and blue tags. No time limit to answer was specified, however the written instruction suggested to the participants not to think too much about their answer and rather follow their first impression. In addition to subjects' answers, their response time was also recorded.

Results

Some subjects reported to have pressed the wrong key at least once during the experiment. For each pair $(\Delta L, \alpha)$, 96 answers were collected (16 subjects × 3 sessions × 2 repetitions). The collected data can be represented for each subject by the percentage of answers for the blue square appearing in front, for each combination of auditory and visual stimuli. In that way a percentage smaller than 50% indicates that the answer "red" was given more often than the answer "blue", and a percentage greater than 50% indicates that the subject answered more often "blue" than "red". To assess the multisensory gain of combining redundant multisensory information, results were analyzed for each value of ΔL and were described by a psychometric function representing the percentage of answers "blue" as a function of the α value. The expected outcome is a psychometric function having a S shape: theoretically, answers for "blue" should increase as a function of α, from 0% for $\alpha = 0.3$, to 100% for $\alpha = 0.7$, while values of α close to 0.5 should lead to about 50% answers for "blue". The boxplots of Figure 3.4 summarize the distributions of answers for each value of α, and show a S shape as expected. Besides, for all boxplots answers are more spreaded around $\alpha = 0.5$, also suggesting uncertainty in this region. However, comparing in Figure 3.4 the case where there is no audio level difference, i.e. no audio cue, with cases where there are audio level differences, does not show any noticeable differences in people's answers: if the audio cues would help them in answering correctly, answers for α slightly smaller than 0.5 should be less spreaded and closer to 0% and answers for α slightly greater than 0.5 should be closer to 100%.

To analyze in more details the influence of auditory cues, paired t-tests have been applied to the observation under a given audio level difference ($\Delta L \neq 0$ dB), and the observation without audio cue ($\Delta L = 0$ dB) for the same range of α values. Results are displayed in Table 3.1. Using the 5% significance level, results of the paired t-tests reveal significant differences for observations made under negative audio level differences: when the word "red" is louder than "blue", people use the audio cue to evaluate which square appears in front.

Table 3.1. t and p values of various paired t-tests.

observation 1	observation 2	t	p
$\Delta L = 0$ dB, $\alpha \leq 0.5$	$\Delta L = -12$ dB, $\alpha \leq 0.5$	2.018	0.04699
$\Delta L = 0$ dB, $\alpha \leq 0.5$	$\Delta L = -6$ dB, $\alpha \leq 0.5$	2.6273	0.01033
$\Delta L = 0$ dB, $\alpha \leq 0.5$	$\Delta L = -2$ dB, $\alpha \leq 0.5$	2.0831	0.04048
$\Delta L = 0$ dB, $\alpha \geq 0.5$	$\Delta L = +12$ dB, $\alpha \geq 0.5$	-1.9091	0.05988
$\Delta L = 0$ dB, $\alpha \geq 0.5$	$\Delta L = +6$ dB, $\alpha \geq 0.5$	-0.4561	0.6496
$\Delta L = 0$ dB, $\alpha \geq 0.5$	$\Delta L = +2$ dB, $\alpha \geq 0.5$	-0.5697	0.5705

Another way to investigate the impact of auditory cues is to compare results at $\alpha = 0.5$ when ΔL is positive or negative. Figure 3.5, which displays the distributions of answers for the various values of ΔL at $\alpha = 0.5$, also shows the effect of the word "red" when it is the loudest signal, which is corroborated by paired t-tests

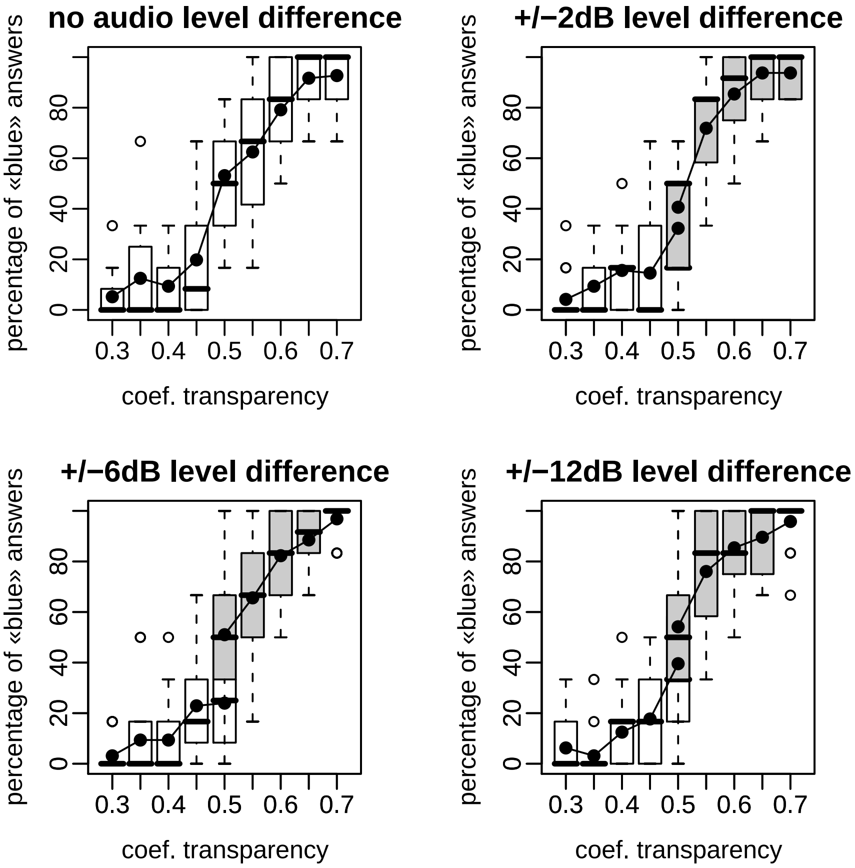


Fig. 3.4. *Experiment 1: Percentage of answers "blue" as a function of the coefficient of transparency. The boxplots depict the median and the 25%/75% percentiles. White boxplots represent results for negative or null audio level differences, and grey boxplots represent results for positive audio level differences. Means are represented by solid circles and are connected by a line. Outliers are marked by 'o'.*

between the distributions for $\Delta L \neq 0$ and the distribution for $\Delta L = 0$. In particular, the most significant difference is found between the distributions $\Delta L = 0$ and $\Delta L =$ -6 dB (the paired t-test gives a mean difference of about 29.17%, with $t = 5.2175$ and $p = 0.0001042$).

Further investigations on the influence of the audio signals on the answers have been performed. The psychometric functions where the average percentage of "blue" answers is plotted against α have shapes close to ogives, it is thus expected that they should become linear when the average percentages are expressed as z scores [30]. This transformation enables to quantify the slopes of the resulting linear functions and compare them. Figure 3.6 illustrates the z scores calculated from p values for $\Delta L = 0$ dB and $\Delta L = \pm 12$ dB. The shapes of the z scores are not

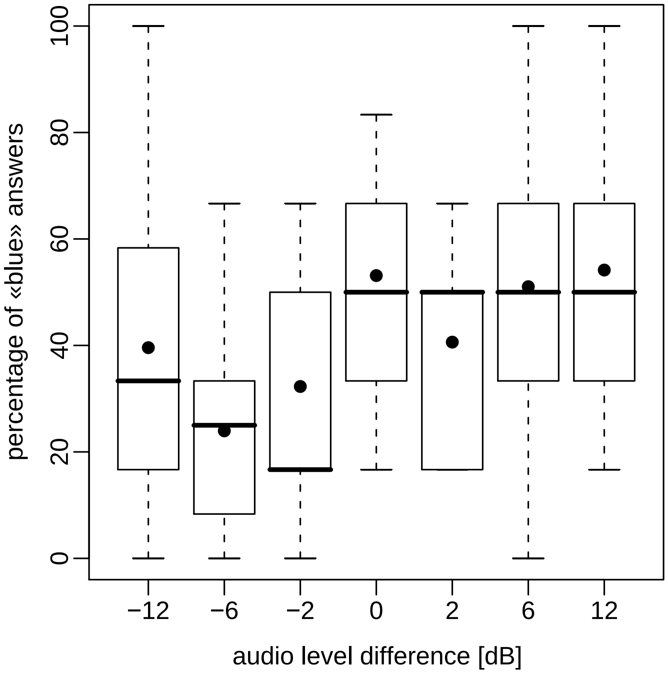


Fig. 3.5. *Experiment 1: Percentage of answers "blue" at $\alpha = 0.5$ for the various values of auditory level differences. The boxplots depict the median and the 25%/75% percentiles. Solid circles represent the means of the distributions.*

linear over the whole range of α values, however they approach a linear behavior if the range of α values is restrained to [0.4 ; 0.6]. Linear regressions performed on the distributions for $\Delta L = 0$ [slope = 10.07 with $s.d. = 1.11$, $F(1,3) = 82.49$, $p = 0.003$] and for $\Delta L = \pm 12$ dB [slope = 10.48 with $s.d. = 1.31$, $F(1,3) = 63.65$, $p = 0.004$] do not show a significant difference between the two slopes (angle $\simeq 0.22°$). This result suggests that in average auditory cues have no effect on the participants' answers, and consequently that the increase of answers "red" when the word "red" is louder than "blue" is an isolated effect.

Finally, analysis of subjects' response time do not show any significant difference between the case $\Delta L = 0$ and $\Delta L \neq 0$ (a paired t-test between the distributions for $\Delta L = 0$ and for $\Delta L = \pm 12$ dB gives a mean difference of 0.04 s with $t = 0.9766$ and $p = 0.3574$). Response times are shown in Figure 3.7 and simply suggest that some people take more time to answer when the coefficient of transparency is close to 0.5 because it is visually more difficult to determine which square is in front of the other.

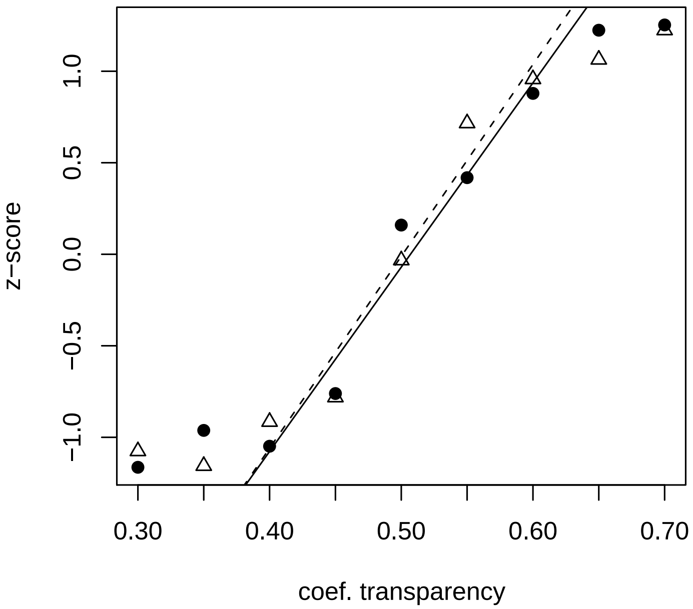


Fig. 3.6. *Experiment 1: Z scores of average percentage of answers "blue" as a function of transparency coefficient. Solid circles: results for $\Delta L = 0$. Triangles: results for $\Delta L = \pm 12$ dB. Solid line: linear regression of the means for $\Delta L = 0$ and $\alpha \in [0.4; 0.6]$. Dashed line: linear regression of the means for $\Delta L = \pm 12$ dB and $\alpha \in [0.4; 0.6]$.*

Discussion

Results of the experiment did not reveal any significant influence of auditory cues on visual perception of depth. However an isolated effect of the word "red" has been found when it is louder than the word "blue". An equal loudness contour applied to both audio signals should prevent from perceiving one signal louder than the other, even with different temporal dynamics. It is therefore difficult to find a plausible explanation.

For signals coming from different senses to be integrated, the brain has to establish a correspondence between these signals and decide whether they come from the same object or event. This auditory-visual integration depends on the level of abstraction of the auditory and visual representations that are involved. In the specific case of auditory-visual speech perception, one may argue that speech would more easily fuse with the vision of the mouth generating the words, like in the case of the ventriloquism effect. According to recent studies [32, 78], two hypotheses could explain the integration of the senses in this case and are based on low-level processes: first there is a strong temporal correlation between the audi-

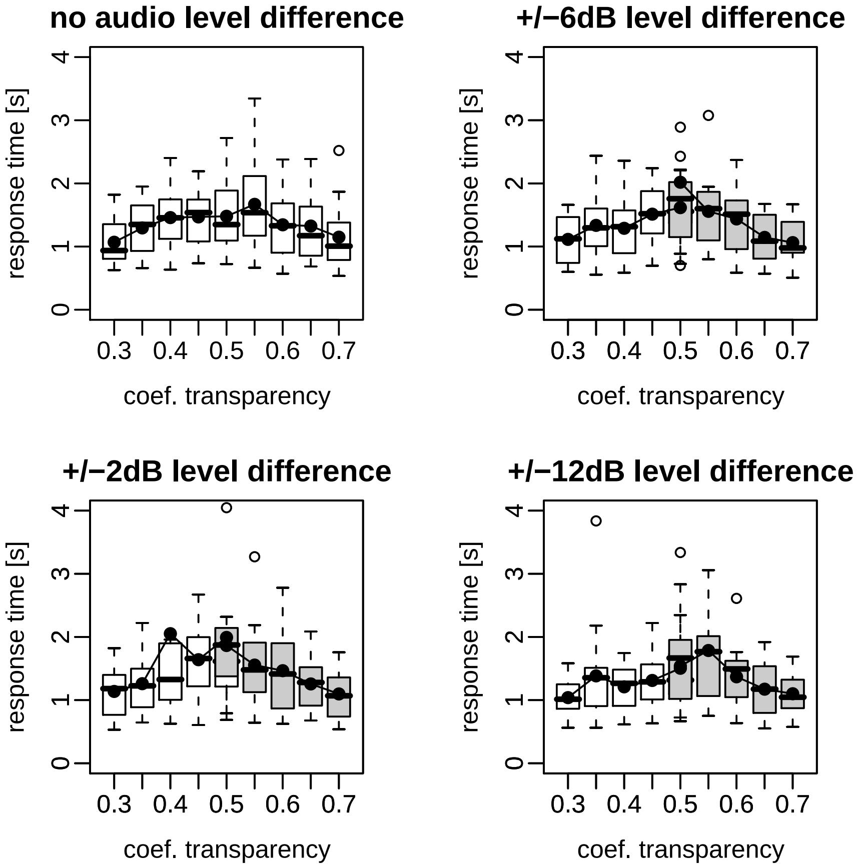


Fig. 3.7. *Experiment 1: Response time values as a function of the coefficient of transparency. The boxplots depict the median and the 25%/75% percentiles. White boxplots represent results for negative or null audio level differences, and grey boxplots represent results for positive audio level differences. Means (represented by solid circles) are connected by a line. Outliers are marked by 'o'.*

tory and the visual signals (e.g. of the sound level with the degree of lip aperture), and secondly a coherence of movement (a spectro-temporal variation of the audio signal may be correlated with a movement of the lips). Thereby several mechanisms of auditory-visual integration may cooperate, at different levels of processing in the brain. For auditory-visual perception of speech, the integration process seems to be based on low-level processes which could explain the robust integration of audio and visual information. In our experiment on the contrary, a higher semantic level of processing in the brain is necessary to give a meaning to the words "red" and "blue ", which could explain why our experiment does not reveal any auditory-visual unity: the content of the auditory signals is not taken into account in the integration process and is therefore irrelevant for judging the visual order

of the two squares.
Nevertheless, results of the present experiment conflict with previous investigations by Ecker and Heller [18]. Auditory stimuli used in the two experiments were of different nature: Ecker and Heller used recorded sounds of rolling and impacting whereas we used speech signals. However Ballas and Howard [3] suggested that speech and everyday sounds, including rolling and impacting, are similar in several aspects, in particular everyday sounds may be thought of as a form of language because they are integrated on the basis of cognitive processes similar to those used to perceive speech. Causes explaining the difference between results from Ecker et al and ours are not obvious. However differences in the design of the two experiments might give some clue. First, the instructions given to the subjects were different: while Ecker and Heller instructed their subjects to "make a judgment about a ball and the path it travels", therefore not specifying on which sense to base their judgment, we gave the participants the instruction to "determine which square appears in front" therefore implying a visual judgment. Besides their experiment dealt with dynamic auditory and visual information whereas ours used static stimuli. Therefore in addition to temporally and spatially coincident auditory-visual cues, dynamic information from both senses may reinforce the auditory-visual unity. In order to verify this assumption, a proposed follow-up of the present experiment is to introduce a dynamic factor by delaying one visual square and its corresponding auditory stimulus.

3.3.2 Experiment 2

Method

Participants

Eleven Italian volunteers (3 women and 8 men) participated in the experiment. None of them had participated in the first experiment. Their ages ranged from 21 to 33 years and all of them had at least basic knowledge of the English language. All reported to have normal or corrected-to-normal vision and normal hearing. All studied or worked at the University of Verona, Italy. None of them worked in the field of crossmodal interactions and they were all naive as for the purpose of the experiment.

Stimuli and Apparatus

Audio and visual stimuli as well as apparatus were those used in experiment 1.

Design

The whole experiment consisted of three sessions per test subject, separated with breaks of about ten minutes. For each of the two colors red and blue, the auditory stimulus was synchronized with the corresponding visual stimulus. Parameters of the auditory and visual signals were combined like in experiment 1, so that auditory and visual information never conflicted. Unlike experiment 1, a 200 ms delay was introduced between the two squares. The order of the two squares was randomized, however each subject received the same number of stimuli with the red square appearing first and with the blue square appearing first. Besides, the visual blue and red squares were not exchanged between two consecutive trials like in experiment 1, but their orientation was randomized, with the relative angle between the two squares remaining at 45°(see Figure 3.8 for an example of visual stimulus). A total of 78 auditory-visual stimuli were presented per session.

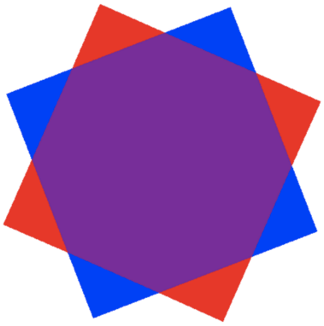


Fig. 3.8. *Experiment 2: An example of the visual configuration of the two squares. The orientation of the pair of squares was randomized for each trial.*

Procedure

Before the experiment, a written instruction was given to each subject. Participants sat at a viewing distance of about 70 cm from the computer screen and wore headphones which played back the auditory stimuli. Possible auditory-visual interactions were not suggested to the subjects. For each stimulus people were asked to determine which square appeared to be in front of the other, and it was emphasized that the goal was not to detect which square appeared first. To answer, the “Q” and “/” keys of the MacBook Pro keyboard were covered respectively with red and blue tags. The keys used were changed after the first experiment because it was found that a larger distance and a dissymmetry between the two keys limited the number of wrong keys pressed. No time limit to answer was specified, however the written instruction suggested to the participants not to think too much about their answer and rather follow their first impression. In addition to subjects’ answers, their response time was also recorded.

Results

The boxplots of Figure 3.9 summarize the distributions of answers as a function of α and for each value of audio level difference. Similar analyses to those performed on the results of experiment 1 did not show any significant difference between answers to stimuli without audio cue ($\Delta L = 0$) and answers to stimuli with audio cues ($\Delta L \neq 0$). Results of the paired t-tests are summarized in Table 3.2.

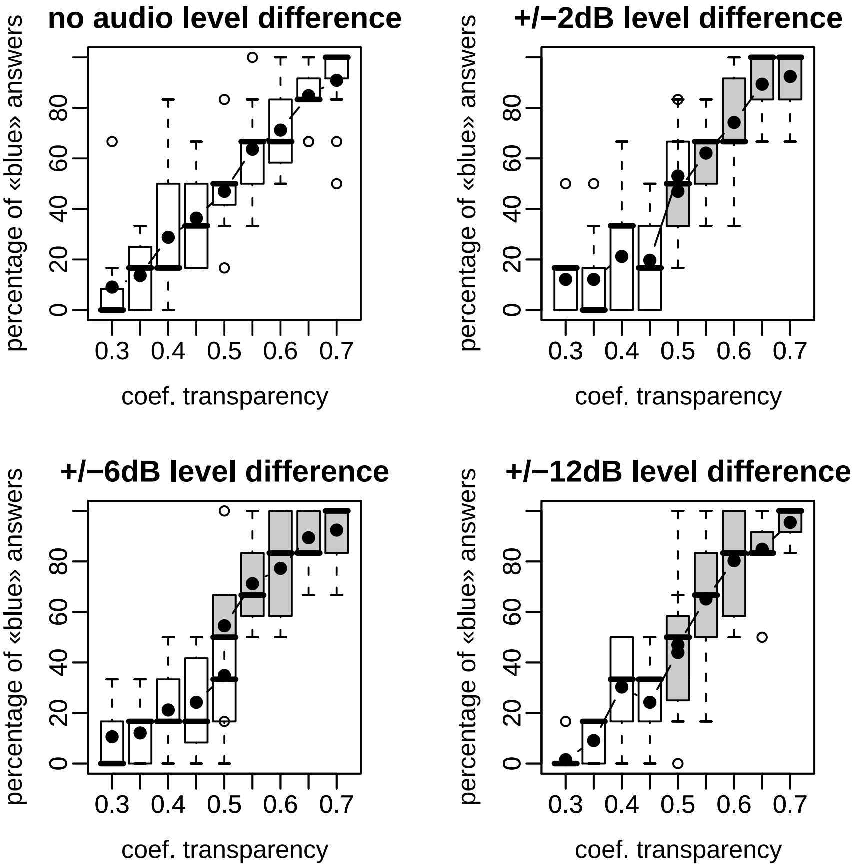


Fig. 3.9. *Experiment 2: Percentage of answers "blue" as a function of the coefficient of transparency. The boxplots depict the median and the 25%/75% percentiles. White boxplots represent results for negative or null audio level differences, and grey boxplots represent results for positive audio level differences. Means (represented by solid circles) are connected by a line. Outliers are marked by 'o'.*

For $\alpha = 0.5$, the distributions of answers for the various values of ΔL are shown in Figure 3.10. Paired t-tests to compare the distributions for the different values of ΔL at $\alpha = 0.5$ did not reveal any significant difference. In particular, unlike experiment 1, the paired t-test between the distributions $\Delta L = 0$ and $\Delta L =$ -6 dB did not show a significant difference (mean difference of about 11.36%, with $t =$ 1.0764 and $p =$ 0.3070), which rejects an influence of auditory cues at the point of bistable transparency.

A linear regression applied to the z scores for $\Delta L = 0$ dB and $\Delta L = \pm 12$ dB in the range $\alpha \in [0.4\ ;\ 0.6]$ gave the following results: for $\Delta L = 0$: slope = 7.49 with $s.d. = 0.09$, $F(1,3) = 171.1$, $p = 0.0009$, and for $\Delta L = \pm 12$ dB: slope = 8.18 with $s.d. = 0.25$, $F(1,3) = 25.87$, $p = 0.015$ (See Figure 3.11). The angle between the

Table 3.2. t and p values of various paired t-tests.

observation 1	observation 2	t	p
ΔL = 0 dB, $\alpha \leq$0.5	ΔL = -12 dB, $\alpha \leq$0.5	1.7473	0.08155
ΔL = 0 dB, $\alpha \leq$0.5	ΔL = -6 dB, $\alpha \leq$0.5	1.8172	0.07012
ΔL = 0 dB, $\alpha \leq$0.5	ΔL = -2 dB, $\alpha \leq$0.5	0.9364	0.3498
ΔL = 0 dB, $\alpha \geq$0.5	ΔL = +12 dB, $\alpha \geq$0.5	-0.6206	0.5353
ΔL = 0 dB, $\alpha \geq$0.5	ΔL = +6 dB, $\alpha \geq$0.5	-1.4511	0.1477
ΔL = 0 dB, $\alpha \geq$0.5	ΔL = +2 dB, $\alpha \geq$0.5	-0.3587	0.72

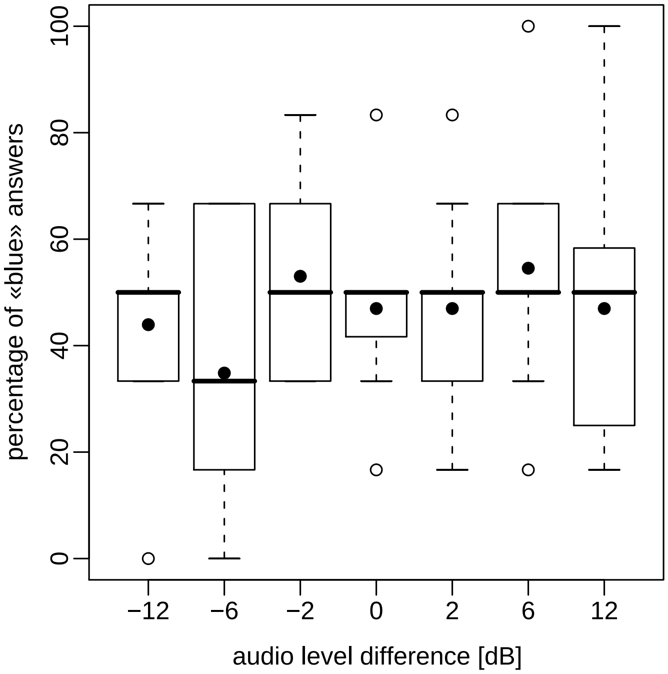


Fig. 3.10. *Experiment 2: Percentage of answers "blue" at α = 0.5 for the various values of auditory level differences. The boxplots depict the median and the 25%/75% percentiles. Solid circles represent the means of the distributions.*

two lines was slightly greater than in experiment 1 (angle ≃ 0.63°).

