Application of Wave Field Synthesis in Videoconferencing

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof.dr.ir. J.T. Fokkema,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op

4 oktober 2004 om 13.00 uur

door

Werner Paulus Josephus DE BRUIJN

natuurkundig ingenieur
geboren te Tilburg.

Dit proefschrift is goedgekeurd door de promotor:
Prof. dr. ir. A.J. Berkhout

Samenstelling promotiecommissie:

Rector Magnificus	voorzitter
Prof. dr. ir. A.J. Berkhout	Technische Universiteit Delft, promotor
Prof. dr. ir. A. Gisolf	Technische Universiteit Delft
Prof. dr. A.G. Kohlrausch	Technische Universiteit Eindhoven
Prof. dr.-ing. K. Brandenburg	Technische Universität Ilmenau
Prof. dr. W. Woszczyk	McGill University
Dr. ir. M.M. Boone	Technische Universiteit Delft
Dr. R. Nicol	France Télécom R&D

ISBN 90-9018438-4

SUPPORT
The research presented in this thesis was financially supported by France Télécom R&D, Lannion, France.

Cover design: Carolien Postma.

Printed in The Netherlands by Febodruk BV, Enschede.

Contents

CHAPTER 1 *Introduction*

1.1 Teleconferencing

In today's globalized world, telecommunication has become an indispensable tool. With the rise of multinational companies with offices all over the world (globalization) and the increase of the number of cooperations between companies or institutions based in different locations, there has become a need for ways to have meetings without all parties involved actually having to travel to a single location. Instead, it would be preferred if a virtual meeting could be held with all parties remaining at their own local sites by using some technical means. This general concept of having a meeting with people, either individuals or groups, located at different locations through technical means, is what is commonly referred to as a *teleconference*. Essential in the definition of a teleconference is that in terms of the number of people involved, the communication involves more than just one-to-one communication between two individuals. This means that either people at more than two locations take part in the conference, or more than one person takes part at one location, or a combination of these two possibilities. Also, the communication has to be two-way, meaning that participants at both (or in the case of more than two locations: all) locations are able to receive the contributions of the participants at the other locations, while at the same time being able to send their own contribution to those remote participants.

As such, the general term ‘teleconference’ encompasses a broad range of different telecommunication schemes. An *audioconference* is a teleconference in which only audio information, i.e., the voices of the conference participants, is transmitted. An example is the common telephone conference. A *videoconference* combines audio with motion-video images of the participants. The concepts of both audio- and videoconferences can be extended by the addition of other communication tools which support the primary auditory and visual information, like, for instance, graphic boards, on which drawings can be made, or additional audio- or video streams. This could be referred to as a *multi-mediaconference*. And finally, thanks to the rise of the internet, in more recent years another concept of teleconferencing has come up which has gained much popularity, especially among young people: the *internet chat*.

The concept of teleconferencing is not new at all. Since the early days of radio, people have been using telecommunication means to have virtual meetings involving more than just one-to-one communication between two individuals and since decades people are able to have telephone conferences with people participating at multiple locations simultaneously. The first commercial example of a teleconferencing system that combined audio and video was AT&T’s *Picturephone*, introduced in 1970, basically a combination of a telephone with a small video image of the face of the person at the remote site, updated only every two seconds, which by the way never became very successful ([AT&T]). High quality two-way videoconferences between remote locations first became feasible, both technically and commercially, with the launch of the ATS-6 satellite by NASA in 1974, which had been developed specifically for this purpose ([Cowa84]).

1.1.1 Videoconferencing

Videoconferencing systems can be divided into two basic categories, based on the number of people that are able to participate at each location: systems that are intended for communication between *individual* users located at different sites and systems that are intended for communication between *groups* of people located at different sites.

For the first category, which might be labeled ‘one-individual-at-each-location’ videoconferencing systems, cheap and rather easy-to-use solutions are now available in the form of ‘desktop conferencing’, in which each individual conference participant sits behind his computer, which is equipped with a webcam to capture the face of the user and a microphone to pick up his voice. The signals are transmitted over the internet by IP to the remote participant’s com-

puter, where the image and voice are reproduced.

For the second category, which might accordingly be labeled ‘one-group-at-each-location’ or simply ‘group’ conferencing systems, a range of solutions exist, which vary in quality and cost. The simplest ones use television-sized screens and simple audio recording and reproduction means, with the signals being transmitted to the remote sites through a number of normal ISDN telephone lines. Such systems are commercially available and can also be rented from specialized companies. At the highest-quality end of the range, there are systems which consist of a ‘video wall’ on which the remote room and its occupants are visualized ‘life-sized’, as if the video screen were a transparent visual window through which the local and remote room were connected, combined with higher quality, possibly multi-channel, sound recording and reproduction facilities. Such a system is usually permanently installed in a dedicated room and its signals, because of the high bandwidth that is required, are transmitted through dedicated communication channels. It will be clear that these systems are very expensive and therefore only large multinational companies can afford to install such a system. However, a market has emerged for specialized companies that have set up their own network of this type of high-end videoconferencing facilities at various locations. These facilities can be rented on a per-hour basis, to accommodate the need of also smaller companies to occasionally have high quality group videoconferences.

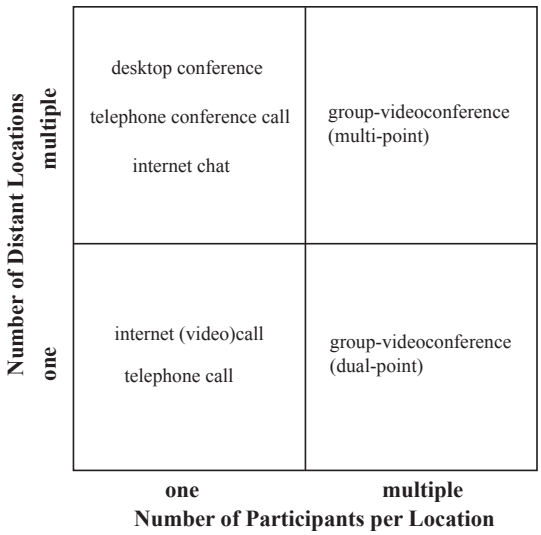


FIGURE 1.1. Classification of some teleconferencing systems in terms of the number of participants present at each location and the number of distant locations connected to the local system.

Figure 1.1 shows a schematic classification of some of the teleconferencing systems that have been mentioned in the previous, in terms of the number of participants present at each location and the number of distant locations that are connected to the local system. (Note that, strictly speaking, the systems in the left-lower quadrant are not really teleconferencing systems, according to the definition.)

1.1.2 Telepresence

Preferably, the quality of the virtual meeting should be as high as possible, in order to maximize the efficiency of the meeting in terms of human communication. In the ideal case, all the participants of the meeting should have the feeling as if they are all located in the same room together and should in fact not be aware anymore that they are communicating 'through a system', since this will inevitably distract attention from the actual conversation to some extent. This scenario is strongly related to a concept called *telepresence*. Telepresence is often defined as the convincing illusion of actually being in another, remote or virtual, location ([Sher92], [Fisc91]). In the ideal case of a virtual meeting as described above, the local participants should have the illusion that they are in the same room as the remote participants, while those remote participants should have the same illusion. This situation can be realized by creating a system which virtually connects both rooms to form a new, virtual, room, in which all participants from both locations are 'telepresent' simultaneously (figure 1.2).

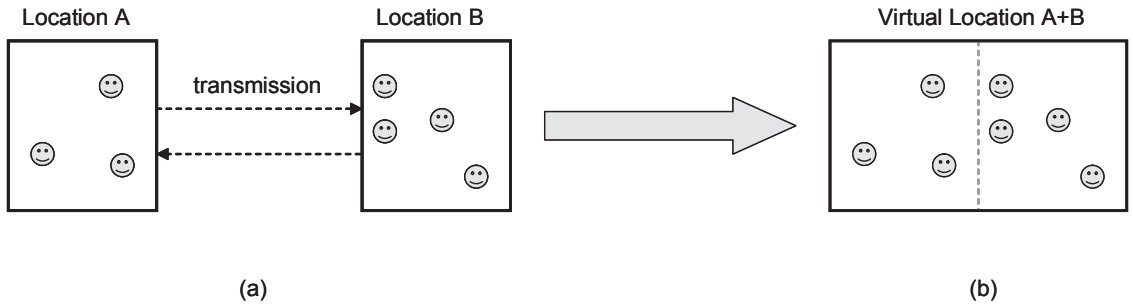


FIGURE 1.2. The concept of telepresence. (a) Separate locations A and B are connected to each other by telecommunication means in such a way that each location is able to make a virtual reproduction of the other one. (b) The result is that the participants of both locations have the impression that they are all present in a common virtual location A+B.

1.2 Spatialized Audio In Videoconferencing Systems

1.2.1 Why Spatialized Audio In Videoconferencing?

Since the first days of stereo, it has been recognized that spatialized recording and reproduction of sound has important advantages over non-spatialized recording and reproduction. If done well, it is able to provide the listener with an impression of a sound event that is much more natural, involving, and convincing, regarding both the perceived spatial locations of the sources and the impression of the space in which the sources were located. Also in audio-visual presentations, spatialized reproduction of sound can greatly enhance the overall impression of space. In the world of cinema, the use of ‘surround sound’ results in the fact that the audience feels much more involved in the movie, and the added value of spatial sound reproduction is also, at least partly, responsible for the huge success of home theatre systems, which provide ‘surround sound’ in the living room. In recent years, this importance of spatialized audio on the overall impression of space in audio-visual systems has been investigated and confirmed in several studies (see for example [Bech95]).

Acoustical Telepresence. In section 1.1.2 the concept of ‘telepresence’ was introduced. In the ideal teleconferencing system, a perfect illusion of telepresence would be achieved in all human senses. It will be clear however, that the main priority of a high-quality conferencing system will be that a convincing illusion of telepresence is achieved at least in the auditory and visual senses. In the context of a teleconferencing system, we introduce the term *acoustical telepresence* to describe the convincing *auditory* illusion that all participants from all conference locations are present in the same virtual location. As such, acoustical telepresence is a subcase of overall telepresence.

To achieve acoustical telepresence, accurate spatialized sound reproduction is essential. In particular, when the reproduction of the sound recorded at the remote site is such that the reproduced voice of a remote participant is perceived by the local participants as being located at the same position as where the corresponding participant is visually displayed on the screen, this will make the audio-visual communication much more natural and realistic. This is especially the case for audio-visual systems with a large projection screen, as it has been shown ([Komi89]) that the larger the screen is, the more important accurate spatialization of sound becomes in order to achieve a natural reproduction.

It should be realized that in a videoconferencing system it is not only important to have a natural audio reproduction for ‘esthetic’ reasons alone. Just as important is the fact that the presence of any artifact in the reproduction of the audio, the video or their combination will distract the participants’ attention from the most important thing: the conversation between the local and remote participants. Generally speaking, artifacts will reduce the overall ‘comfort level’ of the users of the system, which can eventually lead to fatigue and loss of concentration, resulting in a less effective virtual meeting in terms of human communication. Especially for videoconferencing systems which are intended to be used for executive meetings which may last for several hours, this should be avoided as much as possible. An accurate spatial reproduction of the voices of the conference participants, in such a way that they are localized by the participants at the other sites at the same positions as the corresponding images on the screen, is an important aspect of this, since it is known that a noticeable mismatch between the voices and the corresponding images results in a significant reduction of the perceived audio-visual quality of the system ([Chat99]).

In addition to its importance for achieving acoustical telepresence, there are additional advantages of using spatialized reproduction of the voices instead of non-spatialized reproduction that are specific to the application of life-sized videoconferencing and have to do with the ‘effectiveness of communication’ provided by the system. These are:

Improvement of Speech Intelligibility. One of the most important properties of a speech communication system is the intelligibility of the reproduced speech signals. It is well known that when speech signals are perceived as coming from different directions relative to each other or relative to interfering sound sources, this can greatly enhance the intelligibility of the speech signals (see for example: [Plom81], [Merk00]).

Facilitation of speaker identification. When in a videoconference several people are participating simultaneously at the remote site, reproducing their voices from the same spatial locations at which their images are displayed on the projection screen can make it easier for the local participants to instantly identify a specific speaker out of all the remote participants that they see on the screen. This is especially the case when a large projection screen is used and the individual remote participants are located relatively far apart.

1.2.2 Audio Recording- And Reproduction Techniques In Commercial Systems

When a market overview was done at the start of the project described in this thesis to see what was then the current state-of-the-art regarding the audio part of videoconferencing systems, it became very clear, even when we restrict ourselves only to the larger videoconferencing systems for group-to-group communication, that the audio recording and reproduction facilities of all systems were very simple, regarding sound spatialization. Most group communication systems, ranging from the simplest mobile, television-sized systems to the largest life-sized systems, provide nothing more than a single omnidirectional microphone for sound pick-up and a single loudspeaker for sound reproduction. In the more expensive systems there is usually also the option of connecting additional microphones, the signals of which are simply added together into a mono signal. In other words, most systems provide no sound spatialization at all (figure 1.3, left).

Looking for the state-of-the-art at the start of this project, this was found to be the ‘Varèse’ system developed at France Télécom R&D ([Emer98]). Although this was not really commercially available and still under development, it had already been installed in three different locations of France Télécom and was used for official meetings. The Varèse system is a group videoconferencing system with a videowall on which participants, which are seated behind a conference table facing the videowall, are displayed life-sized. Each participant has an individual directional microphone which is placed right in front of him/her, on top of the table, with a maximum total number of six. At the remote site, the sound is reproduced by five equidistant loudspeakers, located underneath the screen (figure 1.3, right). The microphone signals are distributed over these five loudspeakers by a simple input-output matrix in such a way that the lateral position from which a participant’s voice is reproduced (the position of the loudspeaker to which the participant’s voice signal is sent) corresponds as closely as possible to the position on the screen at which the same participant is displayed. Later on in this thesis, this reproduction method will be referred to as a *discrete loudspeakers* reproduction. While this reproduction method provides a stable localization of the sound sources for listeners at any position in the reproduction room, it does not provide any sensation of acoustic source distance and also the source resolution in the lateral direction of the reproduction is limited by the distance between the loudspeakers. At the recording side, the use of individual microphones at fixed positions means that the participants have no freedom to move around in the room.

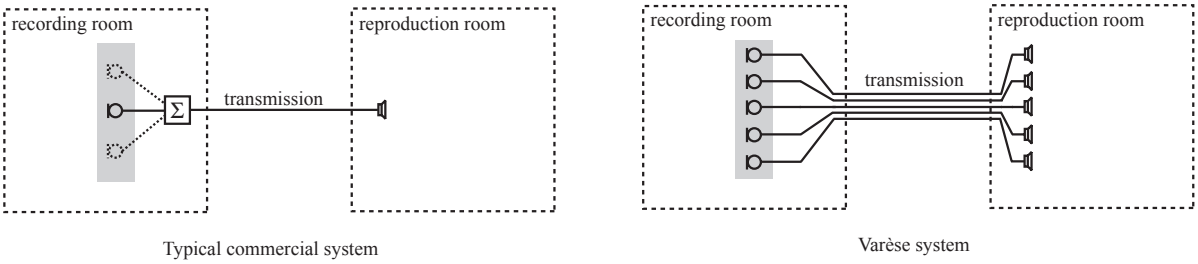


FIGURE 1.3. Audio recording- and reproduction system of a typical commercial videoconferencing system (left) and France Télécom's 'Varèse' system (right). In the commercial system, the dashed microphones are optional (or not present at all) and their signal is simply added to the main microphone signal. In the Varèse system, an equal number of five microphones and loudspeakers has been indicated for simplicity.

1.2.3 The TU Delft-France Télécom Project

From the previous subsection, it is clear that none of the commercial systems existing at the start of the project described in this thesis was able to provide the very high degree of acoustical telepresence, as defined in section 1.2.1, that is desired for the 'ideal' videoconferencing system.

This same notion triggered the research and development department of France Télécom (FTR&D) located in Lannion, France, to start investigations to improve sound reproduction for videoconferencing systems. Nicol ([Nico98], [Nico99]) did an extensive study in which she compared various techniques for spatial audio reproduction for their suitability for application in life-size videoconferencing. The techniques she compared included various stereophonic systems, binaural and transaural systems, ambisonics and holophony. Her comparative study, which was mainly done from a theoretical point of view, used as main criterion that the listening area of the reproduction, i.e.: the area in which users of the system perceive a spatial sound image which closely resembles the original acoustical scene, should be as large as possible. This criterion was used, because for a high-end, life-size videoconferencing system, the following requirements were considered to be essential:

- Multiple participants located at different positions in the room should be able to perceive an accurate spatial sound image.
- Participants should be able to move around in the room, rather than being restricted to fixed positions, while still perceiving an accurate spatial sound image.

The main conclusion of her study was that of all the techniques that were investigated, only ‘holophony’ was able to meet the requirement of providing an extensive listening area. Holophony ([Jess73]) is a general concept for sound reproduction, which aims at producing a perfect physical reconstruction of an original sound field[1], based on the Kirchhoff-Helmholtz representation theorem for the description of sound fields.
A practical implementation of the concept of holophony is *Wave Field Synthesis* (WFS), a multi-channel sound reproduction technique developed at TU Delft, which uses arrays consisting of a large number of small closely-spaced loudspeakers to make a reproduction of a desired sound field that is spatially correct over an extensive listening area ([Berk88]). This means that the localization of the reproduced sound sources is correct for all people in the reproduction area, not just at a single ‘sweet spot’ or a very limited ‘sweet area’, as is the case with conventional stereophonic techniques. For this reason, Wave Field Synthesis seemed to be ideally suited to achieve the aim of acoustical telepresence.
Based on these conclusions, a simple prototype of a WFS sound reproduction system was developed at FTR&D, the design of which was mainly based on the results of research carried out previously at TU Delft. First objective measurements on the sound fields reproduced by this prototype yielded promising results ([Nico99]), which resulted in an interest to continue the investigations on WFS with the intention to optimize the concept to the application of life-size videoconferencing.

Therefore, in 1998 FTR&D started a joint research project with the Laboratory of Acoustical Imaging and Sound Control of Delft University of Technology (TU Delft) where the technique of Wave Field Synthesis had been developed ([Berk88]). The general aim of this project, called ‘*Acoustic Wave Field Synthesis For Teleconferencing*’, was to develop the audio part of a dual-site, high-quality, life-sized group-videoconferencing system, based on WFS, in such a way that the auditory scene of the remote site is reproduced as natural as possible, so that participants at both sites have, at least as far as the sound reproduction is concerned, the feeling that they are all in the same room together. In other words, the aim is to achieve acoustical telepresence in a videoconferencing system that, in terms of the classification matrix of teleconferencing systems of figure 1.1, is of the category of ‘one distant location-multiple participants per location’ systems (lower-right quadrant of figure 1.1).

1. The term ‘holophony’ has been introduced in analogy with the term ‘holography’.

As has been discussed in section 1.2.1, an important requirement to achieve this is that the spatial aspects of the original remote sound field are accurately reproduced by the local system, ideally also including the acoustic sensation of depth, so that sound sources are not only reproduced as being located in the correct direction, but also at the correct distance. As explained above, this requirement was the motivation for using WFS for the sound reproduction. Additional requirements to achieve acoustical telepresence are, among others ([Emer98]):

- high-quality transducers (microphones and loudspeakers).
- large frequency bandwidth (>7 kHz for voice signals).
- transparent coding algorithms in transmission.
- avoidance or efficient suppression of echoes.

If all these requirements are met, this should result in a situation in which, acoustically, the two rooms are virtually connected to each other to form a single virtual room.
In the videoconferencing application, there are additional requirements for the combination of the audio and the video reproduction that have to be met to achieve *audio-visual telepresence*, the most important ones being that the perceived positions of the reproduced sound sources coincide with the perceived positions of the corresponding visual sources on the screen and that the audio is properly time-synchronized with the video. If these additional conditions are also met, then audio-visual telepresence can be achieved, with the video projection screen being perceived as a transparent virtual window to the remote room.
In the ultimate telepresence experience, users should not be aware at all that they are communicating through a system. This leads to the requirements discussed at the start of this subsection that users at the reproduction site should be able to be located anywhere in an extensive listening area and should be free to move around, while perceiving a natural spatial sound image at all times. Likewise, at the recording site, a perfect sense of telepresence is not really possible when conference participants are forced to use microphones that are located at fixed positions, for instance, on top of a conference table, whenever they want to speak. For this reason, a further aim of the project was to develop the audio system in such a way that conference participants at both sites are free to move around in the room, while the quality of the reproduction regarding sound quality and source localization should be preserved at all times at both sites. This means that the system to be developed is, what might be called, a *double-dynamic system*. Preferably, it should be possible for participants to approach the projection screen up to

a close distance, so that people located at different sites can really have a ‘tête-à-tête’ conversation. This puts special demands on especially the sound recording techniques to be used in the system.

In order to achieve a perfect sense of telepresence in a videoconferencing system, also *visual telepresence* would be required. This means that, ideally, the high-quality spatial audio reproduction system that was the aim of this project, should be combined with high-quality, life-sized, three-dimensional video images. Unfortunately, systems that are able to provide such images are not yet available. Of course, a variety of 3D video systems exists, for instance, systems in which all participants wear special glasses with electronic shutters, polarized glasses, or a helmet with an individual LCD display for each eye, in order to provide individual images to the left- and the right eye, so that a 3D image is perceived. In other systems, special displays with integrated lenses (‘lenticular displays’) provide individual images to both eyes without the need to wear glasses or a helmet. However, all these systems suffer from various serious disadvantages, most notably the fact that only a single or a few stereoscopic viewpoints are provided, so that the perspective is distorted when the viewer is located at any other position. Also, the overall quality and ‘level of user comfort’ are usually not sufficient to achieve a real experience of telepresence. Much research is being carried out in this field, also for the specific application of high-quality videoconferencing systems ([PANORA], [VIRTUE]), but they are still very much in the development phase. In Chapter 3, systems for 3D visualization will be discussed in some detail.

Since the focus of the project described in this thesis was on the development of the audio part of the videoconferencing system, not on the development of the video part, using three-dimensional video in the project was not considered to be an option and a decision was made at the start of the project to use conventional two-dimensional video projection instead. The approach that was chosen was to capture, at each site, a single two-dimensional visual image covering the entire room. This can be achieved by using either a single camera that covers the entire room, or by using multiple cameras that each cover a part of the room, after which the images of the individual cameras are combined into a single image of the whole room. This image is then projected life-sized on the projection screen at the remote site, so that the remote participants are presented with a life-sized image of the other site that is a two-dimensional projection of the actual three-dimensional room.

It was realized from the start of the project that this choice of combining spatialized audio

reproduction that includes a correct reproduction of depth with conventional two-dimensional video projection that does not contain depth, could have undesirable side-effects, caused by the fact that the visual perspective and the auditory perspective of the total audio-visual reproduction are not identical. Since not much was known yet about this subject, it was decided that the project should also include a study whether or not any such effects occur in practice.

This thesis describes the research that was carried out at TU Delft in the context of the project that has been described in this subsection. In the next subsection, we will look in some more detail at how the concept of Wave Field Synthesis can be applied for sound reproduction in a videoconferencing system.

1.2.4 Wave Field Synthesis In Videoconferencing

As mentioned in the project description in the previous subsection, Wave Field Synthesis is a sound reproduction technique which is able, by using arrays of closely spaced loudspeakers, to generate a sound field which is spatially identical to some desired sound field, which can be an existing sound field as well as a non-existing virtual sound field, within an extensive listening area. This makes it a very attractive candidate to implement the high-quality spatial audio reproduction that is required to achieve acoustical telepresence in a life-sized videoconferencing system and which, as was shown in section 1.2.2, existing systems are unable to provide. We will now look at how WFS could be implemented in a life-sized videoconferencing system from a conceptual point of view. The theory of WFS will be discussed in Chapter 2 of this thesis.

Imagine the situation in which we want to establish audio-visual telepresence between room *A* and room *B*. Let us first look at the one-way situation of capturing the audio-visual scene in room *A* and reproducing it in room *B* (figure 1.4, left). At the top we see room *A*, in which three conference participants are present. A camera is located such that it is able to capture a picture of the whole area of interest of the room. We also assume that there is some, at this point unspecified, audio recording system, which captures the sound field in the room or the individual voices of the participants. The different methods that can be used for recording will be discussed below. The captured audio and video data is transmitted to room *B*. In room *B*, the video image is projected on the projection screen (indicated by the thick black line). Since we are dealing with a 2D projection, there is a unique viewpoint for which the visual perspective

of the video projection is identical to the visual perspective of the original visual scene in room *A*, as seen from the position of the camera. The location of this viewpoint is determined by the combination of the location and properties of the camera in room *A* and the projection system in room *B* and is indicated in the picture by the small cross in the center of room *B*. An observer located in this viewpoint will see an image which, regarding perspective, is the same as if the part of room *A* that is captured by the camera would have been transferred completely to the location of room *B* and connected to it, indicated by the dashed virtual copy of room *A*, labeled *A'*. The projection screen acts as the virtual visual interface between room *B* and room *A'* and each object in room *A*, including the conference participants, is displayed on the screen in room *B* at the intersection of the line from the viewpoint to the position of the object in virtual room *A'* (indicated by the dashed lines) and the screen.

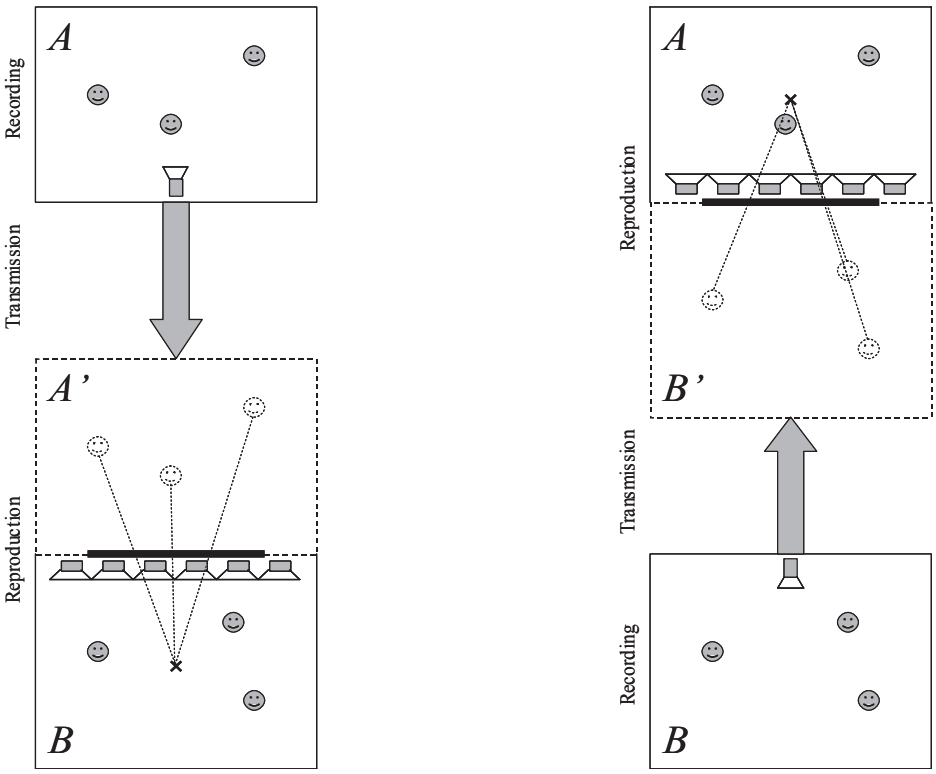


FIGURE 1.4. Concept of using WFS in videoconferencing. Left: An audio-visual recording of room *A* is transmitted to room *B*, where a virtual reconstruction of room *A*, room *A'*, is made, such that for participants in room *B*, it seems as if room *A* has been connected to their room. Right: the reverse situation: a virtual reconstruction of room *B*, room *B'*, is made in room *A*.

The sound signals that have been recorded at location *A* are processed by a Wave Field Synthesis processor, which calculates the driving signals for the loudspeaker array in room *B*, which is indicated in the picture as a linear array in front of the screen for reasons of visualization only. Ideally, it would be a planar array of closely-spaced loudspeakers covering the wall. The loudspeaker array is fed with the calculated WFS driving signals and generates a sound field which is identical to the sound field that would have been present in room *B* if room *A* would indeed have been connected to it as shown in the picture. All participants in room *B* will now perceive the voices of all the participants in room *A* as being located at the positions of the virtual participants in the virtual room *A'*.
In the right of figure 1.4 the complementary situation is shown, in which an audio-visual recording of room *B* is transmitted to room *A*, where a virtual reconstruction of room *B*, room *B'*, is made, such that the participants in room *A* have the impression that room *B* has been connected to their own room.
Effectively, we have now established the desired situation of audio-visual telepresence in a common virtual room *A+B* for all participants at both sites.

Regarding the recording of the sound sources for WFS, two different general approaches can be distinguished. Which of these two approaches is the most appropriate depends on the application, while in some cases it is convenient to use a combination of both. In section 2.4 several applications of WFS will be discussed, including the most appropriate recording approach for each.
The first recording approach is to place a microphone very close to each individual source ('spot microphones'), so that the direct sound of each source is recorded as a separate audio channel (figure 1.5, left). In the reproduction room, the individual audio channels are processed by an operator $\mathbf{W_1}$, which calculates the driving signals for all the individual loudspeakers of the loudspeaker array according to the WFS theory, with the spatial coordinates of the sources as parameters. These driving signals are fed to the loudspeaker array and a sound field is generated which corresponds to the sound field that would have been obtained if the recording room had been connected to the reproduction room, in the way that was explained earlier in this subsection. An observer in the reproduction room, regardless of his listening position, will now perceive virtual sound sources at the correct positions, as if the remote room has been connected to his own room. Since the source coordinates are required by operator $\mathbf{W_1}$ to calcu-

late the loudspeaker driving signals, this approach requires that the positions of the sound sources are known. If the sources should be allowed to move, this means that the position of each source has to be tracked in some way.

The second approach is to capture the sound field in the recording room as a whole instead of recording each source individually, by using an array of closely spaced microphones that spatially samples the sound field (figure 1.5, right). In the reproduction room, the signals of the microphone array are processed by an operator $\mathbf{W_2}$, which calculates the driving signals for the individual loudspeakers in the loudspeaker array. Then, when these driving signals are fed to the loudspeaker array, the original sound field is reconstructed in the reproduction room, again as if the two rooms have been connected. Note that in this approach the spatial information about the sources is present implicitly in the ensemble of the signals recorded by the microphone array. This means that the coordinates of the sound sources do not have to be known explicitly for operator $\mathbf{W_2}$ to be able to calculate the loudspeaker driving signals, only the positions of the microphones of the microphone array, which are fixed, have to be known.

In principle, both recording approaches can be used in the videoconferencing application, both with their own advantages and disadvantages. We will look at the different recording possibilities in some more detail in section 1.3.1.

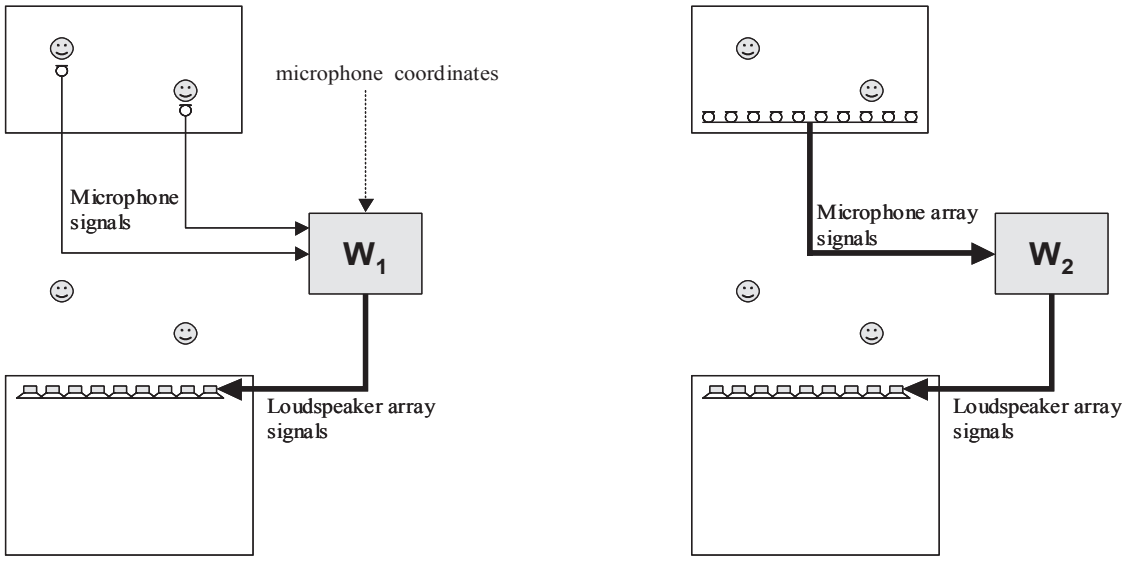


FIGURE 1.5. **The two general WFS recording approaches. Left: spot microphones. Right: microphone array.**

1.3 Audio Chain Of A Two-Way Videoconferencing System

The audio chain of one way of a two-way videoconferencing system basically consists of the following stages, from the local to the remote site: the *recording* stage, which deals with the recording of the voices of the conference participants at the local site, the *transmission* stage, which deals with sending the recorded audio signals from the local site to the remote site and the *reproduction* stage, which reproduces the voices of the local site's participants at the remote site.

In order to completely achieve the general project aim of 'acoustical telepresence' as described in section 1.2.3, many different issues in all of these stages of the two-way videoconferencing system should be investigated, ranging from investigating different techniques for recording the voices of participants, avoidance of echoes, efficient coding techniques needed to be able to transmit the recorded audio signals to the remote site within a given transmission bandwidth, to reproduction of the transmitted signals in the remote room. This range of topics was considered too broad to be investigated in-depth in a single-thesis'-worth of research, so it was decided, as stated in the objectives of this thesis which will be presented in section 1.4, that the main focus of this thesis would be on the reproduction stage of the audio chain and in particular on the requirements for the loudspeaker array that is used for WFS reproduction of the voices of the participants.

Although the other stages of the audio chain described above were not explicitly part of the investigations described in this thesis, they will be discussed briefly in this section anyway, as their respective requirements and limitations and those of the reproduction stage are not independent. We will start with the recording stage (section 1.3.1), then the 'acoustic echo cancelling' stage (section 1.3.2), which is positioned in the audio chain between the recording and the transmission stage and finally, very briefly, the transmission stage (section 1.3.3).