Difference between subjects' response time for $\Delta L = 0$ and $\Delta L \neq 0$ was not significant neither. A paired t-test between the distributions for $\Delta L = 0$ and for $\Delta L = \pm 12$ dB gives a mean difference of 0.019 s with $t = 0.2877$ and $p =$ 0.7809. Finally delaying the second audio-visual square introduced a new factor that could influence people's answers. Indeed several persons reported that they had the impression that the second square systematically appeared in front of the first square. Figure 3.12 shows the distributions of the answers according to the order of appear-

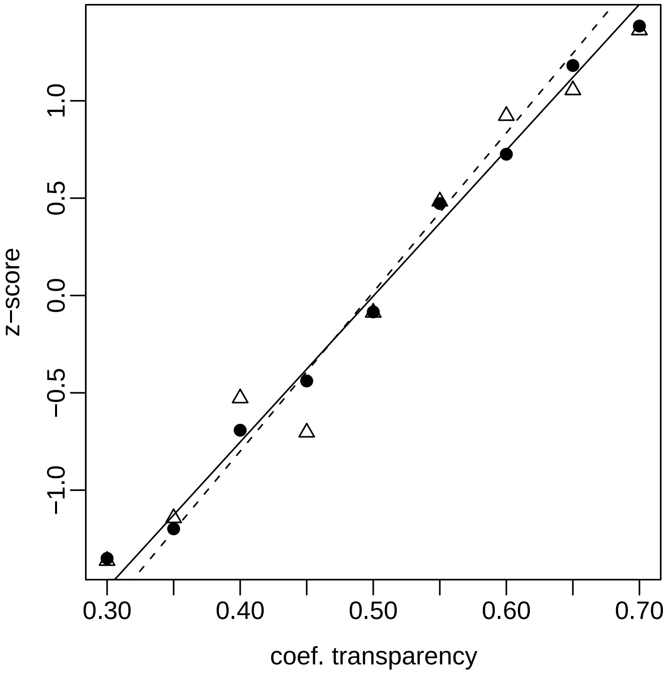


Fig. 3.11. *Experiment 2: Z scores of average percentage of answers "blue" as a function of transparency coefficient. Solid circles: results for $\Delta L = 0$. Triangles: results for $\Delta L = \pm 12$ dB. Solid line: linear regression of the means for $\Delta L = 0$ and $\alpha \in [0.4; 0.6]$. Dashed line: linear regression of the means for $\Delta L = \pm 12$ dB and $\alpha \in [0.4; 0.6]$.*

ance: "red/blue" or "blue/red". The curve connecting the means for the temporal configuration "red/blue" is slightly higher than the curve for the "blue/red" configuration, suggesting that in the first configuration more answers are given for the color blue whereas in the second configuration more answers are given for the color red. A paired t-test between the two distributions effectively revealed a significant mean difference of 22.6% with $t = 13.136$ and $p < 2.2 \times 10^{-16}$. Therefore it can be concluded that the temporal configuration of the two squares has a significant effect on the answers: the second square tends to appear in front of the first square.

3.3.3 Conclusion

Two experiments were carried out on the influence of auditory cues on the perception of visual order in depth. In the first experiment, visual stimuli consisted in a layered 2D drawing of two squares respectively blue and red using semi-transparency. Auditory signals of the two words "red" and "blue" were presented simultaneously to the images. Subjects were required to determine which square appeared in front of the other in these cross-modal conditions. The coefficient of

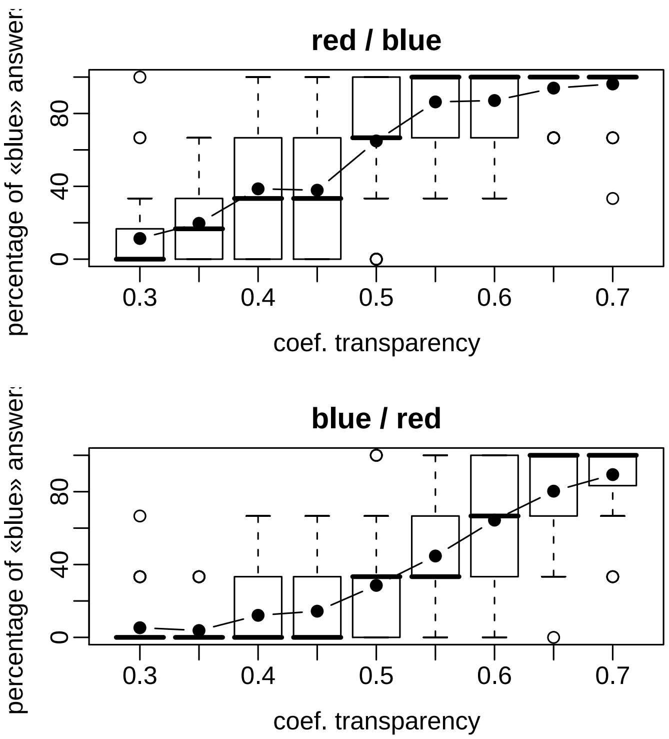


Fig. 3.12. *Experiment 2: Comparison of the two temporal configurations "red/blue" and "blue/red". The boxplots depict the median and the 25%/75% percentiles. Means are connected by a line. Outliers are marked by 'o'.*

transparency as well as the audio level difference between the two speech signals "red" and "blue" were systematically varied. No significant influence of auditory cues on perceived order in depth was found, but a surprising effect of the solely word "red" when it was louder than "blue".

In the second experiment, one visual square and its corresponding auditory stimulus were delayed. Unlike experiment 1, no effect of the word "red" was found, although the same auditory stimuli were used in both experiments. Auditory cues had no significant effect on the perception of visual depth, meaning that the dynamic cue did not improve the auditory-visual integration. As it was mentioned earlier the auditory and visual signals refer to different levels of representation, and most of all the process of sensory integration is still largely unknown. In particular, if one distinguishes between structural (e.g. spatio-temporal correspondence) and cognitive (e.g. semantic congruency) factors, further research is needed to understand their respective contributions in the process of sensory integration, in addition these two types of factors may not be clearly separated [88].

On the side of the main issue under study, the second experiment revealed an unexpected effect of the temporal presentation of the two visual stimuli: subjects tend to see the second square in front of the first one. No literature relating that phenomenon was found by the author, and it would be interesting to investigate further this visual effect. A possible direction of study is to consider the temporal dynamics in bistable perception, that is the temporal variations in dominance periods of one percept over another.

Obviously, using the video of a person talking synchronized with the corresponding recorded speech would lead to the perception of a unitary event because both visual and audio displays have the same source (e.g. a person talking). In this context, we believe that techniques based on auditory-visual cues of depth could be very valuable for applications such as mobile TV, in particular to exaggerate perspective effects and improve thereby the perceived quality of interaction and content fruition. Indeed, the recent development of mobile TV opens the way for new audio-visual rendering techniques especially because of the limited size of the screen and reduced budgets. Sasse and Knoche [74] have demonstrated that the requirements for audio and video quality depend on the context of use. For mobile TV, factors on the perceived quality include the shot types, the audio quality and the legibility of text if present. Watching a football match on a mobile phone is very illustrative: people expect to be able to recognize the players and see the ball, which is not as obvious as on a normal TV screen. For this kind of applications, simple techniques such as transparency could be used to give a sense of depth. In the auditory domain, similar techniques such as the manipulation of the sound level and/or the direct-to-reverberant energy ratio would as well contribute to a consistent 3D rendering of audio-visual contents and improve the quality and efficiency of multisensory products, as highlighted by Spence and Zampini [89].

4

The Digital Waveguide Mesh as a model to render depth cues

Starting from observations made on the acoustics of a real pipe, this chapter will describe a virtual audio space modeling a tubular environment and which reproduces some interesting characteristics of the real pipe for rendering depth information. Real-time implementations of the model in 2D will also be presented.

4.1 Previous work

This work was initiated by Fontana and Rocchesso who observed an exaggeration of reverberation in real tubular environments [22] . Measurements in a 10.2 m long pipe having a 0.3 m round section validated that observation. Furthermore, a listening experiment showed that people were able to perceive depth in the pipe. Given these results, the authors proposed the design of a virtual pipe for depth rendering in auditory displays. Their design strategy follows a physics-based modeling approach: the model is not driven by acoustical parameters such as reverberation time, air absorption, doppler effect, source presence, etc... but by the tube's geometry and dimensions, positions of the listener and the sound source(s), and reflection properties of the boundaries. The proposed model makes use of a 3D Digital Waveguide Mesh (DWM), which had already shown its ability to provide artificial reverberation [76] and simulate complex, large-scale acoustical environments [65]. Besides, this method allows an accurate simulation especially at low frequencies [75], and may be used to simulate real-world acoustical rooms with an adequate modeling of the boundaries by means of properly tuned absorption filters ([40, 45]). The DWM and its different geometries have largely been described in past literature, see [21, 46, 75], and recent developments in the simulation of frequency-dependent boundaries may be found in [48]. Fontana et al. used a 3D rectilinear DWM to simulate the acoustic wave propagation in a tubular environment, and showed its ability to render robust depth cues and in particular an exaggeration of the reverberation [23].
The motivations for using such a model are driven by its usability in human-computer interfaces (HCI). First, the proposed model renders monaural depth cues and therefore does not rely on the reproduction hardware configuration. It can be reproduced by most consumer systems even under non-ideal listening conditions.

Furthermore, in HCI contexts, spatialization of sound sources in depth is meant to be a tool, for example for sound navigation, menu selection or to provide different (dynamic) layers of attention to the listener. Therefore the objective is very different from virtual reality environments which aim at the illusion of presence by recreating a true-like listening scenario, and usually make use of auralization techniques based on Head-Related Transfer Functions to reproduce with high fidelity the acoustic wave propagation in the near field of the listener. On the other hand, the physics-based model provides the only relevant pieces of information needed to perceive depth in the far field, namely the intensity and the direct-to-reverberant energy ratio. In vision, an easy analogy is the use of semi-transparency, like in boxing with the Nintendo Wii, which was already mentioned in Section 3.1. This technique does not reproduce what happens in real life but is a tool that provides a representation of the scenario that is immediately understood by the users.

4.2 The 2D DWM

The 3D DWM resulted to be too resource consuming for real-time simulations of 3D environments of decent dimensions. While downsampling may help in reducing the CPU processing load, a more efficient alternative is to use a model in 2D, consequently simulating a membrane. Although sounding less natural than 3D simulations, the resulting bidimensional audio space presents similar properties, especially for depth rendering.

4.2.1 Theory

Our sound propagation model consists of a two-dimensional rectilinear DWM. An example of a 2D DWM modeling a rectangular membrane is represented in Figure 4.1. Other geometries may be approximated using this scheme, such as a trapezoid presented in Chapter 5. Nevertheless it seems natural to use a rectangular geometry to approximate (in 2D) the acoustics of a tube. This scheme may be seen as a regular grid of scattering junctions. Each internal junction is connected to four other junctions via waveguides providing acoustic wave transmission. Each waveguide models the wave decomposition of a pressure signal p into its wave components p^+ and p^-, and each lossless junction scatters 4 input signals coming from orthogonal directions, p_1^+, ..., p_4^+ into corresponding output signals p_1^-, ..., p_4^- (see Figure 4.2). Then, Kirchoff's laws yield to the computation of the pressure signal p at each junction as a sum of incoming wave variables p_i^+:

$$p = \frac{1}{2}\sum_{i=1}^{4} p_i^+ \tag{4.1}$$

Moreover, the digital waveguide theory states that the sampling frequency f_s of a bidimensional mesh is determined as [46]:

$$f_s = \frac{c\sqrt{2}}{d_W} \tag{4.2}$$

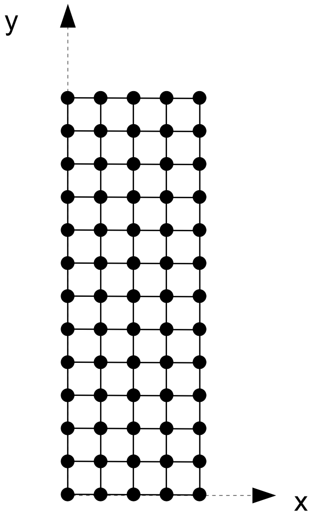


Fig. 4.1. *The two-dimensional rectilinear waveguide mesh modeling a rectangular membrane.*

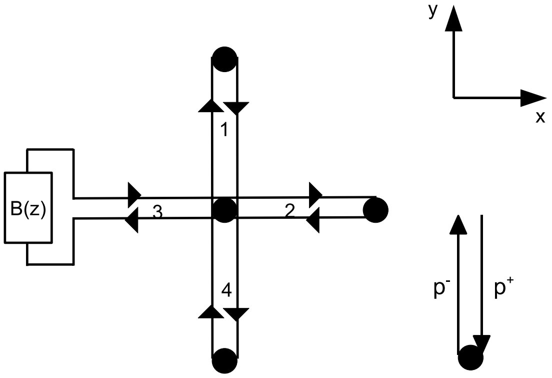


Fig. 4.2. *Zoom on a junction which is connected to other junctions via waveguides 1, 2 and 4. Waveguide 3 leads to a partially absorbing section, modeled using a digital waveguide filter. Triangles filled in black represent oriented unit delays.*

where c is the nominal wave propagation speed and d_W the spatial sampling interval or the length of each waveguide. Note that this relation is exact only for diagonal propagation and at $f = 0$ for all directions, because of the dispersion error (some spatial frequencies travel slower along the mesh). Figure 4.3 displays the dispersion error (represented by the relative phase velocity) in the rectilinear DWM as a function of frequency and propagation angle. It can be seen on this figure that near the center of the plot, corresponding to frequencies up to $f_s/8$ (i.e. $\simeq$ 5.5 kHz at 44.1 kHz sampling frequency), the dispersion error is very limited and may therefore be neglected in that frequency range. We assume that dispersion error in not critical for depth perception, although a proper evaluation of the influence of this error should be conducted. From equation (4.2) we can deduce

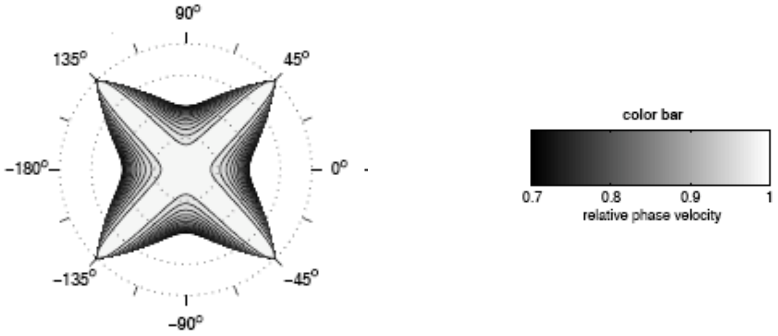


Fig. 4.3. *Dispersion error as a function of frequency and propagation angle. Dotted line circles indicate* $f = (\frac{1}{8}, \frac{1}{4}, \frac{3}{8}, \frac{1}{2}).f_s$. *Taken from [92].*

the length d_W of each waveguide. For $c = 343$ m/s the speed of sound in air and $f_s = 44.1$ kHz, a waveguide corresponds to the propagation of sound pressure waves on a distance of 1.10 cm. Given the length and width of a virtual membrane to be modeled, it is then possible to calculate the number of waveguides necessary for both dimensions.

The properties of the wall materials contribute to the acoustics of a space. This is also the case for a bidimensional acoustic environment since horizontal waves interact with the surface's boundaries. Reflections from surfaces are modeled by Digital Waveguide Filters (DWF), whose coefficients may be tuned to model specific reflective properties of surfaces [21] (See Figure 4.2). The conversion from measurements of physical parameters of real surfaces to DWF coefficients follows the method described in [40]. For a given material, its absorption coefficients typically measured for octave bands (125, 250, 500, 1000, 2000 and 4000 Hz) are converted to the reflection transfer function using the following relationship:

$$|R(j\omega)| = \sqrt{1 - \alpha(\omega)} \tag{4.3}$$

where $R(j\omega)$ is the reflection transfer function and $\alpha(\omega)$ are the frequency dependent absorption coefficients. Then the `invfreqz` function in `Matlab` may be used to design a stable first-order IIR filter fitting the data.

4.2.2 Acoustical properties of the model

Given that model, it is then possible to simulate a virtual environment in which the position of the audio input corresponds to a virtual sound source (or a virtual microphone), and the position of measurement (i.e. the audio output) corresponds to the listening position (or a virtual loudspeaker). From now on, the listening position as well as the position of the sound source(s) are assumed to be pointwise for the sake of simplification, and are always positioned on the main axis of the membrane. The alignment of the listening and sound source(s) positions is intended to produce cues of distances straight ahead of the listener (i.e. depth cues). Moreover, numerical artifacts in DWM simulations may affect the responses especially in the proximity of the mesh boundaries, therefore the listening and sound source(s) positions are always set at least a few junctions away from the mesh boundaries.

Measurements in the virtual environment

Simulations of the listening environment were primarily carried out in `Matlab` in order to investigate the main auditory monaural depth cues, the intensity (or overall magnitude) and the direct-to-reverberant energy ratio, and their variation with the model's parameters. The overall magnitude of a signal s is given by $10 \log \frac{1}{k} \sum_{k=1}^{N} s^2(n)$ where N is the signal length.

Computation of the direct-to-reverberant energy ratio is not straightforward. In binaural sound reproduction using Head Related Impulse Responses (HRTFs), the direct energy is usually integrated among the first 2.5 ms of the delay-free impulse responses. This duration approximates the duration of the HRTFs (measured in anechoic conditions) and therefore captures the direct path of the sound signal [62]. Our spatial rendering system does not make use of HRTFs, therefore the direct energy should only capture the direct signal component, and the reverberant energy should correspond to the remaining part of the signal, starting from the first reflection. However, due to the dispersion error in the DWM (refer to Section 4.2.1), the time of arrival of the first reflection is frequency-dependent. As shown in Figure 4.3, low frequencies travel at the nominal wave propagation speed (343 m/s), but high frequencies travel slower along the axes of the mesh (down to $c/\sqrt{2}$). Therefore, when the distance d between the sound source and the listener increases,

- the arrival time of the direct low frequencies is equal to d/c
- the delay of the first reflected low frequencies decreases and is equal to $\sqrt{d^2 + W^2} - d$, where W is the width of the DWM. Note that if W is very small, e.g. 0.3 m, the reflected low frequencies even merge with the direct ones.
- the delay of the direct high frequencies increases ($= \sqrt{2}d - d$)

Therefore the boundary between the direct signal and the first reflection is not obvious. In particular, both signals can merge at low frequencies when the mesh is narrow. We therefore choose to define the direct signal as the first 2.5 ms after the arrival time of the direct low frequencies. This allows to make comparisons with values from other studies based on the use of HRTFs, and to propose a standardized measure in the style of C50 and C80, which define the ratio of the energy in the early sound (the first 50 ms and 80 ms respectively) compared to that of the reverberant sound. To sum up, the direct-to-reverberant energy ratio in the DWM is computed as follows:

1. For each impulse response the delay of the direct sound is deduced from the distance between the sound source and the listening point. It is then removed from the impulse response.

2. The direct energy is integrated among the first 2.5 ms of the delay-free impulse responses.
3. The reverberant energy is calculated from the tail of the delay-free impulse responses.

Let's consider the simulation of a 10 × 0.3 m^2 rectangular membrane. The listening point is located at $y = 1$ m, while the sound source is successively positioned at one-meter steps ranging from $y = 2$ to $y = 9$ meters (See Figure 4.4).

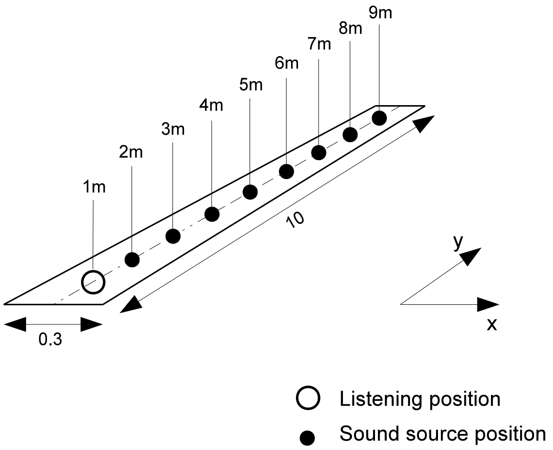


Fig. 4.4. *The virtual membrane showing the listening and measurement positions. All sizes are in meters.*

Examples of time and frequency responses computed at 2 and 9 m are displayed in Figure 4.5.

Influence of the boundary conditions

In order to investigate the influence of the boundary conditions at the membrane sides on depth cues, three different materials are applied to the sides of the membrane (along the y-axis): hard surface (bricks, plaster), carpet and acoustic plaster. Refer to Table 4.1 for their octave-band absorption coefficients. As for the membrane ends (along the x-axis), they are modeled as total reflective surfaces.

Table 4.1. Absorption coefficients α at different frequencies according to the material (from [49]).

Material	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz
hard surface (bricks, plaster....)	0.02	0.02	0.03	0.03	0.04	0.05
carpet	0.02	0.03	0.05	0.10	0.30	0.50
acoustic plaster	0.08	0.15	0.30	0.50	0.60	0.70

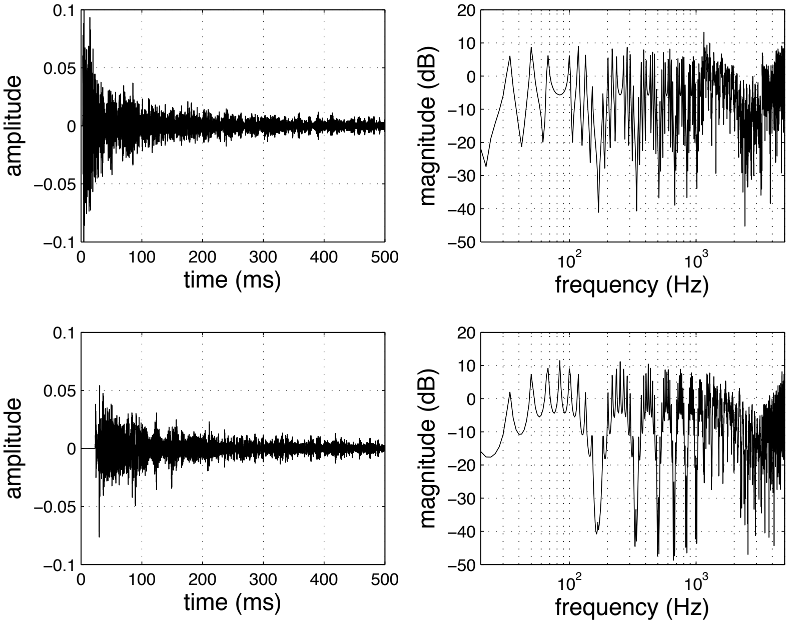


Fig. 4.5. *Time responses (left) and frequency responses up to 5 kHz (right) at positions $y = 2$ m (top) and $y = 9$ m (bottom).*

Figures 4.6 and 4.7 respectively show the intensity and the direct-to-reverberant energy ratio cues as a function of distance, for the aforementioned materials. Note that the intensity cue is displayed as the variation of the overall average magnitude of the signal, in Figure 4.6. The intensity variation in open space (i.e. in anechoic conditions), that is characterized by a reduction of 6 dB per distance doubling [47], is also shown. Besides, all magnitudes are normalized such that they are all equal at a distance of 1 m from the measurement position. It is then easier to compare the intensity variations for the different materials and with the free field condition. As for the variation of the direct-to-reverberant energy ratio v shown in Figure 4.7, it is also computed for a semi-reverberant auditorium as a function of distance r, modeled by Zahorik with the function $v = -3.64 \log_2(r) + 10.76$ [105]. From Figure 4.6, the intensity decreases with distance, and the dynamic range greatly depends on the material used to model the membrane boundaries: while intensity is almost constant with hard surfaces, the intensity change with distance using acoustic plaster or carpet is more similar to the open space case. Depending on the application, it is therefore possible to shape the intensity variation with distance. As for the direct-to-reverberant energy ratio in the mesh, all curves computed in the DWM follow a similar trajectory to the one computed in the semi-reverberant auditorium but with lower values, in particular when the mesh sides simulate hard surfaces. It means that the reverberant energy is exaggerated in the virtual environment, which validates the observation made in a 3D DWM by Fontana et al. [23]. This property may allow to enhance depth perception and

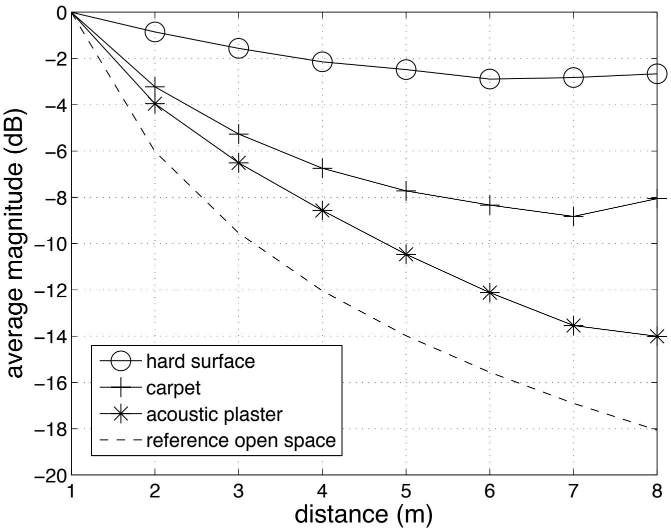


Fig. 4.6. *Average magnitude as a function of distance for various boundary materials.*

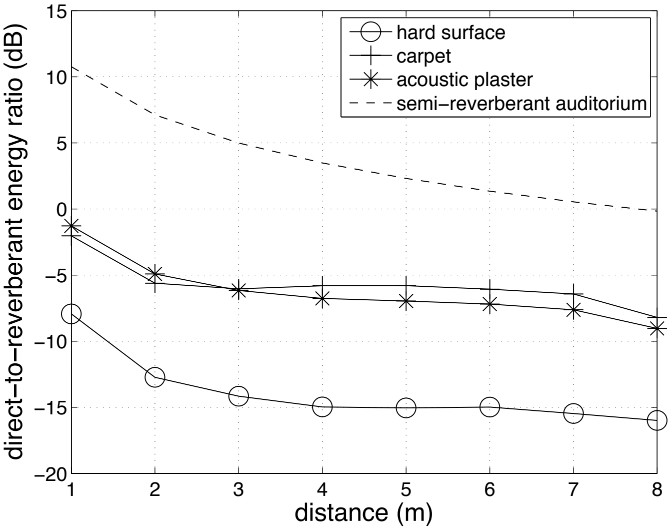


Fig. 4.7. *Direct-to-reverberant energy ratio as a function of distance for various boundary materials.*

as a consequence the usability of the auditory interface.
Figure 4.8 shows the effects of the properties of the mesh ends. In this example, the mesh sides simulate the absorption properties of acoustic plaster, and the ends simulate either reflecting surfaces, acoustic plaster or totally absorbing surfaces.

Even if the impact of the absorption properties of the mesh ends on depth cues is limited, it is consistent with Figures 4.6 and 4.7: the more reflecting the boundaries, the less the total energy decreases with distance and the more the environment is reverberant.

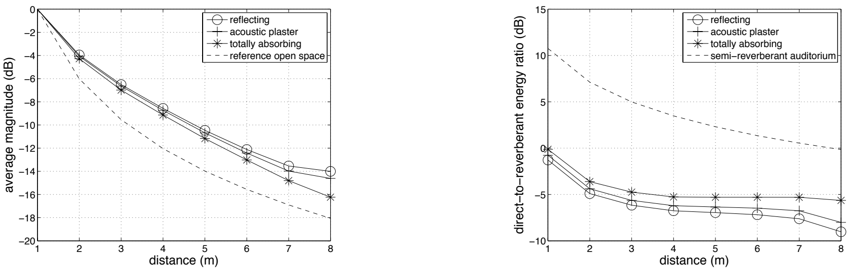


Fig. 4.8. *Intensity cue (left figure) and direct-to-reverberant energy ratio cue (right figure) for various boundary conditions at the mesh ends.*

Influence of the membrane width

Varying the dimensions of the mesh has also an effect on depth cues. Reducing the width allows to get a wider dynamic range of intensity as well as of direct-to-reverberant energy (see Figure 4.9). It can be deduced from these observations that the narrower the membrane, the easier distance discrimination should be.

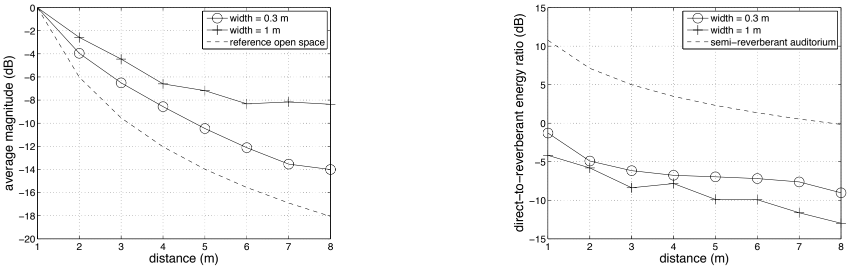


Fig. 4.9. *Intensity cue (left figure) and direct-to-reverberant energy ratio cue (right figure) for two mesh widths.*

4.2.3 Externalization

Listening to sound sources spatialized in the 2D mesh through headphones may lead to in-the-head localization because of the lack of appropriate individual spectral cues [36]. Nevertheless, in the listening experiment conducted by Fontana et. al and using the 3D mesh, only one out of twelve participants reported some lack of externalization for the closest stimulus. As already suggested in Section 2.6, reverberation most probably helps in externalizing sound sources. Further research is obviously required to ascertain this assumption.

4.3 Real-time implementation

Practical implementations of the 2D DWM have been realized as *Max/MSP*[1] external objects, written in C++ to comply with *flext*[2] (The initial code for the 3D DWM was written in C by Federico Fontana and Stefania Serafin). This programming layer is dedicated to cross-platform development of PureData and Max/MSP externals. The resulting Max/MSP externals have been included in the *Sound Synthesis Tools for Sound Design*[3], developed for the EU project *CLOSED - Closing the Loop of Sound Evaluation and Design*[4].

4.3.1 Parameters of the DWM

The user of the externals controls the membrane dimensions, and the y-coordinates of the listening position and of the sound source(s) (all dimensions are in meters). Since the system is meant to render depth cues, i.e. distance information about sound sources straight ahead of the listener, the model sets both listener and sound sources on the main axis of symmetry of the rectangular mesh (parallel to the y-axis in Figure 4.1).

Properties of the boundaries may also be set by varying the coefficients [$\mathrm{b_0}$,$\mathrm{b_1}$,$\mathrm{a_1}$] of the first-order filters modeling the boundaries along the two axes of the mesh, implemented as first-order filters:

$$OUT(z) = \frac{b_0 + b_1 z^{-1}}{1 + a_1 z^{-1}} IN(z) \tag{4.4}$$

4.3.2 Algorithm

During the initialization phase, the external variables are declared:

- length L and width W of the mesh
- coefficients [b_{0X},b_{1X},a_{1X}] and [b_{0Y},b_{1Y},a_{1Y}] of the first-order filters modeling the boundaries along the x and y axes respectively.

Then, the update cycle at each time step n for the discrete-time system consists of the following schedule:

1. update of the pressure signal p at the position of the sound source

[1] http://www.cycling74.com/products/maxmsp

[2] `http://grrrr.org/ext/flext/`

[3] `http://closed.ircam.fr/deliverables.html`

[4] `http://closed.ircam.fr`

2. update of the output signals at the boundaries:

$$\begin{aligned}
p^-_{4,n}(x,0) &= b_{0Y} * p^+_{4,n}(x,0) + b_{1Y} * p^+_{4,n-1}(x,0) - a_{1Y} * p^-_{4,n-1}(x,0)\\
p^-_{1,n}(x,L) &= b_{0Y} * p^+_{1,n}(x,L) + b_{1Y} * p^+_{1,n-1}(x,L) - a_{1Y} * p^-_{1,n-1}(x,L)\\
p^-_{3,n}(0,y) &= b_{0X} * p^+_{3,n}(0,y) + b_{1X} * p^+_{3,n-1}(0,y) - a_{1X} * p^-_{3,n-1}(0,y)\\
p^-_{2,n}(W,y) &= b_{0X} * p^+_{2,n}(W,y) + b_{1X} * p^+_{2,n-1}(W,y) - a_{1X} * p^-_{2,n-1}(W,y)
\end{aligned}$$

3. pressure signal p at each junction

$$p(x,y) = \frac{1}{2} * [p^+_1(x,y) + p^+_2(x,y) + p^+_3(x,y) + p^+_4(x,y)]$$

4. update of the transmission

$$p^-_i(x,y) = p(x,y) - p^+_i(x,y) \qquad \text{for i=1,...,4}$$

5. delay pass:

$$\begin{aligned}
p^+_1(x,y) &= p^-_4(x+1,y)\\
p^+_4(x+1,y) &= p^-_1(x,y)\\
p^+_2(x,y) &= p^-_3(x,y+1)\\
p^+_3(x,y+1) &= p^-_2(x,y)
\end{aligned}$$

6. the output of the system is the pressure signal $p(\frac{W}{2}, y_{listening})$ at the listening position.

Note that prior to computation, all dimensions are converted in number of waveguide lengths. In the event of non integer values, dimensions of the mesh are rounded and the positions of the listener and the sound source(s) are de-interpolated over the two adjacent junctions along the y-axis.

4.3.3 Max/MSP externals

Two Max/MSP externals are provided, depending on the application:

1. `mesh2D_static~`: allows to render one or more static sound sources located at different positions from the listener on the y-axis of the membrane (see Fig. 4.10). In fact, positions of both sound sources and the listener may be modified, but only at control rate (typically below 1 kHz).
2. `mesh2D_dynamic~`: allows to render one moving sound source along the y-axis of the membrane, i.e. to simulate one approaching/receding source. In this case, the computation in the mesh is made more efficient by switching the sound source and the listening positions, such that the measurement position varies at runtime (i.e. at signal rate) and not the position of the source, which is updated at control rate (see Fig. 4.11).

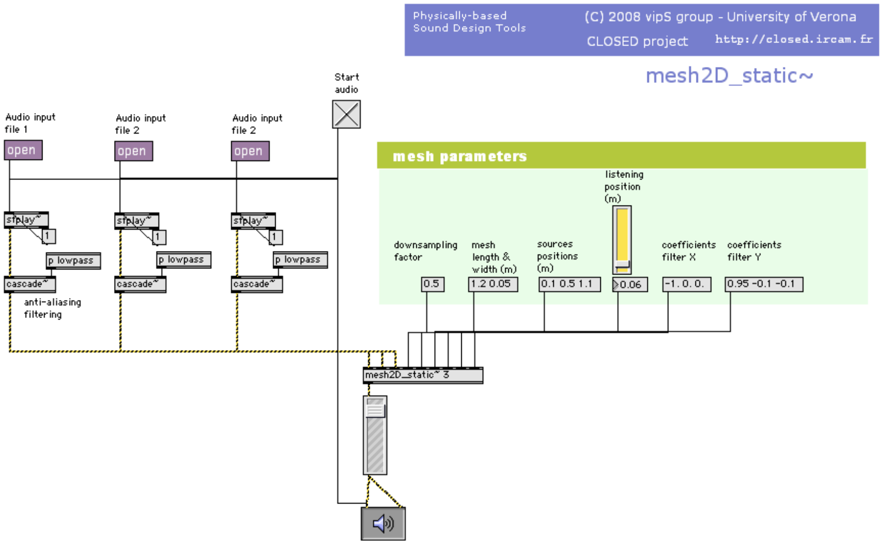


Fig. 4.10. Patch example using the `mesh2D_static~` external.

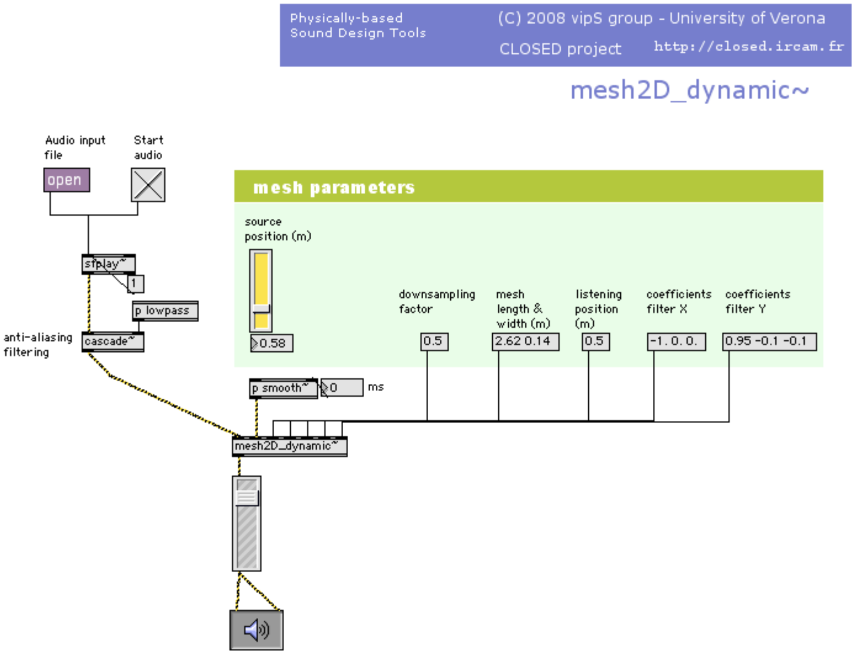


Fig. 4.11. Patch example using the `mesh2D_dynamic~` external.

4.3.4 Sampling rate

Due to the heavy computational load of the model, downsampling may be required prior to the computation in the mesh, set by the downsampling factor. Inside the external, only decimation is performed such that an anti-aliasing filter is required at the audio input(s). After computation in the mesh, an upsampling is performed inside the external using the library *libsamplerate* (also known as *Secret Rabbit Code*), by Erik de Castro Lopo[5]. Downsampling the simulation yields to the use of a coarser mesh, meaning that the modeling is only valid at low frequencies.

4.4 Implications

With the real-time implementation of the rectangular bidimensional DWM, it is then possible to investigate the potentiality of the simulated auditory virtual environment to provide dynamic depth cues in response to human actions. The two different externals are optimized for two scenarios: either the aim is to provide multiple sound sources spatialized in depth, or to simulate a dynamic approaching/receding sound source. In addition, all physical parameters of the model - except the geometry of the enclosure - may be controlled. Prototypes of auditory interfaces using these externals will be presented in Chapters 6 and 7.

[5] See the libsamplerate home page: `http://www.mega-nerd.com`

5

Linearizing auditory distance estimates by means of a Digital Waveguide Mesh

5.1 Introduction

5.1.1 Objective

As seen in Chapter 2, the relationship between perceived distance and physical distance in a real reverberant room is not linear but rather logarithmic. Our question is: *Can we create a virtual environment in which the psychophysical function between perceived and physical distances is linear?* Creating virtual auditory environments may indeed offer new opportunities to manipulate auditory distance cues. Besides, a model proposed by Bronkhorst and Houtgast and described in Section 5.1.2 may assist the design of the environment by predicting the perceived distance of a sound source based on the computation of the corresponding room impulse response. From a practical point of view, such a virtual environment may become a tool for interactive sound spatialization. The ability to manipulate the relative depth of sound sources may enrich the content and quality of auditory displays, or offer novel human-computer interactions.

In this chapter we present a spatialization model using a 2D DWM which is devoted to render the sound source depth, and in particular to provide a linear mapping between the physical distance of the source in the virtual environment and the estimate of the distance perceived by the listener. This study has been published in Acta Acustica united with Acustica (Paper G).

5.1.2 Bronkhorst and Houtgast's model

A model for distance perception, proposed by Bronkhorst and Houtgast [11], has shown to converge with other research results used by Zahorik in its power function fit analysis [107]. Bronkhorst and Houtgast's model relates the estimate of perceived distance to the reverberation cue. The amount of reverberation is commonly assessed by measuring the ratio between the direct sound energy and the reverberant sound energy. More precisely, the direct sound energy is generally considered to lie within the first 2.5 ms of the sound signal. This time window was chosen because it approximates the duration of Head Related Impulse Responses measured in anechoic conditions and therefore captures the direct path of the sound signal [62]. Bronkhorst and Houtgast [11] proposed the use of a larger integration window (6 ms) for determining the energy of the direct signal. This choice is supported by the observation that listeners underestimate far physical distances (effect called "auditory horizon"). By increasing the window size, the overall energy contained in the integration window may decrease more slowly at far distances if the energy of the first reflections integrated in the window prevails against the energy of the direct signal. Furthermore, the choice of 6 ms is in line with studies conducted on the *precedence effect* involved in localization tasks. Indeed Houtgast and Aoki [38] showed the dominance of the first 5 to 10 ms of the sound signal in cues used in sound localization. These values may also be related to the 8 ms temporal resolution of the auditory system, found by Moore et al. [63]. Coming back to Bronkhorst and Houtgast's model, the integration time window W is defined as:

$$W = \begin{cases} 1 & \text{for } 0 < t \leq t_W - \frac{\pi}{2s} \\ 0.5 - 0.5\sin[s(t - t_W)] & \text{for } t_W - \frac{\pi}{2s} < t < t_W + \frac{\pi}{2s} \\ 0 & \text{for } t \geq t_W + \frac{\pi}{2s} \end{cases}$$

with $t_W \simeq 6.1$ ms and $s \simeq 400$ s^{-1}. Finally, the perceived distance d_s of a sound source is expressed as:

$$d_s = Ar_h \left(\frac{\hat{E}_r}{\hat{E}_d} \right)^j \tag{5.1}$$

where r_h represents the critical distance (also called reverberation radius), A and j are parameters determined empirically through experiments, $\hat{E}_r$ is the modified energy of the direct sound calculated using the integration window W, and $\hat{E}_d$ is the modified energy of the reverberant sound calculated using the integration window $1 - W$.

Thanks to this quantitative model, it is possible to predict the perceived distance of a sound source in a given acoustical space. We will make use of the same model for synthesis rather than analysis purposes, i.e. to shape a virtual acoustical environment in order to render sound source distances.

5.2 The trapezoidal membrane

After some informal experimentation and assessment of the reverberation cue provided by several 3D and 2D shapes, our attention was finally captured by a 2D vibrating surface (i.e. a membrane), whose shape resembles that of a trapezoid. Such simulations have been conducted by varying the shape and boundary absorptions in a DWM modeling the environment under analysis. In the following we provide a summary of the main features of our DWM model.

5.2.1 Geometry of the DWM

Our sound propagation model consists of a two-dimensional rectilinear DWM (see Chapter 4 for the description of the DWM) having a trapezoid shape as shown in Figure 5.1. For $c = 343\ m/s$ the speed of sound in air and $f_s = 44.1\ kHz$, a

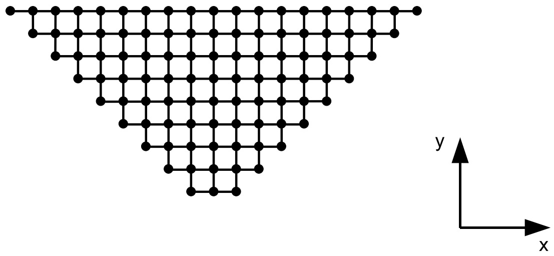


Fig. 5.1. *The two-dimensional rectilinear waveguide mesh representing a trapezoid geometry.*

waveguide corresponds to the propagation of sound pressure waves on a distance of 1.10 cm. In order to reproduce distances up to 10 meters, the total length of the mesh along the y-axis is chosen to be 10.05 m, which amounts to 914 waveguides. Since the mesh is symmetric, the top base of the trapezoid equals 10.05 × 2 = 20.1 m.

5.2.2 Boundary conditions

The boundaries of our trapezoidal DWM exhibit the following properties (refer to Figure 5.2 for the definition of the surfaces A, B and C):

- Surfaces A and B provide the simulation of hard surfaces, such as bricks or plaster (See Table 4.1 in Section 4.2.2 for the values of the absorption coefficients).
- Each surface of the group C is modeled to behave as total absorber. We chose to cancel reflections from these panels in order to avoid audible echoes created from waves traveling back and forth between surfaces A and C.

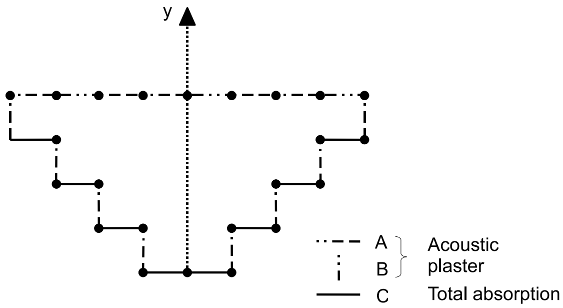


Fig. 5.2. *Properties of the mesh boundaries.*

5.2.3 Acoustical properties of the listening environment

Listening and sound source locations

The listener and the sound source are assumed to be point-wise and are positioned on one of the mesh junctions on the y-axis. The listening point is located at the junction no. 2 near the smaller base of the trapezoid, while the sound source is successively positioned at junctions no. 93, 184, 275, 366, 457, 547, 638, 729, 820 and 911 in order to simulate one-meter steps ranging from 1 to 10 meters away from the listening point. Simulations of the listening environment for the ten sound source positions were carried out in `Matlab`, and ten corresponding impulse responses were computed. Figure 5.3 shows the impulse responses and the frequency responses up to 5 kHz, measured at 1 and 6 m.

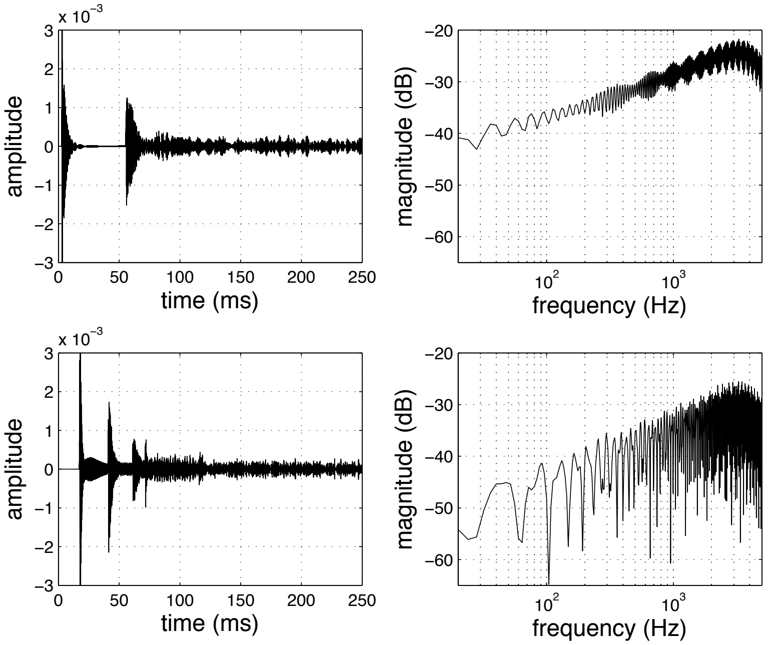


Fig. 5.3. *Impulse responses (left) and frequency responses (right) inside the trapezoid mesh. Top: 1 m. Bottom: 6 m.*

Overall intensity level

Figure 5.4 shows how the total energy of the impulse responses varies with distance. The energy inside the trapezoid mesh decreases significantly less than in open space, characterized by the well-known 6 dB law. The present 2D trapezoid DWM exhibits a decrease of only 6 dB from one to ten meters, in particular the intensity

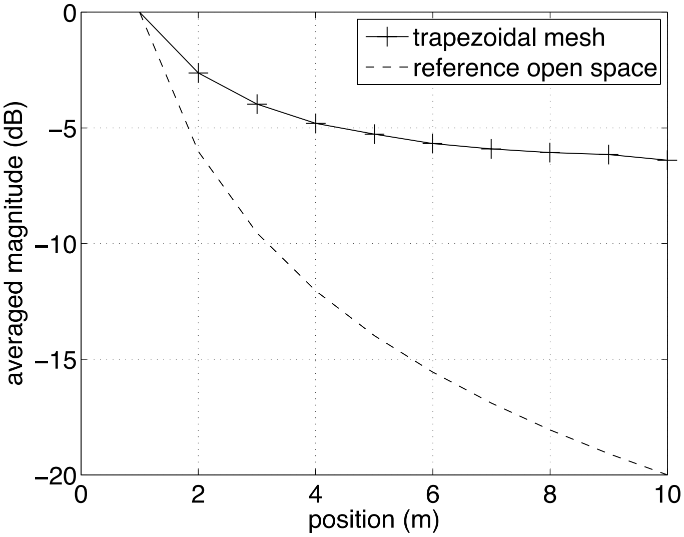


Fig. 5.4. *Average magnitude of the impulse response as a function of distance.*

change over the range 6–10 m is less than known just-noticeable differences for intensity (~ 1 dB) [43]. In that distance range, it may be concluded that the overall intensity of the sound signal may hardly provide a reliable source of information to judge distance.

Direct-to-reverberant energy ratio

The virtual acoustic space we have designed aims at rendering depth information mainly through the direct-to-reverberant energy ratio cue. Figure 5.5 shows the values of the direct-to-reverberant energy ratios for the ten distances in the trapezoidal mesh. The method used to compute these ratios is explained in Section 4.2.2. For comparison the direct-to-reverberant energy ratio v was computed for a semi-reverberant auditorium, modeled by Zahorik with the function $v = -3.64 \log_2(r) + 10.76$ [105]. The two curves follow the same trajectory for distances up to 9 m, suggesting that on this distance range the direct-to-reverberant energy in the virtual environment exhibits a natural profile. Numerical artifacts in Digital Waveguide Mesh simulations, which affect the responses especially in the proximity of the mesh boundaries, may be responsible for the increase of the direct-to-reverberant energy ratio between 9 and 10 m.

5.2.4 Application of Bronkhorst and Houtgast's model

To apply Bronkhorst and Houtgast's model, the critical distance r_h is to be estimated. As seen in Section 2.1.1, it corresponds to the distance at which the sound pressure level of the direct and the reverberant sound fields are equal, providing a

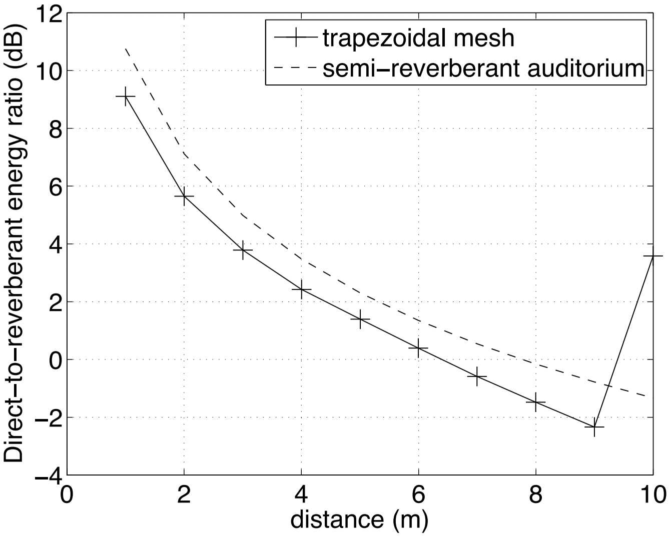


Fig. 5.5. *Direct-to-reverberant energy ratio as a function of distance.*

direct-to-reverberant energy ratio of 0 dB. From Figure 5.5, this condition arises at a distance of about 6 m. A more accurate computation of the direct and reverberant energies focused around this position leads to a value of 6.3 m.
Finally, the modified direct-to-reverberant energy ratios for the ten distances are calculated from eq. 5.1 and are shown in Figure 5.6. Note that the explanation given for the increase of the direct-to-reverberant energy ratio at the far end of the membrane also applies to the behavior of the distance estimate at the greatest distances. Before the analysis, it should be noted that what is called *physical distance* on this plot relates in fact to a distance inside the modeled environment which is derived from the digital waveguide theory. An almost linear behavior arises for physical distances from 1 to 8 m along the mesh. This behavior means that Bronkhorst and Houtgast's model predicts that in the trapezoidal mesh the perceived distance is a linear function of the physical distance in that range.