1.3.1 Recording Of Participants' Voices

The first stage in the conferencing system's audio chain is the recording of the voices of the conference participants. There are various methods to record the voices, all with specific advantages and disadvantages. In the context of the videoconferencing concept that is desired in this project, i.e., a system in which the participants are relatively free to be located anywhere in the room while taking part in the conference, rather than being bound to a fixed position and in which the auditory scene is to be reproduced at the remote site as natural as possible, some

of these methods are more suitable to be implemented than others.
More specifically, in the concept which is investigated in this thesis, in which Wave Field Synthesis is applied to reproduce the voices in a natural spatial way, the recording method has to be able to provide to the reproduction part of the remote system both an accurate spectro-temporal representation of the speech signals of the participants as well as spatial information about the locations of the participants. The spatial information could be provided explicitly for each individual source, in the form of spatial coordinates as a function of time, or they could be present implicitly in the signals recorded by an array of microphones. Likewise, the spectro-temporal representation of the speech signals could be provided either on an individual-source basis, with each voice being available to the reproduction system as a separate audio channel, or in a more holistic way as the ensemble of signals recorded by a microphone array. In this subsection, an overview is given of the different recording methods and their respective advantages and disadvantages are discussed.

Tie-clip microphones. One approach is to record the voices of the participants as close to the source as possible, so that the gain that is required to obtain a good speech signal level is minimized and the recorded signal consists of almost only the direct sound from the sound source to the microphone and almost no room reverberance or other acoustic disturbances that might be present, such as other voices, noise from equipment, etcetera. This means that the recorded signal has an optimal signal-to-noise and direct-to-reverberant ratio, almost regardless of the movements of the person wearing the microphone. Additionally, it reduces the risk of echo problems, caused by sound from the remote site that is reproduced by the loudspeakers and picked up by the microphone, either directly from the loudspeakers to the microphone or indirectly through room reflections (section 1.3.2).
The main disadvantage of tie-clip microphones is that no spatial information about the location of the person wearing the microphone is present in the recorded signal, so an additional *source-tracking* system is needed to provide the spatial information that is needed for the reproduction. We will briefly discuss source-tracking systems at the end of this subsection (section 1.3.1.1).
An additional disadvantage of tie-clip microphones is that the positions of the microphones relative to the positions of the loudspeakers of the system are not restricted, since the persons wearing the microphones are able to walk around in the room, with the risk of the microphones getting too close to the loudspeakers, which increases the chance of occurrence of echoes.

Also, the fact that the positions of the microphones are non-fixed complicates the cancelling of these echoes (section 1.3.2).
Another (minor) disadvantage is that people have to be ‘clipped’ to participate in the conference. This seems to be no major drawback, but it might somewhat reduce an otherwise perfect illusion of telepresence.
In this recording approach, each participant’s voice is recorded as an individual audio signal and the data to be transmitted consists of a number of audio channels that is equal to the number of participants, plus source coordinates obtained from the source-tracking system.

Fixed microphones. In this approach, which is used in many existing conferencing systems, a number of (directional) microphones is placed at fixed positions in the conference room, for instance, on top of a conference table, with each participant having his own personal microphone right in front of him. In case of a conference without a table, one could think of several microphones located at different fixed positions within the room, which people have to approach whenever they want to speak. The number of audio channels to be transmitted equals the number of active microphones in this case.
An advantage of this scenario is that the positions of the sound sources are well-defined: they coincide with the positions of the microphones. Therefore, there is no need for source-tracking in this approach. Another advantage is that a good signal-to-noise ratio can be achieved if the microphones are used correctly by the participants.
However, since the positions are fixed, this approach is not flexible at all with regard to freedom of movement of the participants: they have to be close to a microphone every time they want to speak. Since one of the requirements of the system to be developed in this project was that participants should be free to walk around in the room during the conference, this is undesirable. Also, the obligation to actively approach a microphone in order to be able to participate seriously degrades the sensation of telepresence, which was the general aim of the project. Furthermore, since the fixed microphone approach requires participants to actively use microphones, this approach potentially suffers from microphone handling problems, since many users of a conferencing system are not used to handle microphones and are not trained to use them properly, resulting in large level variations and unwanted noises caused by touching, breathing, etcetera.

Directional microphones. In this approach, several highly-directional microphones are mounted at carefully chosen positions on the ceiling, in such a way that each of them ‘covers’ a certain part of the room area. If a participant is located somewhere in the area covered by a certain microphone, the participant’s voice will be picked up mainly by this microphone. The position of the sound source, from which it will be reproduced at the remote site, is in this case determined as simply the position that the microphone is aiming at. By having the ‘covered areas’ of adjacent microphones slightly overlap, a smooth transition from one microphone to a neighboring one can be achieved, so that at the remote site the reproduced voice does not suddenly ‘jump’ from one position to another. As with the fixed microphones approach, the number of channels to be transmitted equals the number of active microphones. A source-tracking system (section 1.3.1.1) can be used to control the on/off-switching of the individual microphones, so that at any given moment only the appropriate microphones are active.
An advantage of this approach, compared to the ‘fixed microphones’ approach, is that participants are free to move around and do not have to handle any microphones themselves.
Disadvantages are a bad signal-to-noise ratio, due to the rather large distance from the sources to the microphones and bad spatial resolution, due to the limited directivity of the microphones. Additionally, the conflicting requirements of no occurrences of sudden ‘jumps’ from the area covered by one microphone to that of another and a low amount of cross-talk between the microphones in order to have good spatial resolution, makes the installing of a microphone system according to this concept very critical and quite complicated.

Microphone array (sound field sampling). Ideally, in order to achieve a perfect spatial reproduction of a sound field, we would like to be able to capture all spatial information it contains. This is possible, at least in one plane and up to a certain temporal frequency, by using a line-array of closely spaced microphones, instead of a set of individual microphones, that ‘sample’ the pressure and normal component of the particle velocity of the sound field along the array. The recorded signals together contain all the spatial information of the sound field. If the signals recorded by the microphone array are then transmitted to the remote site and reproduced by an array of closely spaced loudspeakers according to the WFS theory (Chapter 2), a perfect spatial reproduction can be achieved. Also, this recording method would give conference participants complete freedom to move around in the room.
However, this approach is unfeasible in practice, at least in the videoconferencing application, since it results in too many audio channels that have to be transmitted to the remote site.

Microphone array (delay-and-sum beamforming). An interesting alternative to the ‘ideal’ microphone array approach described above is to use a so-called ‘beamforming microphone array’. In this approach, very high sensitivity is achieved for sounds arriving at the array from a very narrow region around a selected direction[2] by applying proper delays to the individual microphone signals and then summing them, while sound coming from all other directions is suppressed. By applying parallel delay-and-sum processing paths, each with a different set of delays, it is possible to create multiple ‘beams’ simultaneously.

In the videoconferencing application, a beamforming array could be installed below the screen. In the simplest case, the number and the directions of the beams are fixed. Whenever a person is located within one of the beams, his voice is picked up by this beam, while sounds from other directions are suppressed. This means that a very good direct-to-reverberant and signal-to-noise ratio can be achieved. Also, it is possible to optimize the directivity pattern of the array such that the response is minimal in the direction of the loudspeakers, which reduces the risk of echoes. Since the direction of the beam is known, only the single-channel output of the delay-and-sum algorithm and the direction of the beam have to be transmitted to be able to reproduce the voice from the correct direction at the remote site. The disadvantage is that when a participant is not located within one of the beams, his voice is not picked up by any of them. Furthermore, when a person is walking around while talking, his voice will jump from one direction to the next while he crosses different beams. To avoid this, a source-tracking system (section 1.3.1.1) can be used to localize the participants and to dynamically steer individual beams in their directions. An even more sophisticated solution is to use *adaptive beamforming*, an approach in which the room is ‘scanned’ by one or several beams to find the directions in which sound sources are located, by varying the delays in the delay-and-sum algorithm. Obviously, the disadvantage is the much more complicated processing that is required.

A disadvantage of all beamforming approaches described above is that only the directions of the participants are known and not their distance to the array. If required, techniques are available to estimate the distance as well (for instance, by analyzing the cross-correlation of neighboring microphones in the array), but it makes the processing much more complicated and, besides, it may not be that important to reproduce the sound sources at exactly the correct distances.

2. It is also possible to achieve high sensitivity only for sound from a small range of *positions* around a certain *position*, so-called *focal beamforming*.

Conclusion. As has been shown above, several approaches are available to achieve spatial recording of the voices of the conference participants, each with their own specific advantages and disadvantages. As explained in the introduction of this section, investigating which of the different recording methods is the most suitable one for our application was not really part of the research project described in this thesis, but it might be expected that an optimal solution might consist of a combination of two or more of the methods that have been described.

1.3.1.1 Source-tracking systems

A variety of source-tracking systems, based on various principles, is available. The most often used types are electro-magnetic based, using either radio- or infrared signals, and video-based systems. The electro-magnetic systems have the disadvantage that the participants have to wear a special device. The video-based and infrared systems have the disadvantage that they have difficulty dealing with (partial) occlusion of the objects to be tracked by other objects.
It is also possible to do acoustic source-tracking, by correlating the signal from a tie-clip microphone carried by the person to be tracked with the signals recorded by at least 3 remote microphones ([Atma02]). This is attractive because the participants do not need to wear a special device in order to be tracked.

1.3.2 Acoustic Echo Cancelling

A problem in many two-way electro-acoustic systems is the occurrence of acoustic echoes. The situation is sketched in figure 1.6. A sound signal s_A, for instance, the voice of a conference participant at location *A*, is picked up by microphone M_A and transmitted to location *B*, where the signal is reproduced by loudspeaker L_B. The sound coming from the loudspeaker is unintentionally picked up by microphone M_B, which is intended to pick up the voice of a conference participant at location *B*, indicated by s_B, and the microphone signal is transmitted back to location *A*, where it is reproduced by loudspeaker L_A. This results in the fact that the person at location *A* hears an echo of his own voice, with some delay that is determined by the transmission part of the system (and possibly by processing steps between recording and reproduction of the signal). The echo signal, in turn, is subsequently picked up by microphone M_A and sent back to location *B* again. In other words, there is a so-called acoustic loop. This is a highly undesirable situation, which can result in severe deterioration of speech intelligibility

and making it almost impossible to communicate. Therefore, the occurrence of acoustic echoes should be avoided as much as possible. One way of doing this is to use highly-directional microphones or close-miking techniques, as explained in section 1.3.1. If echoes do exist, a simple way to reduce their negative effects is to use 'dynamic gain control'. This is a technique which monitors the sound that is picked-up by the microphones and varies the gain of a microphone such that it is only active when a speech signal of a nearby person is detected. Dynamic gain control is easy to implement, but does not completely solve the problem, since at the moments when a microphone is active, it still picks up all other sounds that are present at that moment, including sound from the remote site that is reproduced by the loudspeakers. Furthermore, dynamic gain control can introduce 'pumping' or switching artifacts.

A more elaborate way to avoid echoes is to use *echo cancelling*. If the impulse response h_B of the acoustic path from loudspeaker L_B to microphone M_B is known, then it is possible to make an accurate estimation of the echo signal by convolving the signal x_A that is transmitted from location *A* with (a truncated version of) the impulse response h_B. Then, it is possible to remove the echo from the signal picked up by microphone M_B before it is sent back to location *A*, by subtracting the calculated echo signal y_B from the microphone signal z_B. The problem in practice, however, is that the acoustic path from L_B to M_B is in general not static but varying with time. For instance, in the case that a conference participant is wearing a tie-clip microphone, it changes when the participant walks around in the room. Besides this possible variation of the direct path from loudspeaker to microphone, also other changes in the room can change the impulse response h_B significantly. This means that it is in general not possible to use a fixed filter to remove the echo, but instead h_B has to be estimated in real-time by an adaptive filter, which tries to make an optimal estimate $\hat{h}_B$ of h_B by minimizing the difference between the estimated echo $\hat{y}_B$ and the echo y_B that is really picked up by M_B. The problem is that any acoustic signal that is picked up by M_B apart from the signal radiated from L_B, including the voices of conference participants at location *B*, acts as a disturbance for the adaptation process, which can easily cause stability problems in the adaptive filter. However, elaborate signal processing schemes exist to cope with these problems (see for example [Gay00]).

In multi-channel systems, such as WFS, the situation becomes even more complicated. In the case of a system with N_L loudspeakers and N_M microphones, there is a separate acoustic path

for each combination of loudspeaker $L_{B,i}$ and microphone $M_{B,j}$, leading to (N_L•N_M) different impulse responses $h_{i,j}$, which all have to be estimated simultaneously. This makes multi-channel echo cancellation a very complicated matter and an in-depth analysis of the problem is outside the scope of this thesis. Much research is carried out on the subject, however, and the reader is referred, for instance, to [Buch02], which describes research that was carried out in the context of the EC project CARROUSO ([CARROU]).

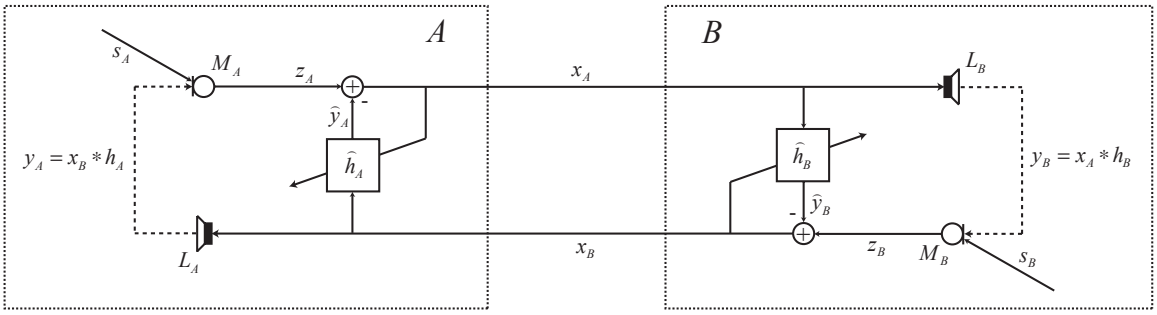


FIGURE 1.6. Schematic drawing of a single-channel Acoustic Echo Cancelling circuit.

1.3.3 Transmission Of Audio Signals

Given the fact that transmission bandwidth is always limited, it will be clear that there is a general requirement to keep the total data-rate of the audio signals that have to be transmitted within certain limits. Therefore, although in this thesis we will not be concerned with the details of the transmission of the recorded signals and will not take any fixed maximum number of channels or any fixed maximum data-rate into account, a large part of the investigations that will be presented aim at finding the *minimum* requirements for a given performance level, which is directly motivated, besides general considerations of manufacturing and computational costs, by this general transmission requirement. Specific transmission issues, such as bit-rate reduction techniques and latency, were not included in the research project of this thesis. They are however investigated extensively in the EC project CARROUSO, which has as key objective ‘to provide a new technology that enables to transfer a sound field, generated at a certain real or virtual space, to another usually remote located space’ ([CARROU]). More specifically, techniques are developed in the project to transmit Wave Field Synthesis data, using the object-oriented MPEG-4 standard for multimedia transmission.

1.4 Objectives Of This Thesis

This thesis has two main objectives. The first one is:

> *To investigate whether application of Wave Field Synthesis results in an audio reproduction system for videoconferencing that enables a more effective virtual meeting than can be achieved when conventional audio reproduction systems are used, with emphasis on acoustical telepresence, speech intelligibility and speaker identification and within the restriction of using two-dimensional video projection for the visual part of the system.*

The use of two-dimensional video projection is mentioned explicitly, because, as was mentioned in section 1.2.3, it was a restriction of this project, which was imposed by the fact that three-dimensional video projection is not feasible yet for a practical system and will probably not become feasible in the near future. As also mentioned in section 1.2.3, it was realized at the start of this project that combining two-dimensional video projection and spatialized audio reproduction that includes depth could potentially introduce problems. Since not much research had been done about this issue yet, investigating these effects was also considered an important part of the project.

The second main objective of this thesis is:

> *To optimize the Wave Field Synthesis reproduction system in the context of the videoconferencing application, such that all participants of the local room have a natural sound perception of all the voices of the remote participants with correct localization, also when the local participants walk around in the room.*

The most important element of this optimization process is to find out what is the minimum number of loudspeakers in the WFS array that is required to achieve this natural reproduction of the voices of the remote participants, such as to minimize the required amount of processing power, transmission bandwidth and the costs related to those issues. Some important issues that will be investigated in order to be able to answer this question are:

- What spatial resolution is needed in the reproduction of participants' voices with regard to sound source localization, in both the horizontal and vertical direction, taking into account the sensory interaction with the visual information that is received from the video projection?
- What maximum distance can be allowed between the individual loudspeakers of the reproduction array, such that no annoying coloration of the reproduced sound field is perceived by users of the videoconferencing system?

1.5 Outline Of This Thesis

First, the concept and theory of Wave Field Synthesis will be summarized in Chapter 2.
Chapter 3 first gives an overview of the aspects of auditory-, visual- and audio-visual perception that are most relevant to the research presented in this thesis and then studies their possible consequences for the design process of the audio part of the videoconferencing system. Special attention is given to the issues that are related to the combination of two-dimensional video projection and spatialized audio reproduction that includes a realistic reproduction of auditory distance.
Chapter 4 describes a series of audio-visual perception experiments that were carried out to investigate the spatial requirements of the audio reproduction and the effects of combining two-dimensional video with WFS audio reproduction. Audio-visual interaction plays an important role in these experiments. Also, an experiment is described in which the speech intelligibility of a WFS-based system is compared to that of a conventional system.
In Chapter 5 a study is made of the coloration that is introduced in the reproduced sound field because of the finite distance between the loudspeakers of the WFS array. The strength of this coloration increases for increasing distance between the loudspeakers, so the maximum amount of coloration that is allowed determines, together with the requirements for localization accuracy that are investigated in Chapter 4, the maximum distance between the array loudspeakers that can be allowed in this application. The coloration strength is analyzed both in an objective way and in a perception experiment.
Chapter 6 investigates the application of a fairly new type of loudspeaker, the Distributed Mode Loudspeaker (DML) panel, in WFS sound reproduction and the extension of its concept to so-called Multi Actuator Panels (MAP). MAP Panels have some clear advantages that facilitate the implementation of multi-channel sound reproduction systems such as WFS, namely

being lightweight, cheap and non-intrusive in the room interior. In the specific application of videoconferencing, they have the additional advantage of being flat and stationary (in contrast to a conventional cone loudspeaker which moves back and forth), so that they can be used as projection screen and loudspeaker system simultaneously.

In Chapter 7, the results of all the preceding chapters are collected and related to the objectives that were defined in section 1.4.

Finally, in Chapter 8 the general conclusions of this thesis are presented.

CHAPTER 2 *Wave Field Synthesis*

2.1 Introduction

It has always been the ultimate goal of audio engineers to be able to fully recreate the auditory experience that is perceived at a given location and time at another location and time. A perfect recreation of an arbitrary auditory event requires a perfect recreation in both a spectro-temporal and a spatial sense, the spectro-temporal part encompassing characteristics of the auditory event such as ‘timbre’, ‘dynamic range’, ‘attack’ etcetera, while the spatial part includes characteristics such as ‘source localization’, ‘spaciousness’ and ‘apparent source width’.

In a spectro-temporal sense, it is safe to say that this goal has been reached to a very high level of perfection. Very high quality microphones, loudspeakers and amplifiers have been available on the market for decades now and since the introduction of the Compact Disc in the early eighties, it has become possible to store audio almost without any loss in quality, up to the limits of the capabilities of the human hearing system. And with the new high resolution, high sampling rate formats that have been introduced in more recent years, such as Super Audio CD and DVD-Audio, it seems that a satisfactory level of spectro-temporal perfection in audio reproduction has been reached.

In the spatial sense, however, the story is very different. Although the earliest efforts to record and reproduce sound in such a way that the spatial characteristics are preserved date back to

the first experiments on stereophony at the end of the 19th century, the capabilities of almost all techniques that have been developed since that time to make a truly realistic spatial reproduction of an original auditory event are very limited. A problem of most techniques is that they are, at best, only able to provide a reproduction with accurate source localization at, or within a very limited area around, a certain position, so the listener must be located at that position to get the correct spatial impression. In stereophony, this is commonly referred to as the 'sweet spot'. When a listener moves out of this 'sweet spot', the spatial image breaks down very quickly, so it is impossible to speak of a perfect recreation of the original acoustical event in this case, even though the spatial impression, in terms of the localization of sound sources, might be quite good at the sweet spot itself. The same holds for the stereophony-based 'surround sound' systems, such as the popular '5.1' standard.

An additional limitation of stereophony is that it is unable to recreate the sense of spaciousness and being enveloped by the sound of the original acoustical event and although the multi-channel 'surround sound' systems referred to above are able to provide a sense of spaciousness and envelopment by reproducing reverberant and ambient sound from the back and possibly also the sides, the impression of spaciousness that is perceived is usually an artistically pleasing effect rather than a truly convincing recreation of the original experience.

It is clear that in order to obtain a perfect, place-independent recreation of an original auditory event, the ultimate solution would be to reconstruct the original physical sound field of that auditory event. This, as will be shown in this chapter, is exactly what the concept of Wave Field Synthesis does.

Intuitively, Wave Field Synthesis (WFS) can be thought of as being based on the well-known "Huygens' principle", formulated by the Dutch scientist Christiaan Huygens and published in 1690 ([Huyg90]). He stated that the propagation from time instant t to time instant $t+\Delta t$ of a wave front that is emitted by a source[1] Ω at time instant t_0, can be described by imagining a continuous distribution of so-called secondary point sources along the wave front at t, which all emit a new circular wave front simultaneously (figure 2.1). The wave front at instant $t+\Delta t$ is constructed as the envelope of all the secondary wave fronts at $t+\Delta t$. In the case of sound waves, we could think of the primary source Ω as a sound source which emits an acoustic

1. in Huygens' work he is talking about a light source, but the principle applies to any phenomenon that is described by waves.

pulse at t_0. If we know the location and shape of the wave front at t, we can imagine the secondary sources along the wave front as being small loudspeakers, which all emit a pulse at time instant t. The wave front at $t+\Delta t$ is then built up by all the individual contributions of the loudspeakers. According to this concept, the sound field which is received by a listener at a position R can arbitrarily be thought of as either being produced directly by the primary sound source Ω itself or by the secondary sound sources. Consequently, for the listener at position R, it does not matter whether the primary sound source Ω is really present or only the loudspeakers, provided that they emit the correct signal at the correct time. In other words, we are able to *synthesize* a copy of the sound field of source Ω by the loudspeakers, not only at listening position R, but in the whole space beyond the location of the wave front of source Ω at time t.

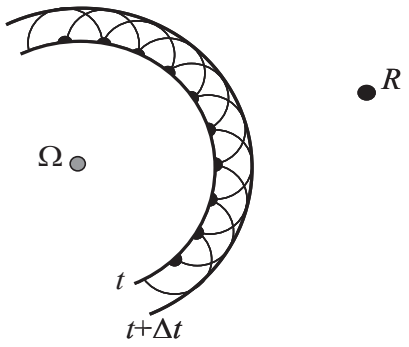


FIGURE 2.1. Huygens' Principle.

Although the Huygens principle is a very strong intuitive way to describe the way in which wave fronts propagate, it does not hold mathematically. The real theoretical basis for the concept of Wave Field Synthesis is the so-called Kirchhoff-Helmholtz theorem, which mathematically describes the sound field *within* a closed surface S, which does not contain any sources itself, in terms of the sound field *on* S and it shows that the sound field within S can be interpreted as being generated by a certain continuous distribution of secondary monopole and dipole sources on S. If there is a primary sound source Ω outside S, this means that the sound field *within* S due to Ω is equivalent to the sound field of this secondary monopole and dipole source distribution on S and a listener located within S is unable to distinguish between the sound field due to the real primary source Ω and the copy of its sound field, synthesized by the secondary sources (figure 2.2).

Although the concept of making a perfect spatial reproduction of a sound field based on the

Huygens principle and the distribution of secondary monopole and dipole sources from the Kirchhoff-Helmholtz integral had been suggested before (e.g.: [Jess73]), the theoretical and practical steps that were necessary to enable a practical implementation of the concept were never taken, until Berkhout at Delft University of Technology proposed the concept of Wave Field Synthesis in the late eighties ([Berk88]). Since then, the concept has been developed further at TU Delft, first in general ([Voge93]), then for specific applications, such as direct sound enhancement ([Star97]), sound reproduction ([Verh97]) and variable acoustics and auralization ([Sonk00], [Huls04]) and has been implemented in several real-time demonstrators. This thesis continues this line of research, by investigating the application of WFS in videoconferencing.

In this chapter, the most important aspects of WFS, both theoretical and practical, will be reviewed. It is basically a summary of work that has been done before by others, and for details on each subject that is discussed, the reader will be referred to the relevant publications.
First, in section 2.2, the theoretical basis of WFS is presented, starting with the aforementioned Kirchhoff-Helmholtz integral (section 2.2.1) from which two special cases, the 3D Rayleigh I and II integrals, are derived (section 2.2.2). In section 2.2.3, an approximation of the Rayleigh I integral is presented, which enables implementation of the concept by using a line array of secondary sources instead of a planar array, which is required by the 3D Rayleigh integrals.
The theory described in section 2.2.2 and section 2.2.3 requires a continuous distribution of secondary sources on an infinite plane or an infinitely long line, respectively. This, of course, is unfeasible in practice, where we always have to deal with a finite number of loudspeakers of finite size, distributed over a finite amount of space, so in practice the WFS operators described in section 2.2 have to be discretized and truncated. Both these practical limitations introduce certain artifacts in the reproduced sound field. The effects of discretization and truncation are described in section 2.3.
The chapter closes with a discussion of some applications of WFS and the specific practical implementations of the WFS concept for those applications (section 2.4).

2.2 Theory

2.2.1 Kirchhoff-Helmholtz Integral

The mathematical starting point in our development of the WFS theory is Green's second theorem ([Berk87]):

$$\int_V \left(\Phi\nabla^2\Psi - \Psi\nabla^2\Phi\right) dV = -\int_S \left(\Phi\nabla\Psi - \Psi\nabla\Phi\right)\cdot\vec{n}dS, \tag{2.1}$$

in which *S* is some closed surface, *V* the volume that is contained in *S* and $\vec{n}$ is the inward pointing normal vector on *S* (figure 2.2). Φ and Ψ are some scalar functions that are defined inside and on *S*, with continuous second-order derivatives inside *S*.

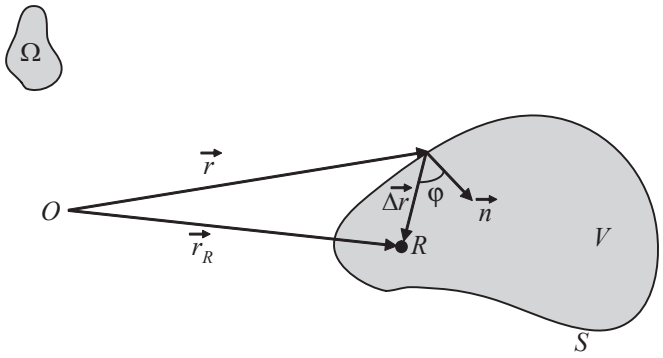


FIGURE 2.2. Definition of variables in the derivation of the Kirchhoff-Helmholtz integral.

We can choose for Φ the pressure field in the frequency domain $P(\vec{r},\omega)$ of some source Ω, located somewhere outside *S*, which satisfies the homogeneous wave equation:

$$\nabla^2 P(\vec{r},\omega) + k^2 P(\vec{r},\omega) = 0, \tag{2.2}$$

for all points inside *S*, with *k* the wave number and ω the angular frequency.

For Ψ we may choose the pressure field $G(\vec{r}\,|\,\vec{r}_R,\omega)$ of an impulsive point source located at a position *R* with position vector $\vec{r}_R$ inside *S* (figure 2.2), which satisfies the inhomogeneous wave equation:

$$\nabla^2 G(\vec{r}\,|\,\vec{r}_R,\omega) + k^2 G(\vec{r}\,|\,\vec{r}_R,\omega) = -4\pi\delta(\vec{r}-\vec{r}_R), \tag{2.3}$$

inside *S*. Due to reciprocity, we can interchange the roles of source and receiver in (eq. 2.3), so that:

$$G(\vec{r}_R\,|\,\vec{r},\omega) = G(\vec{r}\,|\,\vec{r}_R,\omega). \tag{2.4}$$

Inserting $P(\vec{r},\omega)$ for Φ and $G(\vec{r}\,|\,\vec{r}_R,\omega)$ for Ψ in (eq. 2.1) and using (eq. 2.2), (eq. 2.3) and (eq. 2.4), we obtain the so-called Kirchhoff-Helmholtz integral:

$$P(\vec{r}_R,\omega) = -\frac{1}{4\pi}\int_S \left[G(\vec{r}_R\,|\,\vec{r},\omega)\nabla P(\vec{r},\omega) - P(\vec{r},\omega)\nabla G(\vec{r}_R\,|\,\vec{r},\omega)\right]\cdot\vec{n}\,dS. \tag{2.5}$$

If for $G(\vec{r}\,|\,\vec{r}_R,\omega)$ we choose the pressure field of a *monopole* point source located at position *R* inside *S*:

$$G(\vec{r}\,|\,\vec{r}_R,\omega) = \frac{e^{-jk|\vec{r}-\vec{r}_R|}}{|\vec{r}-\vec{r}_R|}, \tag{2.6}$$

which satisfies (eq. 2.3) and insert this in (eq. 2.5), we obtain:

$$P(\vec{r}_R,\omega) = \frac{1}{4\pi}\int_S \left[j\omega\rho_0 V_n(\vec{r},\omega)\frac{e^{-jk\Delta r}}{\Delta r} + P(\vec{r},\omega)\frac{1+jk\Delta r}{\Delta r}\cos\varphi\,\frac{e^{-jk\Delta r}}{\Delta r}\right] dS, \tag{2.7}$$

in which ρ_0 is the static density of mass, V_n is the (inward pointing) component of the particle velocity normal to the surface *S*, Δr is short for the distance $|\vec{r}-\vec{r}_R|$ between integration point $\vec{r}$ on *S* and receiver position $\vec{r}_R$ inside *S* and φ is the angle between $\overrightarrow{\Delta r}$ and $\vec{n}$ (figure 2.2) and in which we have made use of the relationship between pressure and particle velocity:

$$\nabla P(\vec{r},\omega) = -j\omega\rho_0 V(\vec{r},\omega). \tag{2.8}$$

The Kirchhoff-Helmholtz integral (eq. 2.7) forms the basis of WFS theory. If we look at the integral more closely, we see that the first term of the integrand can be interpreted as the pres-

sure at position $\vec{r}_R$, located *inside S*, that is caused by a *monopole* source at position $\vec{r}$ *on S*, with a source strength that is proportional to the normal component of the particle velocity at that position $\vec{r}$. Likewise, the second term of the integrand can be seen as the pressure at position $\vec{r}_R$ that is caused by a *dipole* source at position $\vec{r}$, with a source strength that is proportional to the pressure at that position $\vec{r}$. In other words, (eq. 2.7) describes the sound pressure at *any* position $\vec{r}_R$ *inside S* in terms of the pressure and normal component of the particle velocity *on S*, provided that the volume *V* enclosed by *S* does not contain any sources itself. This means that if we have some arbitrary distribution of sound sources *outside S*, of which we know, either by calculation or by measurement, the pressure and normal component of the particle velocity *on S*, then we can use (eq. 2.7) to calculate the pressure at any position *inside S*. It can be shown that the integral of (eq. 2.7) is identical to zero for any observation position $\vec{r}_R$ located outside *S*.

From the above, it is only a small step to interpret (eq. 2.7) in the following, slightly different, way. Suppose we want to replicate, or *synthesize*, *inside the entire volume V that is enclosed by surface S*, the sound field of a sound source, or distribution of sound sources, Ω, located somewhere outside *S*. The Kirchhoff-Helmholtz integral tells us that this is possible by having a continuous distribution of monopole and dipole sources on *S*, with the source strength of each monopole and dipole source being proportional to the local value of, respectively, the normal component of the particle velocity and the pressure caused by Ω, which, again, can be obtained by either calculation or measurement. This means that also in the absence of the actual source Ω, it is possible to synthesize the sound field that would have been obtained if Ω had actually been present, *inside the whole volume enclosed by S*. It is this interpretation of (eq. 2.7) that describes the general concept of Wave Field Synthesis.

From this point on, we will refer to the monopole and dipole sources on *S* as *secondary sources*, while Ω will be called the *primary source (distribution)*.

2.2.2 3D Rayleigh Integrals

Implementation of WFS by direct application of the Kirchhoff-Rayleigh integral (eq. 2.7) requires a continuous secondary source distribution of both monopole and dipole sources over an entire surface enclosing the desired listening area, as discussed in the previous subsection. It would be convenient if it was possible to obtain the same result by using either monopole or

dipole sources only. It will be shown in this subsection that, within certain restrictions, this is indeed possible.

Let us first look at the case of using monopole sources only. Comparing (eq. 2.7) to the general form of the Kirchhoff-Helmholtz integral (eq. 2.5), we see that the dipole term in (eq. 2.7) corresponds to the second term of the integrand of (eq. 2.5). In the derivation of (eq. 2.5) we have only assumed for $G(\vec{r}|\vec{r}_R,\omega)$ that it satisfies (eq. 2.3). So if we can find a function $G(\vec{r}|\vec{r}_R,\omega)$, instead of (eq. 2.6), which satisfies (eq. 2.3) and which satisfies:

$$\nabla G(\vec{r}_R|\vec{r},\omega)\cdot\vec{n}=0, \qquad \vec{r} \text{ on } S, \tag{2.9}$$

then this second term in the integrand of (eq. 2.5) vanishes and only the first term, which corresponds to the monopole term in (eq. 2.7), remains.

It is easy to verify that if we add a function Γ that has no singularities on or inside S to $G(\vec{r}|\vec{r}_R,\omega)$ of (eq. 2.6), so that $G(\vec{r}|\vec{r}_R,\omega)$ becomes:

$$G(\vec{r}|\vec{r}_R,\omega)=\frac{e^{-jk|\vec{r}-\vec{r}_R|}}{|\vec{r}-\vec{r}_R|}+\Gamma, \tag{2.10}$$

and which satisfies the homogeneous wave equation:

$$\nabla^2\Gamma+k^2\Gamma=0, \tag{2.11}$$

then the new $G(\vec{r}|\vec{r}_R,\omega)$ of (eq. 2.10) still satisfies (eq. 2.3), so that it is still a valid $G(\vec{r}|\vec{r}_R,\omega)$ to use in (eq. 2.5).