5.3 Experiment I

5.3.1 Subjective assessment of the distance linearity

Common procedures used to estimate psychophysical distance functions include magnitude estimation (on explicit scales such as meters or on dimensionless scales), or other methods such as perceptually directed action [53] consisting of walking to the perceived location. They generally aim at evaluating the absolute egocentric distance of a sound source, one at a time, and may suffer from adaptation effects. Subjects may indeed readjust their responses due to the subsequent stimuli as they settle on a range of scale values. The objective of our listening test consists of

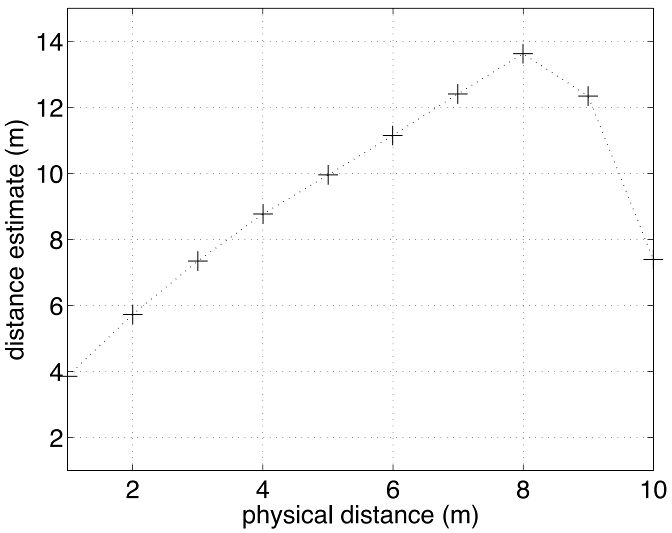


Fig. 5.6. *Distance estimate according to Bronkhorst and Houtgast's model.*

assessing the linearity of depth perception, therefore we are interested in evaluating relative depth perception of a set of sound sources. Considering possible bias due to learning effect and high variability of responses across trials, we thought that the common evaluation methods were not adequate for the present issue. Preferably a method allowing direct multiple comparison of the different stimuli would provide a more robust estimate of the relative distance perception of a set of sound sources. Such a methodology, the MUSHRA test, exists for the subjective evaluation of audio quality, and is defined by ITU-R recommendation BS.1534-1 [91]. MUSHRA stands for MUltiple Stimuli with Hidden Reference and Anchor. This methodology was developed for the subjective evaluation of audio coding systems and allows the comparison of high quality reference sounds with several lower quality test sounds. The aim of each experiment is to compare a high quality reference sound to several test sounds sorted in random order, including the reference. According to the procedure, each subject is asked to assess the quality of each test sound (relative to the reference and other test sounds) by grading it on a quality scale between 0 and 100. It is not necessary that one test sound be graded 0, however at least one must be graded 100 (because the reference sound is among the test sounds). The test sounds must also include one or several *anchor sounds* computed similarly for all experiments using simple signal processing operations. In particular, one of these anchor sounds must be the reference sound low-pass filtered at 3.5 kHz.

5.3.2 Method

Participants

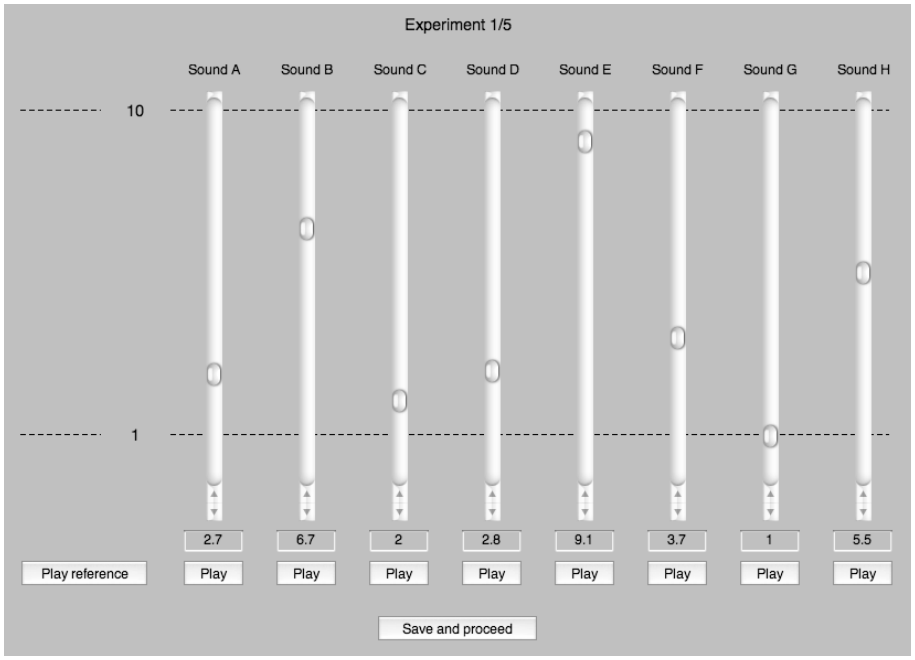
Eleven Italian volunteers (4 women and 7 men), studying or working at the University of Verona, Italy (except one teenager aged 15), aged 15–41 years old, participated in the experiment. All reported to have normal hearing.

Stimuli and Apparatus

The objective of the listening test was to assess the linearity of distance perception for physical distances ranging from 1 to 8 m in the trapezoid mesh, as predicted by Bronkhorst and Houtgast's model. The 8 corresponding impulse responses, computed at 44.1 kHz and 0.25 s long, were convolved with a 4 s speech signal. The latter consisted of a female voice saying the words "Kroklokwafzi? Sememememi" (opening of Christian Morgenstern's poem *Das grosse Lalula*) already used in a previous study [22]. The stimuli were generated by an interactive `Matlab` script and presented over an AKG K240 open headphone set. Subjects gave their answers on a computer terminal with a numeric keypad.

Procedure

Distance judgments were collected using a modified version of the MUSHRA test [91] with hidden reference. This reference was the sound source located at 1 m in the virtual environment. Prior to the experiment, listeners received a written instruction explaining the two phases of the test. First a training phase allowed the user to browse through the 8 stimuli, ordered randomly on the computer screen, and get acquainted with the auditory distance range. In the second phase, namely the evaluation phase, subjects had to rate the perceived distance of the 8 stimuli against the reference sound, on a scale ranging from 0 to 10 dimensionless units, through steps of 0.1 units. Each sound was graded by moving the corresponding slider on the computer screen, and the grading value appeared in the corresponding text field below. Again, the 8 stimuli were randomly ordered for each trial. The hidden reference was also included in the test, and listeners were instructed to rate this particular sound to the value "1". Besides, they were instructed to use the value "0" for sound sources that appeared to originate from their head. They could listen to the signals in any order, as many times as they wanted. When they were satisfied with their grading of all signals, they clicked on the "save and proceed" button on the screen to go to the next trial. In total each subject had to compute the task five times, which took approximately 15–20 minutes. Figure 5.7 shows a snapshot of the user interface. The whole procedure was carried out on `Matlab` using a modified version of the MUSHRAM package [93], a set of `Matlab` routines dedicated to MUSHRA listening tests. In both phases of the experiment, the sequence of the stimuli was randomized.



Experiment 1/5

Sound A	Sound B	Sound C	Sound D	Sound E	Sound F	Sound G	Sound H
2.7	6.7	2	2.8	9.1	3.7	1	5.5

Fig. 5.7. *Snapshot of the user interface during the evaluation phase.*

5.3.3 Results

Results for each subject are displayed in Figure 5.8. None of the participants used the value “0”. However at this stage, we cannot conclude that they perceived all the simulated sound sources as external to the head since they had to grade the closest sound source a “1”.

The observation of the individual psychometric functions on the whole range of sound source distances does not reveal evident similarities. An analysis of the results for each subject shows that the individual relationships between perceived and physical distance are better described with linear functions [$r^2 = 0.7071$, $F(1, 38) = 91.73$ and $p = 1.117 \times 10^{-11}$ for the worst fit] than with power functions [$r^2 = 0.5894$, $F(1, 38) = 54.56$ and $p = 7.424 \times 10^{-9}$ for the same data]. Power functions are of the form $r' = kr^a$, where r' is an estimate of perceived distance, and r is the physical distance of the sound source. Parameters of each individual fitted linear and power functions are reported in Table 5.1 which also shows the goodness-of-fit (r^2) for each fit. Figure 5.8 suggests that linearity is more apparent for near sound source distances. In order to assess more precisely this indication, a linear regression was applied to each individual psychometric function for 5 different distance ranges, namely 1-8 m, 1-7 m, 1-6 m, 1-5 m and 1-4 m. The box plot of Figure 5.9 summarizes the distribution of the residual standard errors of the linear fits for each distance range. By taking the mean of the individual residual standard errors of the linear regressions for each distance range, the minimum is found for the distance range of 1 to 4 meters (mean residual standard error $\simeq$

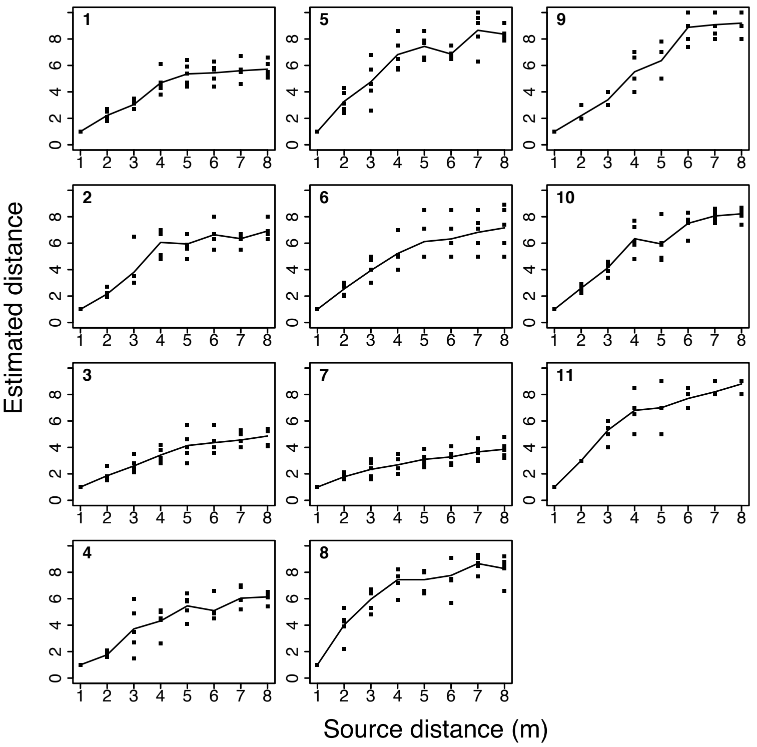


Fig. 5.8. *Exp. I: Individual apparent distance judgments as a function of physical distance. Each plot displays data for one subject. Means are connected by a line.*

0.73). The analysis of the best linear fit for this distance range gives the following results: $r^2 = 0.8268$, $F(1, 18) = 85.92$ and $p = 2.834 \times 10^{-8}$. From these results we may conclude that the relation between subjective and physical distances up to 4 m in the proposed virtual environment is well described with a linear function. Please note that in Paper G, the analysis of the distance range providing the best linear fit considered standard errors of the slope instead of residual standard errors, which was a mistake.

5.3.4 Test Repeatability

Since the modified version of the MUSHRA test is not a common procedure in auditory distance perception, we first propose to check the repeatability of the test, i.e. the consistency of individual responses over repetitions of the same experiment under the same conditions. This may be easily done by estimating the standard deviation of the five repeated evaluation tasks for each subject. An analysis of variance was performed with the number of repetitions as a within-subjects factor. No statistically significant differences between the different sessions were found, meaning that the procedure is repeatable [Mean square= 2.71, $F = 1.1724$, $p = 0.2795$].

Table 5.1. Exp. I: Individual fitted linear and power function parameters for each subject.

	Linear fit			Power fit		
subject n.	slope	intercept	r^2	a	k	r^2
1	0.69	1.04	0.78	0.86	1.15	0.74
2	0.84	1.07	0.74	0.96	1.15	0.72
3	0.55	0.86	0.78	0.77	1.06	0.77
4	0.75	0.84	0.74	0.92	1.05	0.73
5	1.02	1.33	0.77	0.99	1.33	0.68
6	0.86	1.00	0.72	0.93	1.21	0.71
7	0.39	0.96	0.75	0.63	1.08	0.75
8	0.95	2.06	0.71	0.95	1.51	0.59
9	1.30	-0.14	0.90	1.13	1.02	0.85
10	1.04	0.79	0.86	1.01	1.20	0.77
11	1.05	1.26	0.83	1.01	1.32	0.70

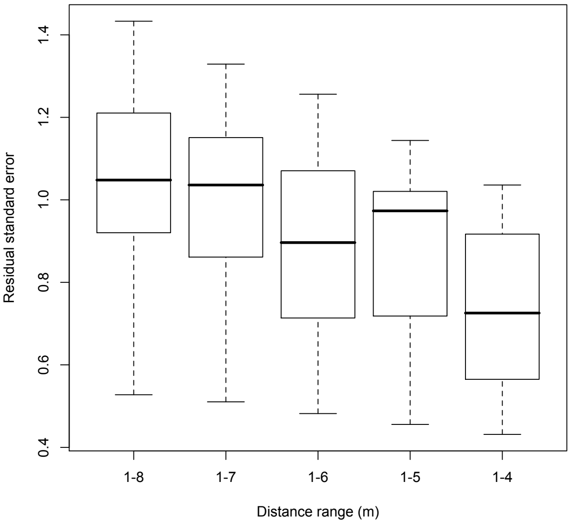


Fig. 5.9. *Exp. I: Distributions of the individual standard errors of linear regressions applied to 5 distance ranges.*

5.4 Experiment II

In order to validate the scaling procedure described in Section 5.3, we propose to compare it to another procedure commonly used to assess distance perception. Although it was shown that all common procedures produce similar results [107], the magnitude estimation procedure was chosen since it was used in a reference study [104].

5.4.1 Method

The same eleven volunteers as for experiment I participated in the second experiment. The stimuli and apparatus were those already used in experiment I. Only the procedure differed: after listening to the 8 stimuli ordered according to increasing distance, the listeners were asked to judge the distance of the 8 different stimuli presented in a random order via a unitless magnitude estimation procedure. Subjects were free to choose their own standard, assigning any number to the first stimulus and all subsequent ones. Responses were made after listening to each stimulus on the computer terminal with the numeric keypad, and participants could use a precision of one decimal. As in experiment I, a grade of "0" was used for sound sources that appeared to originate from their head. Subjects had to perform the task 3 times, leading to a total of 24 trials.

5.4.2 Results

Figure 5.10 shows the results for each subject. A grade of "0" was used by five

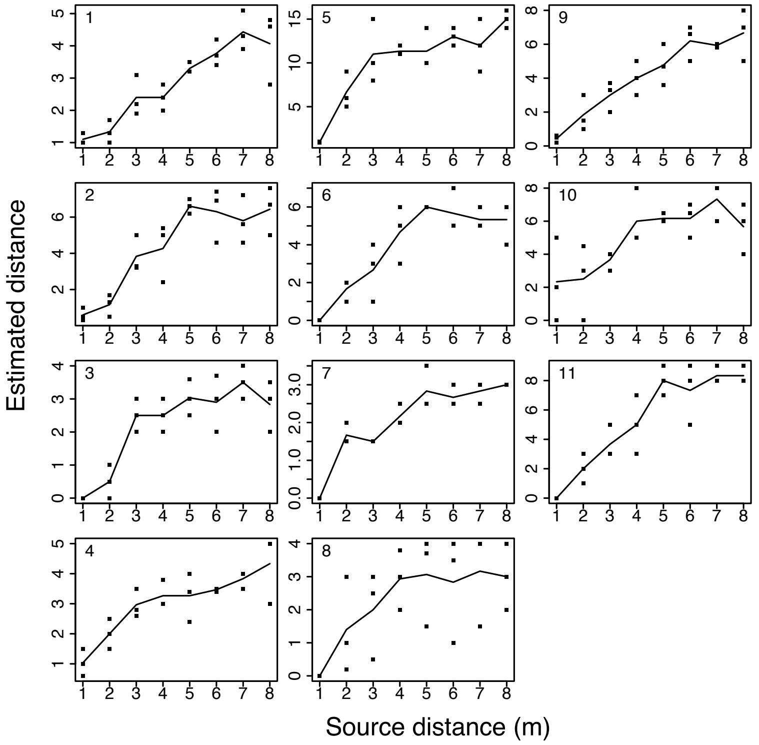


Fig. 5.10. *Exp. II: Individual apparent distance judgments with the magnitude estimation procedure. Each plot displays data for one subject. Means are connected by a line.*

subjects for rating the 1 m stimulus (subject 3 also rated once the 2 m stimulus to "0"). This did not happen in experiment I, because the participants were constrained to use the grade "1" for the closest sound source corresponding to the reference source (therefore being easily identified). If we discard the zero attributed by subject 3 to the 2 m sound source, we may deduce that some listeners may hear the closest sound as originating from their head. Nevertheless we think that this fact does not prevent from applying the model to an auditory interface, as long as users can discriminate between sound sources located at different distances.

Except for subjects 1 and 8, the relation between perceived and physical distance is better described with a linear function [$r^2 = 0.4778$, $F(1, 22) = 20.13$ and $p = 1.838 \times 10^{-4}$ for the worst fit] than with a power function [$r^2 = 0.393$, $F(1, 22) = 14.24$ and $p = 1.045 \times 10^{-3}$ for the same data]. These data agree with those of experiment I, and parameters of each individual fitted linear and power functions may be seen in Table 5.2. For subjects 1 and 8, the logarithmic fit gives slightly better results, however the data for subject 8 are still not well approximated by a power function since the variance accounted for is low [$r^2 = 0.4276$, $F(1, 22) = 16.43$ and $p = 5.29 \times 10^{-4}$]. It can also be seen from Figure 5.10 that subject 8 might have reversed the distance scale for some judgments. Besides, his/her results are not consistent through the three repetitions (An analysis of variance with the number of repetitions as a factor gives: Mean square= 18.2756, $F = 28.724$, $p = 2.578 \times 10^{-5}$). Subject 8 is therefore discarded in the following.

Table 5.2. Exp. II: Individual fitted linear and power function parameters for each subject.

	Linear fit			Power fit		
subject n.	slope	intercept	r^2	a	k	r^2
1	0.49	0.64	0.81	0.49	1.91	0.82
2	0.88	0.43	0.70	0.82	1.56	0.69
3	0.44	0.26	0.62	0.73	1.09	0.61
4	0.40	1.21	0.70	0.43	2.16	0.66
5	1.56	3.15	0.65	0.88	3.01	0.56
6	0.79	0.38	0.66	0.92	1.24	0.65
7	0.37	0.42	0.75	0.61	1.28	0.65
8	0.39	0.56	0.37	0.65	1.23	0.43
9	0.89	0.11	0.83	0.82	1.52	0.78
10	0.66	2.02	0.48	0.59	2.44	0.39
11	1.24	-0.24	0.80	1.09	1.22	0.74

Like in experiment I, a linear regression was applied to each individual psychometric function for the 5 different distance ranges 1-8 m, 1-7 m, 1-6 m, 1-5 m and 1-4 m. The box plot of Figure 5.11 summarizes the distribution of the residual standard errors of the linear fits for each distance range. By taking the mean of the individual residual standard errors of the linear regressions for each distance range, the minimum is found for the 1-4 m distance range (mean residual standard error = 0.97). This result is in good agreement with the linearity found in experiment I for the 1-4 m range.

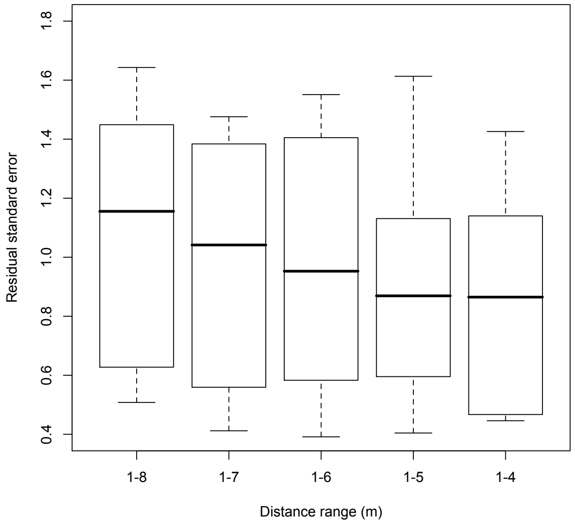


Fig. 5.11. *Exp. II: Distributions of the individual standard errors of linear regressions applied to 5 distance ranges.*

By comparing the shape of the psychophysical functions in Figure 5.8 and Figure 5.10, data of many subjects look qualitatively similar for both response procedures, except for subjects 3 and 7. However, by looking at the individual responses, data in Figure 5.10 seem to be more scattered. In order to quantify this difference, standard scores were calculated since both experiments used different scales (in experiment II the scale was free whereas in experiment I subjects could rate the perceived distance up to 10). More precisely, the absolute value of the z-scores was calculated for each value of distance estimate, each experiment, and each of the remaining 10 subjects. Then the mean was calculated for each subject and for each experiment. Figure 5.12 shows the results. It can be seen from this plot that the dispersion is greater in experiment II, and this is validated by calculating the mean of the absolute values of the z-scores over all participants (2.50 for experiment I, and 4.65 for experiment II). The high variability in distance judgment for a fixed sound source distance is well known [107], and it seems that the procedure derived by the MUSHRA test in experiment I allows to reduce it and provide more consistent responses.

5.5 Experiment III

The linearity observed in the proposed virtual environment with the procedure used in experiment I may be validated by carrying out the experiment on stimuli spatialized in a different virtual environment in which the linearity of the psychophysical functions is not expected. For this purpose, a 2D rectangular DWM

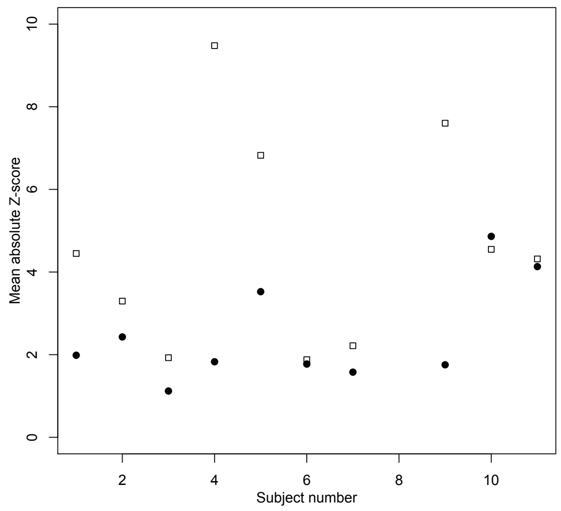


Fig. 5.12. *Exp. II: Mean absolute z-score as a function of subject number. Solid circles: experiment I. Blank squares: experiment II.*

is proposed as the comparison environment, and has the following characteristics (See Figure 5.13):

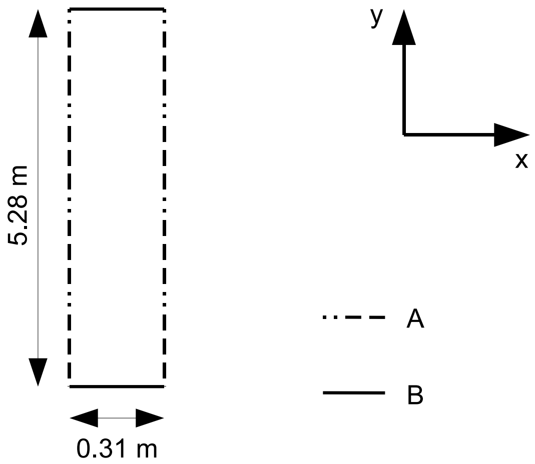


Fig. 5.13. *The two-dimensional rectangular waveguide mesh.*

- Dimensions: 5.28×0.31 m^2

- Boundary conditions: reflections from surfaces are modeled by first-order digital filters which simulate
  1. surfaces A as strongly vibrating walls (modeling for instance a panel of wood mounted in front of a rigid wall [49])
  2. surfaces B as total reflectors

Please note that the length of the new environment is about half the length of the trapezoid, thus we expect that listeners will use smaller values to rate the estimates of the perceived distances. Figures 5.14 and 5.15 show respectively the intensity and the direct-to-reverberant energy ratio cues in the rectangular DWM. While the intensity decrease is more linear than in the trapezoid mesh (Compare Figures 5.4 and 5.14), the direct-to-reverberant energy ratio shows a sudden decrease in the beginning of the distance range and remains almost constant for distances greater than about 2 m.

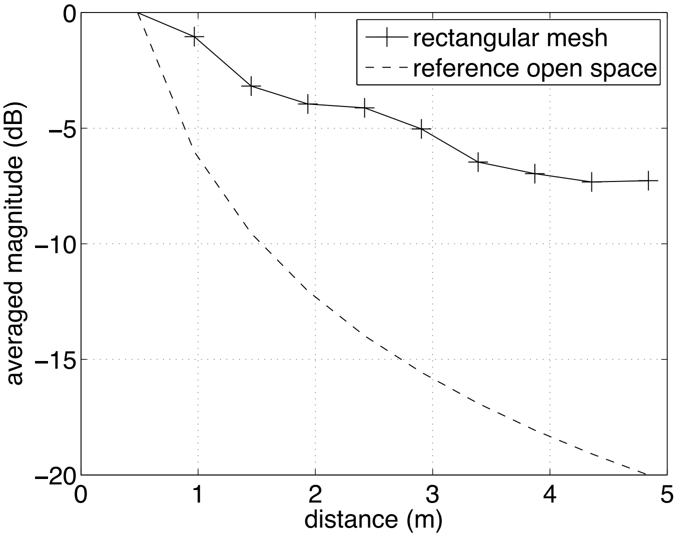


Fig. 5.14. *Average magnitude of the impulse response as a function of distance in the rectangular mesh.*

Applying Bronkhorst and Houtgast's model to this model leads to the psychophysical function shown in Figure 5.16. The resulting curve is not linear and is better described by an exponential function: it seems that this particular virtual environment should produce an expansion of perceived distance, in opposite to the compression observed in most real-world listening environments.

5.5.1 Method

This experiment was run on the same subjects as for experiments I and II. Apparatus and procedure were identical to those used in experiment I. Only the

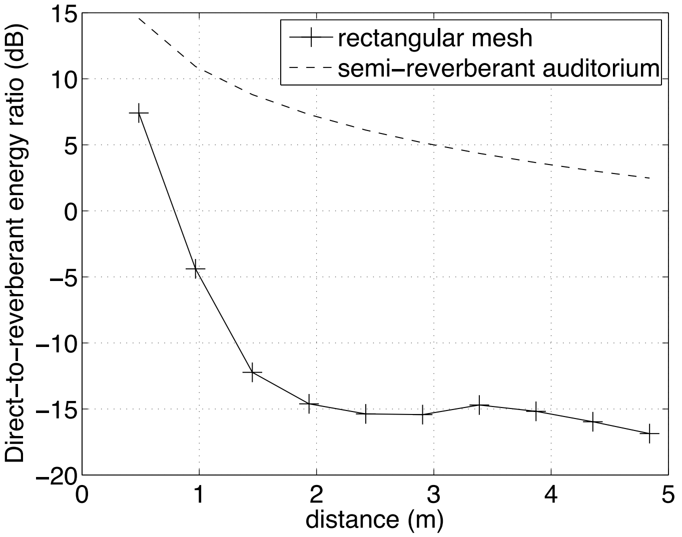


Fig. 5.15. *Direct-to-reverberant energy ratio as a function of distance in the rectangular mesh.*

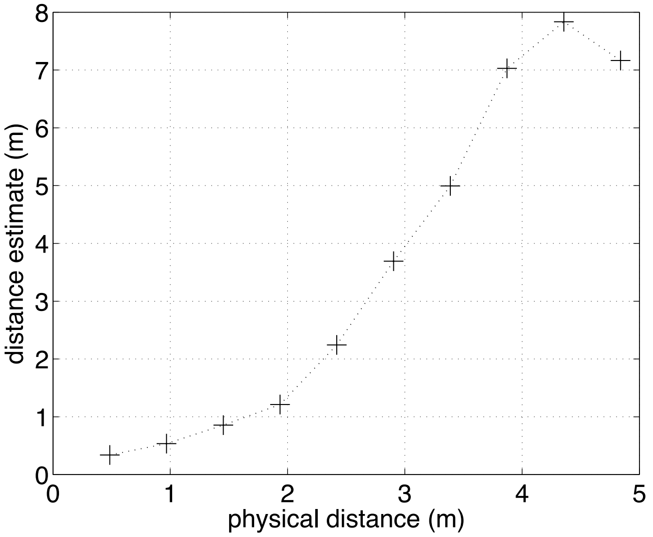


Fig. 5.16. *Distance estimate according to Bronkhorst and Houtgast's model.*

spatialization of the stimuli was different. The 4 s speech signal used for experiments I and II was convolved with 8 impulse responses computed every half-meter from 0.5 to 4 m.

5.5.2 Results

Figure 5.17 shows the estimates of perceived distance in the rectangular DWM for each subject. A comparison with the scales used by the same participants in experiment I (Figure 5.9) shows that the distances are effectively estimated as closer. Indeed subjects were not forced to use the whole scale during the experiment. They had the freedom in grading the furthest distance (up to 10), and they spontaneously reduced the scale for sounds rendered in the half-length mesh.

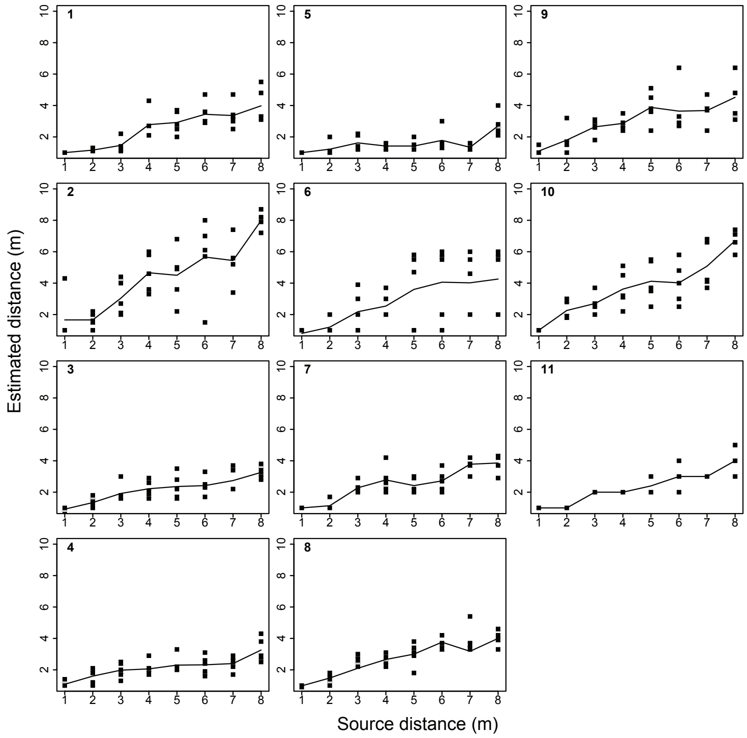


Fig. 5.17. *Exp. III: Individual apparent distance judgments in the rectangular DWM as a function of physical distance. Each plot displays data for one subject. Means are connected by a line.*

An analysis of the results for each subject was carried out by fitting the estimates with a linear fit, a logarithmic fit and an exponential fit. Except for subjects 2 and 8, the relation between perceived and physical distance is better described with a power function, however the variance accounted for is quite low for some subjects [$r^2 = 0.352$ for subject 5]. Estimates given by subjects 2 and 8 are rather fitted by a linear regression but again with a quite low variance accounted for [$r^2 = 0.6542$ for subject 2, and $r^2 = 0.554$ for subject 8] (See Table 5.3 for details on the fitted function parameters). As a consequence, it is rather difficult to draw conclusions from the analysis. It seems however that for most subjects the

Table 5.3. Exp. III: Individual fitted linear and power function parameters for each subject.

	Linear fit			Power fit		
subject n.	slope	intercept	r^2	a	k	r^2
1	0.45	0.48	0.69	0.48	1.75	0.76
2	0.84	0.53	0.65	0.61	2.11	0.64
3	0.30	0.81	0.63	0.36	1.88	0.67
4	0.24	1.03	0.55	0.28	2.08	0.58
5	0.15	0.87	0.31	0.18	1.96	0.35
6	0.54	0.42	0.40	0.54	1.65	0.43
7	0.41	0.67	0.73	0.44	1.85	0.75
8	0.42	0.77	0.55	0.43	1.91	0.40
9	0.44	1.01	0.56	0.45	2.10	0.62
10	0.70	0.56	0.72	0.57	2.02	0.75
11	0.41	0.46	0.83	0.44	1.76	0.86

psychophysical functions are not linear, meaning that unlike the trapezoidal membrane, the virtual auditory environment modeled by a rectangular DWM does not produce linear auditory distance estimates.

5.6 Summary

According to the distance perception model of Bronkhorst and Houtgast, the trapezoidal DWM allows a linear scaling between perceived and physical distance straight ahead of the listener. An experiment was conducted to assess the model, with stimuli in the 1-8 m range. Results reveal that the trapezoidal membrane can effectively render linear depth perception, at least in the 1-4 m distance range.

The procedure used in the experiment derives from the MUSHRA test and was proposed as a method suitable for judging the relative distance of multiple sound sources. In a second experiment the direct scaling procedure was applied to the same stimuli. Results allowed to validate the modified version of the MUSHRA test as well as to reveal the lower variability of responses across trials with the latter method. Finally, a third experiment was conducted using again the modified version of the MUSHRA test on stimuli spatialized in a rectangular DWM and ranging from 0.5 to 4 m. Participants accordingly perceived the sound source distances as being shorter. Furthermore, unlike the trapezoidal mesh, the rectangular mesh did not lead to a linear relationship between estimated distance and physical distance.

The distance range of the stimuli rendered in the trapezoidal membrane (i.e. 1-8 m) was kept constant through the experiments I and III, and one may argue that the effect of linearityg might be dependent on the distance range under consideration. One answer to this issue is the recent study by Zahorik [106] which shows that for most subjects, apparent distance biases do not depend on the range of source distances presented within a block of trials.

Coming back to the issue of externalization already mentioned in Section 4.2.3, it may be questionable as to whether the simulated sound sources are perceived as external to the head. It seems that individual Head-Related Impulse Responses lead to the largest externalization [99]. However, to the authors' knowledge, past literature has never proposed auditory depth experiments comparing stimuli providing only intensity and reverberation cues versus spatialized stimuli containing additional binaural information, in terms of sound source externalization in the median plane. Moreover, the effect of binaural cues on distance perception of sound sources in the median plane is rather limited, at least in the far-field [104]. Furthermore, in the experiments subjects seldom graded the stimuli a "0" which was reserved for those sound sources that appeared to originate from a location within their head. In fact, only the closest sound source received a grade of "0" a few times. In our opinion, externalization is not an issue in the current context as long as listeners are able to discriminate different sound source distances. Although the realism of the simulated acoustic environment would ask for using more sophisticated acoustic rendering techniques, the proposed spatialization technique allows to render depth with a monaural signal and does not rely on the reproduction hardware configuration, unlike techniques using Head-Related Transfer Functions. Finally, virtual environments designed for rendering auditory distance may offer novel tools for interactive sound spatialization. In particular, a linear mapping between perceived and physical distance may enable a more transparent user manipulation of sounding objects displayed along the depth dimension and thereby improve the perceived quality of interaction.

6

An Audio-tactile interface based on auditory depth cues

6.1 Using spatial audio in audio-tactile interfaces

Rendering sound sources in depth allows a hierarchical display of multiple audio streams and therefore may be an efficient tool for developing novel auditory interfaces. The rendering of multiple spatialized sound sources is supported by a study carried out by Pitt and Edwards [69]. While investigating the ability of using the cursor as a virtual microphone for pointing tasks, the authors found that rendering simultaneous sound sources discriminated by stereo panning allows a faster target search in comparison with presenting only one sound at a time. To explain this result, they suggested that the hypothesized complexity of the multi-target task is overtaken by its naturalness in everyday life situations where people use spatial information to discriminate between different sound sources. Furthermore, the search speed was higher with a one-dimensional display (where sound sources are arranged on a horizontal line from left to right in front of the user) than with a two-dimensional display (where sound sources are arranged on a circle centered on the user).

Another audio-tactile interface, developed by Zhao et al. [109], proposes a translation of the iPod visual menu selection to the auditory domain. It makes use of a circular touchpad for radial menu selection, and the audio feedback consists of human voices spatialized by implementing interaural time and intensity differences. Unlike HRTFs, this left-right spatialization technique does not provide a full reproduction of sound sources all around the user, however the authors claim that the illusion is achieved when interacting with the circular touchpad, suggesting cross-modal interactions between the tactile and auditory modalities. If corroborated, this observation has great implications for the spatialization of sound sources in audio-tactile interfaces whose both spaces have similar shapes, because it allows to use simplified techniques in order to provide the only relevant pieces of information needed to represent the audio space. In a user study comparing the circular auditory feedback with a linear visual feedback, results reveal that the speed and accuracy of the audio feedback catch up with the performance achieved with the visual feedback after some practice time. Moreover an equal number of participants preferred the visual and the audio menus, which shows that this kind of auditory interface is promising.

In this chapter we present an audio-tactile interface for audio navigation based on discrimination of sound sources in depth. This work is also described in Paper C and Paper F (the latter reports the user study as well).

6.2 The user tactile interface

6.2.1 Design approach

This section presents an interface meant to navigate among multiple audio streams ordered in depth and using a tactile input from the user. Auditory depth information is provided by a Digital Waveguide Mesh modeling a bidimensional rectangular membrane and already described in Chapter 4. The user tactile interface consists of a ribbon controller, the Infusion Systems *SlideLong*[1], inspired by music controllers.The choice of a linear position touch sensor is validated by its common use for browsing through playlists on portable music players equipped with touch sensors. In addition, the *SlideLong* has an active area of $210 \times 20\,\mathrm{mm}^2$, and therefore exhibits a main dimension along which depth information may be explicitly understood. Since the ribbon has a rectangular geometry, it also allows an easy analogy with the geometry of the rectangular DWM simulating the auditory environment. The continuous interaction is performed by a mapping of the position of the user's finger on the ribbon onto the position of a virtual microphone on the membrane.

6.2.2 Prototype development for audio browsing

The position tactile sensor gives a value corresponding to the touch position on the ribbon. A gamepad plays the role of sensor interface and is connected to the USB-port of the computer. As underlined by Jensenius et. al [42], such a game controller has several advantages, in particular it is cheap and has analog inputs which comply to the 0-5 volt sensor outputs. Besides connecting the sensor outputs on the motherboard is easy and the device uses the Human Interface Device driver supported in Max/MSP, which allows to make a fast, simple and low-cost interface sensor out of the game controller. The incoming values in Max/MSP are read by the built-in *human interface* object and are scaled to float numbers between 0 and 255. Therefore, after rescaling, the incoming value from the ribbon controller provides the listening position input to the computation of the auditory signals in the DWM. In this way, a coherent mapping is performed between the touch position on the ribbon and the position of a virtual microphone in the DWM, and by moving the finger on the ribbon the user may explore the virtual environment where different audio streams are being attributed different positions. Like music controllers, this touch sensor intends to provide an interface that is intuitive to use with immediate and coherent response to user's gesture.
Due to restrictions of hearing discrimination capability, it was found by informal listening tests that at most four audio streams could be reproduced in the virtual environment. Therefore a second tactile interface is added in order to allow to

[1] `http://infusionsystems.com/catalog/product_info.php/products_id/52`

browse among more sound files. For this purpose, we use one of the rotary encoders of a keyless MIDI controller (the Novation *ReMOTE ZeRO SL*), which is available at our laboratory. The controller is connected to the USB-port of the computer. The encoder has discrete steps and therefore may be easily manipulated to switch between discrete levels. Like the position tactile sensor, its output value is read in Max/MSP, and determines the four audio streams processed in the virtual environment. Figure 6.1 shows the complete setup.

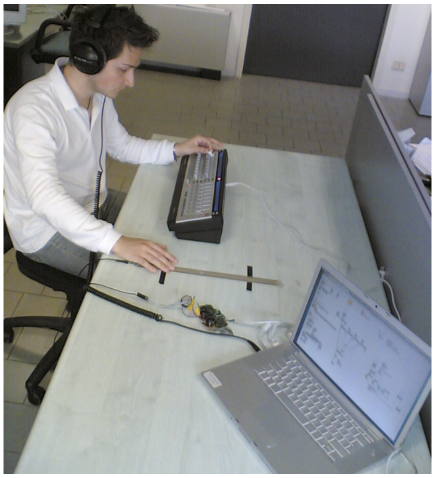


Fig. 6.1. *Picture of the experimental setup.*

6.3 Modeling the audio space

6.3.1 The Acoustic Environment

Our proposed virtual acoustic environment consists of a rectilinear two-dimensional mesh. Due to software restrictions, the mesh dimensions are chosen to be 110×5 nodes, which correspond to a $120 \times 5\,\mathrm{cm}^2$ membrane. The ends of the mesh simulate totally reflective surfaces, while the side edges are partially-absorbant. Although the audio space is quite small, the model allows to produce noticeable differences in the depth cues, as it will be shown in Section 6.3.2.

6.3.2 Acoustical properties of the membrane

In order to investigate the auditory depth cues in the virtual environment, impulse responses are computed at different distances on the membrane, corresponding to the positions of four audio streams. The sound sources are assumed to be point-wise and are equally-spaced on the y-axis of the membrane. Measurements are carried out at 5 cm from the boundary. Refer to Fig. 6.2 for the sources and measurement positions. Simulation of the listening environment was carried out in `Matlab`. Fig-

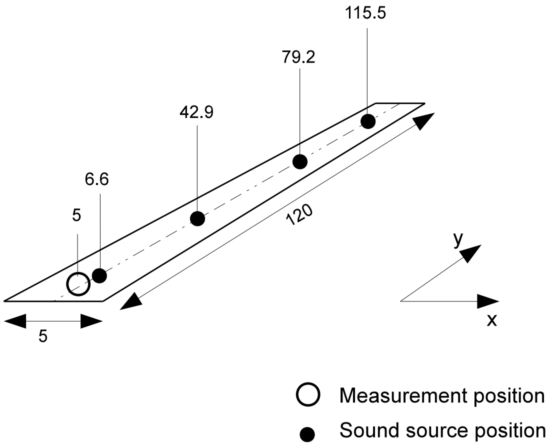


Fig. 6.2. *The virtual membrane with the sound source and the measurement positions. All sizes are in centimeters.*

ure 6.3 shows the frequency responses up to 5 kHz, measured respectively at 6.6 cm and 115.5 cm on the virtual membrane.

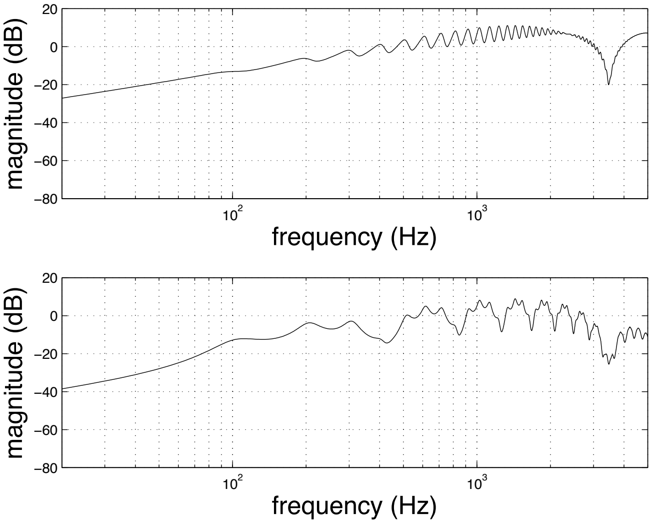


Fig. 6.3. *Frequency responses up to 5 kHz on the membrane. Top: 6.6 cm. Bottom: 115.5 cm.*

Intensity cue

Figure 6.4 shows the variation of the total energy with distance. Based on the smallest just-noticeable differences for the intensity cue (about 1 dB [43]), the decrease of energy with distance in the present auditory environment should be perceived, in particular for the near distances. By comparing with the energy decrease in open space, it can be seen that the overall intensity on the membrane decreases significantly less above 0.4 m. This behavior allows to hear even the farthest sound sources at any location on the membrane.

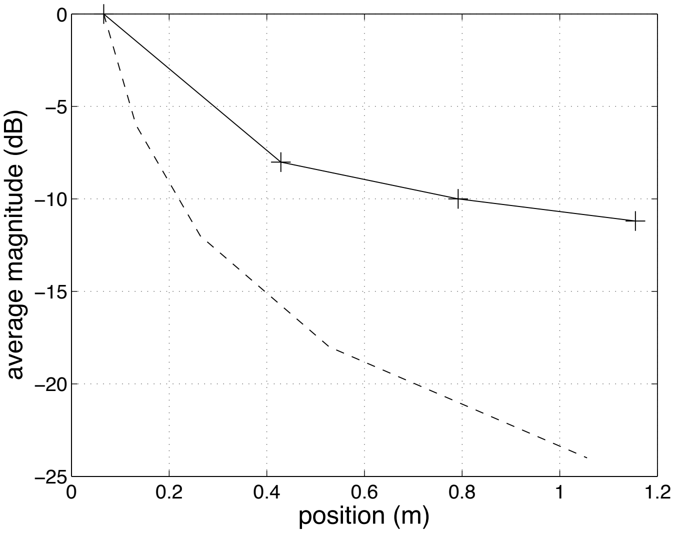


Fig. 6.4. *Average magnitude of the impulse response as a function of distance. Solid line: 2D mesh. Dashed line: Reference open space.*

Direct-to-reverberant energy cue

The virtual acoustic space we have designed aims at rendering depth information mainly thanks to the direct-to-reverberant energy ratio cue. Figure 6.5 shows the values of the direct-to-reverberant energy ratios for different distances on the mesh, computed according to the method described in Section 4.2.2. If one refers to the sensitivity thresholds for this cue (2 to 6 dB, depending on the study [50, 105]), the ratio differences between adjacent sound source positions (marked by "+" in Figure 6.5) should allow to discriminate between them, in particular between the adjacent sources. Besides, the values of the ratios are much lower in the 2D mesh than in the semi-reverberant auditorium modeled by Zahorik [105], as it can be noted by comparing the solid line with the dashed line in Figure 6.5. This means that the amount of reverberation is exaggerated in the virtual environment, as a conceivable consequence estimated distances might be greater than physical

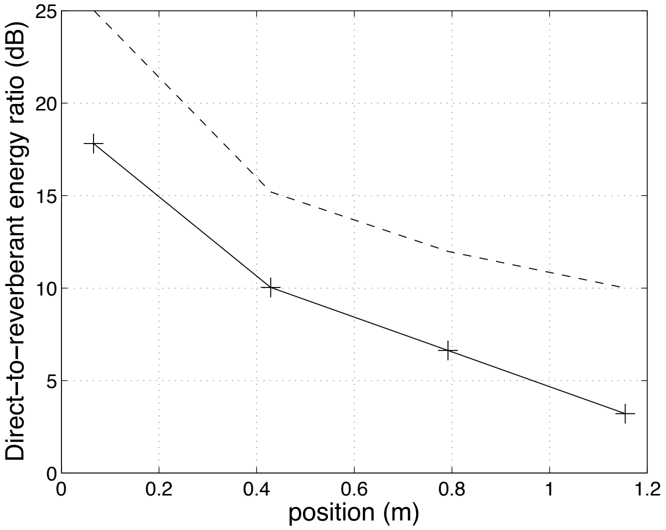


Fig. 6.5. *Direct-to-reverberant energy ratio as a function of distance. Solid Line: 2D mesh. Dashed line: Semi-reverberant auditorium.*

distances. Both curves are however almost parallel, suggesting that changes of the reverberation cue with distance are consistent with those observed in a familiar environment.

6.4 User study

An experiment was designed to evaluate the interface described in Section 6.2 in a task of audio browsing.

6.4.1 Method

Participants

Twelve Italian students (one woman and eleven men) from the University of Verona, Italy, aged 22-26 years old, participated in the experiment. All reported to have normal hearing. One is a DJ in his free time and two are musicians.

Stimuli and apparatus

The playlist consisted of twelve loop samples intended for DJs found on Internet[2]. Most of them had a bpm of 130, and the others were also set at 130 bpm by time stretching. Presenting all the music pieces at the same bpm prevents from creating a cacophony in the virtual environment which makes the identification task more

[2] `http://www.audiobase.com`

difficult and less pleasant. This observation was made during a preliminary user study with various pieces of pop music. In addition, we believe that the present interface may be particularly appropriate for DJs who handle loop samples and must in real time look for new ones with the same bpm. Finally, the twelve wave files were RMS equalized to avoid bias due to intensity differences between the samples.

The ribbon and the knob described in section 6.2 were used for the experiment and connected to the USB-ports of an Apple MacBook Pro computer. Data from the ribbon were rescaled to fit the length of the DWM. Four sound files were equally spaced on the y-axis of the 120 cm virtual membrane, placed at 6.6, 42.9, 79.2 and 115.5 cm respectively. The objective of the knob was to be able to switch among the four sound files processed in the mesh and simultaneously available to the user. In a preliminary evaluation test, incrementing the value of the encoder enabled to move one sound file forward or backward in the playlist. It was however found to be not very efficient. As a consequence, it was chosen to jump over more soundfiles when incrementing the encoder's value.

Processing of the incoming MIDI data as well as spatialization in the DWM were carried out in Max/MSP. The output mono audio signal was presented over a pair of Beyerdynamic DT 770 headphones, and the sound level control of the computer was kept constant for all users. The experimental setup may be seen in Fig. 6.1.

Conditions

Advancing of 3 and 4 sound files with one knob increment were retained as the two conditions studied in the experiment, and are respectively named *condition 3* and *condition 4*. Figure 6.6 shows how the knob increment operates on the selection of the four concurrently playing audio files spatialized in the DWM.

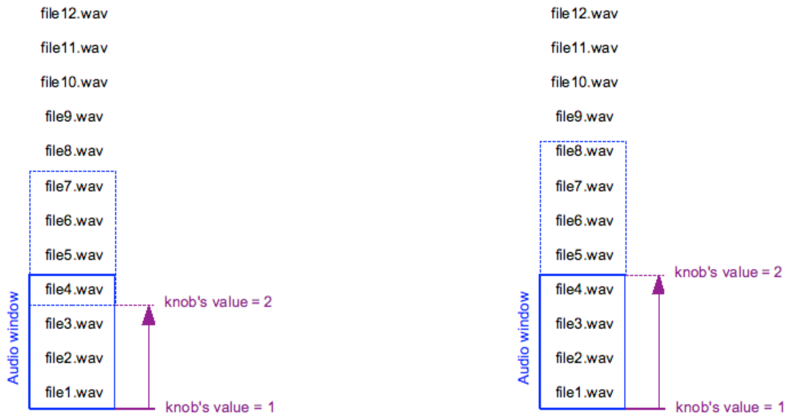


Fig. 6.6. *Four sound files are simultaneously processed in the DWM and the increment of the knob's value enables to move the mesh 3 sound files forward (left figure) or 4 sound files forward (right figure).*

The reason for choosing these two conditions is to study the two different mental representations that may be induced by the two implementations. In condition 3, the overlap between consecutive windows is intended to model advancing in a linear space. This condition is therefore meant to make the user explore the whole set of sound files as a monodimensional space, which may be accessed through a window of 4 sound sources. As for condition 4, a different mental representation of the space should arise: this condition allows to find each file in only one block of four sounds, and the blocks are independent from one another. As a consequence, the mental representation is no meant to be linear since one could explore the first block, then the third one and finally the second one. In this case the different blocks are rather represented as parallel spaces containing different sound files.