Given an arbitrary closed surface S, it is very hard to find a function Γ that both satisfies (eq. 2.11) and for which (eq. 2.10) satisfies (eq. 2.9). However, it appears to be possible to find such a function Γ by a special choice of surface S. Consider the configuration of figure 2.3, in which surface S consists of plane surface S_1 and hemisphere S_2 with radius R_0 which is closed by S_1. If we now assume that all primary sources are located in the halfspace left of S_1 and we increase the radius R_0 of hemisphere S_2 to infinity, then the contribution of S_2 to the integral of (eq. 2.5) vanishes for any finite time interval, so that only the contribution of plane surface S_1 remains. For this special configuration, it can be shown ([Berk87]) that the conditions

described above are satisfied if we choose for Γ the field of a monopole point source located at a position R' that is the mirror image of position R in the plane surface S_1 (figure 2.3):

$$\Gamma(\vec{r}\,|\,\vec{r}_{R'},\omega) = \frac{e^{-jk|\vec{r}-\vec{r}_{R'}|}}{|\vec{r}-\vec{r}_{R'}|}. \tag{2.12}$$

Inserting (eq. 2.10) with this choice of Γ in the Kirchhoff-Helmholtz integral (eq. 2.5), together with the notion that only S_1 contributes to it, results in the so-called *3D Rayleigh I integral*:

$$P(\vec{r}_R,\omega) = \frac{j\omega\rho_0}{2\pi}\int_{S_1} V_n(\vec{r},\omega)\frac{e^{-jk\Delta r}}{\Delta r}dS. \tag{2.13}$$

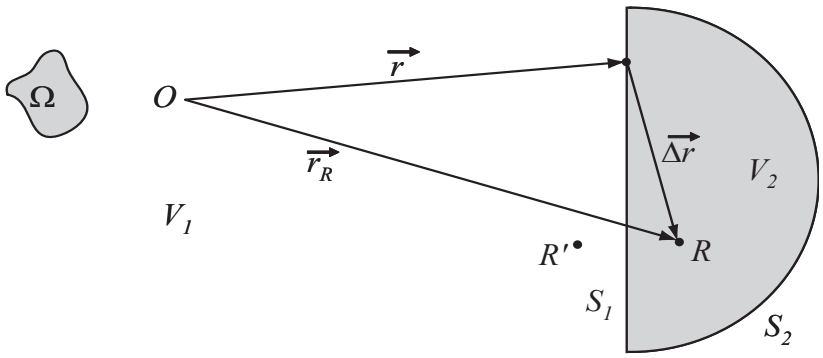


FIGURE 2.3. Geometry for the derivation of the Rayleigh I integral (eq. 2.13).

We have now reached a configuration in which the source space is separated from the receiver space by an infinite plane S_1 with a continuous distribution of secondary monopole sources. The Rayleigh I integral (eq. 2.13) states that the sound field in the receiver space due to a primary source distribution Ω located somewhere in the source space can be synthesized by this infinite, continuous distribution of only monopole sources on plane surface S_1. The strength of each secondary monopole source is proportional to the normal component of the particle velocity caused by Ω at the position of that monopole.

A difference between the results that are obtained with the Rayleigh I integral (eq. 2.13) and the Kirchhoff-Helmholtz integral (eq. 2.7) is that while the latter results in a total field equal to zero in the source space, this is not the case for the Rayleigh I integral. Instead, the secondary

monopole distribution on S_1 radiates a field into the source space which is the mirror image of the field radiated into the receiver space.

An expression similar to (eq. 2.13) can also be derived for the case in which we want to use dipole sources only. The result is known as the Rayleigh II integral and will only be stated here (see e.g. [Berk87] for the derivation):

$$P(\vec{r}_R,\omega) = \frac{1}{2\pi}\int_{S_1} P(\vec{r},\omega)\frac{1+jk\Delta r}{\Delta r}\cos\varphi\frac{e^{-jk\Delta r}}{\Delta r}dS. \tag{2.14}$$

2.2.3 2½D Rayleigh I Integral

To synthesize the sound field of a given source distribution, the 3D Rayleigh integrals (eq. 2.13) and (eq. 2.14) require a plane of secondary sources, separating the source- and receiver space. This requirement makes a practical implementation of the WFS concept quite difficult. It would be much more convenient if it would be possible to achieve approximately the same result with only a linear array of secondary sources. As will be shown in this subsection, it appears that under certain circumstances this is indeed possible, especially if accurate reproduction is only required in one plane, for example the horizontal plane through the listeners' ears. The derivation of the result below is a summary of the derivations in [Star97] and [Verh97]. For details, the reader is referred to those publications.

Consider the situation of figure 2.4, where the primary source Ω and the receiver *R* are in the same plane *z*=0, perpendicular to the plane *S* of secondary monopole sources at *y*=0. The Rayleigh I integral over surface *S* (eq. 2.13) can be rewritten as a line integral over lateral coordinate *x* of contributions of vertical lines of secondary sources to the total pressure in *R*:

$$P(\vec{r}_R,\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} P_{line}(\vec{r}_R,x,\omega)dx, \tag{2.15}$$

in which P_{line} is the pressure at position *R* due to the vertical line of secondary monopole sources at lateral position *x* in *S*. It can be shown, by using the so-called ‘stationary phase approximation’, that the main contribution to the pressure at position *R* from the line at $x=x_L$ comes from the secondary sources around (x_L,0,0), i.e., the intersection of the vertical line with the horizontal plane *z*=0 in which both Ω and *R* are located (figure 2.4). Using this approxima-

tion, it is possible to replace the vertical line of secondary monopoles at lateral position x_L by a single secondary monopole source at the 'stationary point' (x_L,0,0), which yields approximately the same pressure at position R as the original vertical line of monopoles. Verheijen ([Verh97]) derives the following expression for P_{line} for the case of a primary source at a large distance from the array (kr>>1) with source spectrum $S(\omega)$ and a far-field directivity characteristic $G(\varphi,\theta,\omega)$ (see figure 2.4 for the definition of the variables):

$$P_{line}(\vec{r}_R, x_L, \omega) = S(\omega)\sqrt{2\pi jk}\sqrt{\frac{\Delta r}{r+\Delta r}}G(\varphi,0,\omega)\cos\varphi\frac{e^{-jkr}}{\sqrt{r}}\frac{e^{-jk\Delta r}}{\Delta r}. \tag{2.16}$$

Substituting:

$$V_n(x,\omega) = \frac{S(\omega)}{\rho_0 c}G(\varphi,0,\omega)\cos\varphi\frac{e^{-jkr}}{r}, \tag{2.17}$$

for the normal component of the particle velocity $V_n(x,\omega)$ at (x,0,0), due to Ω (for kr>>1), with c the speed of sound, (eq. 2.16) can be written in terms of $V_n(x,\omega)$ as:

$$P_{line}(\vec{r}_R, x_L, \omega) = \rho_0 c\sqrt{2\pi jkr}\sqrt{\frac{\Delta r}{r+\Delta r}}\frac{e^{-jk\Delta r}}{\Delta r}V_n(x_L,\omega). \tag{2.18}$$

Finally, the total pressure at position R now follows from (eq. 2.15) and (eq. 2.16) as:

$$P(\vec{r}_R,\omega) = \int_{-\infty}^{\infty} Q(x,\omega)\frac{e^{-jk\Delta r}}{\Delta r}dx, \tag{2.19}$$

with:

$$Q(x,\omega) = S(\omega)\sqrt{\frac{jk}{2\pi}}\sqrt{\frac{\Delta r}{r+\Delta r}}G(\varphi,0,\omega)\cos\varphi\frac{e^{-jkr}}{\sqrt{r}}, \tag{2.20}$$

the driving function of the secondary monopole sources, which, using (eq. 2.17), can alternatively be written in terms of the normal component of the particle velocity at x due to Ω, as:

$$Q(x,\omega) = \rho_0 c \sqrt{\frac{jk}{2\pi}} \sqrt{\frac{\Delta r}{r + \Delta r}} \frac{1}{\sqrt{r}} V_n(x,\omega). \tag{2.21}$$

Because the wave field of a source that radiates sound waves in three dimensions is reconstructed by (eq. 2.19) in a two-dimensional plane only, it can be considered to be something in between a full 3D and a strict 2D solution. Therefore, (eq. 2.19) is referred to as the *2½D Rayleigh I integral*.

The different terms in the driving function (eq. 2.20) can be interpreted as follows: $S(\omega)$ is the Fourier transformation of the time-domain source signal. The first square-root term acts like a high-pass filter with a slope of 3 dB/octave. Note that it does not depend on the secondary source position, so it can be incorporated in the source signal $S(\omega)$. The complex exponential corresponds to a delay in the time-domain of (r/c) seconds, with c the speed of sound, and accounts for the propagation time from the primary source to the secondary source. The remaining terms together form the amplitude of the secondary source.

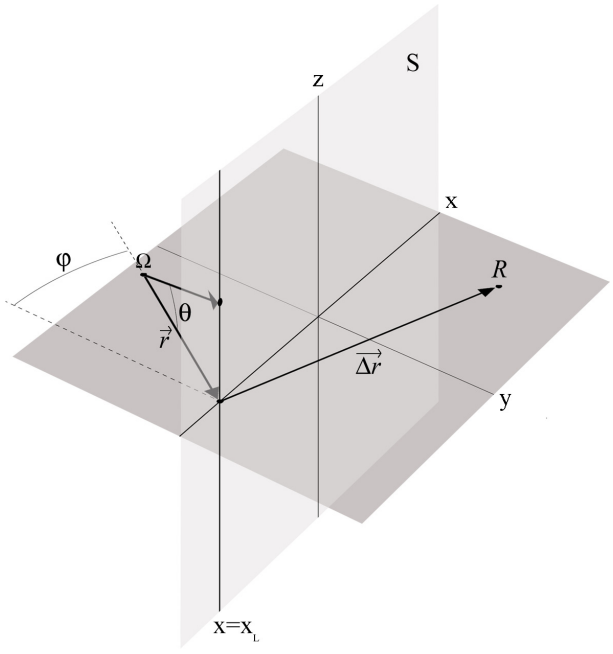


FIGURE 2.4. Geometry for the derivation of the 2½D Rayleigh I integral (eq. 2.19).

If we take a close look at (eq. 2.20), we observe that, while the cosine term, the source directivity term $G(\varphi,\theta,\omega)$ and the exponential term depend only on the position of the primary source relative to the position of the secondary source, the square-root term depends also on the

receiver position *R*. This is undesirable, since this means that the driving functions of the secondary sources have to be adapted to the receiver position, contrary to the goal to obtain a position-independent solution in the whole listening area. Note that this is a result of the approximation that was used to replace the contribution of an infinite vertical line of secondary sources by a single monopole, as expressed by (eq. 2.16).

Although it is not possible to reconstruct the correct amplitude in the whole listening area at the same time using a linear array as an approximation of a planar array, it appears to be possible to reconstruct the correct amplitude on a reference *line*, instead of only at one reference position, as the discussion of (eq. 2.20) above suggests. This can be achieved by performing another stationary phase approximation, but now along the secondary source line. Again, the detailed derivation will not be given here (see [Star97] or [Verh97]), but the intuitive interpretation of the approximation is that we consider that for each receiver position *R* on a certain reference line L_{ref}, the main contribution will come from the secondary source at position x_0, which is located at the intersection of the secondary source line *L* and the line from the primary source Ω to the receiver position *R* (figure 2.5). Conversely, for a given primary source position, each secondary source position x_0 is the main contributor to the sound field at a unique receiver position *R* on reference line L_{ref}, which is located at the intersection of that reference line and the line from the primary source through the secondary source position x_0. By fixing Δr in the square-root term in (eq. 2.20) to Δr_0 for each individual secondary source position x_0, as indicated in figure 2.5, the driving function of each secondary monopole no longer depends on the receiver position, but only on the fixed distance between the secondary source and the intersection point of the line from the primary source Ω through that secondary source and the receiver line L_{ref}. Using this approximation, (eq. 2.20) is modified to:

$$Q(x,\omega) = S(\omega)\sqrt{\frac{jk}{2\pi}}\sqrt{\frac{\Delta r_0}{r+\Delta r_0}}G(\varphi,0,\omega)\cos\varphi\frac{e^{-jkr}}{\sqrt{r}}. \qquad (2.22)$$

Note that when the receiver line L_{ref} is a straight line parallel to the secondary source line, the modified square-root term is a constant.

Application of driving functions (eq. 2.22) introduces amplitude errors in the synthesized sound field at positions that are not located on the selected reference line. More specifically, a comparison between (eq. 2.20) and (eq. 2.22) and the corresponding geometries shown in

figure 2.4 and figure 2.5, respectively, shows that between the secondary source line and the reference line the synthesized amplitude will be too high, while beyond the reference line it will be too low, compared to the true sound field of source Ω. These amplitude errors, however, are usually quite small. For example, in the synthesized sound field of a point source located 1 m behind the array, the maximum deviation from the amplitude of a real point source located at the same position is only about 1.5 dB within a large listening area extending from 1.1 to 4.5 m from the array ([Verh97]).

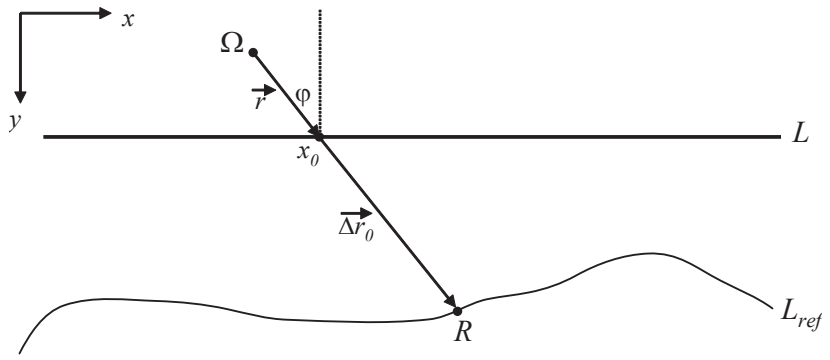


FIGURE 2.5. Geometry for the derivation of (eq. 2.22).

The '2½D with a reference line' approach with driving function (eq. 2.22) is the WFS approach that is used for all experiments described in this thesis. Driving function (eq. 2.22) can also be applied to non-straight lines of secondary sources, in which case the integration is no longer along the *x*-axis, but along the secondary source line. If there are discontinuities in the curvature of the secondary source line, such as corners, this introduces diffraction artifacts. However, these are usually quite small ([Star97], [Verh97]).

2.3 Practical Implementation Issues

In the previous section, the 2½D Rayleigh I integral was derived, which is able to synthesize a desired sound field in the horizontal plane by an infinite, continuous, linear array of secondary monopole sources. In practice, however, this concept is inevitably implemented by means of loudspeakers of finite size, separated by a finite distance, which consequently transforms the continuous array into a discrete array. This introduces artifacts in the synthesized wave field, known as *spatial aliasing*. Furthermore, in practice the array can not be infinitely long, so it is truncated to a finite length. This limits the aperture of the array and also introduces artifacts

caused by diffraction at the edges of the array. The effects caused by discretization and truncation of the continuous linear array are discussed in this section.

2.3.1 Discretization Of The WFS Array

Spatial Sampling. Just like time sampling causes a phenomenon called *temporal aliasing* if the signal is sampled with a sample rate which is less than two times the highest temporal frequency that is present in the original signal, spatial sampling of a continuous wave field at discrete positions by a microphone array with finite equidistant spacing Δx introduces a similar phenomenon, called *spatial aliasing*, if the spacing Δx with which the field is sampled is too large with respect to the highest spatial frequency component that is present in the original wave field. To be precise, spatial aliasing will occur if the original source field contains spatial components for which ([Berk87]):

$$\left|k_x\right| = \left|k \sin\theta\right| > \frac{\pi}{\Delta x}, \tag{2.23}$$

in which k_x is the wave number in the x-direction of the original wave field and θ is the angle of incidence of a plane wave component of the original field, relative to the array (with θ=0 corresponding to normal incidence). Given the relationship between the wave number k and the temporal frequency f, we can derive from (eq. 2.23) that spatial aliasing will be present in the sampled wave field for temporal frequencies above:

$$f_{\max} = \frac{c}{2\Delta x \sin\theta_{\max}}, \tag{2.24}$$

in which $\theta_{\max}$ is the maximum angle from which sound field components reach the array[2].

2. Strictly speaking, (eq. 2.24) holds only for source fields which have a symmetrical spatial amplitude spectrum, i.e., wave fields which contain spatial components in the range $-\theta_{\max} \le \theta \le \theta_{\max}$. In the case of source fields with asymmetrical spatial amplitude spectra, the factor $2\sin\theta_{\max}$ has to be replaced by $\sin\theta_{\max}-\sin\theta_{\min}$, with $\theta_{\min}$ the minimum angle from which sound field components reach the array ([Star97]).

Discretization Of The WFS Array. Discretization of the continuous reproduction array, i.e., replacing the continuous line of secondary sources by a linear array of individual secondary sources spaced by an equidistant spacing Δx, can be seen as a spatial sampling process as described above. The effects of the discretization on the synthesized wave field are most conveniently analyzed in the spatial Fourier domain. Let us start with the continuous driving function $Q(x,\omega)$ (eq. 2.22). By performing a spatial Fourier transform on (eq. 2.22), we decompose the source field at the position of the array into plane wave components from all possible directions. In figure 2.6.a, the amplitude of the spatial Fourier transform $\tilde{Q}(k_x,\omega)$ of (eq. 2.22) is shown for the wave field of a monopole point source at some distance from the array. We now sample the continuous driving function at intervals Δx, by multiplying (eq. 2.22) with a periodic dirac function $\delta(x\text{-}n\Delta x)$, resulting in the sampled driving function $Q_\Delta(x,\omega)$. In the spatial Fourier domain, this corresponds to convolving $\tilde{Q}(k_x,\omega)$ with the spatial Fourier transform of the periodic dirac function, which is a periodic dirac function itself, with period $2\pi/\Delta x$. This means that the spatial amplitude spectrum of the sampled driving function $Q_\Delta(x,\omega)$ is a periodic repetition of the spatial amplitude spectrum of $Q(x,\omega)$, shown in figure 2.6.a, with the same period $2\pi/\Delta x$. Depending on the value of Δx, consecutive repetitions of the spatial amplitude spectrum of $Q(x,\omega)$ will or will not overlap in the spatial amplitude spectrum of $Q_\Delta(x,\omega)$. If overlap occurs, this means that spatial aliasing is present in the sampled driving function $Q_\Delta(x,\omega)$. In figure 2.6.b and figure 2.6.c, the spatial amplitude spectrum of $Q_\Delta(x,\omega)$ is shown for two sampling distances Δx. Figure 2.6.b shows the situation for a value of Δx which is sufficiently small to ensure that (eq. 2.23) is not true for any spatial component that is present in $Q(x,\omega)$, so that no spatial aliasing is present in $Q_\Delta(x,\omega)$. Figure 2.6.c shows the situation when Δx is too large, so that (eq. 2.23) is indeed true for some spatial components of $Q(x,\omega)$. As can be seen, overlap does occur in this case and therefore spatial aliasing is present in $Q_\Delta(x,\omega)$.

Now we look at the synthesis process. For the discrete array, the synthesis integral (eq. 2.19) can be written as:

$$P(\vec{r}_R,\omega) = Q_\Delta(x,\omega) * W(x,\vec{r}_R,\omega), \tag{2.25}$$

in which * denotes convolution and $W(x,\vec{r}_R,\omega)$ is the response at receiver position $\vec{r}_R$ of the secondary source at lateral position x. Spatial Fourier transformation of (eq. 2.25) yields:

$$\tilde{P}(k_x,\omega) = \tilde{Q}_\Delta(k_x,\omega)\tilde{W}(k_x,\omega), \tag{2.26}$$

in which the ~ symbol denotes the spatial Fourier transform of the corresponding variable. If the secondary sources are monopoles, then the sound field of $W(x,\vec{r}_R,\omega)$ contains spatial components in all directions between $-\pi/2$ and $+\pi/2$, so the spatial amplitude spectrum of $W(x,\vec{r}_R,\omega)$ is non-zero for all values of k_x for which $|k_x|<k$. The boundaries of this range are indicated by the dashed lines in figure 2.6.b and figure 2.6.c. From (eq. 2.26) it now follows that in the case of figure 2.6.c, the spatial amplitude spectrum of the synthesized wave field is corrupted by the overlapping of consecutive repetitions of the spatial amplitude spectrum of the continuous driving function $Q(x,\omega)$ in the periodic spatial amplitude spectrum of its sampled version $Q_\Delta(x,\omega)$, caused by the fact that the sampling distance Δx is too large. In the case of figure 2.6.b, Δx is sufficiently small, so the spatial amplitude spectrum of the reproduced field is equal to the spatial amplitude spectrum of the original source field, shown in figure 2.6.a.

From looking at figure 2.6.c, it can be seen that spatial aliasing could have been avoided if the range of spatial components of the continuous driving function $Q(x,\omega)$ picked up by the array would have been limited to components having a spatial frequency $|k_x|<\pi/\Delta x$ (indicated by the solid vertical lines), while rejecting all components with larger values of $|k_x|$. In that case, the consecutive repetitions of $Q(x,\omega)$ in the spatial amplitude spectrum of the periodic sampled driving function $Q_\Delta(x,\omega)$ would no longer overlap and no aliasing would be present in $Q_\Delta(x,\omega)$. Verheijen ([Verh97]) shows that this can be achieved by using directional microphones for recording the source field. If additionally the spatial components that are reproduced by the secondary source field $W(x,\vec{r}_R,\omega)$ are limited to the same range, then no spatial aliasing will be present in the reproduced sound field. In [Verh97] it is shown that this can be achieved by using directional loudspeakers. Finally, it can easily be shown that if the recorded source field contains components from angles between $-\theta_{\max,source} \le \theta \le \theta_{\max,source}$ and the sound field from the individual loudspeakers contains components with angles between $-\theta_{\max,LS} \le \theta \le \theta_{\max,LS}$, then the maximum temporal frequency for which no spatial aliasing is present in the reproduced field is given by:

$$f_{\max} = \frac{c}{\Delta x(\sin\theta_{\max,source} + \sin\theta_{\max,LS})}, \tag{2.27}$$

which in that case replaces (eq. 2.24).

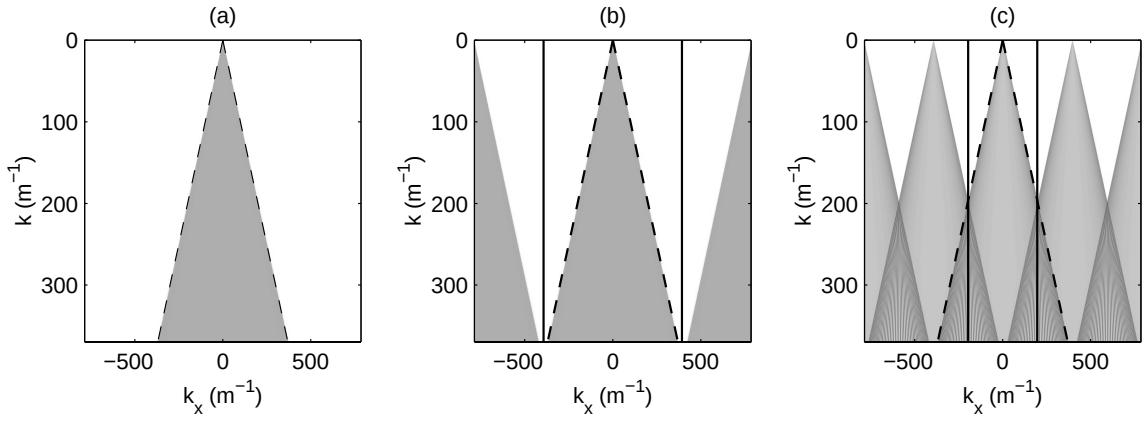


FIGURE 2.6. Spatial aliasing in the spatial Fourier domain. (a) Spatial amplitude spectrum of the continuous driving function (eq. 2.22) for a monopole point source, plotted as function of k_x (horizontal axis) and k (vertical axis). Level of grayness indicates amplitude. (b) Spatial amplitude spectrum of the sampled driving function, with Δx sufficiently small. Consecutive repetitions of the spatial amplitude spectrum of the continuous driving function do not overlap, so no spatial aliasing is present in the sampled driving function. (c) Spatial amplitude spectrum of the sampled driving function, with Δx too large. Consecutive repetitions of the spatial amplitude spectrum of the continuous driving function overlap, so spatial aliasing is present in the sampled driving function.

The two vertical solid lines in figures (b) and (c) are at $k_x=\pm\pi/\Delta x$ (half the period of the spatial amplitude spectrum). Also indicated in all three plots are the lines $k_x=k$ (dashed lines).

Now we will look at the effects of spatial aliasing of the synthesized wave field in the time- and frequency domains. The left plot of figure 2.7 shows the synthesized wave field of a point source located 1 m behind an array of secondary monopole sources with Δx=12.5 cm, which emits a broadband pulse. The synthesized field is recorded along a line parallel to and at 3 m distance from the array. We observe that first the intended wave front, corresponding to a point source at the virtual source position, reaches the receiver line, around t=12 ms at the center position x=0. Then, clearly, this first, intended, wave front is followed by additional, unintended, wave fronts. These are the result of the spatial aliasing. In the right plot of figure 2.7, the frequency response of the recorded signal is shown along the same receiver line. We see

that the fact that the intended wave front is followed by additional unintended events in the time domain, results in a significant distortion of the frequency spectrum for frequencies above the spatial aliasing frequency. Note that the spectral fluctuations are highly place-dependent.

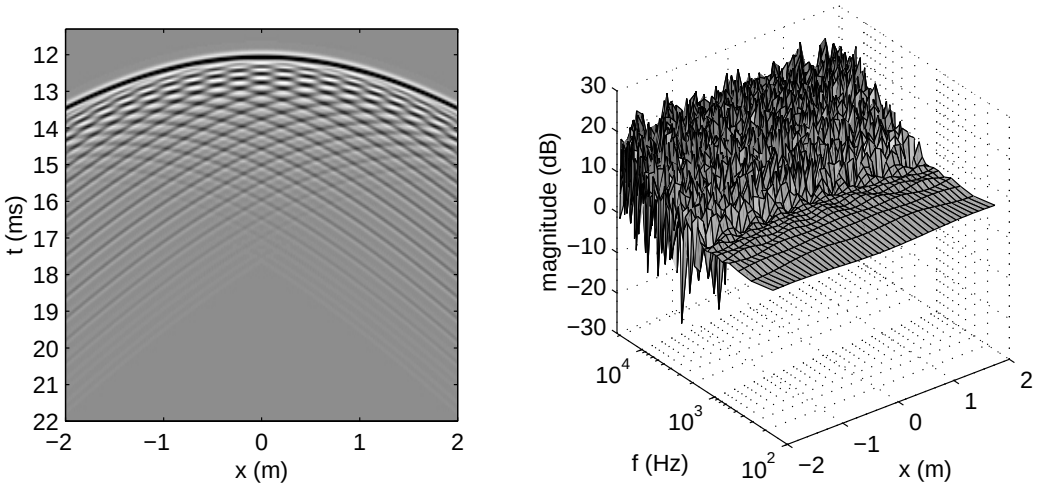


FIGURE 2.7. The effect of spatial aliasing on the synthesized wave field of a virtual point source located 1 m behind an array with a loudspeaker spacing of 12.5 cm. Left: time response along a line parallel to the array (distance 3 m). Right: frequency response along the same line.

It has been shown experimentally that as long as no spatial aliasing is present for frequencies below about 1.6 kHz, discretization of the loudspeaker array will not significantly degrade the localization of the synthesized virtual sound sources ([Voge93], [Star97]). This can be explained from the dominant role of interaural time cues, which are mainly effective for frequencies below about 1.6 kHz, in the localization of sound sources when both interaural time and intensity cues are available (see section 3.1.1). According to (eq. 2.24), this corresponds to a distance between the loudspeakers of about 11 cm. However, it has been shown experimentally that also for loudspeaker spacings that are considerably larger, localization of virtual sources is still quite accurate, although the virtual sound sources seem to become broader in this case, mainly for higher frequencies ([Voge93], [Star97]). As long as the spacing of the loudspeakers in the WFS array does not exceed the value of 11 cm too much, the main perceptual effect of spatial aliasing will therefore be a possible place-dependent coloration of the reproduced sound field, resulting from the distortion of the source spectrum, as shown in the right plot of figure 2.7. The coloration of the reproduced wave field due to spatial aliasing will be studied extensively in Chapter 5.

An additional effect of the discretization of the array is the fact that the $\sqrt{jk}$ filter in (eq. 2.22) is no longer valid for frequencies above the spatial Nyquist frequency. This filter can be interpreted as compensating for the frequency dependency of the response of a monopole line source, which is inversely proportional to $\sqrt{jk}$ ([Berk87]). For signal components below the spatial Nyquist frequency the array can be regarded as a continuous line of secondary monopole sources, so the filter is valid, but above that frequency, the loudspeakers act more like individual point sources, having a frequency-independent response. Thus, applying the $\sqrt{jk}$ filter to those frequencies overemphasizes them. This effect can be seen in the right plot of figure 2.7. We will also come back to this issue in Chapter 5.

2.3.2 Truncation Of The WFS Array

As was discussed in section 2.2.3, in the stationary phase approximation the synthesized wave field at a certain receiver position *R* is determined mainly by the secondary source at the intersection point of the line from the primary source to the receiver point and the array (figure 2.5). In case of an infinitely long array, there is always such an intersection point. However, when the array is truncated, there is only a limited receiver area for which there is a corresponding ‘stationary secondary source’. For receivers outside this area, no such stationary secondary sources exist, so the wave field is not synthesized correctly. Equivalently, it can be stated that the primary source position relative to the array and the array aperture together define the receiver area in which correct reconstruction can be achieved. The situation is sketched in figure 2.8 (left) and is analogous to an optical aperture that is illuminated by a light source (figure 2.8, right). The aperture and the position of the light source relative to the aperture together define the receiver area that is illuminated, while outside this area there is shadow.
Another analogy with the case of the optical aperture is the fact that diffraction effects occur at the boundaries of the aperture, i.e., at the edges of the loudspeaker array, which distort the synthesized wave field in the receiver area. The effect is shown in figure 2.9 (left), in which a time recording is shown of the wave field, synthesized by an array of length 1 m, of a point source located 1 m behind the array. The diffraction is seen as two wavelets arriving at the receiver line after the arrival of the main wavelet, which appear to be emanating from the edges of the array, as if there is an additional point source located at each edge. In [Vrie94] it is shown that this interpretation also holds mathematically.

FIGURE 2.8. **Left: truncation of the array limits the receiver area in which the sound field of the source (indicated as a point) is reconstructed. Right: the analogous optical situation, in which the receiver area that is illuminated by the light source is limited by the optical aperture.**

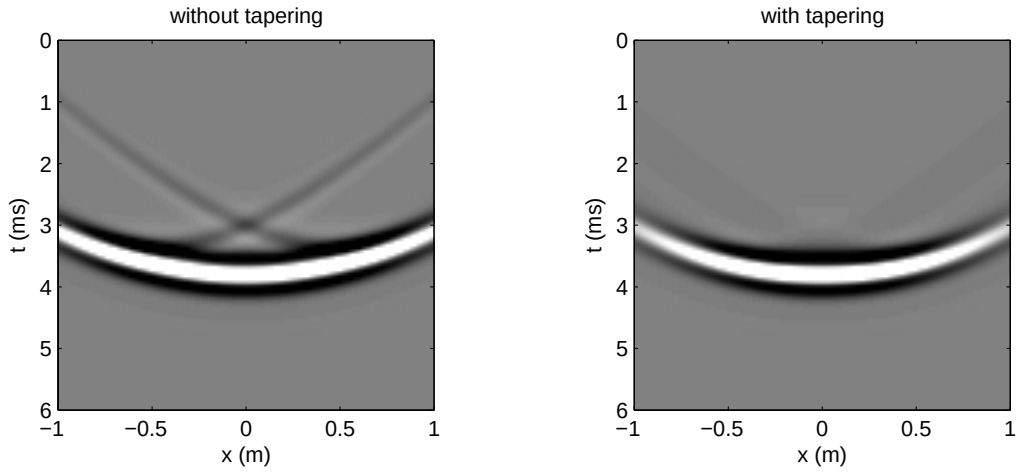


FIGURE 2.9. **Reduction of diffraction by tapering. The plots show the synthesized wave field of a point source located at 1 m behind an array with a length of 1 m, which emits a band-limited pulse. The wave field is recorded along a line parallel to and at 1 m distance from the array. Both plots have the same gray-level scale. Left: without tapering. Diffraction artifacts are clearly visible. Right: with a squared cosine taper applied, which reduces the amplitude of the driving function over one-third of the total array length at each side.**

A simple yet effective solution to avoid diffraction is to apply a spatial window to the loudspeaker driving functions, which smoothly reduces the amplitude to zero over the outer parts of the array, a technique known as *tapering*. Using such a taper, the discontinuities that occur at the edges of the array, which are the cause for the diffraction effects, are reduced, which reduces the diffraction artifacts. The wider the transition range of the taper, the more the diffraction effects are reduced. However, applying a taper also results in a reduction of the effective array length, so a trade-off must be made. A squared-cosine taper usually gives

satisfactory results ([Verh97]). In figure 2.9 (right), the effect is shown of applying a squared-cosine taper to the array, with a transition range at each side of one-third of the total array length. The reduction of the diffraction artifacts is clear, as well as the slight reduction of the amplitude of the intended wave front in the left- and right edges of the picture.

2.4 Applications Of Wave Field Synthesis

Much research has been done on the application of WFS in various specific areas of sound reproduction since the concept's introduction in the late 1980's. This section gives a brief overview of several of these applications. For each application, the specific implementation of the WFS concept, including aspects such as the recording of source signals, will be discussed.

2.4.1 Direct Sound Enhancement

In large room situations, for instance, in theaters and concert halls, there is often a need to increase the level of the direct sound of the sound sources on stage, for instance, the voices of actors or musical instruments, so that everyone in the audience is able to hear them properly. For this purpose, most theaters are equipped with a so-called public address system, which amplifies the signals picked up by microphones and sends them to loudspeakers. Typically, the loudspeaker system consists of a stack of loudspeakers on each side of the stage. This often introduces problems, especially for people who are close to the stage, as they will see the actors on stage while their voices are clearly coming from the loudspeakers. This can be very disturbing. Furthermore, such a set-up provides very little, if any, spatialization of the sound sources. For the audience, it would be much more natural if the amplified voices would seem to come from the positions of the corresponding people on stage. This can be achieved by using WFS for reproduction of the voices.

Vogel ([Voge93]) did the first experiments on WFS and designed a prototype WFS reproduction system, of which the main intended purpose was speech amplification. This resulted in the installation of the first WFS system in a theater. With the first experiences with WFS now gained, Start ([Star97]) continued these investigations and designed and implemented a WFS system for direct sound enhancement. The basic set-up is shown in figure 2.10. A horizontal linear array of loudspeakers is mounted above the front of the stage. Depending on the shape of the hall, the shape of the array can be adapted to optimize the aperture in relation to the audience area. Since the array is located several meters above the stage, there is the risk that the

audience will localize the sources at the height of the array. This risk is reduced by making use of the well-known ‘precedence’ (or ‘Haas’) effect. By slightly delaying the signals that are reproduced by the WFS array, it is ensured that the direct sound from the sound sources on stage always reaches the listeners before the sound from the array, so that it dominates the vertical localization.

Recording of the source signals in this application is preferably done by using a close-miking technique, i.e., each source has an individual microphone, very close to that source, which almost exclusively picks up the direct sound of that single source. In the case of static sources, the microphones can be at fixed positions (the ‘fixed microphones’ concept, explained in section 1.3.1). The microphone signal of each microphone is then reproduced by the WFS array as a ‘notional source’ located at the position of the microphone. When the sources are moving, as is usually the case in theater situations, wireless tie-clip microphones are used. As explained in section 1.3.1, this introduces the need of a source-tracking system that provides the WFS system with the current source locations, so that the WFS system can reproduce the microphone signals as notional sources located at those current positions.

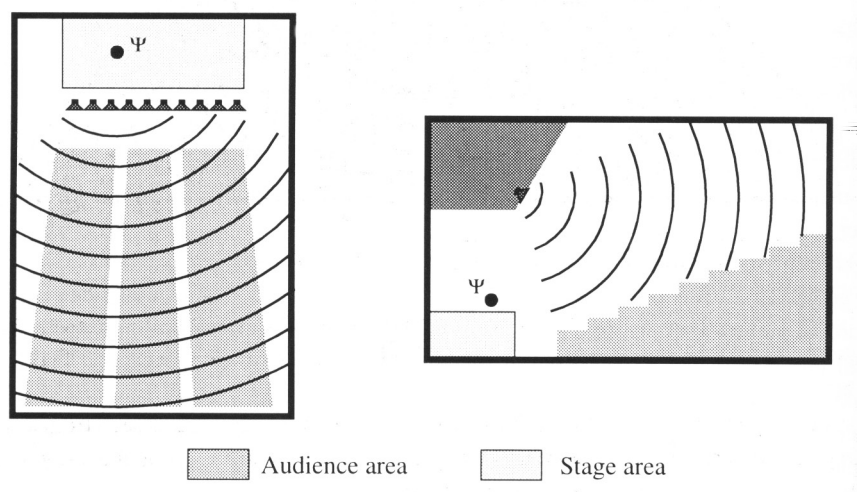


FIGURE 2.10. Schematic illustration of application of WFS for direct sound enhancement in a theater. Left: top view. Right: side view (from: Start ([Star97])).

2.4.2 Sound Reproduction

A second main area of application of WFS is in *sound reproduction*, by which is meant the reproduction of an acoustical event recorded at another location and/or time. As explained in the introduction of this chapter, conventional sound reproduction systems, such as the common stereophonic and stereophony-based surround systems, are incapable to make a truly convincing reproduction of a real acoustical event, especially regarding its spatial properties. With WFS, on the other hand, it is possible to achieve this, as will have become clear in this chapter. Verheijen ([Verh97]) did a thorough study to optimize the WFS concept for this application. The basic sound reproduction concept is shown in figure 2.11. On the left, the recording site is shown, with a stage area and an audience area. In the reproduction room, shown right, a rectangular linear loudspeaker array encloses the listening area in which we wish to reproduce an equally-sized part of the original sound field, indicated as a shaded rectangular subarea of the audience area in the recording room. In other words, the aim is to make such a reproduction within the listening area that listeners within this listening area perceive the same sound field as they would have perceived if they had been present in the dashed area of the recording room at the time of the recording.

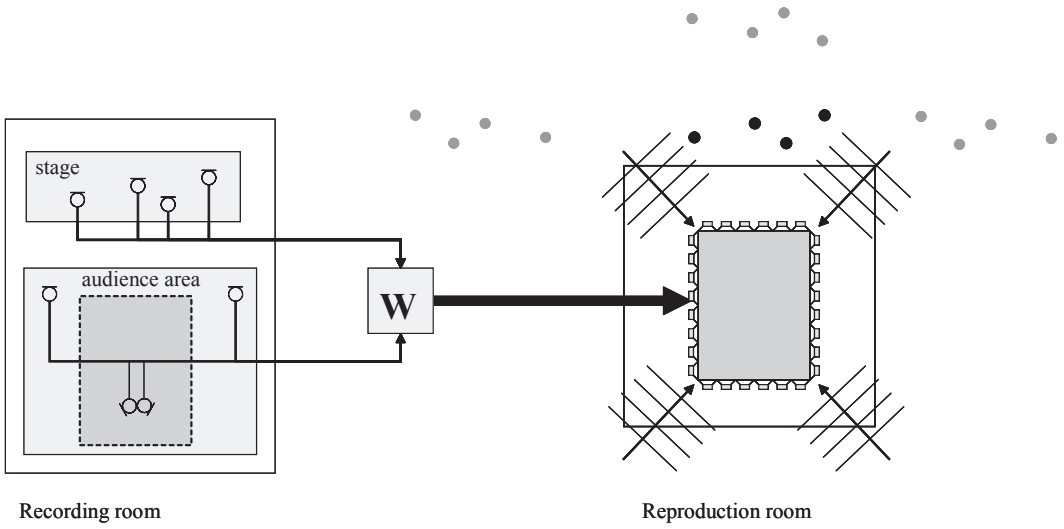


FIGURE 2.11. Schematic representation of a WFS system for sound reproduction (after Verheijen ([Verh97]))

In Verheijen's approach, the sound field that is to be recorded and reproduced is divided into three temporal parts: the *direct sound field* of the sound sources, the *early reflections* and the *reverberant sound field*.

The direct sound of each source is recorded using a spot microphone, i.e., a (usually directional) microphone very close to each individual source. These microphones are indicated in the stage area of the recording room in figure 2.11. The signals of these spot microphones are sent to the reproduction room as individual channels, together with their spatial coordinates. In the reproduction room, each spot microphone signal is processed by the WFS processor (indicated by the block labelled '**W**') and reproduced as a virtual point source by the array. This way, a very realistic reproduction can be made of the original source configuration on the stage of the recording room. However, since the source signals were recorded directly at the sources, it is also possible to reproduce the signals as virtual point sources at any other desired position, by simply changing the source coordinates in the calculation of the loudspeaker driving functions (eq. 2.22). This opens up the possibility of changing the configuration of the sound sources as desired, which makes this concept very flexible and very attractive for interactive sound reproduction applications. In the case of many sound sources, for instance, when recording an orchestra, recording each instrument on a separate channel is not practical. In that case, Verheijen proposes to record instruments by section rather than by individual instrument. The basic scheme of figure 2.11 remains the same, with the remark that each spot microphone on the stage now covers a certain part of the stage area. In the reproduction, the signal of the microphone is reproduced as a virtual point source located at the virtual center of the corresponding stage area.

Recording of individual reflections is not feasible in practice. However, when the geometry of the recording room and the positions of the sound sources are known, reflections can be simulated by applying the well-known mirror image source model and synthesized as virtual point sources in the reproduction room. In figure 2.11 some of these synthesized reflections are indicated as light-gray copies of the direct sources. A simpler alternative is to record and reproduce the reflections together with the reverberation, as will be explained below.

For the recording of the reverberant part of the sound field various microphone configurations can be used. In figure 2.11 a configuration is shown with two microphones in the front of the audience area and two microphones more to the back. The main requirements for the reverberation microphones are that they pick up as little as possible of the direct sound and that the sig-

nals of the microphones are highly uncorrelated, to avoid coloration in the reproduced sound field. In [Brui98a] and [Brui98b], some microphone techniques are proposed for recording the reverberant part of the sound field. By careful placement of the reverberation microphones, they can also be used to simultaneously pick up some early reflections. The recorded reverberation signals are reproduced in the reproduction room as plane waves from different angles, as shown in figure 2.11. Alternatively, the reverberation signals can be generated by electronic reverb devices instead of recording them live.
According to the concept described above, Verheijen designed and implemented a 128-channel demonstration system for sound reproduction, with a listening area of 4 x 6 m, which was permanently installed in the laboratory of Acoustic Imaging and Sound Control and has been used for many WFS demonstrations. Details of this demonstration system can be found in Verheijen's Ph.D. thesis ([Verh97]).

An alternative approach to sound reproduction by WFS, in which the sound field is recorded as a whole rather than discriminating between direct sound, early reflections and reverberation, has recently been suggested by Hulsebos et al. ([Huls02a]). They propose a circular microphone array consisting of 288 small electret cardioid microphones, which is placed in the recording room at the center of the area which is to be reproduced in the reproduction room. Before conversion to the digital domain, the 288 individual channels are reduced to 24 channels, which together cover the full 360 degrees, by combining the individual microphone signals in an analog way, which can be seen as a circular spatial low-pass filtering operation. After acquisition, these 24 channels can be combined to obtain the signals of virtual microphones aiming in any desired direction and having a desired directivity pattern. These synthetic microphone signals are then reproduced in the listening room as plane waves from the corresponding directions. In this way, it is possible to obtain a very realistic spatial reproduction of the original sound field, including the reflections and reverberation. Of course, the flexibility of Verheijen's concept to position the direct sound sources at any desired position at the time of reproduction is lost using this technique. Which of the two concepts is preferred depends on the situation.

Compatible reproduction. A special case of WFS sound reproduction is the reproduction of audio material recorded for a conventional multi-channel format such as 5.1 surround, using the concept of *virtual speakers (*[Boon99]). In typical living room situations, it is often not possible to meet the standardized requirements for placement of all five loudspeakers, due to lack of space, furniture etcetera. Using WFS, it is possible to solve this problem by reproducing the individual channels as virtual point sources located at the standardized positions. Furthermore, it has been shown ([Boon99]) that by reproducing the surround channels as plane waves from the standardized directions rather than as point sources, the 'sweet spot' of the surround reproduction can be enlarged significantly. First of all, because plane waves are not localized as coming from a certain position but only as coming from a certain direction, the exact listening position on the front-back axis becomes less critical, as the surround signals are always perceived as coming from the standardized directions. Furthermore, because of the $1/\sqrt{r}$ decay of the amplitude of the plane waves instead of the 1/r decay of point sources, the level balance between left- and right surround becomes more homogeneous in the left-right direction, so also the exact listening position on the left-right axis becomes less critical. This concept has recently been implemented for the first time in a small commercial cinema in Ilmenau, Germany ([IOSONO]), where besides the surround channels also the left- and right front channels are reproduced as plane waves, resulting in the same benefit of a more homogeneous level balance for those channels. The center channel is always reproduced as a point source, since it is usually associated with the dialog and should therefore be localized at the center of the screen by all listeners.

2.4.3 Variable Acoustics And Auralization

Today, there are many so-called multi-purpose halls which are used for a wide range of applications, ranging from lectures to theater and orchestra performances. This leads to a desire to be able to change the acoustics of the hall, since every type of performance requires different acoustics. The conventional passive means, like absorber- and reflector panels, are not able to provide the needed flexibility. Using electro-acoustics, however, much more can be done, especially if the room itself is relatively 'dry' so that the main intention will be to *add* reflections and reverberation.

A related but more complicated desire is the wish to be able to realistically recreate the acoustics of a certain hall in another room, a concept known as *auralization*. For a variable acoustics

system, it is sufficient to generate a sound field which fulfills some global objective and subjective requirements, such as reverberation time, clarity, homogeneity, etcetera and which does not suffer from artifacts such as strong coloration and so-called ‘flutter’ echoes. In the auralization case, however, the generated acoustics should resemble the acoustics of the original hall. This requires detailed information about the sound field in the original hall.

Both the variable acoustics and auralization concept can be implemented very well by WFS. The general concept is shown in figure 2.12. In the original hall, on the left, a sound source is placed at a representative source position and the impulse response of the hall is measured at many closely spaced positions along a line or lines in the hall, by both pressure and velocity microphones. Since usually the acoustic properties of a hall can be considered to be time-invariant, it is possible to use a single microphone mounted on a rail that is automatically moved along the measurement line. This makes the measurements time-consuming, but realizable in practice. In figure 2.12 two perpendicular measurement lines are shown, but other configurations are possible, see [Huls02a] for a detailed study. From these measurements, by using WFS theory, it is possible to extrapolate the measured impulse responses to any desired position in the hall, so that effectively, the impulse response in the whole hall is known. Now imagine we want to recreate the area indicated by the dashed rectangle in the reproduction room, shown right. In order to achieve this, the measured impulse responses are extrapolated to the boundaries of this rectangle (the extrapolation operator is indicated by the block labeled ‘**W**’). In the reproduction room, the desired listening area is surrounded by a rectangular linear loudspeaker array. The driving signal for each loudspeaker is calculated by convolving a dry source signal with the corresponding extrapolated impulse response. The result is a close reproduction of the sound field that would have been perceived in the dashed rectangular area in the original room, had the reproduced source been present in that hall at the measurement source position.

A disadvantage of the concept described above is the enormous amount of data that has to be stored and the computational power that is needed to perform a convolution with a full impulse response for each individual loudspeaker. Fortunately, the measured impulse responses contain a lot of perceptually redundant information, so a significant amount of data reduction is possible.

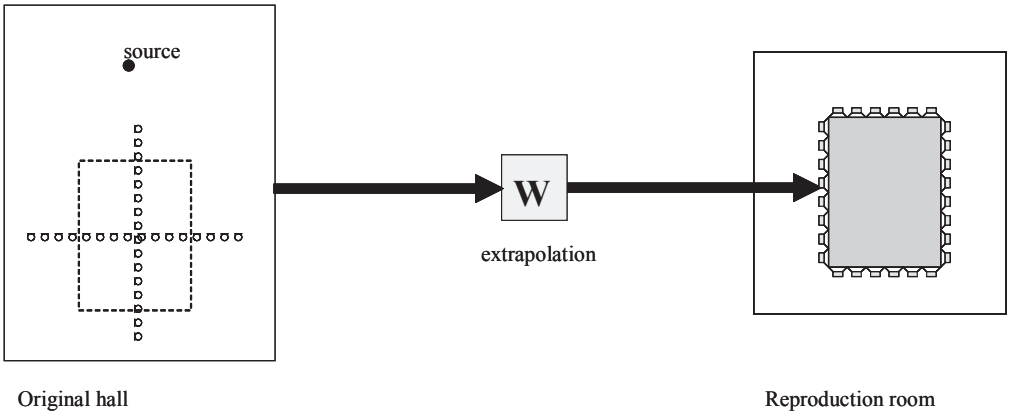


FIGURE 2.12. Schematic representation of the concept of auralization.

Sonke ([Sonk00]) describes two methods to reduce the amount of data. One is spectro-temporal parameterization of the individual impulse responses and the other is spatial parameterization of the reverberant field. The spectro-temporal parametrization means that the impulse response is divided into two parts: the direct sound plus early reflections and the reverberation. The former part is reproduced accurately, as measured in the original room, since direct sound and early reflections contribute individually to the perceived sound field. The reverberant part is considered to be stochastic and is described using a limited number of global parameters, for instance, the decay rate of the energy. In this way, the amount of data needed for a single impulse response can be reduced significantly, while perceptually it is almost indistinguishable from the original response.

The spatial parametrization consists of reproducing the reverberant part by only a limited number of plane waves from different directions. A true reverberant field is considered to be isotropic, which means that the same amount of reverberant energy is reaching the listener from all possible directions. In theory, this means that an infinite number of plane waves would have to be reproduced to achieve a natural reproduction of a reverberant field. However, Sonke shows that for most listeners no more than 11 plane waves, equally distributed over all directions in the horizontal plane, are needed to achieve a reverberant field, which, with restriction to the horizontal plane, is perceptually indistinguishable from a true isotropic reverberant field.

Hulsebos ([Huls02a]) has refined these techniques, by using a circular microphone array to capture the impulse responses of the original hall and by using an improved way to parameterize the measured responses.

A complicating factor in the auralization concept as described above is that a set of measured impulse responses is in principle representative for a single source position only. When the position of the source in the original hall changes, the reflection pattern changes as well, so in principle this means that a set of impulse responses has to be measured for each individual source position that is to be used in the reproduction. For the reverberant part of the impulse response this will in general not be necessary, since it is assumed to be diffuse, so that the same reverberant ‘tail’ of the impulse response can be used for a wide range of source positions. However, as was stated above, early reflections can contribute individually to the perceived sound field, so in order to achieve a convincing spatial reproduction of the original sound field for various source positions, the early part of the impulse responses has to be measured or calculated with sufficient accuracy for all those source positions. An interesting question is how large variations, in terms of angles of incidence and time delay relative to the direct sound, in early reflection patterns have to be, before a change in the sound field is perceived. When all reflection patterns corresponding to sources located within a certain range of source positions are perceived as being identical or almost identical, then the same set of early reflections can be used to auralize the sound field corresponding to all those source positions. Boone ([Boon04a]) describes a first experiment that addresses this issue, in which the sensitivity for changes in angle and delay of a single reflection is investigated.

CHAPTER 3 *Spatial Perception and Audio-Visual Interaction*

When an audio system is to be designed, it is essential that sufficient knowledge about the human perception of sound is included in the designing process. Knowing the properties of the hearing system, it is possible to determine which attributes of the sound field that is to be reproduced should be presented accurately to the listener's ears, so that the listener experiences the auditory sensation that is intended. For an efficient design, it is equally important to understand which attributes of the sound field do *not* have to be reproduced with high accuracy to achieve the intended auditory sensation, since this might significantly reduce the complexity and cost of the system in terms of number of audio channels required in recording, transmission and reproduction, the required bandwidth of the audio signal, etcetera.

Likewise, it is essential that sufficient knowledge about the human visual system is included in the designing process of a visualization system.

Consequently, in the designing process of a system that combines audio and video, like, for instance, a videoconferencing system, it is necessary to have sufficient knowledge about the perceptual mechanisms of both modalities. This, however, is not sufficient, because in general the combination of audio signals and video images results in an overall audio-visual perception that is not just the sum of the perceptions in the two individual modalities, but includes effects

that are the result of sensory interactions between the auditory and visual modality, so-called *audio-visual interaction* effects.

This chapter gives an overview of the most important aspects of both auditory- and visual perception of space and of spatial audio-visual interaction. These spatial aspects of auditory-, visual- and audio-visual perception are of particular interest to this project, since one of its main objectives, as defined in section 1.4, was to develop a videoconferencing system that reproduces the spatial aspects of the voices of conference participants in a natural way.
First, in section 3.1 an overview is given of some important aspects of spatial perception in the auditory modality: horizontal plane localization (section 3.1.1), median plane (or vertical) localization (section 3.1.2) and distance perception (section 3.1.3).
Then, in section 3.2 several aspects of perception of space in the visual modality are discussed that are relevant to this project. The section starts with a brief discussion of the perceptual mechanisms that are involved in 3D vision (section 3.2.1), followed by a review of the history and technologies of 3D visualization systems (section 3.2.2). Section 3.2 ends with an analysis of what happens to the visual perception of space when viewing a 2D perspective projection of a 3D visual scene (section 3.2.3).
Section 3.3 then continues with a discussion of the phenomenon of audio-visual interaction that is of great importance in the design and evaluation of audio-visual systems, such as the life-size videoconferencing system that is the objective of this project, focussing on the interaction of the perceptions of space in the two individual modalities (section 3.3.1). The chapter closes with a discussion of the possible consequences of all the above when true perspective, i.e., including a realistic reproduction of depth, audio reproduction, such as WFS, is combined with 2D video projection (section 3.3.2).

3.1 Overview Of Auditory Spatial Perception

3.1.1 Horizontal Plane Localization

The way humans localize the direction of sound sources that are located in the transverse plane (the horizontal plane through the ears, figure 3.1) has been the subject of many studies since many years, so the characteristics of this horizontal plane localization and the mechanisms that it is based upon are quite well understood.

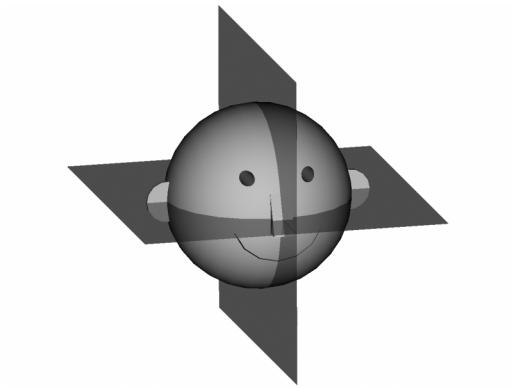


FIGURE 3.1. The transverse (horizontal) and median sagittal (vertical) plane.

The classic localization studies by Blauert ([Blau83]) show that the localization blur, the smallest change in source direction that is perceived by 50% of human subjects, is in the order of 1 to several degrees for sources located directly in front of the subject, depending on the type of source signal. For sources located in other directions in the horizontal plane the localization blur is significantly larger, with maximum values for sources to the sides of the subject, at angles of 90 and 270 degree angles.

Horizontal plane localization is based mainly on two separate physical cues:

- 'Interaural Time Differences' (ITD's).
- 'Interaural Level Differences' (ILD's).

Interaural Time Differences (ITD's). Because of the spatial separation between the two ears, the sound waves from a source do not reach both ears simultaneously but with a small difference in arrival time, corresponding to the difference in distance that the sound waves have to travel from the source to each ear. This time difference is zero for sources located in the median sagittal plane, the vertical plane that divides the head in left- and right half, and has its maximum value for sources at the sides, about 0.6-0.7 ms.

The time difference between the signals received at the left- and right ear is evaluated by the hearing system by performing a sort of cross-correlation process between these two signals. The ITD is estimated from the maximum of this interaural cross-correlation function within a certain range of possible delays, which is related to the maximum interaural delay that occurs for natural sources. For signals with wavelengths shorter than the maximum difference in

travel distance to each ear, about 21 cm, corresponding to frequencies of 1.6 kHz and higher, the interaural cross-correlation function has more than one maximum within the possible range, so the mechanism is not able to provide an unambiguous estimate for the ITD and the ITD cue no longer results in reliable localization. However, for signals that contain only signal components above 1.6 kHz, the hearing system is also able to extract a useful ITD cue from the time difference between the amplitude envelopes of the left- and right ear signals. In this case the hearing system ignores the temporal fine-structure of the 'carrier' signal and only uses the envelope for determining the time difference.

Interaural Level Differences (ILD's). For sound waves of which the wavelength is in the order of the the size of the head or smaller, corresponding to frequencies above approximately 1 kHz, shielding of the sound waves by the head becomes significant. This results in sound pressure level differences between the signals received by each ear that can be as large as several tens of decibels. These Interaural Level Differences (ILD's) are also used by the human hearing system as a cue to localize sound sources.

In headphone experiments it is possible to study the perceived left-right position of a sound source as a function of the time difference between the ears only. In this case, the sound source is usually not perceived as being located at a position outside the head, but instead it is perceived as having a certain position on the line between the two ears, so that it appears to be located inside the head. In this case one speaks of 'lateralization' rather than 'localization'. Provided that the signal contains frequencies below 1.6 kHz, the perceived lateral position between left- and right ear changes almost linearly from completely left to completely right for time differences between the ears ranging from -630 to +630 μs ([Blau83]), which indeed corresponds closely to the maximum difference in travel distance to each ear.
Similarly, it is possible to carry out a headphone experiment to study the lateralization when only ILD cues are available. Also in this case there is a more or less linear relationship between the level differences between the ears and the perceived lateral position, but the relationship is strongly frequency dependent ([Blau83]). It is good to note that although ILD cues do not occur 'naturally' for frequencies below about 1 kHz, an artificially generated interaural level difference for signals with lower frequencies still results in clear lateralization of the source signal, so the effectiveness of ILD cues is not limited to a certain frequency range, as is the case with ITD's. This property is used to great effect in music production, where 'intensity

panning' of source signals is used to position sound sources somewhere between left and right in the stereo image.
Finally, when both ITD and ILD cues are available to subjects, the ITD cue dominates the source localization, which can be concluded from experiments in which conflicting ITD and ILD cues are presented to subjects ([Wigh92]).

Although ITD's and ILD's are the most important cues for horizontal plane localization they are not sufficient. For instance, both the ITD and ILD are zero for sources both in front and at the back of the head, so it is not possible to discriminate between them on the basis of ITD and ILD cues. To resolve this kind of ambiguities, spectral cues and information obtained from small head movements are used.

3.1.2 Median Plane Localization

Localization of sound sources in the median sagittal plane (the vertical plane through the center of the head, perpendicular to the line through the ears, figure 3.1) is essentially a monaural process ([Hebr74]), contrary to localization in the horizontal plane for which, as was explained in the previous subsection, the source direction is evaluated mainly by using interaural cues: interaural time differences (ITD) for frequencies below about 1.6 kHz and interaural level differences (ILD) for higher frequencies.
The main cue for median plane source localization is the spectrum of the signal at the eardrum, which is modified by source direction-dependent filtering by the outer ear, the *pinna*. Reflections on the pinna, having a delay of 10 to 300 μs relative to the direct signal path ([Hebr74]), cause a 'comb filter'-like response with peaks and dips, the positions of which depend on the angle of incidence.
Blauert ([Blau69]) has shown that for narrow-band source signals (1/3 octave-wide noise bands) in the median plane, the source direction that is perceived is determined completely by the center frequency of the signal and does not depend at all on the actual source direction. Specific frequency bands appear to correspond to specific perceived directions, a concept Blauert referred to as 'directional bands'. So, in the case of narrow-band signals it is almost impossible for a listener whose head is fixed to determine in which median plane direction the source is located.
In the localization mechanism of broad-band sources in the median plane, these directional

bands play a crucial role. The source signal is filtered by the pinna, depending on the direction of incidence, such that signal components in certain directional bands are emphasized, while others are suppressed. The directional bands containing the most signal power (or energy, in the case of impulsive signals) after this filtering process determine the elevation that is perceived. This model applies to the majority of 'normal' signals, which explains how it is possible that also signals that are relatively unfamiliar to a listener can be localized with reasonable accuracy, although being familiar with the original spectrum of the source makes median plane localization more accurate in general ([Hebr74]).

Localization blur in the median plane is minimal for directions straight ahead. Under conditions in which the subject’s head is immobilized it is about 4 degrees for white noise, 9 degrees for speech of a familiar person and 17 degrees for an unfamiliar voice. For elevated sources in the median plane the localization blur is larger ([Blau83]).

This mechanism of directional filtering of the source signal by reflections on the ear is only effective for signals containing sufficient energy above a certain frequency, a phenomenon which is related to the physical dimensions of the ear. Extreme values for this critical frequency that have been found are 4 kHz ([Watk78]) and 7 kHz ([Hebr74], [Roff67]). In conclusion, source signals should contain a sufficient amount of high-frequency energy in order to make localization in the vertical plane possible at all.

In addition to the static spectral cues described above, which are available to a subject whose head position is fixed, dynamic spectral cues are provided by small head movements. These head movements can cause changes in the spectra at the eardrums in the order of 0.25 dB/ degree ([Wigh97]). Since unconscious head movements of about 2 degrees are normal and 0.25 dB differences between spectra at high frequencies are detectable ([Wigh97]), detectable spectral changes can easily occur, thus providing an additional cue that enables median plane localization with a somewhat higher accuracy.

Finally, although median plane localization is mainly a monaural process, there is some evidence that also interaural spectral differences caused by small differences between the left- and right ear pinna are of some importance ([Sear75]).

3.1.3 Distance Perception

The human ability to estimate the distance of an audio source is rather limited. The accuracy of distance estimation is low and as sources move further away from the listener, the ratio of the increase in perceived distance and the corresponding true increase of source distance becomes smaller, a phenomenon sometimes referred to as the 'auditory horizon' effect ([Bron99], [Loom98]).

There are several cues that the auditory system can use to estimate distance in the case of a static source and listener. The first cue is the loudness of the source signal: the sound pressure level of the source signal changes according to the $1/r$ law. For this cue to be meaningful, it is necessary that the listener has a reference regarding the relation between the distance and the absolute loudness of the source in question.

A second static cue which becomes relevant for sources at larger distances is the frequency content of the source signal. Due to the frequency dependent absorption in air, the source signal is low-pass filtered during the propagation from the source to the listener. The longer the path, the more the high-frequency content of the source signal will be filtered out. However, as in the case of the loudness cue, the listener must have a reference regarding the relation between the distance and the spectrum of the source signal. It is well known that in static, anechoic conditions, where loudness and spectrum are the only distance cues available, it is not possible for listeners to estimate the distance to an unfamiliar sound source with reasonable accuracy ([Cole62]), while even for familiar sources the distance estimation under these conditions is quite poor. This is illustrated informally in [Gard68], where it is mentioned that in an anechoic environment, people were unable to correctly identify which of several sound sources that were placed on a line straight ahead of the listener was reproducing speech sound. Instead, the subjects consistently indicated in all cases that the sound was coming from the speaker closest to them, which was the only one they could actually see as the other speakers were located behind it.

More accurate distance estimation is possible under non-anechoic conditions. In this case the so-called direct-to-reverberant energy ratio, the ratio of the mean square pressure of the direct sound and reverberant sound at the listener position, is available as a distance cue. If the reverberant field is assumed to be diffuse, the RMS sound pressure level of the reverberant field is homogeneous in the whole room. The sound pressure level of the direct sound, however,

depends on the source-listener distance according to the 1/*r* law. Bronkhorst and Houtgast ([Bron99]) give a model that describes the subjective distance d_s (in m) as:

$$d_s = Ar_r \left(\frac{E_r}{E_d} \right)^j, \tag{3.1}$$

in which *A* and *j* are constants (with experimentally determined values of *A* ranging from 1.2 to 2.7, depending on the listening conditions and a value of approximately 0.5 for *j*), E_r and E_d are the energies of the reverberant and direct sound fields, determined by integration of the squared impulse response over the interval 0-6 ms and 6-∞ ms respectively, and r_r is the 'reverberation distance', the distance at which the RMS sound pressure levels of the reverberant and direct sound field are equal. It is given by ([Kutt73]):

$$r_r = 0.1 \left(\frac{GV}{\pi T} \right)^{\frac{1}{2}}, \tag{3.2}$$

in which *G* is the directivity factor of the source, *V* is the volume of the room and *T* is the reverberation time (in s).

Additionally, besides the above mentioned static cues, there are also dynamic cues available for estimating auditory distance. The first one is 'motion parallax': the fact that the perceived direction of an auditory source and the angles between different auditory sources change when the listener moves in a direction other than straight towards or moving away from the source, or, equivalently, when the source or sources are moving. The second dynamic cue is the rate of change of the perceived loudness of the source signal when listener and source are moving relative to each other. These dynamic cues, however, seem to play only a limited role in auditory distance estimation compared to the static cues ([Spei93]).

3.2 Overview Of Visual Spatial Perception

For an efficient design process of an audio-visual system it is of course necessary to have a good understanding of the human visual perception system. Even in cases when the focus of research is primarily on the audio part of the system and not on the visual part, knowledge of the visual perception mechanism is essential because of the fact that the perception of the reproduced sound is influenced by the perception of the visual image that the audio is combined with and vice versa. This so-called '*audio-visual interaction*' plays an important role in this thesis, particularly in chapter 4, in which several audio-visual perception experiments are described that were performed to investigate the effects that arise from the combination of spatialized audio reproduction and two-dimensional (2D) video projection. The topic of audio-visual interaction is the subject of section 3.3.

While an extensive discussion of the physiological and cognitive aspects of the human visual system is beyond the scope of this thesis (the interested reader is referred to one of the many books that have been written about this subject, for example [Buse97]), the issues of spatial vision that are of particular importance to this research project are reviewed in this section. First, an overview is given of the mechanisms involved in three-dimensional (3D) vision, then some attention is given to 3D visualization systems and finally a discussion is given of the 3D interpretation that is derived by the human cognitive visual system from looking at a 2D projection of a real-life, 3D, situation. Later on in this thesis, this last subject will prove to be of particular importance when spatial audio reproduction that includes a realistic reproduction of depth is combined with 2D video projection.

3.2.1 3D Vision

The mechanisms for human visual perception of depth can basically be divided into two categories: those that are based on so-called binocular (two eyes) cues and those that are based on monocular (one eye) cues.

By far the most important and accurate mechanism of visual depth perception uses the cue of **binocular disparity** (or 'stereoscopic parallax') which results from the fact that humans have two eyes that each provide an image of the visual scene from slightly different angles. Neurological processing of the two images provides vital information about the spatial layout of a scene and enables quite accurate estimation of absolute and relative distances of objects ([Buse97]).

Besides the binocular disparity cue, the human visual system can also to some extent make use of several **monocular** cues to get depth information from a visual scene.

The most obvious monocular cue is the perceived angular *size* of an object, both absolute (for absolute distance estimation) and relative to other objects (to evaluate the spatial layout of objects relative to each other). Prior knowledge of the absolute and relative sizes of the objects is required to extract meaningful depth information from these cues.

Closely related to this is the concept of *linear perspective* of a visual scene. This is the geometrical phenomenon that parallel lines extending away from the viewer appear to converge to one point, the 'vanishing point'. Furthermore, all planes extending away from the viewer that are parallel to each other seem to converge to one line, which for planes that are parallel to the ground is a horizontal line located at eye level, the 'horizon'. The concept of linear perspective is probably the strongest technique that is used to suggest depth in two-dimensional graphic art.

Another static monocular cue that provides basic information about the spatial layout of objects relative to each other is *occlusion* (or overlap) of part of one object by another object, while in everyday life additional cues, such as light and shadow effects, color effects, sharpness, etcetera, help to create an impression of depth.

There is also an important dynamic monocular cue for visual depth perception, namely *motion parallax*. When an observer is moving or moves his head, his viewpoint changes and consequently his view of objects in the visual scene and their orientations relative to each other changes. This is an important cue for resolving the spatial layout of objects in a visual scene. Additionally, when an object is moving with a certain linear velocity in a direction with a lateral component relative to the viewer, the angular velocity of the object, relative to the eye, depends on the distance to the viewer. The closer the distance, the higher the angular velocity of the object seems to be. Of course the same applies when the observer is moving and the visual objects are static. In everyday life this phenomenon is experienced, for instance, when looking out of the window of a moving train: trees that are close to the railway appear to move very fast, while far-away trees appear to be almost static.

Finally, for short range (order of 1 m ([Buse97])) depth perception, also a physiological cue is available in the form of the *accommodation* of the lens of the eye: depending on the distance of the object, the lens has to be accommodated to obtain a sharp image on the retina. This 'level of accommodation' is also believed to be used by the human visual system as a distance cue.