Procedure

The two conditions were assigned to different groups of subjects. For each condition, the twelve sound files of the playlist were randomly ordered for both training and evaluation phases. After being instructed about the functioning of the two tactile interfaces, users put on the pair of headphones and were allowed to browse freely among the twelve sound files in order to get acquainted with the tool. They could spend as much time as they needed for this training phase.
In the second phase, namely the evaluation phase, the experimenter played once a target sound file selected randomly in the playlist for 10 seconds. In the beginning of each search the knob was reset. Then the experimenter gave the go-ahead to the user for searching the sound file in the playlist. When the sound file was found, the user had to press the button above the encoder on the ReMOTE ZeRO SL. Just as the button was pressed, the time required to find the target file, the position of the user's finger on the ribbon and the value of the encoder were recorded. In addition to these data, a log file holding the values of the encoder and the position on the ribbon every 100 ms was saved for each subject. The whole experiment consisted in three consecutive sessions. For each of them, the user had to find each of the 12 sound files selected randomly among the playlist. Therefore each participant had to find a total of 36 target files.
After the experiment, a debriefing phase allowed to get a subjective evaluation of the prototype by asking users questions such as “Was the task difficult?”, “What do you think about the usability of the ribbon and the slider?”, ”What would you suggest to improve the interface?”, etc... These questions were only suggested to the users as they could freely write down their comments.

6.4.2 Results

User testing

The position of the finger on the ribbon and the level of the knob when the user reached the estimated position of the target sound file were used to compute the difference between the estimated position and the actual position of the target file. Results are displayed as a function of the actual sound file position on the ribbon. Figure 6.7 shows the distributions of the users' answers, for both conditions. The

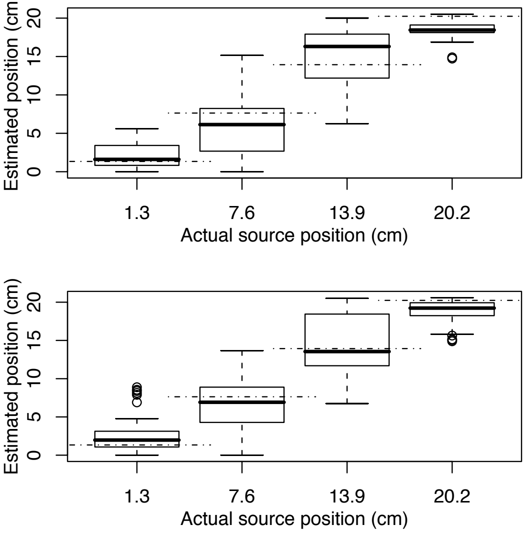


Fig. 6.7. *Estimated position versus actual position of the target file. The boxplots depict the median and the 25%/75% percentiles. Outliers are marked by 'o'. Dot-dash lines represent the actual positions of the sound sources. Top: condition 3. Bottom: condition 4.*

threshold for deciding whether the sound file found by the user was effectively the target sound file was logically chosen to be at half distance between the actual position of the sound source on the ribbon and the position of the adjacent sound source(s). In this way, the percentage of good answers could be calculated. A value of 93.02 % was achieved for condition 3, and 93.75 % for condition 4. Both conditions showing similar results, a linear regression between the estimated positions and the actual positions of the sound files was performed on all subjects. Results give a slope of 0.92 with $s.d. = 0.027$, $r^2 = 0.76$ and $p < 2 \times 10^{-16}$.

Time required to find the target sound file was also analyzed for both conditions, after taking out the wrong answers. Results are shown in Fig. 6.8, as a function of the position of the target file in the playlist. The Shapiro-Wilk test run on the time required to find a target file for each condition excludes that the two distributions are normal [$p < 2.2 \times 10^{-16}$ for condition 3, and $p = 1.434 \times 10^{-14}$ for condition 4]. Therefore, requirements are not fulfilled to test the null hypothesis that the two population means are equal using the Student's t-test. A non-parametric alternative is the two-sample Wilcoxon test, which reveals a statistical difference between the two conditions [$p = 0.048$]. Computation of the mean time for condition 3 and 4 leads to values of 20.47 and 15.03 respectively, meaning that in average, users under condition 4 perform the task faster than users under condition 3.

Through sessions, the average time does not change significantly under the two conditions [paired Wilcoxon tests between each pair of sessions gives $p_{min} = 0.07209$ for condition 3, and $p_{min} = 0.06445$ for condition 4]. Top and bottom plots of Fig. 6.9 represents the average time values for each sound file position in the playlist

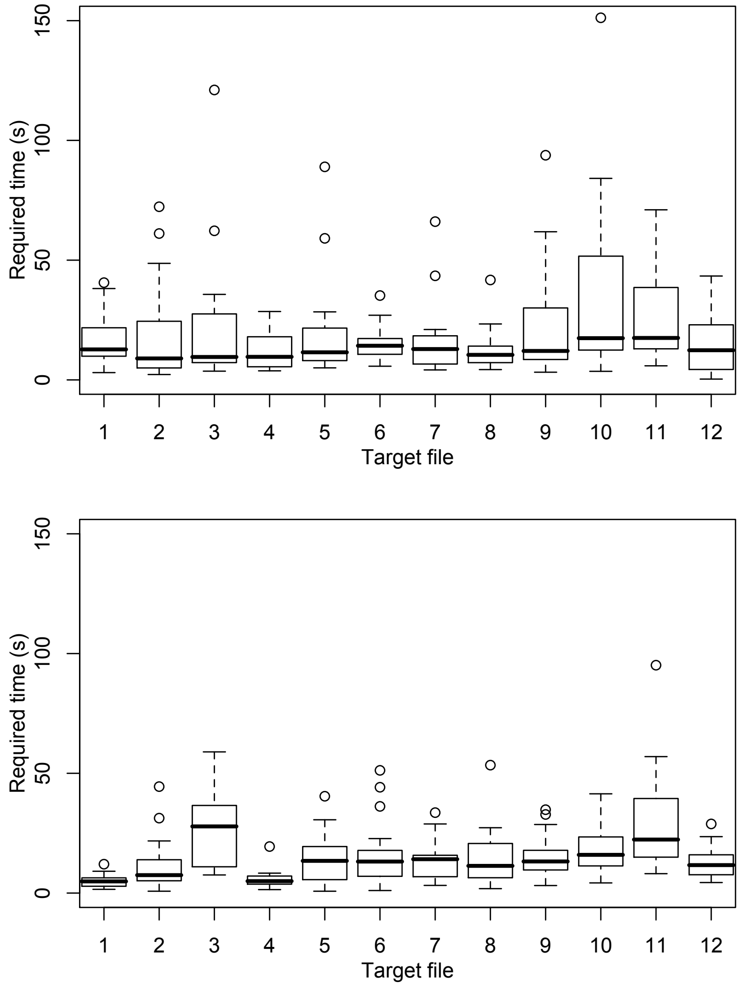


Fig. 6.8. *Time required to find the target file in the playlist. The boxplots depict the median and the 25%/75% percentiles. Outliers are marked by 'o'. Top: condition 3. Bottom: condition 4.*

for the three sessions, under condition 3 and 4 respectively. Another way to investigate the evolution of the performance over the sessions is to compute average time values as a function of the position of the target sound source on the ribbon, as shown in Fig. 6.10. While the evolution of performance under condition 3 does not exhibit any clear behavior, it can be seen that time decreases with training for condition 4, and in particular for the two central positions on the ribbon. At these positions, target sound files might be more difficult to find during the first trials because unlike the two extreme positions, they are surrounded by two adjoining sound sources located at equal distances from the target file.

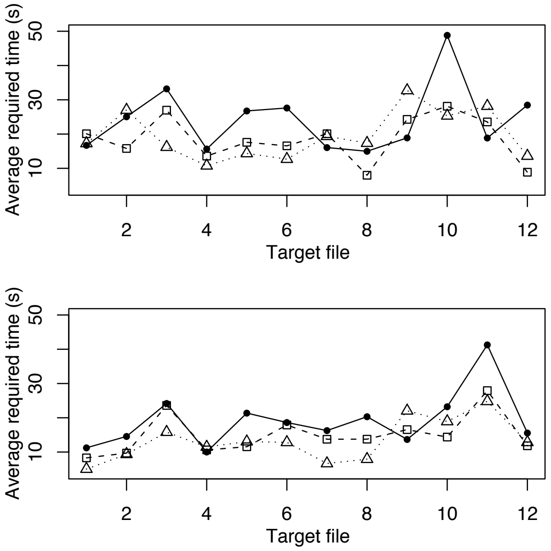


Fig. 6.9. *Average required time to find the target file as a function of the target file's position in the playlist. Solid line: session 1, dashed line: session2, dotted line: session 3. Top: condition 3, bottom: condition 4.*

Qualitative evaluation

All users found the interface entertaining and intuitive. They could quickly use it for searching a sound file as they reported that it was a simple and fast method for audio browsing. A few users however pointed out the confusion that could arise between some sound files, in particular among the rhythmic ones, while the task was easier for more melodic loop samples. In addition, many users suggested a more clear separation between the sound files in the virtual environment.

6.4.3 Discussion

From the performance evaluation, it can be seen that users manage to find the target file in the playlist using the tactile interface. Comparison between the two conditions reveals that in average the time required to reach the target sound source is smaller when the knob increment advances of four sound files in the playlist. This result suggests that the implementation is more efficient when no sound file is repeated from one knob's position to the consecutive one. Besides, efficiency did not improve significantly through the experiment under condition 3. One could expect that some participants could remember the position of some sound files from one session to another since the playlist order remains unchanged. Although not statistically significant, results under condition 4 show a time decrease over the three sessions. The difference between the two conditions may be explained by the more simple implementation of condition 4: each sound file is played for one knob's value only, whereas condition 3 renders some files for two different values of the knob and at different positions in the virtual environment.

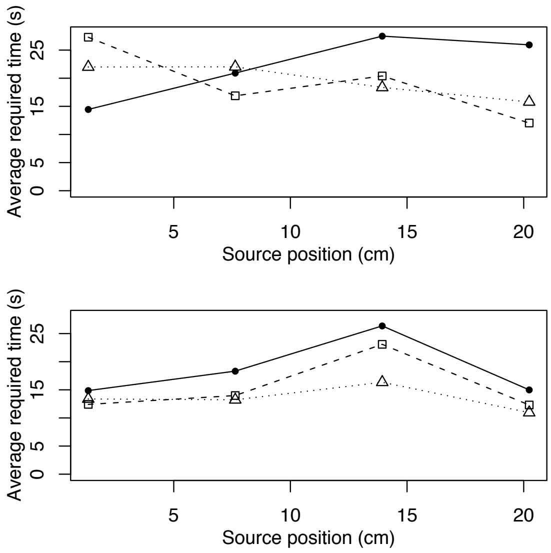


Fig. 6.10. *Average required time to find the target file as a function of the target file's position on the ribbon. Solid line: session 1, dashed line: session2, dotted line: session 3. Top: condition 3, bottom: condition 4.*

Even if condition 3 may provide a more complex interface, the users tested under this condition did not express any particular difficulty, compared to those under condition 4. In fact, all reported that the interface was easy and enjoyable to use.

6.5 Conclusion and future work

We have proposed an interface for audio browsing based on spatialization of audio data in depth. The virtual environment rendering depth cues of sound sources is modeled by a rectangular Digital Waveguide Mesh. Associated with a rectangular ribbon playing the role of a tactile input to the interface, this coherent tool may offer new opportunities for designing auditory interfaces. Originally used as a substitute for a computer mouse, touchpads are now invading the world of portable media players and personal digital assistants. They are used as a control interface for menu navigation on all of the currently produced iPod portable music players. Another illustrative example is the iPhone that may be used as a touchpad to wirelessly control a computer. The proposed spatialization in depth of sound sources for navigation could therefore benefit from the already existing hardware and diversify the interaction modes. Furthermore, as suggested by the audio-tactile integration of spatial cues hypothesized by Zhao et al. [109] (refer to Section 6.1), the perception of the sound sources order in depth may benefit from the user interaction with a tactile interface whose shape is identical to that of the audio space. An evaluation of the interface used for audio browsing was conducted. The limited length of the DWM requires another tool to navigate among numerous sound files. Then the tactile ribbon plays the role of an audio window in the virtual environment, and a knob allows to move this window along the whole audio space. A user study showed the ability of the tool to find audio files, in an entertaining and intuitive manner. The acoustic environment provided by the DWM needs however further investigations to evaluate its “naturalness” and ”pleasantness”. Sounds resulting from audio signals rendered in the DWM are colored but are not annoying. Obviously, the present DWM model can not be compared to artificial reverberators which ideally provide musical performances with realistic ambience and as much transparency as possible. Instead, this thesis has proposed to apply the DWM to human-computer interfaces for which quality criteria (such as Quality of Service [74]) may largely differ from those established by the music industry. Regarding the usability of depth cues for sound navigation, the DWM simulation could be compared to other techniques such as crossfading or stereo panning. Though, spatialization in depth outputs a monaural signal and therefore does not rely on the reproduction hardware configuration, unlike stereo panning as proposed by Pitt and Edwards [69]. Besides, CPU load restricted the available length of the DWM, while a more efficient computation may allow to increase the length and consequently a better sound source discrimination.

Two implementations for the knob’s control on the displacement of the audio window were tested by different groups of users. A statistical analysis revealed a significant time reduction in the case of 4 sound files shift, i.e. when no audio stream is repeated from one audio window to another. While a shift of 3 sound files may induce a linear audio space representation, a shift of 4 splits the space into parallel and independent subspaces that can be directly accessed. Under this assumption, the user under condition 4 should more easily remember which subspace the target audio file belongs to, and therefore increase its performance over time. This reasoning is supported by the time decrease with the number of sessions observed for condition 4. Further analysis may deal with the strategies used by the two groups of participants, by looking at the individual log files which provide the evolution of the audio window position and the listening position inside this window over time. Furthermore, it could be interesting to develop models for the task completion time and compare them to experimental results. For condition 4 where all the blocks are independent, the model might consist in dividing the task completion time into the time needed to find the right audio window (depending on the knob’s value) and the time needed to find the target file in the audio window (depending on the position of the file on the ribbon). This model should no longer fit for condition 3 since the blocks are not independent.

The user study presented in this chapter offers various research directions, both in terms of interaction design and auditory perception. Among the numerous issues, the nature and the number of concurrently playing sound sources in the audio window most probably affect the performance results. A compromise should be found between auditory overload and performance.

Finally, other uses of the interface may be explored. In particular, an auditory menu may originate from carefully sonified menu items, spatialized in depth using the DWM, and accessed through the linear tactile interface.

6.6 A physically-oriented interface

As stated in Chapter 2, listener's familiarity with the sound source signals improves distance perception. Therefore everyday sounds may contribute to the usability of the aforementioned audio-tactile interface. It has indeed been demonstrated that everyday sounds are an intuitively accessible way to convey information to users [29].

As part of a collaborative work, a demonstration using physics-based models of three everyday sounds was realized, resulting in an interface which is entirely physics-based (Paper D, in Italian).

More specifically, the sound models, specified by physical descriptions and actions, were initially conceived within the European project *SOb (Sounding Object)* and further developed within *CLOSED*. From low-level models such as impact or friction, more elaborated synthesis models were developed, and compiled for both PureData and Max/MSP. These perceptually and physically coherent sound models are meant to be a tool for sound design, and may be found in the *Sound Synthesis Tools for Sound Design.*

For the demonstration, sounds of frying, knocking and dripping were spatialized in a 1×0.05 m^2 rectangular mesh. People could freely navigate through the set of sounds using the aforementioned ribbon equipped with the touch position sensor. They could also follow their position on a screen, as shown in Figure 6.11. Participants' feedback was positive. For a proper application, one could think of using this kind of physics-based models to sonify menu items and therefore propose the interface as a tool for menu selection.

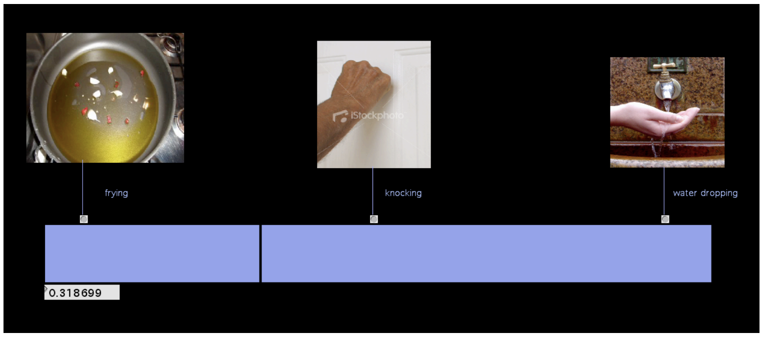


Fig. 6.11. *Visual feedback of the audio-tactile interface.*

7

Depth cues of a moving sound source

7.1 DepThrow: a physics-based audio game

7.1.1 Introduction

In the previous chapter, the proposed interface allowed to move the listening position in real time relatively to the fixed positions of the sound sources. This chapter now considers the case where a sound source is approaching or receding from the listening position which is fixed. The Max/MSP external `mesh2D_dynamic∼` was created for this purpose (Refer to Section 4.3.3 for more details) since it allows to render dynamic variations of the sound source position, and consequently provides a tool for interactive manipulation of sound sources along the depth dimension. The idea was therefore to use the bidimensional rectangular DWM described in Chapter 4 to spatialize a moving sound source as the result of a user action. The sound of a rolling ball fits well the objective of the prototype: it requires that the sound source is moving. Rolling sound was already used by Yao and Hayward [100] and Rath and Rocchesso [71]. In [100] the authors studied the cues exploited by users to locate the position of a virtual rolling or sliding ball inside a tube, and showed the relatively good ability of users to guess the length of the tube just by tilting it and perceiving (from auditory and/or haptic cues) the virtual object's dynamics. A physics-based model of rolling sound was also used in [71] as a continuous auditory feedback for balancing a ball along a tiltable track. To evaluate the interface, subjects had to move a virtual ball until a specific position on the track was reached. Even without any visual display, all subjects managed to solve the task, which means that people were able to locate the virtual ball with the auditory feedback only.

In collaboration with Stefano Papetti, an interactive audio game was designed as an entertaining application, on the one hand to show the potential of physics-based models for the simulations of both the sound source and the acoustical environment, and on the other hand to investigate the perception of dynamic depth. The work presented in this section is also available in Paper B and Paper E (the latter is written in Italian).

7.1.2 Game rules

The game consists in throwing a virtual rolling ball inside a virtual open-ended tube which is inclined. The goal is to make the ball roll as far as possible and therefore adjust the initial velocity applied to the object such that the ball does not fall out at the far end of the tube. Figure 7.1 shows the behavior of the ball inside the tube.

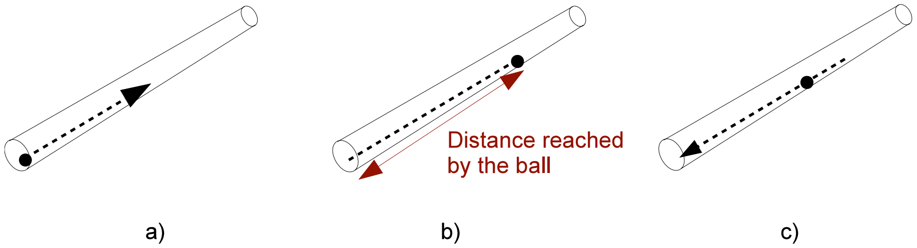


Fig. 7.1. *Movement of the ball in the tube. a) Initial position of the ball. b) The ball recedes and reaches a maximum distance. c) The ball rolls back to its initial position.*

The user throws the virtual ball using a *Wii Remote*[1] mounted on a trolley, as shown in Figure 7.2. The Wii Remote provides a 3D accelerometer, a vibration actuator, several buttons, and Bluetooth connectivity. It is therefore possible to map the acceleration obtained by the accelerometer onto the initial velocity of the virtual ball. Besides, the Wii Remote is also exploited to provide a haptic feedback with the actuator vibrating when the ball gets back to the user's hands.

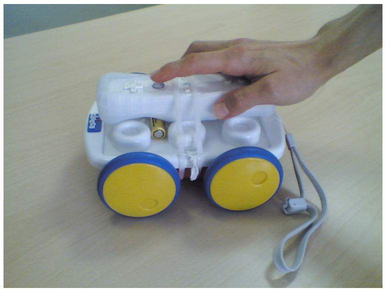


Fig. 7.2. The virtual ball is thrown with the Wii Remote mounted on a trolley.

[1] The handed wireless controller for Nintendo Wii console

7.1.3 Architecture of the interface

The game prototype has been developed as a complex Max/MSP[2] patch where three interacting physics-based models enable to simulate respectively:

1. current displacement and velocity of the ball based on the initial velocity (given by the user) and the tilt angle of the tube
2. real-time synthesis of the rolling sound:
   The ball simulation consists of a physics-based rolling sound model which was originally developed and implemented during the EU-funded project *SOb* by Rath [70], and further developed for the EU-funded project *CLOSED*. As a result of the physical coherence of the model, it is straightforward to map its control parameters to continuous physical interaction. In this case, the model is driven by the current velocity of the ball. In addition to the rolling sound, a bouncing model is used to simulate the ball falling over at the far end of the tube.
3. depth rendering:
   For simulating the tubular environment, a rectangular 2D DWM (as described in Chapter 4) modeling a 10×0.3 m^2 membrane is used. Figure 7.3 shows the variation of depth cues with distance in that environment.

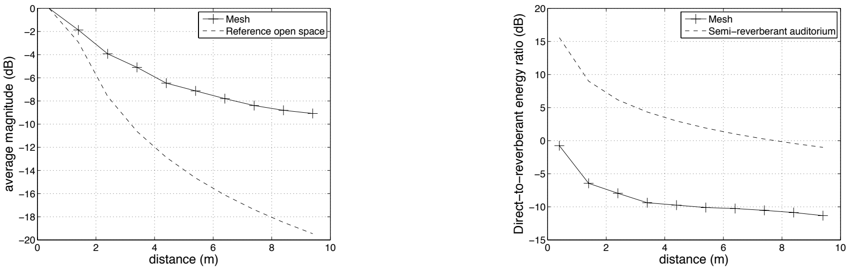


Fig. 7.3. *Intensity cue (left figure) and direct-to-reverberant energy ratio cue (right figure) in the 2D DWM.*

The three physics-based models are connected in a top-down chain structure. Figure 7.4 shows the main Max/MSP patch.

7.1.4 Objectives and Issues

This game represents a potential tool for exploring the usability of auditory depth information in interaction design. Besides entertainment, this physically-consistent prototype allows to evaluate the playability of the game which is the result of a perception-compliant mapping of user gesture (the act of throwing the virtual object) onto the control parameters of the tilted plane and the rolling model. In the present prototype, the user mush push the trolley horizontally and push a button

[2] `http://www.cycling74.com/products/maxmsp`

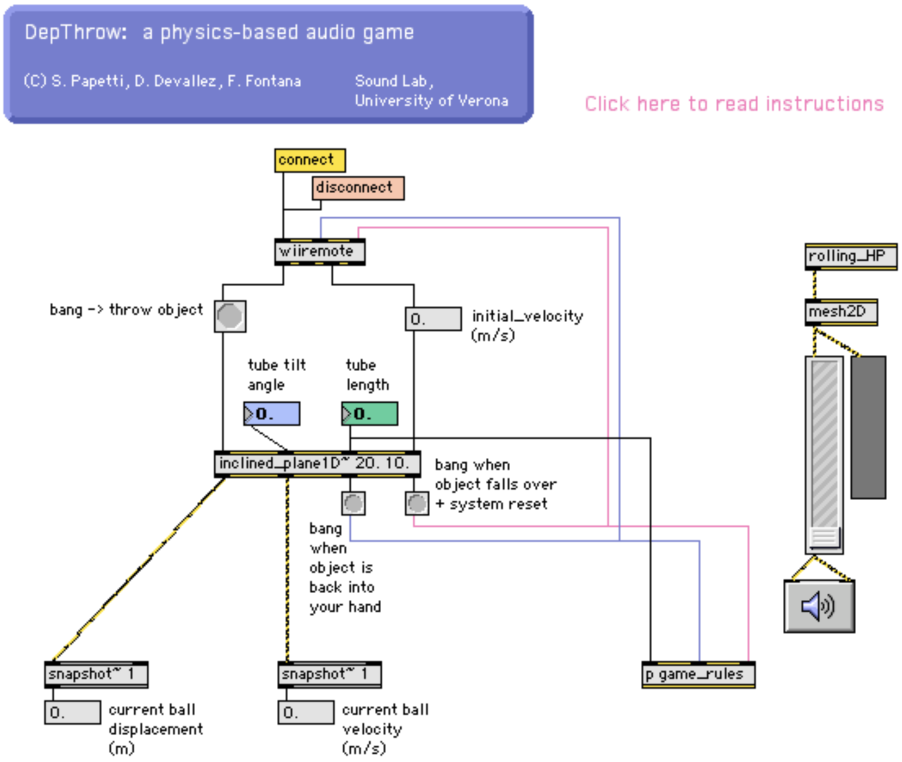


Fig. 7.4. Main Max/MSP patch.

on the Wii Remote during the last part of his gesture, which corresponds to the integration time for the computation of the ball's initial velocity. This restricted gesture was required to obtain a proper estimation of acceleration, but it strongly limits the user's commitment that would otherwise result from a more natural ball throw. During demonstrations of the game, it was therefore difficult to collect comments because people did not feel engaged.