3.2.2 3D Visualization Systems

Since centuries, humans have tried to develop visualization systems that are able to reproduce a realistic sensation of depth. Already as early as in the 3rd century B.C., Euclid described how binocular disparity is the basis for visual depth perception. The first systematic efforts to reproduce depth in visual images were made in the Renaissance during the 16th century, when the concept of perspective was studied extensively and was used by painters like Da Vinci to suggest depth in their work. However, it was Wheatstone who in 1833 demonstrated the first optical system that successfully used the binocular disparity cue to reproduce a convincing sensation of depth by sending slightly different images to each eye ([Buse97]). This concept of simulating the binocular disparity cue by sending separate images to the left and right eye has since then been the basic principle of most 3D visualization systems, up to the present time.

There have been many variations in the implementation of the concept. The earliest 'stereoscopes' relied on providing two individual pictures, each viewed by only one eye. A very popular example of this was, and still is, the 'View-Master' ([ViewMa]): a children's toy which basically is a pair of spectacles which can be loaded with a disc containing a whole series of pairs of stereoscopic pictures.

A little later, techniques were developed that combined the left- and right eye pictures into a single stereoscopic picture. The first technique that was used for this was the so-called 'anaglyph', in which the left- and right eye pictures were combined by printing one in green and the other in red in the same picture, slightly shifted relative to each other. When viewed with special glasses with one red and one green glass, the two pictures were effectively separated and each eye only received its corresponding picture. This technique was soon also used in motion pictures as well and experiments with 3D television broadcasts based on this principle have also been carried out, not to much success because of the unnatural colors and the tiring effect on the eyes (and the resulting headache). Later, a more sophisticated technique was developed that instead of projecting the two images in different colors uses differently polarized light for the two projected images. When viewed with special glasses of polarizing material ('Polaroid' glasses) each eye again only receives its own image. This enabled the production of full-color 3D movies, a technique which is used for example in many of the well-known Imax theatres ([IMAX]) that can be found all over the world.

In more recent years, thanks to the advancements made in micro-electronics, another technique has become popular in which the left- and right eye image are projected not simultaneously but

alternating at a rate that is sufficiently high not to be noticed by the spectators. Viewers wear glasses with electronic shutters that alternatively shut the left-and right eye, synchronized to the projector. This is a technique that is also used in many modern ‘3D cinemas’, including Imax theatres ([IMAX]).

Although all the systems described above succeed to some extent in producing images that suggest depth by simulating the binocular disparity cue, a limitation shared by all these systems is that they do not include the motion parallax cue, but provide an image for a single-viewpoint only, i.e., the projected images (and therefore the perceived 3D image) do not change when the viewer changes position or orientation as is the case in real life and thus should be the case for a truly convincing 3D projection. This means not only that viewers are unable to view a projected object from different perspectives (as in: change position to look at for example the left side of the object), but the single perspective that is provided is also distorted to a certain extent for all but one viewpoint.
With the powerful computers of today it is possible to overcome this problem by tracking the position and orientation of the viewer’s head and dynamically computing the corresponding left- and right eye images. In recent years, much research has been done on the development of such visual ‘Virtual Reality’ (or ‘VR’) systems. In many of these systems the user of the system has to wear a so-called ‘Head Mounted Display’ (HMD), or ‘VR Helmet’: a helmet that includes a separate display (usually of the LCD variety) for each eye and a head-tracking device (figure 3.2). The system is connected to a powerful computer that analyzes the data from the head-tracker, computes the correct images for the two eyes and sends them to the corresponding displays. A market overview of HMD’s can be found at [Ster3D]. An example of an application of this technique in research on treatment of phobias (fear of heights) carried out at Delft University can be found at [VRphob]. The main disadvantage of the HMD approach is that observers have to wear a device on their head which reduces the sense of ‘naturalness’, both because of the physical presence of the device on the head and because it visually isolates them from the ‘outside world’. Especially when several observers want to have the feeling that they are experiencing the same virtual scene together, a shared sense of ‘presence’ so to speak, this is a problem.

FIGURE 3.2. **Example of a Head Mounted Display.**

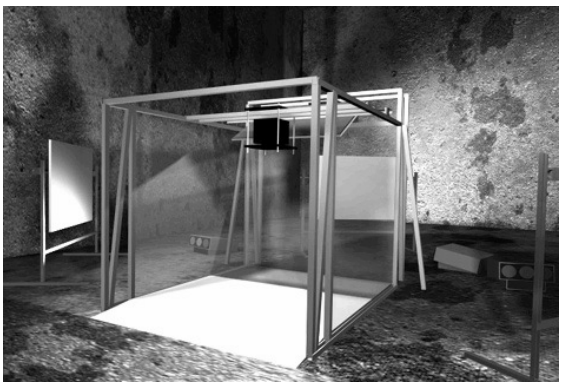


FIGURE 3.3. **The CAVE is a cube-shaped room in which stereoscopic images are back projected on the front- and two side walls, while a fourth image is projected on the floor from above (picture courtesy of SARA, Amsterdam).**

These objections against the use of HMD's in a multi-user environment are removed in systems like the 'CAVE' ([CAVE]). This is a cube-shaped room of which the front wall, the two side walls and the floor consist of projection screens (figure 3.3). Stereoscopic images are projected on the screens, completely immersing the user in the virtual environment. The CAVE uses the 'alternating left- and right eye image' technique described above, so the user is only wearing glasses with electronic shutters instead of the usually much larger and visually isolating HMD's. The projected images are adapted in real-time, based on information from head- and hand tracking devices, allowing the user to walk around. The system allows multiple users to share the same virtual experience by also wearing shutter glasses, so there is a much greater sense of 'shared presence' than when HMD's are used. The limitation however of the CAVE approach is that a single user, the 'active user' who is wearing the tracking devices, determines the dynamic viewpoint of the stereoscopic projections. This means that the other users are

actually sharing the active user's virtual experience, with the additional problem that, depending on how far they are removed from the position of the active viewer, their perceived perspective of the stereoscopic projection is distorted to a certain extent.

A different approach to producing stereoscopic images, but without the need to wear special glasses, is a technique called xography, which uses so-called 'lenticular displays' ([Rede00]). These displays consist of many small vertical lines of lenses which make it possible to project a separate image to the viewer's left- and right eye. With this technique it is also possible to project different pairs of stereoscopic images in different directions, so that depending on the viewer's orientation relative to the display, a different stereoscopic image is received. A well-known example of this technique is the so-called '3D postcard': a card with a rough plastic surface of lines, the image of which changes when it is moved. This means that it is possible with this technique to project stereoscopic images of the same scene recorded from different viewpoints to different directions, so that the viewer can look at the scene from different angles, depending on his orientation relative to the screen. In other words, this is a '*multi-viewpoint*' stereoscopic projection system. This also means that several people can view the virtual scene simultaneously, with each of them receiving a stereoscopic image that is adapted to their personal viewpoint. Most existing systems that are based on this technique have a fixed number of stereoscopic images that are projected to a discrete set of directions, so the virtual viewpoint of the viewer 'jumps' from one direction to the next when he walks around in the room. Recently, research has been done to develop a system with a continuous range of viewpoints, in which the direction of projection is continuously adapted to the viewer's orientation, again by using a tracking device, which requires an adaptive lenticular display. In [Rede00] the design of such an '*adaptive multi-viewpoint*' system is described for the case of a single viewer, a study which was carried out in the context of the European Commission sponsored PANORAMA project ([PANORA]). Extending this concept to a multi-viewer system obviously makes the technical requirements significantly more complex.

The ultimate solution for 3D visualization would be a holographic displaying system, which would provide an infinite number of simultaneous viewpoints in a continuous range without the need to wear special glasses and tracking devices, but the immense processing power that would be needed makes it unrealistic to expect that it will become feasible to develop holo-

graphic displaying systems capable of displaying large-size, full-color, non-synthetic images in real-time, within the next couple of years ([Rede00]).

Finally, it is worth mentioning that there are also systems whose approach is to include the monocular motion parallax cue, so the viewpoint of the projection changes corresponding to the observer's changing viewpoint, but not the binocular disparity cue. The development of such a system for videoconferencing applications is the target of the European Commission project 'VIRTUE' ([Schr00], [VIRTUE]). The approach that is taken there is that a virtual conference table is created by composing for each participant a flat projection of life-size images of the other, remote, conference participants in such a way that they all appear to sit around the same table as the local participant (figure 3.4). The motion parallax cue is included by tracking the participant's orientation and synthesizing the images of the remote participants that correspond to this orientation. This is achieved by making a disparity analysis of the images of several cameras that capture each remote participant from several different viewpoints (figure 3.4). Although, because it lacks the binocular disparity cue, this approach does not give a truly three-dimensional visual impression of the scene, the inclusion of the monocular motion parallax cue does improve the feeling of 'presence', without requiring either the wearing of special glasses or helmets by the participants or using complicated display types and projection devices, contrary to all systems discussed in this section that do include the binocular disparity cue.

FIGURE 3.4. Visualization approach of the VIRTUE project ([VIRTUE]). The VIRTUE station has a large plasma display and four cameras mounted around the display to capture images of the participant from different viewpoints (picture from [Xu02]).

3.2.3 3D Interpretation Of 2D Video Images

As we have seen in the previous subsection, systems for 3D visualization are still in the development phase. Although some of the systems that were described are being applied in special 3D movie theatres and in visualization of computer-generated animations, no system has been developed yet that enables the high quality, life-size, real-time, multi-viewpoint visualization of real-life video images that would be needed for the type of life-size videoconferencing system that was the objective of this research project. In fact it can be expected that it will still take quite some time before any system will be able to meet these requirements, let alone can be applied in commercial systems. Also, apart from the technical feasibility issue of the 3D visualization itself, there is the additional issue in the videoconferencing application of having to transmit the recorded 3D images through telecommunication channels to the remote side, which will require a bandwidth that is significantly larger than when conventional 2D images are used.
For these reasons it is safe to assume that for the near future, conventional 2D video projection systems will remain the only feasible choice for use in life-size videoconferencing systems. The use of a conventional 2D video projection system was therefore one of the boundary conditions for this thesis.

When we are watching a 2D image of a real-life situation, then this image is a two-dimensional representation of a three-dimensional space, in which all visual information is projected on a single plane. In this process, the binocular disparity cue for visual determination of depth is lost, as well as the monocular motion parallax and accommodation cues. What remains for the human visual system to get an impression of depth and distance are the monocular cues of size and linear perspective, plus some contextual information about the scene.
Our visual system will try to interpret the two-dimensional image in a three-dimensional way, but since the most accurate cues for visual depth estimation are now absent, the resulting 3D interpretation will in general not be equal to the real 3D situation. Therefore, it is important at this point to have a look at how the visual system constructs a 3D interpretation from a 2D video image and particularly what the problems are in this process.

Let us first have a brief look at how a 2D projection of a 3D scene is formed. Figure 3.5 (left) depicts the principle of the simplest possible camera model: the pinhole camera. This ‘camera’ only consists of a closed box with a tiny hole in the front surface and a light sensitive imaging

surface on the inside back-surface. It is shown how rays of light reflected from a 3D object travel through the pinhole and hit the imaging surface, resulting in an (upside-down) 2D projection of the object.

It can be stated that for every 2D perspective projection of a 3D scene, there is only a single viewing position relative to the image which, geometrically speaking, results in a perceived perspective that is identical to that of the original scene. This unique position is called the *viewpoint* ([Sedg91]). In figure 3.5 (right) it is shown that when an observer watches the 2D pinhole image with his eye at the pinhole position, the perspective of the image that he sees is identical to the perspective of the original 3D scene as 'seen' by the pinhole during the capturing process. This means that in this most simple case the pinhole position is the correct viewpoint of the image.

In a more realistic situation in which the imaging device is more complicated and contains optical components such as lenses and in which the image is subsequently displayed in some way, the determination of the viewpoint is more complex and is determined by the combination of the optics of the camera and the relation between the center of projection of the displaying device and the projection plane ([Sedg91]), but the principle that there is only a single correct viewpoint remains.

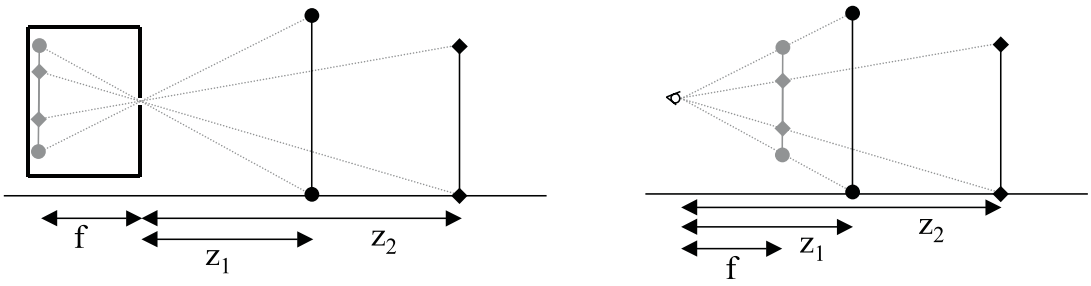


FIGURE 3.5. Principle of a pinhole camera.

Left: a 2D projection of two objects located in 3D space is formed (upside-down) on the imaging surface in the pinhole camera.

Right: when the resulting perspective image is viewed from the pinhole position, the perceived perspective of the virtual scene is identical to that of the original 3D scene, as it was seen from the pinhole when the image was captured.

So, then the next question is: what happens to the spatial perception of a scene when a 2D perspective image is observed from a position *other* than the viewpoint?

Let us first look at the effect of viewing from a position located on the line through the viewpoint orthogonal to the image plane ('on-axis') but too close or too far from the image plane. It can be shown from a geometrical analysis ([Sedg91]) that this results in a compression or expansion, respectively, of the perceived depth of the virtual visual scene, while dimensions parallel to the image plane (width and height) remain unchanged. If we indicate the distance from the viewpoint to the image plane by z and the true distance from a point of the object to the image plane by z_S, then viewing the image from a distance z' instead of z results in a virtual distance from the object point to the image plane z'_S:

$$z'_s = z_s\left(\frac{z'}{z}\right), \tag{3.3}$$

The geometry for (eq. 3.3) is shown in figure 3.6 in which the effect of viewing too close is illustrated for a simple case.

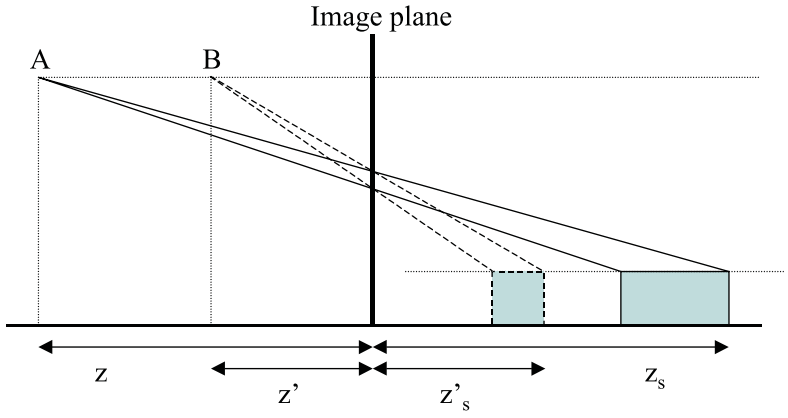


FIGURE 3.6. Example of the effect of viewing from a position too close to the screen.
When the observer is looking at the image from the correct viewpoint, position A, located at a distance z from the image plane, the back end of the box-shaped object appears to be located at a distance z_S behind the image plane. When the same image is viewed from position B located too close to the image plane at a distance z', the back end of the object seems to be located at a distance z'_S behind the image plane which follows from (eq. 3.3). This results in an overall compression of the depth of the virtual scene, as can be seen from the fact that the shape of the object has been compressed in the depth dimension. The height of the object (as well as its width, not visible in this view) remain unchanged.

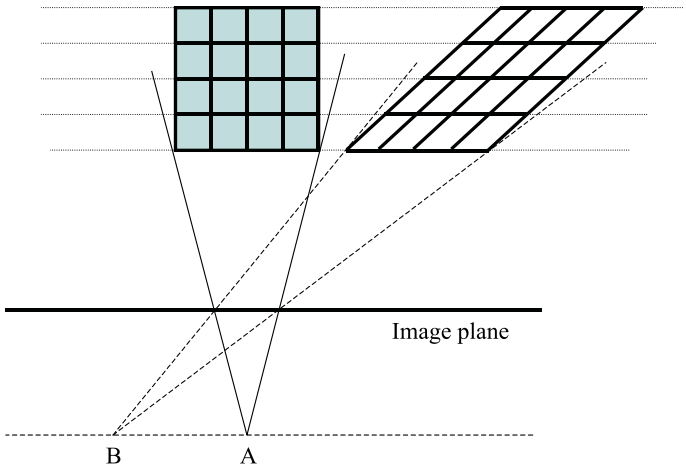


FIGURE 3.7. Schematic drawing of the effect of viewing from a position with a lateral displacement relative to the viewpoint (after [Sedg91]).
The square object appears to be skewed when the image is viewed from a position B that has a lateral displacement relative to the correct viewpoint A. Depth dimensions and dimensions parallel to the image plane remain unchanged, but shapes extending in depth are distorted.

Now let us look at the effect of viewing from a position at the correct distance, but with a lateral displacement relative to the viewpoint. By geometrical analysis it can be shown ([Sedg91]) that this results in a lateral shift of the virtual space to the opposite direction. Since the viewing distance is correct, it follows from (eq. 3.3) that the virtual distance z'_s of each object point is identical to the original distance z_s, leading to the fact that the resulting virtual space is not only shifted, but also skewed, as is shown in figure 3.7. Viewing the image from too high or too low results in a similar distortion of virtual visual space, but now in the vertical direction.

The effect of viewing a 2D perspective projection from a position other than the correct viewpoint can be generalized by combining the effects of viewing from too close or too far and from a position with a lateral displacement, as illustrated for a single object point in figure 3.8. For each original object point *S*, the position of the virtual object point *S'* as perceived by an observer at a position *O* is determined by the virtual distance z'_s as given in (eq. 3.3) and the angle ϕ between the original position *S* and the virtual position *S'*, which is given by:

$$\varphi = \arccos\left(\frac{\vec{s}\cdot\vec{s}'}{|\vec{s}||\vec{s}'|}\right) = \arccos\left(\frac{\vec{s}\cdot\vec{p}}{|\vec{s}||\vec{p}|}\right), \tag{3.4}$$

in which $\vec{s}$ is the vector from the observation point *O* to the object point *S*, $\vec{s}'$ is the vector from the observation point to the virtual object point *S'* and $\vec{p}$ is the vector from the observation point to the position of the projection *P* of *S* on the screen.

However, although important to consider, it should be noted that the purely geometrical analysis given above of distortions of the virtual visual space for non-correct viewpoints does not really give an accurate description of what is actually perceived by human observers. In practice, interpretation of the observed picture by the brain usually causes the distortion that is actually perceived to be considerably less[1], partly due to the presence of extra contextual information about the spatial layout of the scene that the image on the screen represents. This is especially the case for images of real-world situations, for which familiarity with the projected image (for example: a conference room with a table and chairs, people in the room) provides a cognitive frame of reference. In particular, the perceived spatial layout of different objects on the screen (the positions in three-dimensional space of objects relative to each other) seems to remain fairly constant as function of viewing angle and viewing distance, however, the perceived orientation of structures in space can change considerably ([Gold91]). An additional factor that complicates prediction of the 3D spatial interpretation of a 2D picture projected on a screen is the fact that the cognitive visual system is aware of the fact that it is actually looking at a flat surface instead of a real 3D scene, due to the fact that characteristics of the screen itself, like the edge of the screen, the texture of the screen surface, the not completely natural colors and contrast etcetera, are visible. In other words: the image on the screen is not just interpreted as a geometrical 2D projection of a 3D real-life situation, but at the same time it is observed as being a flat object itself, with a certain orientation relative to the observer. This cognitive duality is assumed by some to be one of the main factors that cause the perceived spatial distortions to be considerably smaller than would be expected from a purely geometrical analysis ([Sedg91], [Gold91]).

1. Note that if this were not the case, the concept of a cinema, where many people watch the same perspective image from a wide range of different viewpoints, would be impossible!

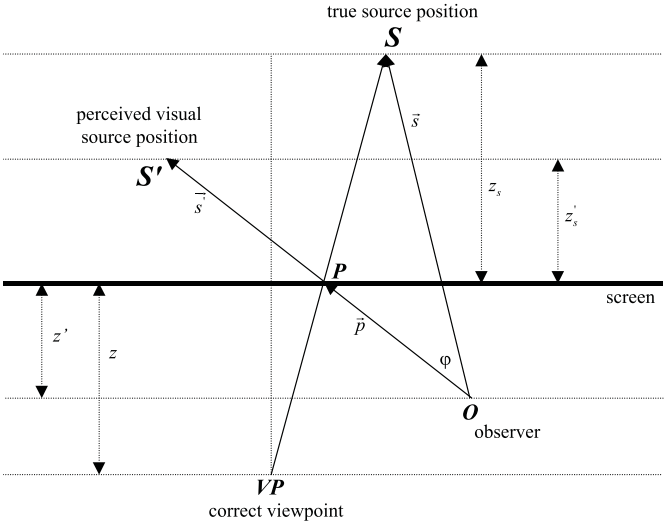


FIGURE 3.8. Illustration of the effects of viewing a two-dimensional projection of a three-dimensional scene from an arbitrary non-viewpoint position. In this example the observer *O* is viewing the projection *P* of an object point *S* from a position that has a lateral displacement relative to the viewpoint *VP* and is too close to the screen, resulting in a virtual object point *S'* and a total virtual scene that is skewed to the direction opposite to the lateral displacement of the observer and compressed in the depth dimension.

3.3 Audio-Visual Interaction

In the previous two sections, an overview of the mechanisms for spatial perception was given for the auditory and visual modalities independently from each other. However, in our daily life we are continuously 'bombarded' by a multitude of stimuli that are picked up by our individual senses simultaneously. The effect this has on our overall perception of the situation we are in is not simply the sum of the contributions of the individual senses as they were discussed in section 3.1 and section 3.2. To the contrary: the stimulation of one of the senses can significantly, sometimes even dramatically, influence the perception in one of the other modalities, a phenomenon referred to as *multi-modal* or *multi-sensory interaction*. A common everyday example, for instance, is the fact that the taste of food is much influenced by smell.

A very strong interaction occurs between the senses of vision and hearing, the effects of which are of particular importance when we are dealing with audio-visual systems, such as cinema, television, flight simulators, or, indeed, videoconferencing systems.

There are many situations in which the simultaneous presence of auditory and visual stimuli causes an interaction between the perceptual processes in the two individual modalities, resulting in the fact that the perception in one or both of the modalities is different than would be the case if only that modality would be active. Various aspects of the relationship between the audio and the video cause specific interaction effects. For instance, the temporal relationship between the audio and video causes specific interaction effects, while the spatial relationship between both causes others.
Some examples of audio-visual interaction effects are:

Temporal interaction. When in a movie the soundtrack is not completely synchronized with the image this can be very irritating. However, when the auditory and visual stimulus clearly belong together, as in speech, the tolerance in how much the audio and video have to be 'out of sync' before it becomes noticeable is larger than would be expected if we would just consider the temporal detection resolutions of the two modalities ([Dixo80]). Apparently, the fact that there are matching stimuli in the two modalities allows the human brain to 'merge' the non-synchronized sound and image to some extent.

Spatial interaction. There are strong interaction effects between the perceived locations in space of auditory and visual stimuli. The perceived location of a source in one modality can significantly influence the localization of a corresponding source in the other modality. An observation in various spatial audio-visual interaction effects is that the perception in the visual modality significantly influences the perception in the auditory modality, much more so than the reverse, so that the visual modality largely dominates the overall perception, a phenomenon that is referred to as *visual dominance* ([Bert76]). This will be discussed in more detail in the next subsection.

Speech perception interaction. Being able to see a speaking person's lips can greatly enhance the intelligibility of speech in situations where the perceived quality of the speech sound by itself is insufficient ([Mass87]). A very striking demonstration of the effect of audio-visual interaction on speech perception is a phenomenon known as the 'McGurk effect' ([McGu76]), referring to the observation that when the video image of the face of a person uttering a certain syllable is combined with the sound of certain other syllable, this leads to the consistent perception of a third syllable (for instance, visual 'ga' plus audio 'ba' leads to perception of 'da').

Quality interaction. The subjective quality of reproduced audio is influenced by the subjective quality of accompanying video material and vice versa. Beerends and De Caluwe ([Beer99]) describe an experiment in which the audio, video and audio-visual quality of AV clips was evaluated. Each clip combined an audio and a video stream that were degraded versions of the original undegraded audio and video streams.Several versions of both the audio and video stream were created with different degrees of quality degradation. They conclude that the presence of high quality video increases the perceived quality of the accompanying sound. The reverse was found to be the case only to a much smaller extent, which suggests that visual dominance exists also in quality interaction. However, in a similar experiment but using audio-visual speech as source material, Rimell et al. ([Rime98]) found no clear visual dominance. In their experiment, a significant quality interaction was observed in both directions.

As can be understood from the above, evaluating or predicting the overall subjective quality of an audio-visual reproduction system is a complicated issue, in which many different aspects of the audio part, the video part and also their combination have to be taken into account. Since audio-visual systems have started to play an increasingly important role in our daily lives during the past few decades, the interest in developing reliable methods to assess their quality has increased as well. However, only since the last couple of years systematic efforts have been made to start to develop standardized methods that take into account the effects of audio-visual interaction ([Wosz95], [ANSI60], [ITU97]) and the subject can to some extent still be considered ‘virgin territory’.

3.3.1 Spatial Audio-Visual Interaction

As already mentioned briefly above, there is a clear interaction between the perception of the location in space of an audio stimulus and the perception of the location of a video stimulus that are presented simultaneously. This is especially the case when the audio and video stimuli are associated with each other, for instance, the sound of speech and the video image of a person making the lip movements that correspond to the speech sound.
A very common effect is that when corresponding auditory and visual sources are physically located at different positions not too far removed from each other, the perceived location of the auditory source is ‘pulled’ towards the location of the visual source, a phenomenon referred to as *visual capture* ([Bert76]). The origin of this visual dominance in localization is believed to

lie in the fact that localization in the visual modality is much more accurate than in the auditory modality ([Mass87]), so that for reasons of survival it is preferable to give more weight to the visual than to the auditory modality when it comes to source localization. Visual capture also applies in perception of motion to some extent: when a visual object is moving, a corresponding static auditory source appears to move along with it ([Kita02]).

The classic example of visual capture is the ventriloquist: a person who is able to produce speech sounds without moving his lips and then, by making a dummy that sits on his lap move its lips in such a way that the lip movements match the speech sound, give observers the strong illusion that the speech is coming from the mouth of the dummy, when it actually comes from the ventriloquist's throat. This illusion works because of the visual dominance effect, which for this reason is often referred to also as the 'ventriloquist effect'. Another example of visual dominance is found in the world of cinema. In most cases, people having a dialog in movies are pictured somewhere in the center area of the screen. Because of the importance of dialog it is highly preferable to have an audio reproduction of the dialog that is spatially stable, i.e., the localization of the dialog sound is satisfactory for every spectator. In order to achieve this, instead of reproducing the voices in a stereophonic individually spatialized way, which would pull the sound image to the side for spectators who are sitting off-center, the complete dialog is typically reproduced from a single loudspeaker in the center of the screen. The visual dominance effect ensures that the, generally slight, disparity of the actual locations of the auditory and visual sources remain completely unnoticed by the average spectator.

Experimental results indicate that in cases where the audio and video clearly belong together, as is the case with audio-visual speech, the audio and video stimulus effectively merge into a single perceptual event up to about 11 degrees of separation between the actual directions of the audio and video source, i.e., no annoying discrepancy between the audio and video source directions is reported, not even by expert observers. For non-expert observers this maximum acceptable discrepancy is even 20 degrees ([Komi89]). This observation is supported by the conclusion from experiments in [Warr83] that a discrepancy of 10 degrees does not weaken subjects' assumption that the auditory and visual stimulus belong to a single event. However, the phenomenon is still effective for larger separations up to 30 to 40 degrees, in the sense that the localization of the audio is still pulled towards the matching video image to some extent. Whether the perceived discrepancy that remains is considered by subjects to be annoying or not is another issue which depends on the situation.

Audio-visual interaction also occurs in distance perception. Even more so than in frontal plane localization, visual estimation of distance is much more accurate than auditory distance estimation, which, as was discussed in section 3.1.3, is quite poor. Experiments on this issue in the context of an audio-visual virtual reality system are described in [Nath97]. It is shown there, by letting subjects rate the perceived auditory distance of a source while looking at a 3D visual image of the source at one of several distances, that subjective auditory distance is influenced significantly by apparent visual distance.

Visual dominance is a phenomenon that can be of great importance in the design of audio-visual systems. For instance, the effects of visual capture can be exploited to reduce the required spatial accuracy of the audio reproduction, which in general can result in a smaller number of audio channels to be recorded, transmitted and reproduced. Specifically, in our case of designing a life-size videoconferencing system, visual dominance has been taken into account in determining the required accuracy for reproduction of vertical sound source location (section 4.1).
In specific situations, in audio-visual systems for which, for some reason, the reproduction of auditory space is actually more realistic than the reproduction of the corresponding visual space, the visual dominance effect can also have a negative effect, in the sense that it effectively degrades the perceived spatial quality of the audio reproduction. This situation possibly arises in our case, where high quality spatial audio reproduction by WFS is, out of necessity, combined with 2D perspective video projection. This issue is the subject of the next subsection.

3.3.2 Combining 2D Visualization With True Perspective Audio

The goal of this project, as was described in section 1.2.3, was to develop the audio part of a life-size group-videoconferencing system in such a way that the voices of conference participants are perceived as coming from their true positions, i.e., where they would be located if the two rooms would be physically connected, with their projection displays forming the interface between the two rooms (figure 1.4). With the Wave Field Synthesis technique that was chosen in this project to achieve this, it is indeed possible to position virtual sound sources at any desired location, including a natural illusion of depth, in such a way that the reproduced sound field is correct within an extensive listening area.

Unfortunately, as was discussed in section 3.2.2, for practical reasons a compromise had to be made for the visualization system. Instead of the ideal situation in which WFS sound reproduction would be combined with a visualization system that is able to reproduce a natural 3D reproduction of the visual scene that also allows multiple observers to view the scene from a wide range of different and possibly dynamically changing viewpoints, a compromise had to be made for conventional two-dimensional perspective video projection.
As was explained in section 3.2.3, this introduces two issues that distort the resulting perception of visual space:

- Inaccurate visual perception of depth because of the absence of several very important cues for visual distance perception, most notably the binocular disparity cue and the motion parallax cue.
- Distortion of perceived visual perspective for observers located at positions other than the single correct viewpoint of the two-dimensional projection, resulting in an incorrect visual perception of either source direction, source distance, or both (figure 3.8).

So, when we combine WFS sound reproduction which preserves the true perspective of the original scene for all observers with 2D video projection which does not, this will in general result in a different perception of space for the visual and auditory modalities when evaluated individually. The situation can be visualized by again looking at figure 3.8 in which the ‘true source position’ *S* now corresponds to the virtual position of the WFS sound source, independent of the observer position *O*, while the perceived visual source position is located at *S’*, which depends on observer position *O*. It is seen that there now is a discrepancy between both the perceived directions and distances of the auditory and visual source.

An important question is: what effect does this have on the overall audio-visual perception of the reproduced virtual space?
Because of the bi-modal nature of the system, this is a complicated question, but given the discussion of the properties of auditory and visual perception of space in section 3.1 and section 3.2 and the discussion of spatial audio-visual interaction in section 3.3.1, the following three main scenarios can be distinguished for the effect on the overall perception of space of combining true perspective (in this case: WFS) sound reproduction with 2D video projection, compared to conventional systems that combine 2D video projection with ‘conventional’ sound reproduction techniques (see section 1.2.2):

Positive. If the discrepancies between the two modalities are not perceived as being annoying, then the fact that the reproduction of auditory space is very natural might enhance the quality of the overall perception of space.

Neutral. If the visual capture effect is so strong in the audio-visual set-up under investigation that the visual modality completely dominates the overall perception of space, then the proper reproduction of auditory space will not significantly add to this overall perception of space. Note however that also in this case it is still possible that there are other beneficial effects of reproducing the audio in a spatially correct way, for instance, improvement of speech intelligibility.

Negative. It is not unimaginable that situations might occur in which the discrepancy between the perception of visual and auditory space is so large that it causes a sensory conflict, i.e., the discrepancy is perceived as being annoying. If this is the case it may even be possible that, given the boundary condition of using 2D video projection, reproducing the audio in a spatially correct way is actually unfavorable to the perceived quality of overall audio-visual reproduction of space. Although even in this case there might still be beneficial effects in terms of, for instance, speech intelligibility, this is a scenario that should be avoided.

As stated, because of the interactions that occur between the perception of the audio and video part of an audio-visual system that combines WFS with 2D video projection, it is difficult to predict which of the above scenarios will be the one that actually applies in our situation. Actually, the subject of combining 2D video projection with ‘true perspective’ audio reproduction techniques in general has not been explored extensively yet, which is not really surprising, since until recently no audio reproduction technique existed that was really capable to achieve ‘true perspective’ reproduction of space. A substantial part of this project therefore consisted of investigating this issue by carrying out several audio-visual perception experiments. These experiments are the subject of the next chapter.

CHAPTER 4 *Audio-Visual Perception Experiments*

The human auditory system is limited in its ability to perceive the details of various properties of sound fields. As we have seen in the previous chapter, the localization of sound sources in the horizontal plane is reasonably accurate for sources in front of us, but gets less accurate for sources at the sides or the back of the head. Localization accuracy in the median plane is much worse, as it has to rely on more ambiguous cues than the interaural time and level differences that are the main cues for horizontal localization. Also the ability to estimate source distance is very limited, especially in environments with little reflections and unfamiliar sound sources. Also in the perception of the acoustics of a room the ability to discriminate between small differences is limited, both in a spatial and temporal sense. Individual reflections only have an individual perceptible influence on the acoustics when they arrive at the listener's ears shortly after the direct sound (roughly speaking: within the first 100 ms, depending on the properties of the room). Also the number of discrete reflections arriving at the listener from different directions that can be discriminated is limited. And finally, concerning the reverberant part of the sound field it has been shown ([Sonk00]) that when reverberant sound reaches the listener from only a limited number of discrete directions (between 8 to 11) the resulting reverberant field can be considered to be perceptually indistinguishable from a true isotropic reverberant field.

The fact that the human auditory system has only limited accuracy relaxes the requirements on the reproduction accuracy of an audio reproduction system. When the limitations of the auditory system are known and understood, this knowledge can be exploited to design an audio system in such a way that its accuracy is sufficient for the application without unnecessarily using resources like production costs of the system itself, required computational power and transmission bandwidth, which is related to the number of audio channels to be transmitted.
Besides the limited accuracy of the human auditory system itself we have the additional effects of audio-visual interaction in the application of videoconferencing. In section 3.3 an overview was given of the spatial audio-visual interaction effects and it was shown that with regard to localization accuracy in both lateral direction as well as source distance, the visual modus dominates the perception of the audio-visual scene. This knowledge can be used to great advantage to further reduce the required accuracy of the audio reproduction part of the videoconferencing system and thus the system's complexity and the number of channels to be transmitted.
Although the existence of these audio-visual interaction effects is well known and experiments have been carried out to investigate their properties at a very elementary level, not much results have been published about the effects in practical systems.
For this reason a series of audio-visual perception experiments has been carried out with the purpose of getting a better understanding of the relevance of these effects in practical audio-visual systems and life-size videoconferencing in particular and to get an indication of the extent to which they can be exploited in the design of such systems.
Additionally, as was explained in section 3.3.2, the combination of 2D video projection and true perspective audio such as WFS also has the potential risk of introducing discrepancies between the visual and auditory modalities, in particular in perceived source directions and perceived source distance. It is important to know if these problems can indeed be expected to occur in practical situations, how serious these problems are and what can be done to prevent them. These issues are also addressed by the perception experiments described in this chapter.

First in section 4.1 an experiment will be described in which the necessary vertical resolution for the reproduction of the voices of videoconference participants was investigated.
Then, section 4.2 describes an experiment in which subjects could interactively position a sound source at the lateral position that they felt best matched the location of a person projected on the video screen, evaluated at different observer positions, to investigate how signifi-

cant the discrepancy between auditory and visual source positions in a system that combines 2D video projection with true perspective audio actually is.
Following this, in section 4.3, is an experiment in which the subjects' task was to identify a specific speaker out of three possible choices to see if, despite the possible problems caused by the discrepancy of the auditory and visual source locations, observers might still have a benefit in identifying individual speakers when WFS is used for audio reproduction because of the effective auditory source separation this provides.
In section 4.4 a simple yet effective solution is presented to avoid or reduce the mentioned discrepancy problems while still maintaining the benefits of the effective WFS source separation.
Finally, section 4.5 describes an experiment in which the speech intelligibility of reproducing conference participants' voices by WFS was compared to a more conventional reproduction method, to investigate if the spatial separation of the voices by WFS leads to a significant improvement of speech intelligibility, which would be a strong argument to indeed apply WFS in a teleconferencing system.

4.1 Vertical Localization

As was discussed in section 3.1.2, the human ability to localize sound sources in the median plane is less accurate than in the horizontal plane. This means that in general the vertical location at which a sound source is reproduced is not too critical. In an audio-visual system there is the additional effect of audio-visual interaction that may be expected to make the vertical placement of audio sources that correspond to a visual source even less critical. Still, one could imagine that in the situation of a life-size videoconferencing system in which participants are free to walk around the room and may be sitting behind a table as well as standing anywhere in the room, the vertical placement of the corresponding sound sources can be of importance, especially when both the observer and the remote participants are allowed to come close to the screen. To find out how critical the vertical localization of sound sources is in the context of this specific application, the experiment described in this section was carried out.
The experiment was done in audio-visual as well as audio-only situations to investigate the influence of the presence of the video image on sound localization.
Additionally, several possible reproduction methods for vertical sound source placement were investigated to determine how suitable they are for application in a life-size videoconferencing system.

Finally, based on the results of this experiment, a proposal is made for the most reasonable strategy for vertical sound source reproduction in the application of life-size videoconferencing.

4.1.1 Reproduction Methods

Three possible reproduction methods for source positioning in the median plane were considered:

Single-speaker reproduction. In a multi-speaker configuration this is the simplest reproduction method, where a source is positioned at a specific vertical position by simply sending the source signal to the speaker closest to the desired source position. This method was considered to be the reference method regarding localization accuracy, since it corresponds to the situation of a real source at the position of the active loudspeaker.
Application of this reproduction method in a videoconferencing system based on WFS would result in a system consisting of several horizontal loudspeaker arrays at different heights behind the screen, in which accurate horizontal localization and depth perception is ensured by applying the WFS technique and sufficient vertical resolution is reached by assigning the virtual source to the horizontal array whose vertical position is closest to the sources' vertical position. The maximum acceptable distance between the horizontal arrays in such a system is to be determined from the experiments.

Wave Field Synthesis reproduction. In this reproduction method a source is positioned at any desired vertical position by synthesizing the source field using a vertical array of closely spaced loudspeakers. In this case a source can also be synthesized as coming from a position behind or in front of the array.
Since WFS will also be applied in the horizontal direction, in which localization is the most critical, application of this concept would result in a plane (or more precise: a matrix) of loudspeakers behind the screen, where each speaker is to be controlled individually.

Intensity-based phantom source imaging. This reproduction method tries to position a source at a position on the vertical line between two loudspeakers by controlling the gain balance between them. This is the analogy in the median plane of standard intensity-based stereophony in the horizontal plane.

From what is known of the working mechanism of median plane localization (see section 3.1.2) it may be suspected that this method is unlikely to work. However, there are conflicting conclusions in literature on this matter. For example: Somerville ([Some66]) concluded that an unambiguous source percept can be produced by controlling the gain balance between the upper and lower speaker, while other research by Start ([Star97]) led to the opposite conclusion, namely that no stable imaging is possible.
The reason that this method was included in the experiments was that it would yield an attractively simple solution for vertical source positioning, requiring only two horizontal loudspeaker arrays: one above and one below the screen. In the resulting system a sound source would be positioned anywhere in 3D space by synthesizing two virtual sources at the correct position in the horizontal plane: one using the array above the screen and one using the array below, while the vertical position is controlled by proper balancing of the intensity of the two virtual sources.

4.1.2 Source Material

As source material for the experiment an audio-visual recording was made of the head of a male person reading a continuous Dutch text in a natural way in front of a neutral white background with the head centered in the video frame and directed towards the camera. The voice was recorded with a spot microphone. The resulting source sequence was captured on the hard disk of a digital video workstation, giving the possibility of looped playback and flexible editing possibilities.

4.1.3 Experimental Set-Up

The experimental set-up is illustrated in figure 4.1. A 15-element electrodynamic loudspeaker array with a speaker distance of 0.127 m was placed vertically at the side of a projection screen (figure 4.1). All 3 reproduction methods could be handled by this single array by appropriately changing the driving signals of the loudspeakers, which were generated by a DSP, controlled through a MATLAB user interface (figure 4.2). The driving algorithms of the different reproduction methods were as follows:

Single-speaker. In this case the source audio signal coming from the digital video workstation was sent to a single speaker at the desired vertical position.

WFS. Each virtual source position was synthesized using a sub-array of 7 speakers (0.89 m total length), with the center speaker located at the desired source position. This configuration was chosen to be able to have symmetrical speaker gain distributions for each source position that was used in the experiment, including the outer positions. Details of the source positions that were used are discussed in the next subsection. To avoid strong diffraction artifacts from the edges of the sub-array, a taper was applied to the outer active speakers, resulting in an effective sub-array length of about 0.76 m. All sources were synthesized as being located 0.5 m behind the array. This combination of sub-array length and source position resulted in a reconstruction angle of about 65 degrees, so that the reconstruction areas of all used source positions contained all listening positions of interest.

Phantom source imaging. For phantom source imaging the two loudspeakers directly above and below the screen were used (see figure 4.1), the distance between them being 1.78 m. Their gain balance for each source position was calculated from the well known 'Law of Sines' for intensity-based stereophony:

$$\frac{R-L}{R+L}=\frac{\sin\theta}{\sin\psi}, \tag{4.1}$$

in which R and L are the amplitudes of the right and left loudspeaker signals respectively, θ is half the angle between the two loudspeakers relative to the sweetspot and ψ is the angle of the desired phantom source location relative to the sweetspot.
The sweet spot was located at 2.25 m in front of the array.

The overall gain levels for the three reproduction methods were balanced to obtain an equal resulting reproduction level for the listeners, comparable to normal conversation level.

Subjects were seated on a chair directly in front of the array with their ears at the height of the central position (position #7). This position coincided with the mouth of the speaking person, who was projected life-sized on the screen (see figure 4.1 and figure 4.2).

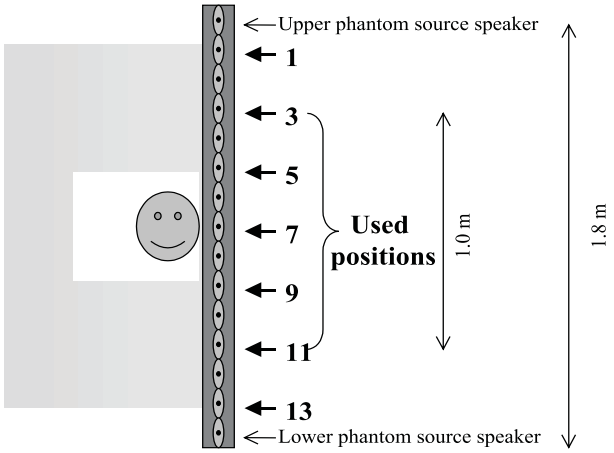


FIGURE 4.1. Schematic drawing of the loudspeaker set-up used in the vertical localization experiment.

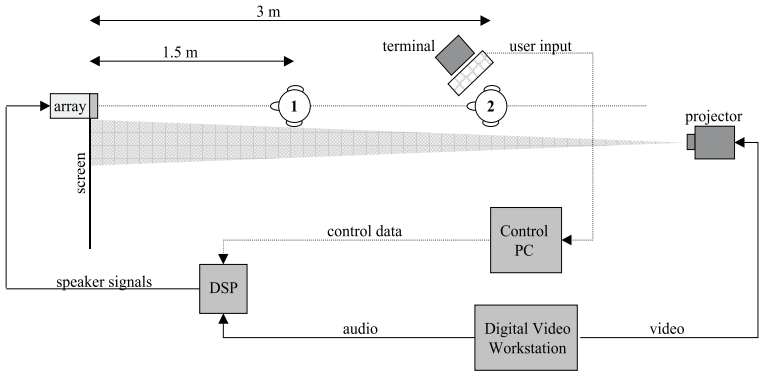


FIGURE 4.2. Schematic drawing of the equipment set-up used in the vertical localization experiment.

4.1.4 Experiment Design

The three main questions to be answered from this experiment were:

- What vertical resolution is needed for the sound reproduction part of a life-size video-conferencing system?
- How suitable are the three different reproduction methods for positioning a sound source in the median plane?
- What is the influence of audio-visual interaction on localization in the median plane?

Because it was expected that vertical localization becomes more critical when videoconference participants approach the screen, due to the larger effective angles between neighboring audi-

tory source positions and between auditory and visual source positions, the experiment included two listening positions situated at 1.5 m and 3 m, respectively, from the loudspeaker array.

From past research it is known that for speech of a known speaker the vertical localization blur is about 9 to 10 degrees for positions in front of the listener ([Blau83]). Therefore it seemed unnecessary to incorporate more source positions in the experiment than are necessary to obtain this separation of about 10 degrees between neighboring source positions at the listening positions. Therefore it was decided to use 7 source positions separated by two loudspeaker distances (.25 m) with the middle position in the center of the array (coinciding with the position of the video image of the mouth of the speaker). For the closest listening distance (1.5 m) this resulted in a separation of 9.6 degrees for neighboring sources on-axis (this is: around the center position) and 8.2 degrees for the outer positions.

To obtain an unbiased response from subjects the two outer positions were not actually used in the experiment. Furthermore, to let the subjects choose from a more continuous range of possible source positions a dummy position was added between each two real positions. Subjects could now choose from 13 equidistant positions with position #7 at the center. The subjects were not informed of the fact that of the 13 possible choices only 5 corresponded to actual source positions being used in the experiment (positions 3, 5, 7, 9 and 11, see figure 4.1).

To investigate the audio-visual interaction, the experiment was carried out both with and without the video image present. It was expected that in the case of single-speaker reproduction without video, the means of the perceived source positions would closely match the true source positions, with observations normally distributed around this true position.

The hypothesis that we wanted to test was that the influence of the matching video image would be to draw the source localization towards the image, thus shifting the means of the perceived source positions significantly towards the image for all but the center position (which already is located at the position of the video image), while for the center position itself the mean would be more firmly anchored to the center as reflected by the mean bias and standard deviation of the observations for the center position.

Twelve subjects participated in the experiment. There were both experienced and non-experienced listeners. Subjects were free to move their head. Numbers on the array indicated the 13 positions from which subjects could choose.

For each subject the experiment was divided into four parts (the order of which was varied randomly between subjects), as shown in the following scheme:

Part	Listening Position	Audio	Video
1	1.5 m (#1)	X	
2	1.5 m (#1)	X	X
3	3.0 m (#2)	X	
4	3.0 m (#2)	X	X

TABLE 4.1. The 4 parts of the experiment. An 'X' indicates that the condition was present in this part.

Each of the 4 parts consisted of 3 sequences of 20 stimuli in which all of the 5 used source positions were presented 4 times in random order. For each sequence the stimulus order was randomized. The first sequence used single-speaker reproduction, the second used WFS and the third phantom source imaging. Each sequence was preceded by a test run of 10 stimuli that were reproduced using the method of that sequence. These test stimuli were coming from random positions. Subjects were already asked to type the number of the position from which they heard the sound, but these inputs were not recorded. Also, subjects did not get feedback about the true positions, since the purpose of the test run was not to train them in recognizing when a sound came from a certain position, but was just intended to familiarize them with the experimental procedure. In the test run, the outer source positions (#1 and #13, excluded from the actual experiment) were also allowed to occur. For the A/V parts subjects were instructed to try to keep their attention directed towards the image, instead of only looking at the numbers on the loudspeaker array. The actual instructions that were given to the subjects before the start of the experiment can be found in Appendix A.1.

Now, the procedure was as follows. The audio-visual source material (see section 4.1.2) was played in a loop from the digital video workstation. In between trials, the audio was muted, whereas the video was visible all the time (only for the parts of the experiment that included video, part 2 and 4 (table 4.1)). At the start of a trial, the audio was switched on and was reproduced from one of the 5 used positions. After 4 seconds, it was switched off again. Then subjects were asked to type the number of their choice on their terminal. Only after they had entered their choice the next stimulus was presented.

Each of the four parts of the experiment took about 10 minutes to complete so that the complete experiment took about 40 minutes per subject. Subjects were free to take a break in between parts to avoid loss of concentration.

4.1.5 Results

Localization Results. In figure 4.3 and figure 4.4 histograms are shown of the pooled observations (i.e., the observations of all subjects are put together). Figure 4.3 shows, for listening position 1, the histograms of the responses obtained with the true source located at the center position (position #7).

The top graphs of figure 4.3 show the results for single-speaker reproduction for both the audio-only (left plot) and audio-visual (right plot) situations. It is seen that localization for the single-speaker method, corresponding to a real source at the center position, is quite accurate in the audio-only case (left plot). Observations are narrowly distributed, with the peak slightly off-center. When the video is added (right plot) the distribution becomes even more narrow, with the peak now located at the center position.

The center graphs of figure 4.3 show the results for WFS reproduction. They show that the spread in the observations is somewhat larger for this method. Also in this case, when the video is added the capturing effect of the image is clearly observed.

The results for phantom source imaging (bottom graphs of figure 4.3) show an interesting phenomenon. In the plot for the audio-only case we observe that the localization for a source between the two loudspeakers is very poor: most observations are located close to either the upper or lower speaker, while very few observations were actually located at the center position. This is consistent with the fact that most subjects reported that most stimuli (for all true source positions) reproduced with phantom source imaging were perceived as coming from either above or below, while an ambiguous source was perceived for the center source position, leaving them no other choice than to `guess'. The strong interaction between the audio and video is particularly striking in this case: when the image is added, the observations are actually distributed around the center position, be it with a larger spread than for the other two reproduction methods. The effect of the video will be quantified later on in this subsection.

Figure 4.4 shows representative results for one of the other (off-center) source positions, position #3, also at listening position 1. The top graphs are for the single-speaker case, the center graphs for WFS and the bottom graphs for phantom source imaging. Also for this source position it is seen that localization is quite accurate for the single-speaker and WFS methods, while it is not for phantom source imaging. Again, in all three cases the capturing effect of the video image at the center position is clear.

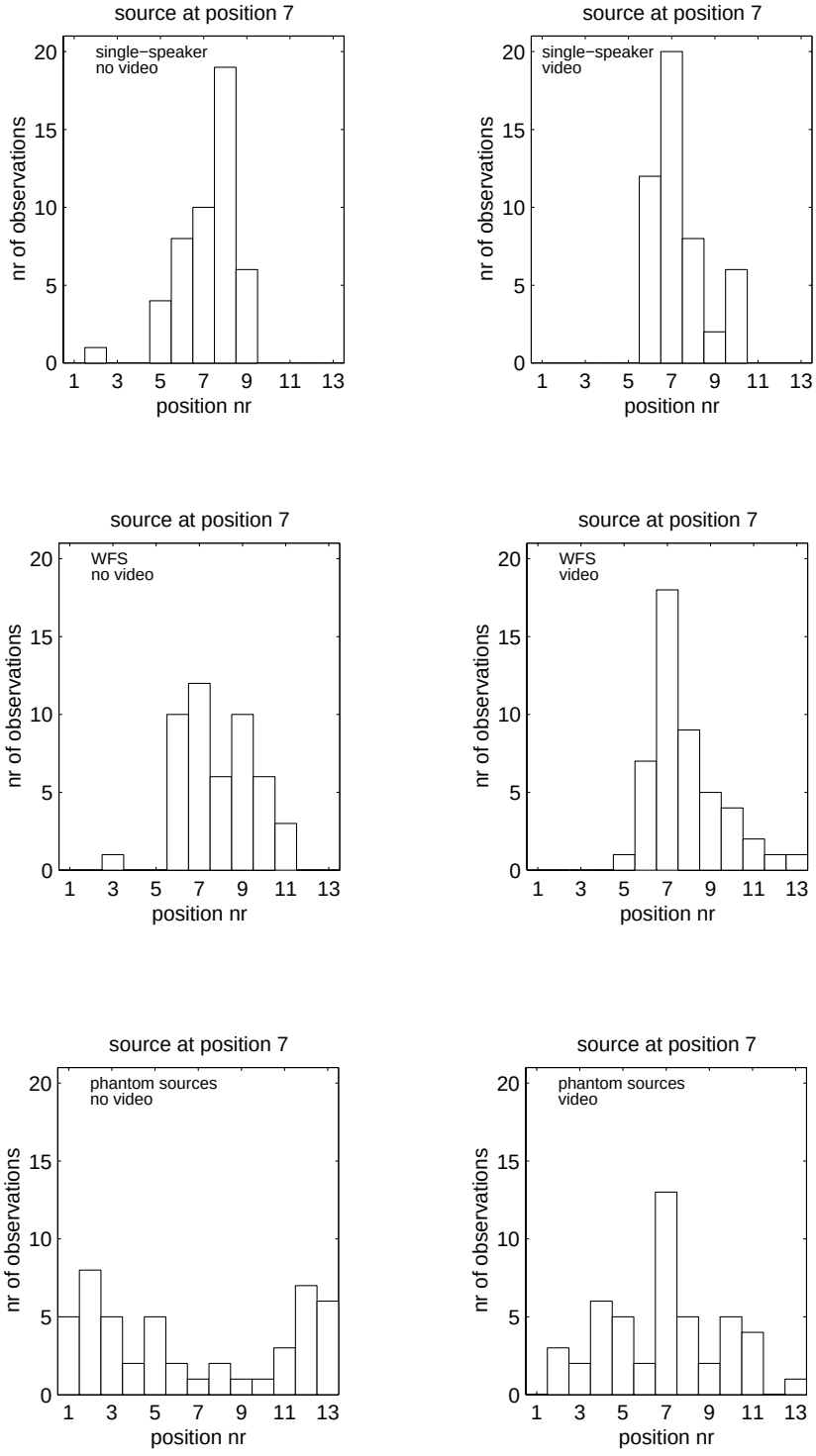


FIGURE 4.3. **Histograms of pooled subject responses for single-speaker (top), WFS (center) and phantom source image (bottom) reproduction at listening position 1 (1.5 m), with the true source at position 7 (center position). For each possible source position the total number of subject responses is given. The left plots are for the audio-only situation, the right plots for the audio-visual situation.**

.

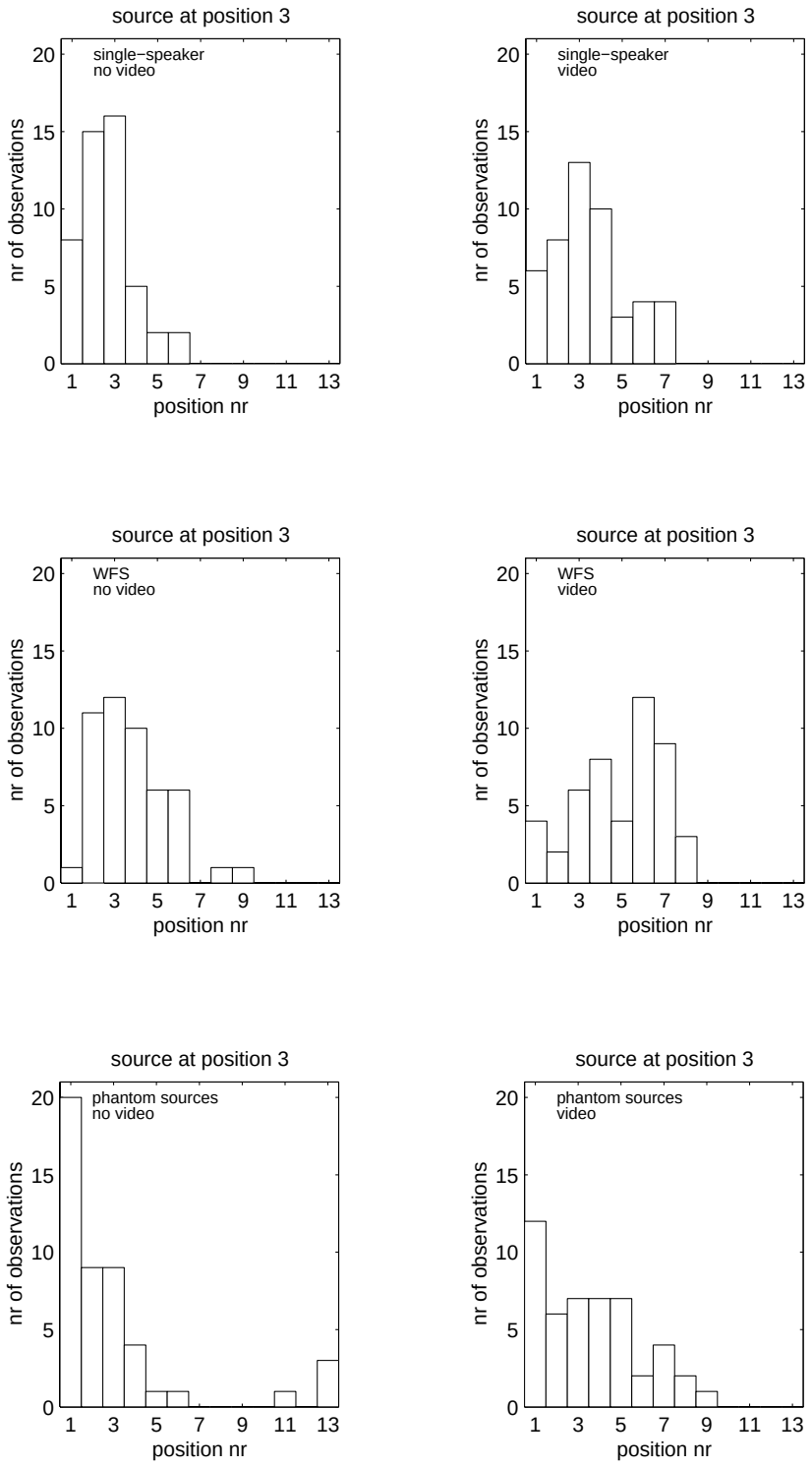


FIGURE 4.4. Histograms of pooled subject responses for single-speaker (top), WFS (center) and phantom source image (bottom) reproduction at listening position 1 (1.5 m), with the true source at position 3. For each possible source position the total number of subject responses is given. The left plots are for the audio-only situation, the right plots for the audio-visual situation.

To get an estimate of the spread in the localization for each source position, an estimate must be obtained for the standard deviation of the distribution of the population of which the observations form a sample. The best estimate for this is the sample standard deviation calculated from all observations for a source position. The standard deviation does not necessarily reflect the accuracy of localization, since a bias may be present in the mean of the observations. It does however give a good indication of the spread in the responses of subjects for a certain source position. The actual localization accuracy is a combination of the bias of the mean and the standard deviation around this mean.

Figure 4.5 shows the localization results for the three reproduction methods, averaged over all subjects. At each position the mean and standard deviation are shown for both the audio-only and audio-visual situation. The left plots are for listening position 1, the right plots for listening position 2. Note that, due to the fact that in the WFS reproduction the sources were synthesized at the same heights as in the other two reproduction methods but now 0.5 m behind the array, the effective source positions on the array, as viewed from the listening positions, lie somewhat closer together in the WFS case. The effective source positions are shown in the center graphs of figure 4.5. In all following results this has been accounted for.

Qualitatively, we see that localization is the most accurate for single-speaker reproduction (top graphs of figure 4.5). Looking at the results for single-speaker reproduction we see that in the audio-only situation the means match the true source positions quite closely, although they seem to be a little exaggerated (high positions localized a little too high, low positions too low). For listening position 1 an average standard deviation of about 0.17 m is observed, corresponding to an angle of 7 degrees. For WFS (center graphs) this is slightly larger, 0.22 m, but this could be expected since the effective listening distance was 0.5 m larger. In the case of phantom source imaging (bottom graphs) the standard deviations are much larger, especially for the center position where it is 0.57 m (21 degrees). This again clearly indicates that no stable image between the speakers is possible with this technique, at least for this short listening distance.

Now, if we look at the audio-visual results for the same situations we see that the means are shifted towards the center, as expected. This audio-visual interaction effect will be quantified later. The standard deviations of the observations are comparable for the audio-only and audio-visual situations in the case of single-speaker and WFS reproduction. For phantom source imaging the standard deviations with the video present are significantly smaller than without

video, indicating that the video provided some stability for the sound image that was lacking without the video.

For position 2 (the right graphs) the results are similar to those for position 1, with one difference: the standard deviations for position 2 are larger than those for position 1 for single-speaker and WFS reproduction, due to the larger distance from and thus smaller angles between the source positions. For phantom source imaging we observe the opposite. This might be due to the fact that for position 1 the angular separation between the upper and lower speaker is so large (61 degrees) that they are always perceived as two individual sources, while at a larger listening distance the angular separation between the two loudspeakers is smaller (33 degrees) so they are more likely to 'merge together' to some extent. This is also suggested by comparing the audio-only results for position 1 (bottom-left histogram of figure 4.3), in which subjects' responses are concentrated near the outer source positions, to the corresponding histogram for position 2 (not shown), in which quite some subject responses are actually located around the center position. However, the spread is still significantly larger than for single-speaker and WFS reproduction.

Table 4.2 summarizes the results quantitatively. In this table the average absolute bias for all combinations of reproduction methods and listening positions is given for the audio-only situation. The averages were obtained by averaging the 5 absolute means biases for audio-only for each of the plots in figure 4.5. Also the average standard deviation for each situation is given. For the audio-visual situation only the average standard deviation is given at this point, not the average absolute bias, since for the audio-visual situation we want to have information about the direction of the bias (towards or away from the image). The quantitative influence of the presence of video on the localization bias will be analyzed now.

configuration	**No Video** $\langle \lvert\text{bias}\rvert \rangle$ **(m)**	**No Video** $\langle \text{s.d.} \rangle$ **(m)**	**Video** $\langle \text{s.d.} \rangle$ **(m)**
Single Speaker, Position 1	0.06	0.17	0.20
Single Speaker, Position 2	0.03	0.22	0.22
WFS, Position 1	0.05	0.22	0.21
WFS, Position 2	0.08	0.31	0.22
Phantom Sources, Position 1	0.11	0.40	0.30
Phantom Sources, Position 2	0.06	0.34	0.28

TABLE 4.2. Localization results, averaged over subjects and source positions

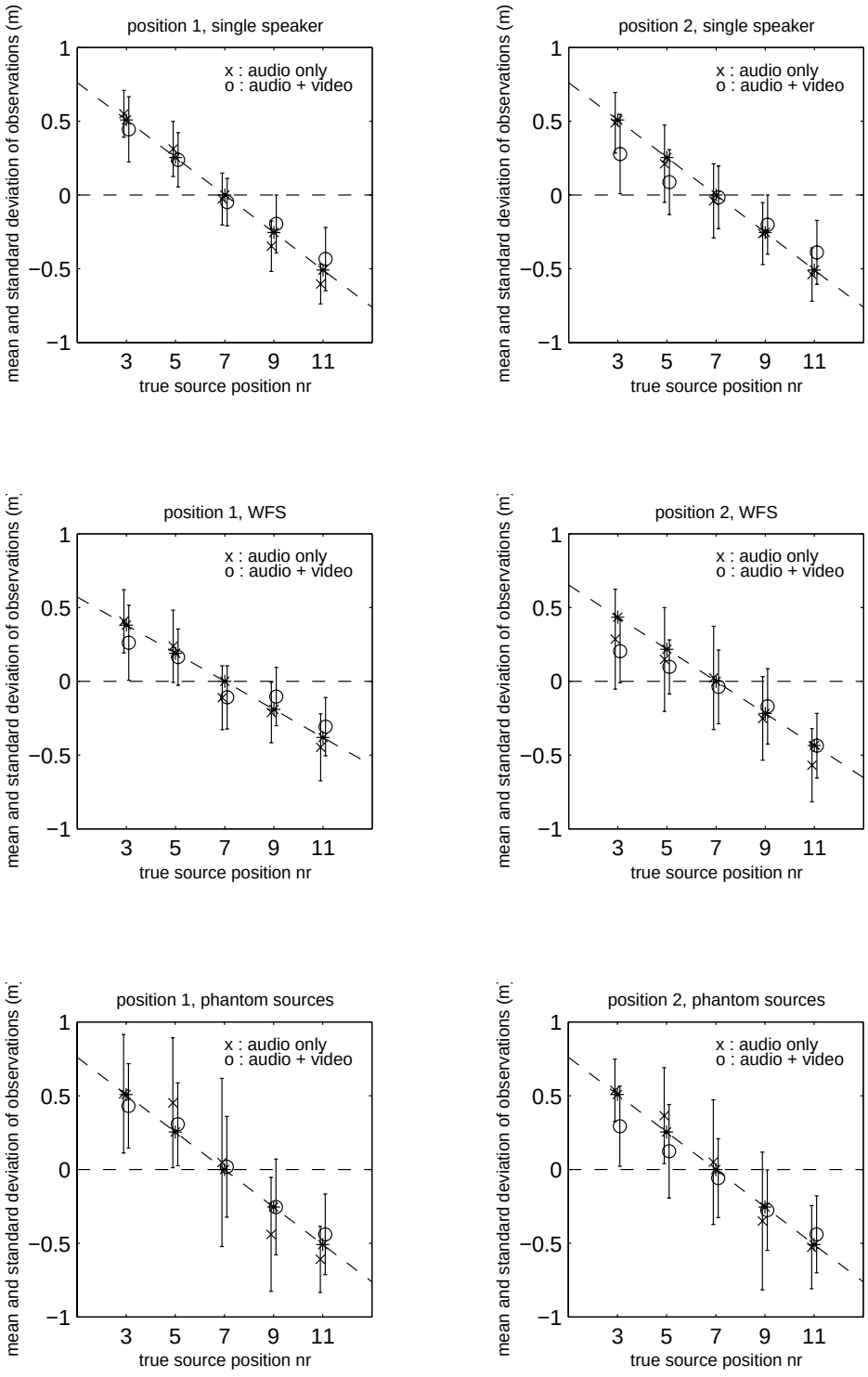


FIGURE 4.5. Localization results for single-speaker (top), WFS (center) and phantom source image (bottom) reproduction, averaged over all subjects. Indicated are locations of observation means for audio-only (x) and audio+video (o) and standard deviation of the observations (error bars). The asterisks on the dashed line indicate the true source positions. The left plots are for listening position 1 (1.5 m), the right plots for listening position 2 (3 m).

Influence of Video on Localization Bias. In order to quantify the influence of video on the localization bias, the errors in the observations have been calculated. Since we wanted to get information to which extent the localization was pulled towards the image, the sign of the errors had to reflect the direction of the error relative to the image position (towards or away from the center) instead of the general qualification ‘too high’ or ‘too low’. It was decided to give a positive sign to deviations towards the image and a negative sign to those away from the image. Therefore, the following definition was used for the errors:

$$\varepsilon = \pm(\text{perceived position - true position}) \text{ (m)}, \tag{4.2}$$

with the plus sign for true source positions below the center and the minus sign for sources above it. Coordinates are positive for positions above the center and negative below it.

Then an analysis of variance (ANOVA) was performed to reveal significant differences between the mean errors in the situations with video and without video.

ANOVA is a widely used statistical method to investigate the probability that two or more sets of experimental data are samples from populations with a common mean (or even samples from the same population) and as such gives information whether the found differences in means between sets of experimental data can be considered as statistically relevant or not. The main output of the method is the so-called ‘*F*’ statistic. Qualitatively, this is a measure that compares the variability *between* the different data sets to the variability *within* the sets. If the data sets are samples from the same population then both variabilities should be equal and the *F*-statistic (which is the ratio of both) should be close to unity. When the sets are actually samples from populations with different means then the variability between sets will be greater than the variability within the sets so that the *F*-statistic is larger than unity. If the *F*-statistic is larger than a certain value then the difference in means can be considered to be significant. This is quantified by the ‘*p*’ statistic, which indicates the chance that the value of the *F*-statistic (which has a so-called ‘*F*-distribution’), given the degrees of freedom in the experiment, is larger than the value that was found. Usually when a value of $p<0.01$ is found, the difference in means between the sets is considered to be highly significant, while often also $p<0.05$ is used as a sufficiently strong criterion. A more detailed description of the ANOVA method is given in Appendix B.

The ANOVA analysis of the difference between mean errors in the ‘video’ and ‘no-video’ conditions was performed separately for all six combinations of reproduction method and listening

distance. Observations were averaged over all subjects and all source positions except the center one, since all deviations for that position are by definition away from the center.

Table 4.3 shows the results for the different situations.

It is seen that in all cases the effect of video is highly significant (p<0.01), with in all cases a mean shift towards the image in the order of 0.1 m.