Yet, the main point of interest for this thesis is the perception of the dynamic depth of the rolling ball. Up to now, most research on auditory motion perception has dealt with sound sources of constant velocity, which may hardly be found in real life situations [44]. The present prototype is therefore a potential research tool to investigate the perception of moving sound sources whose dynamics is physics-based.

While the tubular environment provides information on the intensity and the direct-to-reverberant energy ratio which are two main dynamic distance cues (see Section 2.4), the rolling sound itself may be a carrier of distance information. Typical rolling sounds, including the physics-based model used in DepThrow, exhibit periodic patterns which provide cues on the duration of the roll. In other words, one may be able to count the number of rotations of the ball. Together with the size of the ball, a physical invariant that is also available through the sound of rolling [37], this information may result in a percept of distance, independently of the spatial distance provided by the environment. In our prototype, since the size of the ball is constant, counting the number of rotations of the ball may be sufficient to judge its traveled distance. As a result, two cues for distance perception may be distinguished in our prototype: distance information which is inherent to the rolling sound itself, and distance information provided by the physical distance traveled in the virtual tube. These speculations lead to the following question: do people use spatial information to perceive the distance traveled by the ball in the virtual tube, or only cues provided by the physics-based model of rolling sound? In case that number of rotations of the ball is difficult to assess (for instance a small ball with a high speed lead to a high number of rotations in a short time), another hypothesis may be that people base their estimation of traveled distance simply on the duration of the rolling sound.

7.2 Listening test

7.2.1 Introduction

The present listening test aims at investigating the perception of dynamic depth. For this purpose, the prototype of the audio game previously described allows to simulate a ball rolling in an inclined tube with a physically-consistent movement of backwards and forwards. As explained in Section 7.1.4, several attributes of the rolling sound may contribute to the perception of depth in addition to the cues produced by the depth rendering model. While playing the game, the user's action (i.e. throwing the ball with the Wii Remote) may as well influence his/her perception. In order to restrain the list of candidates for the perception of dynamic depth, the experiment deals with passive listening of stimuli pre-computed in the model. The listening test consequently focuses on the influence of the auditory attributes, and in particular on the duration of the signal and the cues provided by the DWM. In order to evaluate dynamic depth, which involves a movement, listeners are asked to judge the maximum distance reached by the rolling ball inside the tube. The ball starts at the closest end of the tube, recedes, reaches a maximum distance and comes back to its initial position (see Figure 7.1).

7.2.2 Method

Subjects

Nineteen Italian listeners (16 males, 3 females), aged 19-23 years old, voluntarily took part in the listening test. They were all students at the University of Verona, Italy. All of them reported to have normal hearing.

Apparatus and stimuli

Source material

The prototype described in Section 7.1 was used to create the auditory stimuli. By varying two input parameters of the model, namely the initial velocity and the tilt angle of the virtual tube, it was possible to create various stimuli for which the virtual ball would reach different distances inside the tube. In order to investigate the influence of duration and initial velocity of the ball on the perception of the traveled distance, three values were chosen for both parameters, which resulted in nine stimuli. In particular the parameters values were carefully chosen so that people could not estimate the distance traveled by the ball only by counting its number of rotations. Finally, the initial angle and the maximum distance reached by the ball could be easily computed for each stimulus. Table 7.1 summarizes the different attributes of the nine stimuli.

Table 7.1. Parameters of each stimulus. The stimuli are in order of increasing distance.

stimulus no.	duration (s)	initial velocity (m/s)	tilt angle (°)	traveled distance (m)
1	1.4	7	30.6	4.9
2	1.4	8.3	37.2	5.8
3	1.7	7	24.8	5.9
4	1.4	9.6	44.3	6.7
5	2	7	20.9	7.0
6	1.7	8.3	29.8	7.1
7	1.7	9.6	35.1	8.2
8	2	8.3	25.0	8.3
9	2	9.6	29.3	9.6

Experimental setup

The stimuli were generated by an interactive `Matlab` script and presented over an AKG K240 open headphone set. Subjects gave their answers on a computer terminal with a numeric keypad.

Procedure

For the reasons already developed in Section 5.3.1, it was found that direct scaling procedures were not adequate for assessing relative dynamic depths. Procedures that allow a direct comparison of all stimuli such as the modified MUSHRA test described in Section 5.3.1, were also rejected; because of the high number of stimuli and their long duration, the task would have been arduous. It was therefore decided to use pairwise comparisons that requires a simple comparison judgments by the subjects. In such a test, subjects make pairwise comparisons of all stimuli to each other regarding the property under investigation, and make a judgment by deciding which of the two stimuli in each pair has the highest value of that property.

Prior to the experiment, participants received a written instruction. During the experiment, two sounds were available to the subject, who could listen to them in any order and as many times as they wanted by clicking on the corresponding sound, on the left part of the user interface shown in Figure 7.5. It was indeed found in a preliminary informal test that the task was more difficult when playing each sound only once because of the duration of each stimulus. The subject was asked: “For which sound the ball rolls the furthest away in the tube?”. A response was given by clicking on “1” or “2” on the right part of the interface (see Figure 7.5). The experiment consisted of two sessions per subject, conducted on different days or with a break of at least 10 min. For each session, all possible pairs were presented randomly twice to each subject, that is once in each order. Therefore each subject had to evaluate a total of 144 ($9 \times (9-1) \times 2$) pairs.

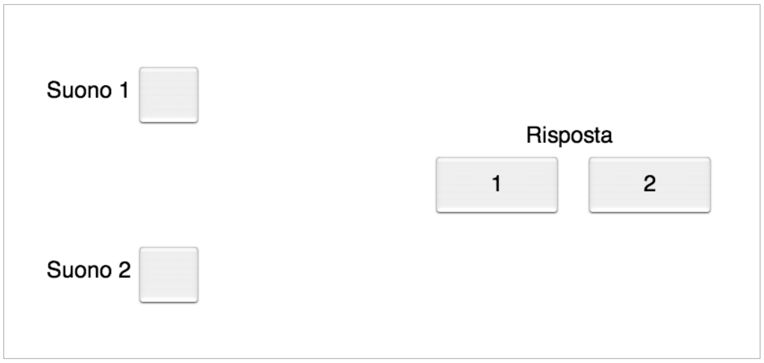


Fig. 7.5. *Snapshot of the user interface (in Italian) during the listening test.*

Statistical analysis

The pairwise choices among the nine stimuli were collected for all subjects, resulting in a *preference matrix* which represents the number of times one stimulus i is chosen over each other stimulus $j(\neq i)$. Converting these values into proportions leads to the *choice frequency matrix.* These fractions are now interpreted as the probability that stimulus i is judged to have a higher value than stimulus j in a direct comparison. It is then possible to order the set of stimuli on a one-dimensional scale by applying a probabilistic choice model to the choice frequency matrix. The Bradley-Terry-Luce (BTL) model is often applied to pairwise comparisons, and was proposed by Bradley and Terry [9] and Luce [54]. In this model, the probability of choosing the stimulus i from the set of simuli S is given by

$$P(i, S) = \frac{u(i)}{\sum_{j \in S} u(j)} \tag{7.1}$$

where u is the ratio scale corresponding to the stimulus i. The use of this model assumes that the choices are made independently of the context introduced by a given pair. It is also important to note that the model provides *relative* scale values, that give information only about the *differences* between the locations of the stimuli on the subjective scale. The analysis relies on the *eba* package[3] provided by Florian Wickelmaier for *R*, a free software environment for statistical computing and graphics. Among others, this package allows to fit the BTL model and to test its goodness of fit using the χ^2 test [98].

7.2.3 Preliminary results

Table 7.3 displays the preference matrix for all subjects. The choice probabilities were estimated based on 19×4 = 76 observations per stimulus pair. Applying directly the BTL model leads to the rejection of the model because the χ^2 test gives a p value less than 10% ($\chi^2 = 48.82$, $p = 0.00872$). While applying a model to the data to derive a scale, it is worth evaluating the consistency of the judgments given by the subjects. This is usually done by analyzing weak, moderate and strong stochastic transitivities defined as follows:
if $p(i > j) > 0.5$ and $p(j > k) > 0.5$, then

Table 7.2. Absolute preference matrix for the nine stimuli in order of increasing distance.

stimulus no.	1	2	3	4	5	6	7	8	9
1	0	24	18	26	8	9	17	13	9
2	52	0	16	37	8	14	12	9	8
3	58	60	0	65	16	35	27	21	16
4	50	39	11	0	10	9	10	7	9
5	68	68	60	66	0	65	51	39	38
6	67	62	41	67	11	0	38	24	17
7	59	64	49	66	25	38	0	25	20
8	63	67	55	69	37	52	51	0	26
9	67	68	60	67	38	59	56	50	0

- Weak stochastic transitivity (WST): $p(i > k) > 0.5$
- Medium stochastic transitivity (MST): $p(i > k) > min[p(i > j), p(j > k)]$
- Strong stochastic transitivity (SST): $p(i > k) > max[p(i > j), p(j > k)]$

The violations of these transitivities have different degrees of severity: violations of weak transitivity are severe, while violations of strong stochastic transitivities may be acceptable. The function *strans* in the *eba* package allows to count the number of violations of the weak, moderate and strong stochastic transitivity per subject, which are reported in Table 7.3. From these data, subjects 3 and 8 both present the highest number of violations of WST, which are critical for fitting a probabilistic model and show a strong inconsistency in their answers. In addition, both subjects exhibit a high number of other violation types. Consequently they will be discarded in the following.
A new fit of the BTL model to the remaining 17 subjects now gives reliable scale

[3] `http://cran.r-project.org/web/packages/eba/`

Table 7.3. Transitivity violations per subject.

subject no.	violations of WST	violations of MST	violations of SST
1	4	20	23
2	2	14	16
3	7	25	33
4	0	3	6
5	3	19	23
6	1	6	24
7	4	32	33
8	7	31	38
9	4	20	24
10	4	23	26
11	1	9	20
12	0	4	4
13	2	20	29
14	6	14	27
15	3	25	26
16	4	30	36
17	0	1	9
18	1	14	19
19	0	7	19

values ($\chi^2 = 37.17$, $p = 0.1152$) which are displayed in Figure 7.6 as a function of the physical distances traveled by the virtual ball in the nine stimuli. The high value of the stimulus no. 5 attracts attention in the first place. The estimated distance for that stimulus is significantly higher than that for the stimulus no. 6, although the corresponding physical distances are very close. A possible explanation for this difference might be that in this case people base their judgment on the duration of the signal, that is 1.7 s for stimulus no. 6, and 2 s for stimulus no. 5. Figures 7.7 and 7.8 allows to have a clearer view on the effect of the duration of the signal and the initial velocity of the ball respectively, by joining the average scale values with a constant parameter value. These figures do not allow to quantify the influence of each parameter, however some interesting observations can be made. On the one hand, from Figure 7.7, the three lines are relatively straight and do not cross each other. For example, stimulus no. 3 is estimated at a higher scale than stimulus no. 4, and stimulus no. 5 at a higher scale than stimuli no. 6 and 7. This means that people tend to use duration to estimate stimuli of constant duration at similar distances. This effect is especially strong for the longest duration where the line joining stimuli no. 5, 8 and 9 is almost horizontal. On the other hand, the initial velocity does not seem to have as much influence on the subjects answers by looking at Figure 7.8.

Coming back to Figure 7.7, the participants evaluated the three stimuli in the correct order for the middle duration (1.7 s). If this effect is proven to be significant, it means that people were able to use distance cues provided by the tubular environment, since the influence of the initial velocity was shown to be limited. For the shortest duration (1.4 s), the order of the three stimuli is reversed. No obvious

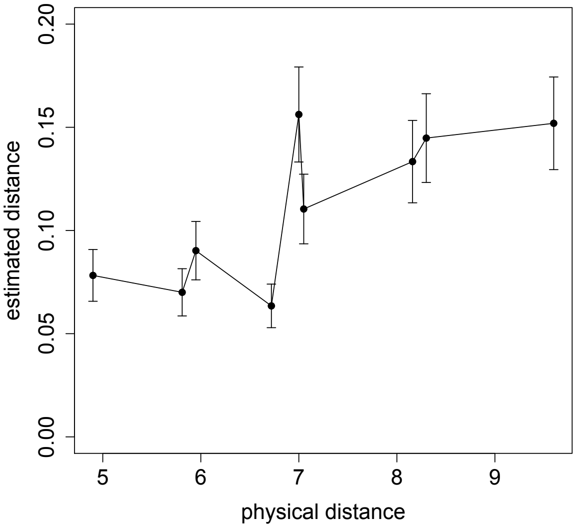


Fig. 7.6. *Ratio scale of distance for the 9 physical distances, derived from pairwise comparison judgments using the BTL model. Error bars show 95%-confidence intervals.*

explanation was found for this strange effect, perhaps the stimulus duration was too short so that subjects could not gather enough information about dynamic cues.

7.2.4 Discussion

A strong influence of stimulus duration was found on the estimation of the distance traveled by the ball inside the inclined tube, especially for stimuli with the longest duration. This effect suggests that the stimulus duration prevails upon the depth cues provided by the spatial rendering model in estimating the distance traveled by the rolling ball. Further analysis of the participants answers are required to correctly evaluate the phenomenon. In particular, the multidimensional scaling technique (MDS) can cope with preference scaling made by subjects on the basis of several acoustic parameters. It maps the stimuli into a multidimensional space, such that the Euclidean distances between the rating scales match the perceived dissimilarities as closely as possible. Then, a preference model with spline transformations can elicit the stimulus physical parameters that are strongly correlated to the perceptual dimensions recovered by the MDS algorithm [90].
Another issue to be addressed is the distance range considered in the listening test. Due to software restrictions, it was not possible to simulate a longer tube, which would have allowed to propose a larger distance range of stimuli.

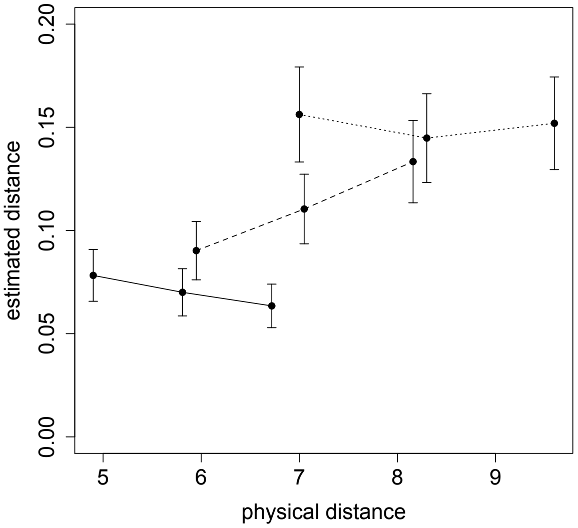


Fig. 7.7. *The ratio scales for stimuli of constant duration are joined by a line. Solid line: 1.4 s, dashed line: 1.7 s, dotted line: 2 s. Error bars show 95%-confidence intervals.*

7.3 On the use of motion in depth in HCI

A conceivable application is the development of an auditory progress bar using depth to represent the progress of the task. In graphical user interface, this tool acts as a notification mechanism on the progress of a task, such as a file transfer or a download. Progress bars may however be disruptive while performing other tasks, and need to be looked at to know the progress of the operation. Some authors have already proposed the design of auditory progress bars, using the sensitivity of the auditory sense to temporal characteristics of sources to convey information on an application's state. While earcons may provide information about the amount of download remaining and its rate [16], spatialization is used in [96] to map the amount of progress onto the position of a sound in space around the listener. In that case, the angular speed also indicates the rate. A user study showed in particular that the spatialized audio progress bar increased performance in a visually demanding foreground task, and that participants preferred the audio bar to the conventional visual one.
In our proposed approach, the action of transferring a file uses the metaphor of throwing a ball (or any other object). Like the audio progress bar described in [96], spatialization along the depth dimension can also provide information about the amount of progress by the distance between the object and the user, and the transfer rate by the velocity. In addition, it can give direct information about the direction of the progress in case of file transfer between the user and an external server. For example, download would be simulated by an approaching sound source, and upload by a receding sound source. In particular, it seems appropriate

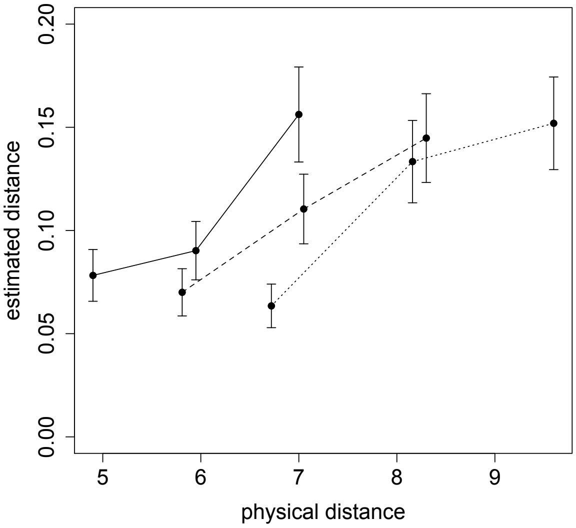


Fig. 7.8. *The ratio scales for stimuli with a constant initial ball velocity are joined by a line. Solid line: 7 m/s, dashed line: 8.3 m/s, dotted line: 9.6 m/s. Error bars show 95%-confidence intervals.*

to use a sound that becomes more intrusive (by approaching the user) to notify of the near completion of the download, such that the new file can be open within a short time.

8

General conclusion

8.1 Synthesis of results and author's contribution

The research presented in this thesis was aimed at contributing to scientific knowledge on depth perception, rendering and applications in interactive auditory interfaces. In this context, the following results have been obtained:

- Two experiments conducted on audio-visual depth perception have outlined the difficulty to create a unified multimodal percept of an object under laboratory conditions. The type of stimuli used in the experiment are therefore to be reconsidered. Nevertheless, results agree with previous studies: vision remains the predominant modality to perceive spatial attributes of objects. In a more natural context leading to multisensory integration, we believe that congruent auditory cues may enhance the perception of perspective.
  The author prepared the stimuli, designed and conducted the listening tests, and performed the statistical analysis of the results. The author has written 90% of Paper A.

- It has been demonstrated that virtual acoustics allows to shape the relationship between physical and perceived depths. In particular, listening tests have revealed the ability of a trapezoidal Digital Waveguide Mesh to linearize depth estimates in a specific distance range, which are usually compressed in natural environments.
  The author has filtered the original sound file through the DWM (filter design performed by Federico Fontana), designed and conducted the listening tests, and performed the statistical analysis of the results. The author has written 90% of Paper G.

- A new procedure has been presented to enable a direct comparison of multiple sound sources rendered at different distances, contributing to the area of human distance perception. It is based on the MUSHRA test, which was originally developed for evaluating the quality of various audio coding systems. A comparison with the direct magnitude estimation method, commonly used for distance perception, shows that it reduces the response variability.

- The bidimensional rectangular DWM exhibits a property similar to that of a real tubular environment, namely an exaggeration of the reverberation, which allows to enhance the perception of depth. For real-time simulations of a tubular environment, the 2D DWM has been implemented as two Max/MSP externals: one for spatializing multiple static sound sources at different distances, and one for simulating a moving sound source in depth.
  The author has adapted the source code of the DWM simulation (originally written by Federico Fontana and Stefania Serafin) to comply with flext, and has created the externals for PD and Max/MSP (as well as help patches in Max/MSP).

- Real-time simulations of the 2D mesh allow to map the user input onto the parameters of the model, creating an interactive system. To illustrate the potential of auditory perspective in human computer interfaces and contribute to the development of auditory interfaces, two concrete realizations were developed.

  - First, a rectangular ribbon featuring a touch position sensor was combined with the aforementioned spatialization model to build an audio-tactile interface for sound navigation. A preliminary user study on target selection has shown that people were able to perform the task, and found the interface intuitive. To search among many files, a knob enabled to move forward and backward the audio window resulting from the virtual tubular environment. Results from two different implementations of the knob showed different searching strategies.
    The author has developed the prototype (the sensor interface made of a gamepad was already available), prepared the stimuli, designed the user study, and performed the statistical analysis of the results. A student conducted the user study. The author has written 90% of Papers C and F.

  - A second illustration of interactive auditory perspective dealt with the rendering of an approaching/receding sound source. This collaborative work has led to an audio game, “DepThrow”, in which a virtual ball is rolling back and forth in a virtual inclined tube. Using physics-based modeling for the sounding object and depth rendering, this ecological interface offers various directions of research in dynamic depth perception. The resulting prototype was used to simulate approaching and receding sounds of a virtual ball with controlled tilt angle and initial velocity for the purpose of conducting a listening test on dynamic depth perception. Based on pairwise comparisons of rolling balls reaching various distances inside the virtual tube, results of the preliminary analysis predict a strong effect of the duration of the stimuli on the perceived distance.
    The author implemented the real time rendering of depth cues, prepared the

stimuli for the listening test, designed and analyzed the listening test (which was conducted by a student). The author has written 50% of Paper B.

8.2 Implications and issues

The perception of egocentric distance, i.e. the ability to quantify the distance of a sound source, is known to be weak. The thesis has shown that auditory interfaces can use distance cues in another way, which is the ability to discriminate between several distances, either to distinguish several static sources (sound source discrimination in depth) or follow the dynamic depth of a single source (dynamic depth perception). Up to now this field has received very little attention, but we envision that this is where auditory perspective can be most effectively exploited in interfaces. In addition, the human ability to improve distance perception with practice in reverberant environments demonstrated by Shinn-Cunningham [85] gives even more support to the use of auditory perspective based on the reverberation cue.

Far from recreating a real life environment to display distance, the objective of the work was to propose another view of the possibilities offered by virtual acoustics, such as:

- to shape the psychophysical function between physical distance in the virtual environment and estimated distance.
- to represent auditory depth with a physics-based model.

In order to quantify the influence of the direct-to-reverberant energy ratio cue on depth perception in the DWM, a proper computation of this quantity should be investigated. As explained in Section 4.2.2, the dispersion error does not allow an easy separation of the direct signal and the first reflection. Besides, it would be worth evaluating the influence of this dispersion error on the perception of depth in the DWM.

Although the 2D DWM allows to render convincing depth cues, it resulted from some informal listening that a 3D model tends to give rise to a more natural and pleasant listening environment. This definitely deserves more investigations because it may affect the usability of the interface and the aesthetics of interaction. Indeed one should not forget that aesthetics results both from the sonification and the means of interaction, therefore including spatialization.

Externalization was mentioned several times in the thesis as an issue related to headphone listening, and may as well affect the "pleasantness" of the interface. It was not directly studied in this work, and very few test subjects reported to hear the sound source as coming from their head in the set of experiments on depth perception. Investigations should therefore be conducted on the ability of reverberation to externalize sound sources in comparison to the common use of Head-Related Transfer Functions for sound spatialization.

Finally, the proposed depth rendering model for sound navigation needs proper formal evaluation by comparing its effectiveness with other existing methods (in audio or in other modalities) in facilitating task performance and usability.

A

List of Articles

- Paper A
  D. Devallez, D. Rocchesso, and F. Fontana. An experimental evaluation of the influence of auditory cues on perceived visual orders in depth. In *Proc. of the 13th International Conference on Auditory Display, Montreal, Canada*, pages 312–318, June 2007.

- Paper B
  S. Papetti, D. Devallez, and F. Fontana. Depthrow: a physics-based audio game. In *Proc. of the 14th International Conference on Auditory Display, Paris, France*, June 2008.

- Paper C
  D. Devallez, D. Rocchesso, and F. Fontana. An audio-haptic interface concept based on depth information. In *Proc. of the third International Haptic and Auditory Interaction Design Workshop, Jyväskylä, Finland*, pages 102–110, September 2008.

- Paper D
  S. Delle Monache, D. Devallez, P. Polotti, and D. Rocchesso. Sviluppo di un'interfaccia audio-aptica basata sulla profonditá spaziale. In *Proc. of the XVII Colloquio Di Informatica Musicale, Venice, Italy*, pages 109–114, October 2008.

- Paper E
  S. Papetti, D. Devallez, and F. Fontana. Depthrow: uno strumento di indagine sulla percezione uditiva della distanza in forma di gioco audio. In *Proc. of the XVII Colloquio Di Informatica Musicale, Venice, Italy*, pages 139–144, October 2008.

- Paper F
  D. Devallez, F. Fontana, and D. Rocchesso. An audio-haptic interface based on auditory depth cues. In *Proc. of the Tenth International Conference on Multimodal Interfaces, Chania, Crete, Greece*, pages 209–216, October 2008.

- Paper G
  D. Devallez, F. Fontana, and D. Rocchesso. Linearizing auditory distance estimates by means of virtual acoustics. *Acta Acustica United with Acustica*, 94:813–824, 2008.

References

1. G.N. Marentakis ad S.A. Brewster. Effects of feedback, mobility and index of difficulty on deictic spatial audio target acquisition in the horizontal plane. In *Proc. of ACM CHI 2006 Conference on Human Factors in Computing Systems, Montreal, Canada*, April 2006.
2. B.L. Anderson. A theory of illusory lightness and transparency in monocular and binocular images: The role of contour junctions. *Perception*, 26:419–453, 1997.
3. J.A. Ballas and J.H. Howard. Interpreting the language of environmental sounds. *Environment and Behavior*, 19(1):91–114, 1987.
4. D.R. Begault. Perceptual effects of synthetic reverberation on three-dimensional audio systems. *J. Audio Eng. Soc.*, 40(11):895–904, November 1992.
5. D.R. Begault. Auditory and non-auditory factors that potentially influence virtual acoustic imagery. In *Proc. of the AES 16th International conference on Spatial Sound Reproduction, Rovaniemi, Finland*, April 1999.
6. E.A. Bier, M.C. Stone, K. Pier, W. Buxton, and T.D. DeRose. Toolglass and magic lenses: the see-through interface. In *Proc. of the 20th annual conference on Computer graphics and interactive techniques*, 1993.
7. F.A. Bilsen. Thresholds of perception of repetition pitch. conclusions concerning coloration in room acoustics and correlation in the hearing organ. *Acustica*, 19:27–32, 1967/68.
8. J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization.* MIT Press, 1997.
9. R.A. Bradley and M.E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39:342–345, 1952.
10. S.A. Brewster, J. Lumsden, M. Bell, M. Hall, and S. Tasker. Multimodal 'eyes-free' interaction techniques for wearable devices. In *Proc. of the SIGCHI conference on Human factors in computing systems, Fort Lauderdale, Florida, USA*, pages 463–480, April 2003.
11. A.W. Bronkhorst and T. Houtgast. Auditory distance perception in rooms. *Nature*, 397:517–520, 1999.
12. D.S. Brungart. Control of perceived distance in virtual audio displays. *Proc. of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 20(3):1101–1104, 1998.
13. D.S. Brungart. Auditory parallax effects in the HRTF for nearby sources. In *Proc. of the 1999 Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, USA*, pages 171–174, October 1999.

14. D.S. Brungart and B.D. Simpson. The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal. *J. Acoust. Soc. Am.*, 112(2):664–676, 2002.
15. P.D. Coleman. Failure to localize the source distance of an unfamiliar sound. *J. Acoust. Soc. Am.*, 34(3):345–346, March 1962.
16. M. Crease and S. Brewster. Making progress with sounds - the design and evaluation of an audio progress bar. In *Proc. of the 2003 International Conference on Auditory Display, Glasgow, UK*, pages 167–177, November 1998.
17. N.I. Durlach, A. Rigopulos, X.D. Pang, W.S. Woods, A. Kulkarni, H.S. Colburn, and E.M. Wenzel. On the externalization of auditory images. *Presence: Teleoperators and Virtual Environments*, 1(2):251–257, 1992.
18. A.J. Ecker and L.M. Heller. Auditory-visual interactions on the perception of a ball's path. *Perception*, 34:59–75, 2005.
19. M. Fernström and C. McNamara. After direct manipulation - direct sonification. *ACM Transactions on Applied Perception*, 2(4):495–499, 2005.
20. J.D. Foley, A. Van Dam, S.K. Feiner, and J.F. Hughes. *Computer Graphics Principles and Practice.* Addison-Wesley, 1990.
21. F. Fontana. *Physics-based models for the acoustic representation of space in virtual environments.* PhD thesis, University of Verona, 2003.
22. F. Fontana and D. Rocchesso. Auditory distance perception in the acoustic pipe. *ACM Transactions on Applied Perception*, 5(3):Article 16, 2008.
23. F. Fontana, D. Rocchesso, and L. Ottaviani. A structural approach to distance rendering in personal auditory displays. In *Proc. International Conference on Multimodal Interfaces (ICMI'02), Pittsburgh, PA*, October 2002.
24. F. Frassinetti. Enhancement of visual perception by crossmodal visuo-auditory interaction. *Experimental Brain Research*, 147(3):332–343, December 2002.
25. G. Furnas. Generalized fisheye views. In *Proc. of ACM SIGGHI'86 Conf. on Human Factors in Computing Systems*, 1986.
26. M.B. Gardner. Proximity image effect in sound localization. *J. Acoust. Soc. Am.*, 43:163, 1968.
27. M.B. Gardner. Distance estimation of 0° or apparent 0°-oriented speech signals in anechoic space. *J. Acoust. Soc. Am.*, 47(1):47–53, 1969.
28. W.W. Gaver. *Everyday Listening and Auditory Icons.* PhD thesis, University of California, San Diego, 1988.
29. W.W. Gaver. The SonicFinder: An interface that uses auditory icons. *Human-Computer Interaction*, 4(1):67–94, 1989.
30. G.A. Gescheider. *Psychophysics: method, theory, and application.* Lawrence Erlbaum Associates, 1985.
31. W. Gogel. Convergence as a cue to absolute distance. *J. Psychol.*, 52:287–301, 1961.
32. K.W. Grant and P.-F. Seitz. The use of visible speech cues for improving auditory detection of spoken sentences. *J. Acoust. Soc. Am.*, 108(3):1197–1208, 2000.
33. R. Guski. Acoustic tau: An easy analogue to visual tau? *Ecological Psychology*, 4(3):189–197, 1992.
34. W.D. Hairston, P.J. Laurienti, G. Mishra, J.H. Burdette, and M.T. Wallace. Multisensory enhancement of localization under conditions of induced myopia. *Experimental Brain Research*, 152:404–408, 2003.
35. S. Handel. *Perceptual Coherence.* Oxford University Press, 2006.
36. W.M. Hartmann and A. Writtenberg. On the externalization of sound images. *J. Acoust. Soc. Am.*, 99(56):3678–3688, June 1996.
37. M. M.J. Houben, A. Kohlrausch, and D. J. Hermes. Perception of the size and speed of rolling balls by sound. *Speech Communication*, 43(4):331–345, September 2004.

38. T. Houtgast and S. Aoki. Stimulus-onset dominance in the perception of binaural information. *Hearing Research*, 72:29–36, 1994.
39. I.P. Howard and W.B. Templeton. *Human Spatial Orientation.* Wiley, London, 1966.
40. J. Huopaniemi, L. Savioja, and M. Karjalainen. Modeling of reflections and air absorption in acoustical spaces – a digital filter design approach. In *Applications of Signal Processing to Audio and Acoustics. 1997 IEEE ASSP Workshop on*, October 1997.
41. U. Ingard. A review of the influence of meteorological conditions on sound propagation. *J. Acoust. Soc. Am.*, 25(3):405–411, 1953.
42. A.R. Jensenius, R. Koehly, and M.M. Wanderley. Building low-cost music controllers. *LNCS*, 3902:123–129, 2006.
43. W. Jesteadt, C.C. Wier, and D.M. Green. Intensity discrimination as a function of frequency and sensation level. *J. Acoust. Soc. Am.*, 61(1):169–177, 1977.
44. B. Kapralos, D. Zikovitz, and S. Khattak. Auditory motion perception threshold. In *Proc. of the IEEE International Workshop on Haptic Audio Visual Environments and their Applications, Ottawa, Canada*, pages 1–4, October 2007.
45. A. Kelloniemi. Frequency-dependent boundary condition for the 3-d digital waveguide mesh. In *Proc. of the 9^{th} Int. Conference on Digital Audio Effects (DAFx-06), Montreal, Canada*, September 2006.
46. A. Kelloniemi. *Room Acoustics Modeling with the Digital Waveguide Mesh Boundary Structures and Approximation Methods.* PhD thesis, Helsinky University of Technology, 2006.
47. L.E. Kinsler, A.R. Frey, A.B. Coppens, and J.V. Sanders. *Fundamentals of Acoustics.* John Wiley & Sons, 2000.
48. K. Kowalczyk and M. van Walstijn. Virtual room acoustics using finite difference methods. How to model and analyse frequency-dependent boundaries? In *Proc. of the Third International Symposium on Communications, Control, & Signal Processing (ISCCSP 2008), Malta*, pages 1504–1509, March 2008.
49. H. Kuttruff. *Room Acoustics.* Spon Press, 2000.
50. E. Larsen, N. Lyer, C.R. Lansing, and A.S. Feng. On the minimum audible difference in direct-to-reverberant energy ratio. *J. Acoust. Soc. Am.*, 124(1):450–461, July 2008.
51. 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. *Exp. Brain Res.*, 158:405–414, 2004.
52. T. Lokki, M. Gröhn, L. Savioja, and T. Takala. A case study of auditory navigation in virtual acoustic environments. In *Proc. of the International Conference on Auditory Display, Atlanta, Georgia, USA*, pages 145–150, April 2000.
53. J.M. Loomis, R.L. Klatzky, J.W. Philbeck, and R.G. Golledge. Assessing auditory distance perception using perceptually directed action. *Perception & Psychophysics*, 60(6):966–980, 1998.
54. R.D. Luce. *Individual choice behavior: A theoretical analysis.* Wiley, 1959.
55. L.F. Ludwig, N. Pincever, and M. Cohen. Extending the notion of a window system to audio. *Computer*, 23(8):66–72, August 1990.
56. R.A. Lufti and W. Wang. Correlational analysis of acoustic cues for the discrimination of auditory motion. *J. Acoust. Soc. Am.*, 106(2):919–928, August 1999.
57. M.K. McBeath and J.G. Neuhoff. The doppler effect is not what you think it is: Dramatic pitch change due to dynamic intensity change. *Psychonomic Bulletin & Review*, 9(2):306–313, 2002.
58. H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1989.

59. D.H. Mershon and E. King. Intensity and reverberation as factors in the auditory perception of egocentric distance. *Percept. Psychophys.*, 18:409–415, 1975.
60. D.H. Mershon and J.W. Philbeck. Auditory perceived distance of familiar speech sounds. In *Proc. of the Psychonomic Society 32nd Annual Meeting, San Francisco, CA*, November 1991.
61. J.C. Middlebrooks and D.M. Green. Sound localization by human listeners. *Ann Rev Psych*, 42:135–159, 1991.
62. H. Møller, M. Sørensen, D. Hammershøi, and C.B. Jensen. Head-related transfer functions of human subjects. *J. Audio Eng. Soc.*, 43:300–321, 1995.
63. B.C.J. Moore, B.R. Glasberg, C.J. Plack, and A.K. Biswas. The shape of the ear's temporal window. *J. Acoust. Soc. Am.*, 83(3):1102–1116, 1988.
64. S. Morein-Zamir, S. Soto-Faraco, and A. Kingstone. Auditory capture of vision: examining temporal ventriloquism. *Cognitive Brain Research*, 17:154–163, 2003.
65. D. T. Murphy and M. Beeson. The kw-boundary hybrid digital waveguide mesh for room acoustics applications. *IEEE transactions on audio, speech and language processing*, 15(2):552–564, February 2007.
66. J.G. Neuhoff. Perceptual bias for rising tones. *Nature*, 395:123, September 1998.
67. J.G. Neuhoff. An adaptive bias in the perception of looming auditory motion. *Ecological Psychology*, 13(2):87–110, 2001.
68. S.H. Nielsen. *Distance Perception in Hearing.* PhD thesis, Aalborg University, Denmark, 1991.
69. I.J. Pitt and A.D.N. Edwards. Pointing in an auditory interface for blind users. In *Proc. of the 1995 IEEE International Conference on Systems, Man and Cybernetics*, pages 280–285, 1995.
70. M. Rath. An expressive real-time sound model of rolling. In *Proc. of the 6th Int. Conference on Digital Audio Effects (DAx-03), London, UK*, 2003.
71. M. Rath and D. Rocchesso. Continuous sonic feedback from a rolling ball. *IEEE MultiMedia*, 12(2):60–69, 2005.
72. L.D. Rosenblum, C. Carello, and R.E. Pastore. Relative effectiveness of three stimulus variables for locating a moving sound source. *Perception*, 16:175–186, 1987.
73. L.D. Rosenblum, A.P. Wuestefled, and K.L. Anderson. Auditory reachability: An affordance approach to the perception of sound source distance. *Ecological Psychology*, 8(1):1–24, 1996.
74. M.A. Sasse and H. Knoche. Quality in context - an ecological approach to assessing QoS for mobile TV. In *Proc. of the 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems, Berlin, Germany*, pages 11–20, Sept. 2006.
75. L. Savioja. *Modeling Techniques for Virtual Acoustics.* PhD thesis, Helsinky University of Technology, 1999.
76. L. Savioja, J. Backman, A. Järvinen, and T. Takala. Waveguide mesh method for low-frequency simulation of room acoustics. In *Proc. of the 15th International Conference on Acoustics, Trondheim, Norway*, pages 637–640, June 1995.
77. C. Schmandt. Audio hallway: A virtual acoustic environment for browsing. In *Proc. of the 11th annual ACM symposium on User interface software and technology, San Francisco, California, United States*, April 1998.
78. J.-L. Schwartz, F. Berthommier, and C. Savariaux. Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*, 93:B69–B78, 2004.
79. R. Sekuler, A.B. Sekuler, and R. Lau. Sound alters visual motion perception. *Nature*, 385:308, 1997.
80. L. Shams, Y. Kamitani, and S. Shimojo. Visual illusion induced by sound. *Cognitive Brain Research*, 14:147–152, 2002.

81. B. K. Shaw, R. S. McGowan, and M. T. Turvey. An acoustic variable specifying time-to-contact. *Ecological Psychology*, 3(3):253–261, 1991.
82. B.R. Shelton and C.L. Searle. The influence of vision on the absolute identification of sound-source position. *Perception and Psychophysics*, 28(6):589–596, 1980.
83. S. Shimojo and L. Shams. Sensory modalities are not separate modalities: plasticity and interactions. *Current Opinion in Neurobiology*, 11:505–509, 2001.
84. B.G. Shinn-Cunningham. Distance cues for virtual auditory space. In *Invited paper, Special Session on Virtual Auditory Space, Proceedings of the first IEEE Pacific-Rim Conference on Multimedia, Sydney, Australia*, 2000.
85. B.G. Shinn-Cunningham. Learning reverberation: Considerations for spatial auditory displays. In *Proc. of the 2000 International Conference on Auditory Displays, Atlanta, GA*, 2000.
86. W.E. Simpson and L.D. Stanton. Head movement does not facilitate perception of the distance of a source of sound. *American Journal of Psychology*, 86(1):151–159, March 1973.
87. J.M. Speigle and J.M. Loomis. Auditory distance perception by translating observers. In *Proc. of the IEEE Symposium on Research Frontiers in Virtual Reality, San Jose, CA, USA*, October 1993.
88. C. Spence. Auditory multisensory integration. *Acoust. Sci. & Tech.*, 28(2):61–70, 2007.
89. C. Spence and M. Zampini. Auditory contributions to multisensory product perception. *Acta Acustica United with Acustica*, 92:1009–1025, 2006.
90. P. Susini, S. McAdams, and S. Winsberg. A multidimensional technique for sound quality assessment. *Acta Acustica United with Acustica*, 85(5):650–656, 1999.
91. ITU (International Telecommunication Union). Recommendation BS.1534-1: Method for the subjective assessment of intermediate quality levels of coding systems (MUSHRA), 2003. http://www.itu.int/rec/recommendation.asp?type=folders&parent=R-REC-BS.1534.
92. M. van Walstijn and K. Kowalczyk. On the numerical solution of the 2D wave equation with compact FDTD schemes. In *Proc. of the* 11^{th} *Int. Conference on Digital Audio Effects (DAFx-08), Espoo, Finland*, pages 205–212, September 2008.
93. E. Vincent. MUSHRAM: a Matlab interface for MUSHRA listening tests, 2005. http://www.elec.qmul.ac.uk/people/emmanuelv/mushram/.
94. G. von Bekesy. The moon illusion and similar auditory phenomena. *American Journal of Psychology*, 62(478):540–552, October 1949.
95. A. Walker, S. Brewster, and S.A. McGookin. Diary in the sky: A spatial audio display for a mobile calendar. In *Proc. Interaction Homme-Machine - Human Computer Interaction (IHM-HCI), Lille, France*, September 2001.
96. V.A. Walker and S.A. Spatial audio in small screen device displays. *Personal Technologies*, 4(2):144–154, 2000.
97. C. Ware. *Information Visualization: Perception for Design.* Morgan Kaufmann Publishers, 2000.
98. F. Wickelmaier and C. Schmid. A Matlab function to estimate choice model parameters from paired-comparison data. *Behavior Research Methods, Instruments, & Computers*, 36(1):29–40, 2004.
99. F. Wightman and D. Kistler. Measurement and validation of human HRTFs for use in hearing research. *Acta Acustica United with Acustica*, 91:429–439, 2005.
100. H.-Y. Yao and V. Hayward. An experiment on length perception with a virtual rolling stone. In *Proc. of the Eurohaptics International Conference, Paris, France*, 2006.

101. P. Zahorik. Scaling perceived distance of virtual sound sources. *J. Acoust. Soc. Am.*, 101(5):3105–3106, May 1997.
102. P. Zahorik. Estimating sound source distance with and without vision. *Optometry and Vision Sciences*, 78:270–275, 2001.
103. P. Zahorik. Assessing auditory distance perception using virtual acoustics. *J. Acoust. Soc. Am.*, 111(4):1832–1846, 2002.
104. P. Zahorik. Auditory display of sound source distance. In *Proc. of the 2002 International Conference on Auditory Display, Kyoto, Japan*, July 2002.
105. P. Zahorik. Direct-to-reverberant energy ratio sensitivity. *J. Acoust. Soc. Am.*, 112(5):2110–2117, Nov. 2002.
106. P. Zahorik. Challenges in the auditory display of distance information. In *Proc. of the 19th International Congress of Acoustics, Madrid, Spain*, September 2007.
107. P. Zahorik, D.S. Brungart, and A.W. Bronkhorst. Auditory distance perception in humans: A summary of past and present research. *Acta Acustica united with Acustica*, 91:409–420, 2005.
108. S. Zhai, W. Buxton, and P. Milgram. The partial-occlusion effect: Utilizing semi-transparency in 3D human-computer interaction. *ACM Transactions on Computer-Human Interaction*, 3(3):254–284, 1996.
109. S. Zhao, P. Dragicevic, M. Chignell, R. Balakrishnan, and P. Baudisch. earPod: Eyes-free menu selection using touch input and reactive audio feedback. In *Proc. of ACM CHI 2007 Conference on Human Factors in Computing Systems, San Jose, CA, USA*, pages 1395–1404, April-May 2007.

Acknowledgments

First of all I owe much gratitude to my supervisor, Davide Rocchesso, for accepting me as his PhD student through the financial support of the European project *CLOSED* and introducing me into the field of human computer interaction. I also thank him for his inspiring ideas. A warm thank to Federico Fontana who has always been very supportive, enthusiastic about my work and has put trust in me. Besides, this thesis would not have been possible without his inventive thoughts.

For additional support and inspiration on the everyday basis, I am very much indebted to my office colleagues, who were always there to patiently answer my - stupid - questions and endure my mood swings. In details, I thank Antonio De Sena for the technical, administrative and linguistic support. I am very grateful to Stefano Papetti *(SteP)* for his help in programming with flext and using Max/MSP, however I must say the real expert in Max/MSP patches is undoubtly Stefano Delle Monache *(SteM)*, who helped me save a lot of time. Also thanks to SteP and SteM for combining our different backgrounds in joint projects that I truly enjoyed. I also kindly thank Carlo Drioli, especially for making me discover structures in Matlab (they changed my life!). Last but not least, many thanks to Pietro Polotti, the wise man of the group, for his kindness and his precious comments.

Of course I thank all the students and employees from the Department of Computer Science at the university of Verona, that kindly volunteered as test subjects in my experiments, for their indispensable contributions to this work. A special thank goes to Andrea Fasoli, who was keen on conducting a pair of experiments I had prepared beforehand.

Many thanks to all the people I met during conferences and who gave me useful feedback on my work. For the good atmosphere of endless meetings which have broadened my horizons, as well as nice social dinners, I have to give credit to the people of the CLOSED project. My one week study visit at the Department of Psychology in Stockholm was also a unique opportunity to meet psychologists and gain knowledge in the design of experiments. Thanks to Birgitta Berglund, Mats Nilsson and Anna Preis for our valuable discussions.

Finally, my thanks go to my parents and friends, for their unconditional support.

Sommario

Nell'apprezzare gli ambienti acustici, la percezione della distanza è cruciale tanto quanto la lateralizzazione. Ancorchè sia stato condotto del lavoro di ricerca sulla percezione della distanza, i moderni display uditivi non traggono ancora vantaggio da ciò al fine di fornire dell'informazione addizionale sulla disposizione nello spazio delle sorgenti acustiche in modo da arricchirsi, di conseguenza, di contenuto e qualità. Quando si progetta un display uditivo si deve tener conto dell'obiettivo dell'applicazione data e delle risorse disponibili al fine di scegliere l'approccio ottimale. In particolare, la resa della prospettiva acustica fornisce un ordinamento gerarchico delle sorgenti sonore e permette di focalizzare l'attenzione dell'utente sulla sorgente più vicina. A parte ciò, quando i dati visuali non sono più disponibili in quanto al di fuori del campo visivo o perchè l'utente è al buio, ovvero perchè è bene non adoperarli per ridurre il carico sull'attenzione visiva, il rendering uditivo deve convogliare tutta l'informazione spaziale inclusa la distanza. Questo lavoro di ricerca intende studiare la profondità acustica (sorgenti sonore dislocate di fronte all'ascoltatore) in termini di percezione, resa, e applicazioni all'interazione uomo-macchina.

Dapprima si propone una rassegna degli aspetti più importanti della percezione uditiva della distanza. Le indagini sulla percezione della distanza sono molto più avanzate nel campo della visione, in quanto hanno già trovato applicazioni nelle tecnologie di visualizzazione. Da ciò, sembrerebbe naturale fornire la stessa informazione nel dominio uditivo per aumentare il grado di realismo del display complessivo. La percezione della profondità di fatto può essere facilitata combinando indizi visuali e uditivi. Vengono riportati alcuni risultati di rilievo della letteratura sugli effetti dell'interazione audio-visiva, e illustrati due esperimenti sulla percezione della profondità audio-visiva. In particolare, è stata indagata l'influenza degli indizi uditivi sull'ordinamento visuo-spaziale percepito. I risultati mostrano che la manipolazione dell'intensità acustica non influisce sulla percezione dell'ordinamento lungo l'asse della profondità, un'evidenza dovuta probabilmente alla mancanza di integrazione multisensoriale. Inoltre, introducendo un ritardo tra i due stimoli audiovisuali, il secondo esperimento ha rivelato un effetto legato all'ordine temporale dei due stimoli visivi.

Tra le tecniche esistenti per la spazializzazione della sorgente acustica lungo la dimenzione della profondità esiste uno studio che ha proposto un modello di tubo virtuale, basato sull'esagerazione del riverbero all'interno di questo ambiente. La tecnica di progetto segue un approccio a modelli fisici e fa uso della Digital Waveguide Mesh (DWM) rettangolare 3D, la quale ha già evidenziato la sua capacità di simulare ambienti acustici complessi in larga scala. La DMW 3D è troppo affamata di risorse per la simulazione in tempo reale di ambienti 3D di dimensioni accettabili. Ancorchè una decimazione possa aiutare a ridurre il carico computazionale sulla CPU, un'alternativa più efficiente è quella di adoperare un modello 2D che, conseguentemente, simula una membrana. Sebbene suoni meno naturale delle simulazioni in 3D, lo spazio acustico bidimensionale risultante presenta proprietà simili specialmente rispetto alla resa della profondità.

Questo lavoro di ricerca dimostra anche che l'acustica virtuale permette di plasmare la percezione della distanza e, in particolare, di compensare la nota compressione delle stime soggettive di distanza. A tale scopo si è proposta una DWM bidimensionale trapezoidale come ambiente virtuale capace di fornire una relazione lineare tra distanza fisica e percepita. Sono stati poi condotti tre test d'ascolto per misurarne la linearità. Peraltro essi hanno dato vita a una nuova procedura di test che deriva dal test MUSHRA, adatta a condurre un confronto diretto di distanze multiple. Nello specifico essa riduce la variabilità della risposta a confronto della procedura di stima di grandezze dirette.

Le implementazioni in tempo reale della DWM 2D rettangolare sono state realizzate in forma di oggetti "external" per Max/MSP. Il primo external permette di rendere una o più sorgenti acustiche statiche dislocate a diverse distanze dall'ascoltatore, mentre il secondo external simula una sorgente sonora in movimento lungo la dimensione della profondità, una sorgente cioè in avvicinamento/allontanamento.

Come applicazione del primo external è stata proposta un'interfaccia audio-tattile. L'interfaccia tattile comprende un sensore lineare di posizione fatto di materiale conduttivo. La posizione del tocco sulla fascetta viene mappata sulla posizione d'ascolto di una membrana virtuale rettangolare modellata dalla DWM 2D, la quale fornisce indizi di profondità per quattro sorgenti egualmente spaziate. In aggiunta a ciò si dopera la manopola di un controller MIDI per variare la posizione della membrana lungo l'elenco dei suoni, permettendo così di passare in rassegna l'intero insieme di suoni muovendosi avanti e indietro lungo la finestra audio costituita dalla membrana virtuale. I soggetti coinvolti nella valutazione d'uso hanno avuto successo nel trovare tutti i file audio definiti come target, così come giudicato l'interfaccia intuitiva e gradevole. Inoltre è stata realizzata un'altra dimostrazione dell'interfaccia audio-tattile adoperando modelli fisici per il suono. Suoni di esperienza quotidiana derivanti da eventi quali "friggere", "bussare", "sgocciolare" sono stati adoperati in modo che sia la creazione del suono che la sua resa in profondità fossero il risultato di una sintesi per modelli fisici, ipotizzando che l'approccio di tipo ecologico potesse fornire un'interazione di tipo intuitivo.

Infine, “DepThrow” è un gioco audio basato sull’utilizzo della DWM 2D per rendere indizi di profondità di una sorgente acustica dinamica. Il gioco consiste nel lanciare una palla virtuale, modellata da un modello fisico di suoni di rotolamento, all’interno di un tubo virtuale inclinato e aperto alle estremità, modellato da una DWM 2D. L’obiettivo è fare rotolare la palla quanto più in là nel tubo senza farla cadere all’estremità lontana. Dimostrato come un gioco, questo prototipo è stato pensato anche come strumento per condurre indagini sulla percezione della distanza dinamica. I risultati preliminari di un test d’ascolto condotto sulla percezione della distanza variabile all’interno del tubo virtuale, hanno mostrato che la durata del rotolamento della palla influenza la stima della distanza raggiunta.