Concerning the center position, it is seen from the histograms in figure 4.3 and the localization graphs in figure 4.5 that the standard deviations of observations for this source position are smaller for the audio-visual situation than for the audio-only situation, while the means biases are about the same (i.e., small), indicating that the center position is indeed more firmly anchored to the image position.

statistic	Single Speaker, Position 1	Single Speaker, Position 2	WFS, Position 1	WFS, Position 2	Phantom Sources, Position 1	Phantom Sources, Position 2
F statistic	43.8	35.1	27.8	9.0	17.7	25.4
p statistic	.0000	.0000	.0000	.0029	.0000	.0000
⟨ε⟩ , no video (m)	-.07	+.01	-.04	+.01	-.12	-.06
⟨ε⟩ , video (m)	+.05	+.14	+.08	+.10	+.02	+.10

TABLE 4.3. Results of ANOVA tests and mean observation errors for comparison of audio-only and audio-visual situations. Positive values for errors indicate a bias towards the center (image position), negative values are away from the center. The center position itself was excluded from this analysis.

4.1.6 Conclusions

The results from this experiment show that sound localization accuracy in the vertical plane is not very high, even for the single-speaker case, which was the most accurate. The localization standard deviation in this case was 0.17 m for the listening position close to the screen, corresponding to an angle of 6.5 degrees (making the localization blur twice this value, i.e., 13 degrees). Since it is desired that conference participants are allowed to approach the screen while still having a natural reproduction, results for positions relatively close to the screen should be used to determine the needed vertical resolution.

Results for WFS reproduction were comparable to those of the single-speaker case.

From the results it is clear that phantom source imaging does not work well for vertical source positioning. Subjects in most cases localized sources as coming mainly from either the upper or lower speaker, while center source positions produced a double, ambiguous image. This confirms the conclusions of Start ([Star97]) on this matter and rejects those of Somerville ([Some66]).

Comparison of audio-only and audio-visual situations showed that the presence of a matching video image significantly shifted the localization towards the image by an average of 0.12 m for the single-speaker reproduction. This, combined with the localization standard deviation that was found, leads to the conclusion that a distance between neighboring source positions as large as 0.6 m is allowed, without the occurrence of distracting discrepancies between the positions of the video image and the auditory source. This minimum separation is based on a minimum distance to the screen of 1.5 m, so it corresponds to a required vertical source positioning accuracy of only 22 degrees; if even closer distances to the screen are to be allowed without the occurrence of noticeable discrepancy between the auditory and visual source, the maximum array separation in terms of distance is lowered correspondingly.

With the conclusions given above, it seems unnecessarily complex to use WFS techniques for the vertical source positioning, since this requires a distance between the loudspeakers that is smaller than when single-speaker reproduction is used and requires more computational power and reproduction channels.

In conclusion, the most suitable configuration for sufficiently accurate vertical sound source positioning in a life-size videoconferencing system appears to consist of several horizontal array bars positioned above each other behind the screen, separated by up to several decimetres, in which sources are assigned to the array closest to the true position.

One additional point to keep in mind: this experiment focused on determining how well people can localize a sound source in the median plane in an audio-visual situation. It does not say anything about whether observed discrepancies are distracting or even irritating. This is a matter that could be investigated further and would probably make the requirements even less critical. A first clue to this might be that several subjects reported that even in cases where they suspected to hear a not too large discrepancy between the positions of the image and the sound, they had no problem accepting that the sound could have come from the image position. When the discrepancy became too obvious, they tended to immediately localize the sound as coming from an extreme position above or below the image.

4.2 Correspondence Of Perceived Source Positions In Auditory And Visual Modalities: Single Source Experiments

As was explained in section 3.3.2, the combination of two-dimensional video projection and true perspective audio reproduction techniques like WFS potentially introduces discrepancies between what we see and what we hear, due to the fact that the perspective of the video projection is only correct in the unique viewpoint, while the audio perspective with WFS is correct in the whole listening area. In particular, it can be expected that discrepancies in source direction are perceived, i.e: the source is heard from another direction than the direction in which we see it. Additionally, even in the viewpoint the perceived depth of the visual scene is in general not equivalent to the true depth of the original scene because of the absence of several important cues for visual depth perception in the 2D projection, so that in general the auditory and visual perceptions of depth also do not match perfectly.

It was argued in section 3.3.2 that it is difficult to predict exactly how severe the problems are in practice, since we are not simply dealing with two independent modalities, but rather with two modalities that interact, so that the perception in one modality is influenced by the perception in the other. Therefore, it is of vital importance to investigate these potential problems, in order to get a clear idea of the question whether we can expect these problems to occur in a practical audio-visual system such as the life-size videoconferencing system that is the subject of this thesis and, if so, how serious these problems are and, finally, if we can prevent these problems from occurring if it is concluded that it is necessary.

This and the next section (section 4.3) describe several perceptual experiments that have been carried out to try to find answers to these questions. They are divided into experiments that use a single sound source (this section) and experiments that use multiple simultaneous sound sources (section 4.3). Furthermore, each source condition (single and multiple) was investigated in both a relatively objective experiment (section 4.2.3 and section 4.3.1 about source lateralization and source identification, respectively) and a highly subjective experiment in which subjects had to grade the discrepancy (section 4.2.4) or ‘naturalness’ (section 4.3.2) of the audio-visual scene.

But first, in the next two subsections, the audio-visual source material and set-up that were used in all these experiments will be described.

4.2.1 Construction Of Audio-Visual Source Material

For the experiments that are described in this and the next section, a visual projection was needed that had a true perspective (scale 1:1 in all dimensions including depth) when viewed from the unique viewpoint (to be defined below). This was done as follows: The camera was positioned in front of the screen (diagonal size 2.54 m) that was going to be used for the projection, with the center of the lens aligned with the center of the screen. The optical system of the camera was adjusted in such a way that the captured image exactly covered the whole screen. This means that when the recorded material is later projected on the same screen with the projector adjusted such that the image again covers the whole screen, objects that were at the position of the screen during recording will be displayed 1-on-1 and the screen becomes a visual window to the imaginary space "behind" it. It is easy to verify that in this special set-up the viewpoint of the projection is simply the position of the camera at the time of recording. After setting up both the recording and projection equipment, the set-up can be checked by measuring for known objects on known positions the position of their projected image on the screen and then determining the corresponding viewoint by geometric construction. The viewpoint resulting from this procedure should then coincide with, or at least be close to, the camera position. Using this method a visual scene was constructed with one person standing at three different positions in the same room:

person #1. at a position with, when seen from the projection side, a lateral displacement of 1 m to the left and 2 m behind the center of the screen,

person #2. standing at a central position 1 m behind the screen and

person #3. at a position having a lateral displacement of 1.5 m right of center and 3 m behind the screen.

This was achieved by capturing the 3 scenes separately with the camera at a fixed position and then superposing them on top of each other, using digital video editing techniques to make the identical parts of each image (except the "base layer" which was a snapshot of the empty room) transparent. This resulted in the visual scene shown in figure 4.6. The viewpoint in this case is on the line through the center of and perpendicular to the screen at a distance of 2.74 m from the screen and 1.37 m above ground level. Table 4.4 shows the coordinates of the source posi-

tions relative to both the viewpoint and the screen.

The audio material that was used was a monaural close-mic recording of a male voice reading a continuous text.

It is important to note that the visual source material that is used in the experiments in this section and section 4.3 is a still image of the room with the three persons, rather than a motion video recording in which the three persons are seen producing the speech that is the audio source material for the experiments. This choice was made for practical reasons. It was realized, however, that the use of a still image rather than motion video may influence the results of the experiments to a certain extent, because it is known that the 'visual capture' effect is particularly strong if the moving lips of the speaker are visible (see section 3.3.1). The implications of the use of a still image instead of motion video for the results and conclusions of the experiments will be discussed in section 4.3.3, when the overall conclusions of the experiments are drawn.

Source Coordinates	1	2	3
Relative to Viewpoint	(-1 , 4.74)	(0 , 3.74)	(1.5 , 5.74)
Relative to Screen	(-1 , 2)	(0 , 1)	(1.5 , 3)

TABLE 4.4. Source coordinates (*x*,*z*) in meter, relative to the viewpoint and the screen.

FIGURE 4.6. The resulting perspective image that was used in the audio-visual experiments. When viewed from the viewpoint, the visual perspective corresponds to the 3 persons standing (from left to right) at a lateral displacement of respectively 1 m to the left, 0 m and 1.5 m to the right of the center and at depth levels of 2 m, 1 m and 3 m from the screen, respectively.

4.2.2 Audio-Visual Set-Up

For all experiments that are described in this and the next section the same hardware was used. The set-up for the WFS audio reproduction consisted of a horizontal array with a total of 32 small loudspeakers with a spacing of 12.7 cm. In the experiments in which besides WFS reproduction also stereophonic and discrete-loudspeaker reproduction were used, a subset of the array was used, as will be described below.

The video material was stored on and played back from the hard disk of a digital video workstation and projected by a ceiling-mounted high-quality video beamer on a projection screen with a 4:3 width:height ratio and a diagonal size of 2.54 m. The screen was made of material that can be considered to be acoustically transparent. For practical reasons the screen was installed a few decimetres in front of the array. This was taken into account in the experiments. The audio material was stored on the same workstation and was sent directly to the input of the audio processing hardware. A schematic drawing of the set-up is shown in figure 4.7.

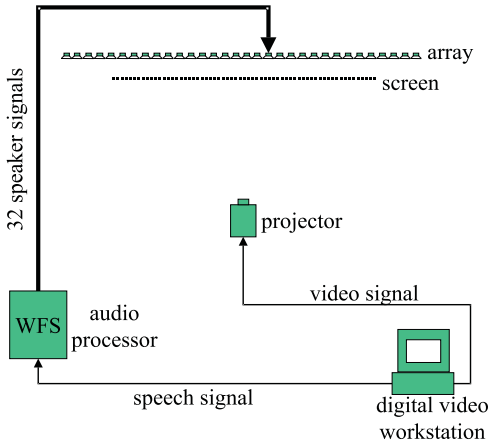


FIGURE 4.7. Hardware set-up used in the experiments.

4.2.3 Experiment A: Lateral Source Positioning

The objective of this first experiment was to investigate the effect on the perceived correspondence of the auditory and visual source positions when a 2D video projection is combined with a sound reproduction having the true (corresponding to the original real-life scene) depth. Therefore, the depth of the sound source was kept fixed to the 'true' value in this first experiment. The degree of freedom in this experiment was therefore the lateral position of the sound source.

4.2.3.1 Experiment description

The experiment was set up in the following way: the perspective visual scene as described in section 4.1.2 (figure 4.6) was projected on the screen. The subject was positioned at a certain observation point, seated on a chair with the eyes and ears at about the same height as the center of the screen. One of the three sources was chosen (at random) by the PC that controlled the experiment. The sound source was then positioned (by WFS) at a random lateral position and at the depth level corresponding to the true source position (as given in table 4.4). The initial lateral position was chosen randomly from a range that depended on the depth level in such a way that a constant angular range was the result when observed from the correct viewpoint.

The subject was told (on his monitor) which was the target position (1, 2 or 3) and was instructed to position the sound source at the position that he/she felt matched the situation pictured on the screen best. To do this, the subject could change the lateral position of the sound source using a graphical user interface on a computer monitor by pressing buttons labeled 'left' and 'right'. The change of lateral position by a button-press equaled 1 degree for a subject at the viewpoint, which is assumed to be approximately the JND for lateral shifts of sound sources. To prevent that the subjects had to press a button many times before reaching the desired position there were also 'coarse' buttons that changed the source position by 5 degrees. When the subject felt that the sound source position corresponded with his perception of the visual geometry he pressed a button labeled 'OK', after which a new source position was chosen and the procedure was repeated. There was no time limit for pressing the 'OK' button. The subjects received no feedback about the position at which the sound source was currently located. Also, there was no indication on the screen at all (for example in the form of a top-view of the situation) what the 'expected' 3D interpretation of the geometry should be, so we actually investigated the complex interaction between subjects' 2D-to-3D interpretation of the geometry of the visual scene and the corresponding sound source position expected by the subjects. The actual instructions that were given to the subjects before the experiment as well as screen shots of the user interface can be found in Appendix A.2.

Using this procedure, we do not get an actual indicator for how 'annoying' a certain discrepancy between visual source position and expected sound source position is (this will be discussed in section 4.2.4), but we get an estimation of the lateral sound source position interval that can be considered to correspond 'naturally' to each combination of source number and observer position. If for a certain source the obtained intervals for different observation posi-

tions have no overlap, then it will be clear that it will be difficult to position the sound source at such a position that it appears as natural for all observers (in the case that the true depth level is maintained, as was the case in this experiment).
The subjects carried out the whole experiment at three different observation positions (figure 4.8):

1. the viewpoint
2. a position 1 m to the right and 1 m closer to the screen relative to the viewpoint
3. a position 1 m closer to the screen relative to the viewpoint

Each subject performed the sound source positioning 5 times for each of the 3 source positions at each observation position, so at one observation position each subject handled 15 stimuli (presented in random order).
This experiment was carried out in combination with experiment B described below (section 4.2.4), so that completion of the whole experiment (A+B) for one observation position took about 20 minutes. Subjects were free to decide whether they wanted to complete the whole experiment (3 observation points) at once or in two or three individual sessions.
Six normal hearing subjects participated in the experiment.
The amplification of the system was set such that for source position 2 (central person on the screen) the SPL at listening position 1 (the viewpoint) of the reproduced source material corresponded to a typical SPL for speech of a person located at the corresponding distance (assuming a typical speech level of 60-65 dB(A) at 1 m distance).

4.2.3.2 Results

In the top graph of figure 4.9 the overall results are shown for observation position 1 (the viewpoint). The asterisks indicate the means of the subjects' responses for each of the three source positions in terms of angle relative to the line through the position of the subject parallel to the axis of symmetry of the set-up. Also shown is the standard deviation of the responses (length of error bars is 2 times the sample standard deviation). In this case, the direction from the observation point to the image on the screen (open circle) and the "true" virtual source position (cross) are the same, as indicated in the figure by the fact that their positions coincide. As we can see, the subjects were quite accurate in positioning the sources at the "correct" position. The means of the responses coincide with the true positions, as expected.

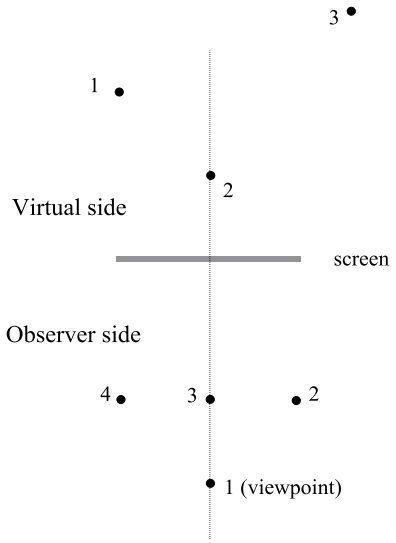


FIGURE 4.8. Observation positions and virtual source positions used in the experiments. Observation position 3 was only used in experiments A (section 4.2.3) and B (section 4.2.4), position 4 only in experiments C (section 4.3.1) and D (section 4.3.2).

The center graph of figure 4.9 shows the results for observation position 2 (off-axis and too close). For this position, the directions of the images on the screen and the true source positions do not correspond. Here we see an interesting phenomenon: the mean responses of the subjects are clearly pulled towards the visual image, but not completely for source positions 1 and 2, indicating that to a certain extent there seems to be some depth interpretation of the visual scene. This conclusion is even more justified by the fact that the 95% confidence intervals of the means (not shown in the figure for reasons of visual clarity) for positions 1 and 2 do not include the position of the visual image. The reason that this effect doesn't occur for source position 3 probably arises from the fact that positioning the sound source somewhere between the direction of the visual image and the true source position required placing it at a position "outside the screen" (see figure 4.8), which was reported by the subjects to be highly unnatural. In this case, the subjects therefore preferred to position the sound source completely at the position of the visual image.

The bottom graph of figure 4.9 shows the results for observation position 3 (on axis, too close to the screen) are shown. They are as expected: the mean of the subjects' responses matches the true position for source 2 (which is on an on-axis position) and is shifted slightly towards the visual position for both position 1 and 3.

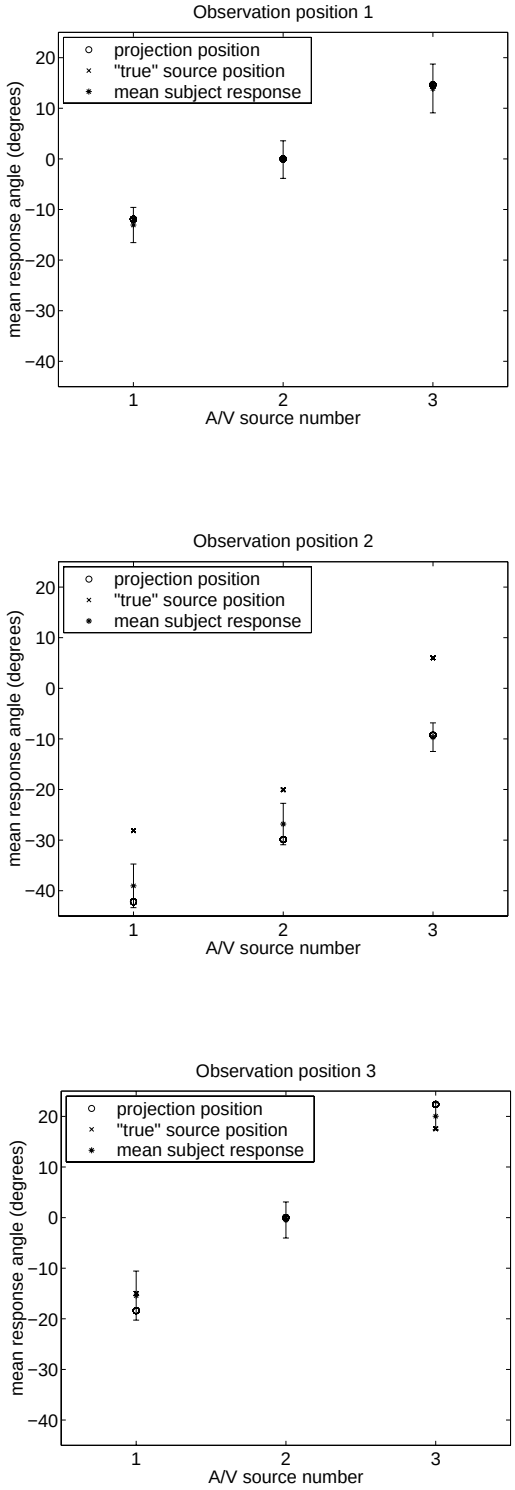


FIGURE 4.9. Localization responses of all the subjects for observation position 1 (top), 2 (center) and 3 (bottom). The asterisks are the means of the subjects' responses as seen from the observation position. Also the standard deviation of the responses is shown. The open circles indicate the direction of the visual image as seen from the observation position and the crosses represent the "true" source positions.

4.2.3.3 Conclusion from experiment A

As expected, the subjects were accurate in positioning the sound sources at the 'true' position when they were seated at the viewpoint. For the non-viewpoint positions we found that the subjects' preferred sound source positions are pulled towards the visual image, but not completely, indicating that some depth interpretation of the two-dimensional image occurs.

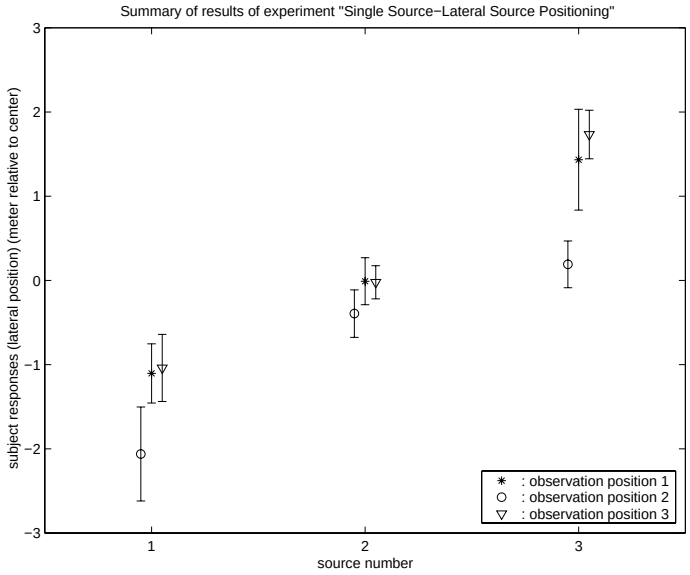


FIGURE 4.10. Summary of the results of experiment A. Horizontal axis: source number, vertical axis: lateral source position intervals preferred by subjects (bars contain 75% of subjects' responses).

Figure 4.10 gives an overview of the results of experiment A. It shows, for the three sources used (horizontal axis), the range in which subjects positioned the sound source (vertical axis, in meters relative to the center of the screen, intervals contain 75% of the subjects' responses) for the three different observation positions (indicated by different symbols).

As can be seen, for visual sources 1 and 3 there is no region where the corresponding preferred sound source intervals of all three observation positions overlap. This means that for those visual sources, it is not possible to position the corresponding sound sources at such positions that the audio-visual reproduction is perceived as being natural by all observers in the room and the audio-visual reproduction will not be perceived as being natural by all observers in the room when the sound sources are reproduced at their corresponding true source positions. For source 2, reproducing the source at the true source position results in less severe problems, due to the central position of this source in the used geometry.

4.2.4 Experiment B: Discrepancy Grading

4.2.4.1 Experiment description

As indicated in the previous section, the results of experiment A give a good indication of the sound source position ranges that are perceived as most natural for several observation positions, but they do not give an actual indication of the 'degree of annoyance' that subjects feel when the sound source is located at a position that they perceive as not being the optimal position. Therefore, combined with experiment A described in the previous subsection, an experiment was carried out to investigate the perceived discrepancy, for different observation positions, between the perceived positions of the auditory and visual sources when the sound source is positioned at the 'true' position.

For this experiment a 5-point impairment scale according to ITU 562-3 ([ITU90]) was used. The meanings of the five points of this scale were that when observing the audio-visual scene the discrepancy between visual and auditory source positions was:

1. ‘imperceptible'
2. 'perceptible, but not annoying'
3. 'slightly annoying'
4. 'annoying'
5. 'very annoying'

After a subject had finished experiment A at a certain observation position, this second experiment was carried out for the same position.

The procedure was similar to that of the previous experiment: the PC selected one of the three visual sources and the voice was reproduced by WFS from the true (virtual) position of that source as seen from the viewpoint (given in table 4.4). Subjects were told which of the three sources was the target source and were then asked to rate the observed discrepancy between what they perceived visually and aurally. The actual instructions to the subjects for this experiment are given in Appendix A.2.

The same 6 subjects participated as in experiment A.

4.2.4.2 Results

In the top graph of figure 4.11 the discrepancy grading results are shown for each of the three source positions with the subjects seated at observation position 1. Indicated are the means of all the subjects' grades (asterisk), the 95% confidence interval of the mean (short error bar) and the standard deviation of the responses (long error bar). As expected, the discrepancy is rated to be very small: the 95% confidence intervals of the means for all three source positions are completely between scores '1' and '2', or in other words: subjects hardly noticed any discrepancy, as should indeed be the case with the subjects sitting at the viewpoint. As can be seen from the standard deviation, subjects were also reasonably consistent in their grading.

The center graph show the results for observation position 2. For this position we see that a significantly larger discrepancy is perceived by the subjects, with the mean score going from 1.4 (for observation position 1) to 3.5 for source position 1, from 1.7 to 2.3 for position 2 and from 1.6 to 3.4 for position 3. Especially for source position 1 the increase in 'annoyance' is quite serious. This can be explained from a geometrical analysis of the situation which shows that indeed the (geometrically) expected discrepancy between the directions from which a subject sitting at position 2 observes the visual image and the sound source is largest for source position 1. Also note that now the standard deviations of the grades are larger than for observation position 1, indicating that subjects were less consistent or did agree less about how annoying the discrepancies were.

In the bottom graph the results are shown for observation position 3. They are comparable to those of position 1, indicating that the distorted depth interpretation of the visual scene (which is the main effect of sitting too close to the screen, the expected discrepancy between visual and auditory source directions is only small) has little effect on the perceived discrepancy.

4.2.4.3 Conclusion from experiment B

Comparing the grading results for observation positions 1 and 3 to those of position 2, we see that a rather serious degradation in perceived correspondence is observed even for this quite moderate lateral distance from the viewpoint. This seems to indicate that indeed in practical situations, where several people will be participating in the conference, sitting or standing at different positions in the same room, annoying effects may occur when the sound sources are placed at their 'true' positions.

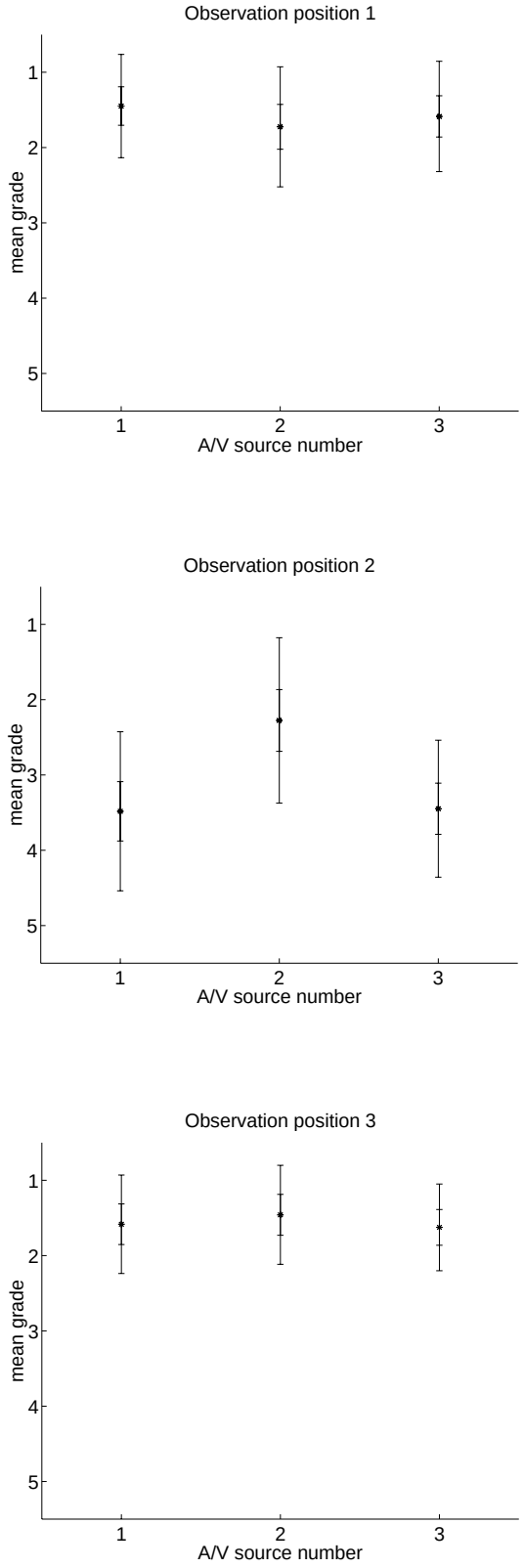


FIGURE 4.11. Discrepancy grading results for observation position 1 (top), 2 (center) and 3 (bottom). Indicated are means (asterisks), 95% confidence interval of the means (short error bars) and standard deviation of the responses (long error bars).

The fact that the results are very similar for observation positions 1 and 3 shows that the main concern should be the directional discrepancy due to lateral shifts from the viewpoint. The distorted depth impression caused by observing from too close or too far from the screen seems to be of less importance.

4.3 Correspondence Of Perceived Source Positions In Auditory And Visual Modalities: Multiple Source Experiments

4.3.1 Experiment C: Speaker Identification In A Multiple Speaker Situation

As explained in section 1.2.1, one of the reasons to start investigating the application of WFS in videoconferencing was the expected improvement (because of the realistic spatial source separation that is associated with WFS) of the ability to identify a specific speaker when several persons on the remote side are talking at the same time. Given the results of the experiments described in section 4.2, however, this may not be so evident any more, since the fact that we are necessarily using 2D video projection introduces some discrepancies between the perception in the auditory and visual modalities, especially in the perception of source direction. Therefore the following experiments with multiple simultaneous sound sources were carried out to investigate this issue.

4.3.1.1 Experiment description

The visual set-up for this experiment was the same as in the experiments described in section 4.2. The purpose of this experiment was to determine whether reproduction of the voice using WFS facilitates the observer's task of identifying which of several persons on the screen is speaking in a multiple-speaker situation, as compared to stereophonic reproduction using two speakers at the sides of the screen and reproduction with a configuration of discrete loudspeakers.

The procedure was as follows: the computer chose one of the 3 projected persons as "the target speaker". By routing in the DSP's, the "target" speech signal played back from the Digital Video Workstation was assigned to this person, while other speech signals were assigned to the two other persons to act as interfering "noise". Then loudspeaker driving signals were calculated according to one of the three reproduction methods used (WFS, stereo, discrete) to reproduce the source signals in such a way that when observed from the viewpoint, the reproduced

sound source positions matched the true source positions (given in table 4.4) as closely as possible (so they corresponded to the projected images on the screen when viewed from this position).
The target speech signal used was a spot microphone recording of a male voice speaking in a normal way, while the two interfering speech signals were recordings of a female voice speaking random sentences. The female voice was the same for both interfering speech signals, but the actual speech content was different.
The reproduction strategies used for the three cases were as follows:

WFS. the three sources were reproduced using the 32-element loudspeaker array (figure 4.7), with the relative levels of the three sources correctly balanced automatically in the WFS processing, according to their respective positions in space.

Stereo. reproduction was done with one array loudspeaker to each side of the screen, the distance between them (stereo base width) chosen such that it corresponded to the viewpoint-array distance in the "standard stereo set-up" sense. For each source the angle under which it was viewed from the stereophonic 'sweet spot' (which coincided with the viewpoint in this case) was calculated and from this and the angles of the two stereo speakers the balance for the left and right speaker was calculated according to the ‘law of sines’ for stereophonic reproduction (eq. 4.1). The source signals were amplitude weighted according to their relative distance from the sweet-spot (1/r law) and the level of the stereo reproduction was globally weighted to match the reproduced levels of the WFS sources.

Discrete. Five equidistant individual loudspeakers were used, covering the width of the projection screen. For each of the sources the angle relative to the viewpoint was calculated and the signal was assigned completely to the individual loudspeaker whose angle was closest to this. In the used source- and loudspeaker configuration this resulted in a negligible spatial mismatch of the images on the screen and their corresponding discrete loudspeakers (maximum error about 1 degree when seen from the viewpoint). As in the case of stereophonic reproduction, the relative source levels were scaled according to the respective distance to the viewpoint and the total level was matched to the WFS level.

The experimental procedure was relatively simple: the three signals (target- and two interfering speech signals) were reproduced using one of the three methods described above and the

task of the subject, who was seated at one of several observation positions, was to indicate which of the three persons on the screen was in his/her opinion most likely to be producing the target speech signal. The instructions can be found in Appendix A.3.

Subjects performed the task at three observation positions: positions 1, 2 and 4 (see figure 4.8). Position 3 was not used in this experiment since it was found from the experiments described in section 4.2 that for this source position the results were similar to those in the viewpoint. Therefore this position was replaced by observation position 4, which is symmetrical with respect to position 2 on the observer side but for which the perspective of the audio-visual scene is different.

Eleven subjects participated in the experiment. Each condition (a combination of a specific target position, reproduction method and observation position) was presented 3 times to each subject, so that the total result for each condition is made up out of 33 subject responses.

4.3.1.2 Results

In the top graph of figure 4.12 histograms are shown of the subjects' responses at observation position 1. In the left column are responses to the sources reproduced by WFS with, from top to bottom, the target source located at source position 1, 2 and 3 respectively. The second column gives the results for stereo reproduction and the third column for discrete loudspeaker reproduction. It is clear from the results that the subjects had no problems whatsoever to correctly identify the target speaker in this case (percentage of correct responses were 98%, 94% and 100% for WFS, stereo and discrete, respectively).

The center graph shows the results for observation position 2. Here we see that several things are different from the results for position 1. In the case of WFS there is a clear shift in identification to the right, especially for a source at position 1 which is in many cases perceived as corresponding to position 2 (top-left histogram). For stereo we see that a source at position 2 was almost always identified as being position 3. This shows clearly that even for this quite moderate deviation from the stereophonic sweet spot, the spatial image breaks down completely and the sound is heard as coming almost entirely from the loudspeaker that is closest to the subject. This will become even more clear when we look at the results of the quality grading experiment for this situation in the next subsection. In the case of discrete reproduction the identification is only slightly less perfect as for the viewpoint. The percentages of correct identification for position 2 were 71% (WFS), 59% (stereo) and 90% (discrete).

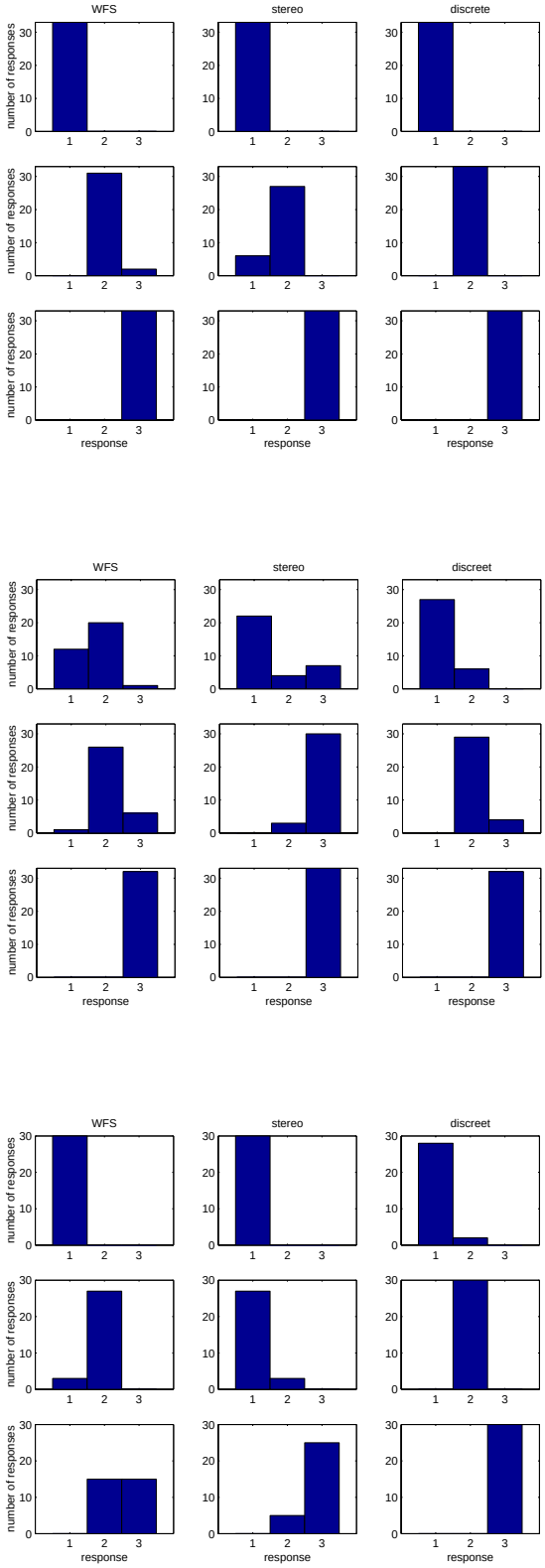


FIGURE 4.12. **Histograms of subject responses at observation position 1 (top), 2 (center) and 4 (bottom). Left column shows responses for WFS for, from top to bottom, target positions 1, 2 and 3. Middle column is for stereo, right column for discrete.**

In the bottom graph the results are shown for observation position 4. This position is located symmetrically with respect to position 2, relative to the central axis, but the positions of the persons on the screen are not, so the perspective is actually different. For stereo we observe the same phenomenon as at observation position 2 and discrete still has almost perfect identification. For WFS there is less misidentification than at position 2, which can be understood by looking at the geometrical situation. Also, in the case of WFS, subjects reported that they found the task easier at this position than at position 2. Percentages of correct identification for position 4 were 80% (WFS), 64% (stereo) and 98% (discrete).

4.3.1.3 Conclusion from experiment C

From this experiment it can be concluded that as far as speaker identification is concerned, discrete loudspeaker reproduction is the most stable method. WFS reproduction with the sound sources located at their true positions in combination with the 2D video image results in some misidentifications which can be explained from a geometrical analysis of the situation. Stereophonic reproduction is very detrimental to identification performance, because the spatial image breaks down completely for off-axis observers.

4.3.2 Experiment D: Multiple Sources: Realism Grading

Although experiment C gave an indication how well subjects were able to identify a specific speaker out of several competing speakers, it says little about how well the perceived auditory scene matched the perceived visual scene. It is not difficult to imagine that situations can arise in which the three sound sources are indeed perceived as being spatially separated so that identification is relatively easy, but with a clearly perceptible discrepancy between the actually perceived directions of the sound sources and their associated visual images. So, although the ability of users to correctly identify individual speakers is an important quality of a videoconferencing system, it is not sufficient for a completely natural communication.

Therefore, also an indication was needed of how well the subjects thought the auditory and visual scenes they perceived matched for all the combinations of reproduction method, source position and observation position that were used in experiment C.

For this reason, experiment C was combined with a discrepancy grading experiment, comparable to the one described in section 4.2.4.

4.3.2.1 Experiment description

After completion of experiment C at a certain observation position, subjects were presented with the same sequence of stimuli and were asked to grade to which extent the spatial lay-out of the three audio sources appeared as being realistic, given the visual presentation of the three persons in visual space.

Again a 5-point scale was used, adapted from ITU 562-3 ([ITU90]). The meanings of the scale numbers were:

1. 'completely realistic'
2. 'realistic, but some noticeable discrepancy'
3. 'moderately realistic'
4. 'hardly realistic'
5. 'completely unrealistic'

Subjects again repeated each condition (combination of target source position, reproduction method and observation position) 3 times. Since in this case the distribution of the three speech signals among the 3 source positions was of no relevance, the results of all observations for a specific combination of observation position and reproduction method were pooled together, so that each condition was actually evaluated 99 times in total.

4.3.2.2 Results

In table 4.5 the grading results are shown for the three reproduction methods and the three observation positions. For each situation the mean grades, the standard deviation of the subjects' responses and half the length of the 95% confidence interval of the mean are given.
If we first look at the results for observation position 1 (the viewpoint), we see that for all three reproduction methods the reproduction is rated as being quite realistic, with the grades being best for the discrete loudspeaker reproduction. We would expect that in this case the grades for WFS would match those of discrete. However, the WFS reproduction seems to be rated as being slightly less realistic than discrete. The reason for this might be an effect of the listening room. Since we are using linear arrays of loudspeakers for WFS, the reproduced sound fields are actually cylindrically symmetrical around the array. This means that also a significant part of the acoustical energy is being radiated towards the ceiling and floor, resulting in strong

reflections in a room with insufficient acoustical damping. This is indeed the case in the listening room where the experiment was carried out. An additional reason might be that the WFS sources are perceived as being slightly broader than the discrete sources due to the use of discrete arrays. This seems to be in agreement with the fact that the grades are about the same as for stereo, of which it can also be said that the observed source width is somewhat larger than for a single loudspeaker.

observation position	statistic	WFS	stereo	discrete
1	mean grade	2.52	2.35	1.77
	standard deviation	1.16	1.05	1.07
	95% conf. int.	±0.23	±0.21	±0.21
2	mean grade	3.21	4.45	1.99
	standard deviation	1.02	0.80	0.89
	95% conf. int.	±0.21	±0.16	±0.18
4	mean grade	2.84	4.66	1.68
	standard deviation	1.07	0.56	0.73
	95% conf. int.	±0.22	±0.12	±0.15

TABLE 4.5. **Results of the ‘Multiple Sources: Realism Grading’ experiment. The table shows the mean grade, the standard deviation of the subjects’ grades and half the length of the 95% confidence interval of the mean for each of the three observation positions and each of the three reproduction methods.**

Now we look at the grading results for observation position 2. As expected, we see that the realism rating for stereo breaks down to being almost 'completely unrealistic': the spatial image is completely lost and the sound is heard as coming from the stereo speaker (just beside the screen) closest to the subject. The realism grades for WFS have degraded compared to position 1, as could by now be expected from the results of the previous experiments. Similarly, the grades for discrete loudspeaker reproduction remain about the same.

Finally, we look at the results for observation position 4. They are comparable to those for position 2, with the subtle difference that the grades for WFS are slightly better for position 4, which is in agreement with the aforementioned fact that subjects also reported that the identification task was a bit easier at this position than at position 2.

4.3.3 Conclusion And Discussion Of Results Of Experiments A-D

From the experiments A-D that have been described in section 4.2 and section 4.3, it can be concluded that adding true perspective (in this case: WFS) audio reproduction to a two-dimensional video projection indeed has the unfortunate side-effect of giving rise to discrepancies between perceived auditory and visual source positions for non-viewpoint observers. The effects of lateral shifts from the viewpoint should be of more concern than shifts from the viewpoint to positions too close or too far from the screen. The effects are noticeable and perceived as annoying, also in situations that can be expected to occur in the practical use of a life-size videoconferencing system. Still, it was expected that applying WFS in a videoconferencing system has positive effects on other aspects of the performance of a videoconferencing system, for instance, the speech intelligibility of the system, which will be investigated later on in this chapter (section 4.5). A way to reduce the observed negative effects is described in the next section (section 4.4).

Motion Video versus Still Images. We end this section with a discussion of the implications of the fact that the video material that was used consisted of a still image instead of motion video, a decision that was made for practical reasons.

In a real videoconference, the reproduced speech sound will be accompanied by moving video images of the person that produces the speech. It was discussed in the overview of spatial audio-visual interaction in section 3.3.1 that, although the ‘visual capture’ effect is generally present in situations in which audio and video stimuli that are considered to ‘belong together’ are present simultaneously, the strength of the effect will be particularly strong in the case of audio-visual speech, i.e., when an audio speech signal and moving video images of the corresponding lips of the person producing the speech are present simultaneously. Effectively, this means that the discrepancy that can be allowed between the location of the audio speech signal and the location of the visual image, before being noticed by observers, can be larger when motion video is used than in the same situation in which the moving video is replaced by a still image. Also, for the same spatial displacement between the audio and video, the perceived discrepancy will probably be smaller in the case of moving video than in the case of a still image. For our experiments, this means that the discrepancies that were perceived were probably larger than would have been found with motion video, so that in a real videoconference the problems caused by the discrepancies between the directions of the audio and the video are

probably not as dramatic as they may seem from the results of the experiments. However, in the configurations of source positions and observation positions that were used in the experiments, discrepancies of up to 15 degrees were present. From the results of Komiyama ([Komi89]) it is clear that also in the case of speech with moving video, such discrepancies are perceptible and to some extent even annoying. Non-expert observers in his experiment reported a perceptible discrepancy in 79% of the cases in which a discrepancy of 15 degrees was present. For expert observers this was even 96%. Furthermore, the non-experts judged the discrepancy as being ‘slightly annoying’ in 21% of the cases. The experts judged the discrepancy as ‘slightly annoying’ in 36% of the cases, while in 20% of the cases it was judged as being ‘annoying’ and even as ‘very annoying’ in 16%. Since the source- and observer positions that were used in experiments A-D can be regarded as being quite realistic for a life-size videoconferencing system such as the one that is the subject of this thesis (note that the maximum lateral displacement of the observers was only 1 m, which is far from extreme), we may conclude that the qualitative conclusions of experiments A-D given above will still hold when motion video is used. However, it is difficult to make quantitative predictions of how the results of each experiment will change when motion video is used. The main relevance of experiments A-D is therefore not to make quantitative predictions of maximum allowable discrepancy, perceived annoyance, degradation of identification performance and naturalness, but rather to call attention to the general problem and its importance for the design of audio-visual systems, including, but not limited to, videoconferencing systems.

4.4 Compression Of Reproduced Auditory Depth

From the experiments described in the previous sections it is clear that WFS reproduction of direct sound sources (voices of conference participants) as virtual sources located at their true positions, in combination with 2D video projection, is likely to result in an unsatisfactory correspondence between the audio and video perspective perceived by conference participants located at positions other than the viewpoint. Especially in the case of lateral displacements relative to the viewpoint, a discrepancy between the perceived directions of visual images and corresponding sound sources easily occurs, even if the lateral distance from the viewpoint is only moderate.

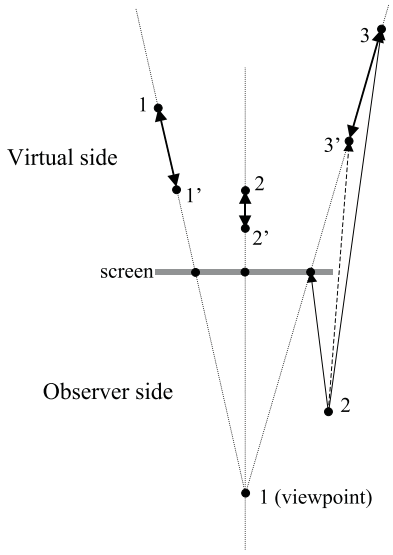


FIGURE 4.13. Principle of compression of the depth of the auditory scene to avoid discrepancies between perceived auditory and visual source directions. Sources 1, 2 and 3 are reproduced from positions 1', 2' and 3', respectively, closer to the viewpoint. For the combination of observation position 2 and source position 3 it is shown how an 'auditory depth reproduction factor' of 0.52 reduces the mismatch from 15 degrees (uncompressed) to the acceptable value of 11 degrees (compressed).

Of course, these problems could be kept to a minimum by limiting the range of positions where participants are allowed to be during the conference, for instance, behind a table not too far from the 'virtual window' to the remote side. In a true two-way system these precautions would have to be taken at both sides. However, this is not a desirable solution for the problem, since the objective of this research project was the design of a conferencing system in which participants are free to be located anywhere in (or at least in a large part of) the room.

A simple yet effective way to reduce the problems, while still retaining the condition that participants are relatively free to be located anywhere in the room, is to compress the depth of the auditory scene to some extent, by pulling the sources closer to the screen along the line from the viewpoint to the original source position, so that the reproduced sound sources are actually closer to the screen than their 'true' position. For an observer at the viewpoint, the perceived directions of the visual and corresponding auditory sources still match, while for all non-viewpoint observers, the angles between the auditory sources and their corresponding visual sources become smaller, thus reducing the chance that annoying discrepancies occur. Of course, when the compression of the depth is taken to the extreme, this solution reduces to the

case of a single loudspeaker on the screen at the position of the projected image. The principle is sketched in figure 4.13 for the source and observer configuration of figure 4.8 that was used in the experiments described in the previous sections.

The optimum amount of compression is a compromise between reducing the discrepancies that are expected to occur to such an extent that they are no longer perceptible (or are at least no longer annoying), while the advantages of reproducing the sources using WFS, such as the more natural spatial separation of sources and better speech intelligibility (to be discussed in section 4.5), are preserved. As a reference, the maximum angle for which the 'ventriloquist' effect (see section 3.3.1) is effective for audio-visual speech, in the sense that even by expert observers no annoying discrepancies are perceived, can be taken, which is about 11 degrees ([Komi89]). With this criterion in mind, the necessary reduction of the reproduction depth of the auditory scene can be determined. Preferably, the 'auditory depth reproduction factor', i.e., the ratio of the distance from the virtual sound source to the projection on the screen (along the line from the viewpoint through the position of the projection on the screen) in the compressed situation to the distance in the uncompressed situation, should be determined for each individual source. This is done by first deciding upon the desired areas of the room in which conference participants should be allowed to be located at both the local and remote side. Then, for each position within the 'source area', the auditory depth reproduction factor is determined that is needed in order to keep the discrepancy below the maximum allowable value within the whole 'observation area'. This results in a 'map' of the source area that assigns an auditory depth reproduction factor to each source position within that area. As an example, figure 4.13 shows how for an observer at position 2 the original discrepancy of 15 degrees for source position 3 is reduced to the acceptable value of 11 degrees by applying an auditory depth reproduction factor of 0.52, i.e., the distance from the sound source to the projection on the screen is reduced to 52% of its original value.

Since from the experiments described in section 4.2 and section 4.3 we know that a mismatch of auditory and visual source *direction* causes problems rather than a mismatch of auditory and (interpreted) visual source *distance*, it seems plausible that it is allowed to apply the proposed depth compression without having to fear for introducing additional problems. To check this the following informal test was carried out.

Alternately, one of the three source positions used in the previous experiments (figure 4.8) was chosen and a speech signal was reproduced from an initial position on the line from the view-

point through the original source position. This compressed source position was chosen randomly from the interval between the point where the line intersects the screen (‘fully compressed reproduction’, corresponding to an auditory depth reproduction factor of 0) to the original source position (‘uncompressed reproduction’, corresponding to an auditory depth reproduction factor of 1). A gain that was randomly chosen from the interval -3 dB to +3 dB, corresponding to a distance cue of 0.7 to 1.4 times the ‘real’ distance of the reproduced source, was applied to the source signal. Subjects, who were seated at either observation position 1 or 2, as used in the previous experiments, had an interface with which they could move the source along the line through the viewpoint. The task of the subjects was to position the sound source at such a position that it seemed to correspond naturally with what they saw on the screen. For this, they had a graphical user interface comparable to the one of experiment A (section 4.2.3) with buttons labelled 'To Front, Coarse', 'To Front, Fine', 'To Back, Coarse', 'To Back, Fine' and 'Ok' to confirm. ‘Coarse’ buttons changed the distance by 1/10th of the total range for each source (so the step-size depends on the original, uncompressed, source distance) while the ‘Fine’ buttons had a step-size of 1/25th of the total range. Subjects were not instructed to pay attention to specific cues such as loudness or source direction, only to adjust the scene so that it seemed natural.

Given the results of the experiments in section 4.2 and section 4.3, it was expected that at non-viewpoint observation positions, subjects would position the sound source such that the direction of the sound source matched that of the visual image, as they did in experiment A. The results in the viewpoint are of principle interest in this experiment. At this position the direction of the sound source always matches the visual image, so the remaining property that subjects could influence was the source distance. It was expected that under these circumstances, the main cue for estimating the distance would be the level of the source signal, so that for the viewpoint the positions at which subjects positioned the sound sources would have a strong correlation with the initial random gain that was applied to the signal.

As this experiment was mainly intended to confirm the seemingly plausible assumption that compression of the depth does not result in the introduction of an additional negative effect for observers in the viewpoint, it was carried out in a rather informal way. Already after a couple of subjects it was clear that, as expected, for non-viewpoint listening positions the preferred sound source position was indeed determined mainly by the perceived direction of the visual image as seen from the observation position, while at the viewpoint the preferred source posi-

tion was determined mainly by the initial random gain of the sound source. Subjects at the viewpoint reported they had great difficulty to determine when to press the ‘OK’ button. This implies that it is indeed possible to compress the perspective of the reproduced sound field in such a way that at non-viewpoint observation positions the perceived discrepancy of visual and auditory source directions is reduced, while at the viewpoint no difference is observed if proper level compensation is applied, meaning that the gain for each sound source is adjusted in such a way that, even though the virtual sources are now closer to the observer than they were originally, the level of each source still correponds to the level of a source at the original, uncompressed, source position.

4.5 Speech Intelligibility

From the audio-visual experiments that were described in section 4.2 and section 4.3 it has become clear that the expected benefits from using true perspective audio reproduction such as WFS in a life-size videoconferencing system with 2D video projection are to a certain extent counteracted by the fact that the perspective of the video projection is only correct in the viewpoint, resulting in noticeable and sometimes even annoying discrepancies between perceived auditory and visual source directions for non-viewpoint participants. In the previous section we have seen that by slightly reducing the perspective of the reproduced audio scene, these problems can be reduced or even prevented, but one could ask the question whether the hypothesis that using WFS in a videoconferencing system leads to a better system still holds.

However, an important aspect of any speech communication system that has not been looked at yet is the so-called ‘speech intelligibility in noise’: the ability of users of the system to understand what the speaker of interest is saying when there is interfering sound from other speakers or noise sources in general.

It is a well-known fact that spatially separating target- and noise sources leads to a better speech intelligibility. The effect is described for example in ([Plom81]). While sources reproduced with WFS are spatially separated in a natural way for all listeners, this is not the case for most conventional systems in which sources are either not separated (single-speaker reproduction) or not correctly separated for non-sweetspot listeners (stereophonic reproduction).

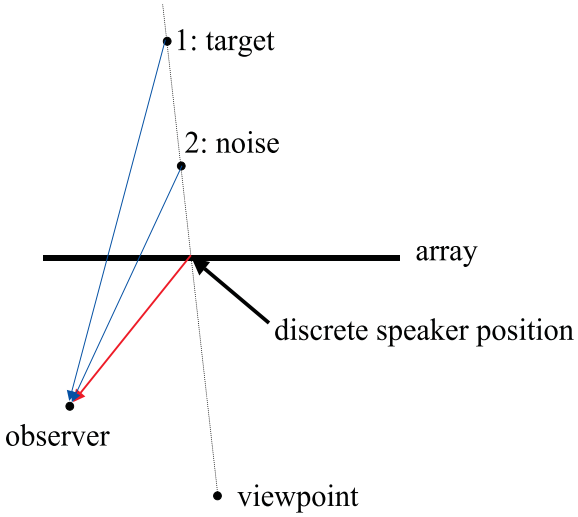


FIGURE 4.14. When voices are reproduced by WFS, the sources are spatially separated, while they are reproduced by the same single loudspeaker when a 'discrete loudspeaker' system is used in situations where the two sources are located on the same line through the viewpoint. For this reason it is expected that the speech intelligibility will be better for WFS.

In the case of the 'discrete loudspeaker' system described earlier, in which each sound source is assigned to the individual loudspeaker that, when seen from the viewpoint, it is closest to, the accuracy of the spatial source separation depends on the configuration of the discrete loudspeaker set-up and on the positions of the sources. Furthermore, the spatial source separation is fixed by the combination of these two factors and does not change with changing listening position as would be the case in real life.

An interesting case is when two remote conference participants are located at such positions in the room that, when seen from the viewpoint, they are standing more or less in the same 'line-of-sight' at different distances from the screen. In this situation, the two voices are reproduced by the same loudspeaker in the discrete loudspeaker system, so there is no spatial source separation, while when WFS is used, the two voices are virtually reproduced at their true positions and their true separation is maintained. For an observer in the viewpoint, the two WFS sources are localized in the same direction, but when the observer moves away from the viewpoint the voices become spatially separated as is the case in real life. This situation is sketched in figure 4.14.

These considerations suggest that reproduction of voices by WFS should result in an improved speech intelligibility of the system. To test whether this is true, the experiment described in this section was carried out.

4.5.1 Description Of Method

The method that was used to investigate the speech intelligibility in noise ('noise' in the sense of any competing sound source; this can be an actual stochastic noise signal but also a second speech signal) for different configurations is the standardized method developed by Plomp & Mimpen ([Plom79]). This method aims at finding the signal-to-noise ratio for which the chance of understanding a sentence correctly is 50%: the Speech Reception Threshold (SRT). This is achieved by a so-called up-down method.

The general procedure is this: the system reproduces a sentence of target speech together with a noise signal and the subject has to try to correctly repeat the target sentence. If the subject fails to correctly repeat the sentence, then the reproduction levels of the next sentences are increased in fixed SPL steps until the subject is able to repeat a sentence correctly. Likewise, if a sentence is repeated correctly, the level of the next sentence is lowered by the same fixed SPL step. After a sufficient number of sentences the method should converge to a state in which the level of the target speech alternates between a level at which the speech can just be repeated correctly and a level at which it can not. Halfway between these two levels lies the level for which the chance of understanding the sentence correctly is 50%: the Speech Reception Threshold.

Essential for the method is the careful construction of the set of sentences that is used in the experiment. These sentences should all have an equal chance of being understood correctly. Also, it is very important that the sentences are in the native language of the subjects. Plomp & Mimpen constructed 10 lists of 13 sentences for the Dutch language which have been recorded and balanced in level in such a way that the resulting recordings have equal chance of being understood correctly. A Compact Disc with these recordings ([TNO88]) was used in this experiment. Each of the recorded sentences on this CD is accompanied by a noise signal that has the same long-term spectrum and the same loudness as the speech signal.

In practice the method will not necessarily converge to a state in which there is a strict alternation between correctly and incorrectly reproduced sentences, so instead after a certain number of sentences the method is assumed to have reached a stable state and the average level of the last *n* sentences is taken as the 50% threshold. Plomp & Mimpen ([Plom79]) found that in their method it is sufficient to use a list of 13 sentences after which the levels of the last 10 sentences, including the 14^{th} one that is not actually reproduced but for which the level is known from the correct or incorrect repetition of sentence 13, are averaged to obtain the 50% level.

The corresponding signal-to-noise ratio is taken as the value of the SRT.

The first sentence of the list is reproduced with a signal-to-noise ratio at which it is impossible to understand the sentence. This first sentence is repeated with increasing level until it is repeated correctly. Note that only this first sentence is reproduced repeatedly until it is understood correctly, all other sentences are only reproduced once.

Plomp & Mimpen have shown ([Plom79]) that an SRT test carried out in this way leads to results that are reproducible and not sensitive to the use of a specific list of sentences, in the sense that the standard deviation of SRT results obtained in multiple runs and by using different lists is only about 1 dB. A typical example of the results of such an experiment is shown in figure 4.15.

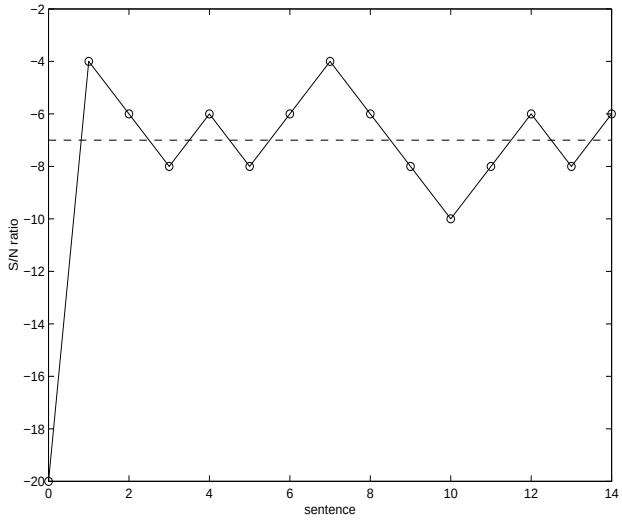


FIGURE 4.15. Example of the result of one series of 13 sentences for one subject. The horizontal axis denotes the number of the sentence in the list, the vertical axis is the SNR at which a sentence was reproduced. The SRT for this series is calculated as the mean of the SNR of the last 10 sentences, which in this case is -7 dB (dashed horizontal line).

4.5.2 Experimental Set-Up

Combinations of two source configurations and two listening positions were investigated in the experiment for both the WFS and discrete loudspeaker system, resulting in 8 individual conditions. In each source configuration there was a target source and a noise source. The target source positions used were the same as source positions 1 and 3 as used in the audio-visual experiments of section 4.2 and section 4.3 (see table 4.4 and figure 4.6) and the listening positions were the same as listening positions 1 and 2 from those same experiments (see

figure 4.8). The positions of both noise sources were chosen such that for an observer at listening position 1 they were located on the same straight line as the corresponding target source, while for an observer at listening position 2 they were spatially separated by an angle of 10 degrees. It was therefore expected that for an observer at position 1 there should be no significant difference in SRT between the WFS and discrete system, while for an observer at position 2 the SRT for WFS is expected to be lower (meaning that the speech intelligibility is better) than both the SRT for the discrete loudspeaker system and the SRT for the WFS system at listening position 1. From an extrapolation of the experimental data reported by Plomp and Mimpen in [Plom81], an SRT difference in the order of 2 dB could be expected.

The source and listening positions that were used are shown in figure 4.16. The 8 conditions that were tested in the experiment are listed in table 4.6. The SPL step size that was used in this experiment was ±2 dB and the initial SNR, measured at the listening position, for sentence 1 was -20 dB, low enough to ensure that it was impossible to understand the target speech.

The target source and noise source levels were both calibrated to a value of 65 dB(A), corresponding to a normal speech level, for each of the 8 conditions separately at the corresponding listening position. It could be argued that since in practice the system will be calibrated at the viewpoint only, it would make sense to also do this in this experiment. The result of this, however, is that when an observer is changing his position from the viewpoint to position 2, the relative levels of the noise and target source also change because their relative distances to the observer change. This has an extra influence on the speech intelligibility, in addition to the effect resulting from the fact that the two sources are spatially separated when moving from the viewpoint to position 2. The reason that it was decided to calibrate the levels at the two listening positions separately is that in this experiment we wanted to investigate solely the effect of the spatial source separation on speech intelligibility that results from using WFS.

The experiment was controlled by a PC on which the audio recordings of the sentences were stored. From this PC the sentences were played as stereo .wav files, with the speech signal in the left- and the corresponding noise signal in the right channel. These two audio signals were fed into the DSP system, which either reproduced them as two individual WFS sources or assigned the signals to the appropriate discrete loudspeaker in case of discrete loudspeaker reproduction. Appropriate gains were applied to the signals to match them to the calibration measurements for the condition that was being tested. The PC that played the audio material also applied the ±2 dB level steps according to the correct or incorrect responses of the subject.

The calculation of the loudspeaker gains and delays for WFS source synthesis was controlled by another PC.

The subjects were seated on a chair at one of the 2 listening positions, facing the front of the loudspeaker array. No effort was made to limit the subjects' head movements, since this experiment was meant to be representative of a real-life use of the system. A picture of the listening set-up is shown in figure 4.17.

For WFS reproduction of the signals, the same loudspeaker set-up was used as in the experiments of section 4.2 and section 4.3, meaning a linear array in front of the subject consisting of 32 loudspeakers with a spacing of 12.7 cm (total length 4 m).

For discrete loudspeaker reproduction the target- and noise source signals were both reproduced by the same individual array loudspeaker since the target- and noise source positions were chosen such that they were on the same line-of-sight when viewed from the viewpoint. Of course, the loudspeaker that was used was a different one for the two source configurations.

16 normal hearing subjects participated in the experiment. Each subject performed the test for 8 lists of 13 sentences which took about half an hour for each subject to complete. The order of the 8 lists was randomized among subjects by dividing the 16 subjects into two groups of 8 for both of which a Latin Square[1] randomization procedure was used to determine the order of the 8 conditions for each subject. The order of the 8 lists of sentences that were used was the same for all subjects.

The experimental procedure was fairly simple: a sentence plus noise signal were played and the subject had to repeat the target sentence. The experiment leader, who had the list of correct sentences, checked the response and indicated on the PC whether the response was correct or incorrect. The judging was strict: a sentence had to be repeated 100% correct to be acknowledged as such. After a correct answer the PC lowered the level of the speech signal by 2 dB and the next sentence was played at this new level. After an incorrect answer the level was increased by 2 dB and the next sentence was played. The level at which each sentence was played was stored on the PC, forming the raw results of the experiment.

1. A Latin Square matrix is an (n x n) matrix containing n different elements arranged in such a way that each element occurs only once in each row and column. Latin Square matrices are commonly used in the design of experiments to randomize the order of conditions between individual runs. In our case the Latin Square technique is used to construct two matrices of (*8 subjects* x *8 conditions*), each element containing one of the 8 conditions, in order to balance out possible systematic effects such as small differences between the individual lists of sentences, subject fatigue etc.

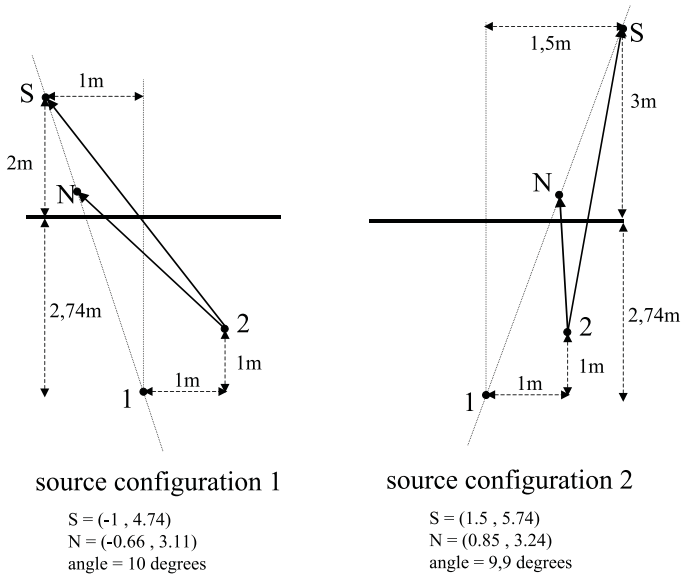


FIGURE 4.16. Source and listening positions used in the Speech Intelligibility experiment. On the left are the target (S) and noise (N) source positions for source configuration 1 and on the right those for source configuration 2.

FIGURE 4.17. Experiment set-up with the two listening positions (chairs). The subject is pictured sitting at position 2, the empty chair is listening position 1 (viewpoint).

condition	1	2	3	4	5	6	7	8
source configuration	1	1	1	1	2	2	2	2
listening position	1	1	2	2	1	1	2	2
reproduction system	WFS	discrete	WFS	discrete	WFS	discrete	WFS	discrete

TABLE 4.6. The 8 conditions that were tested in the SRT experiment.

4.5.3 Results

The results of the SRT experiment are shown in figure 4.18. Indicated are the SRT means averaged over the 16 subjects and the 95% confidence intervals of the means for the 8 conditions that were tested. Table 4.7 shows, for each combination of source configuration and listening position, the difference between the mean SRT for WFS and the mean SRT for the discrete loudspeaker set-up and the value of the *p*-statistic of a one-way analysis of variance (ANOVA) comparing the results of WFS and the discrete loudspeaker set-up.

It is seen that the mean SRT for WFS is lower than that for discrete loudspeakers in all cases. The standard deviation for all 8 conditions was of the order of 1 dB, as Plomp & Mimpen also found for their test ([Plom79]).

Also, it is seen that the SRT's for the discrete loudspeaker system are more or less the same for all four combinations of source configuration and listening position. In fact, an ANOVA comparing the discrete loudspeaker results of the four combinations to each other shows that the found differences between the mean SRT's for the four combinations are not statistically significant (p=0.17). This is as expected, since this indicates that for the discrete loudspeaker set-up the SRT does not change significantly when the listening position is changed, as should be the case, since the target- and noise source are reproduced from the same single loudspeaker.

Now let us look at the results at listening position 1. For the combination of source configuration 1 and listening position 1, the results are not as expected. The SRT that is found for WFS is 2.0 dB lower than the SRT for the discrete loudspeaker system (table 4.7), whereas it was expected that the SRT's would be the same for listening position 1. From an ANOVA of the results it is also clear that this SRT difference is highly significant (p=1.1*10^{-4}, table 4.7). For source configuration 2 and listening position 1, on the other hand, the result is as expected, since only a small difference of 0.4 dB is found between the mean SRT for WFS and the discrete loudspeaker set-up, which is not significant (p=0.30, table 4.7).

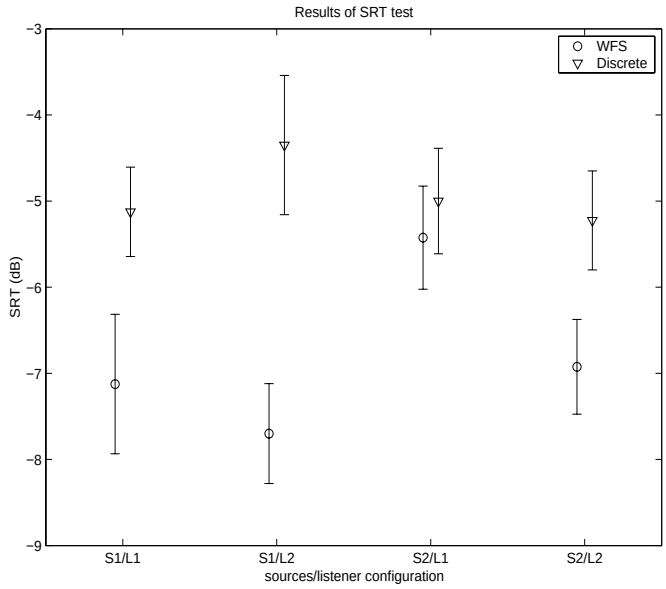


FIGURE 4.18. Results of the SRT experiment. The horizontal axis denotes the combination of source configuration and listening position, 'S1/L2' meaning 'source configuration 1/listening position 2' (see figure 4.16). The vertical axis denotes the SRT (in dB) in terms of SNR. The open circles and triangles indicate the means for the WFS and discrete loudspeaker set-up, respectively, averaged over the 16 subjects. The error bars indicate the 95% confidence intervals of the means.

configuration	S1/L1	S1/L2	S2/L1	S2/L2
SRT difference (dB)	2.0	3.4	0.4	1.7
p-statistic	1.1e-4	5.5e-8	0.30	8.1e-5

TABLE 4.7. Mean SRT differences between WFS and discrete loudspeaker set-ups in dB (discrete-WFS) and *p*-values of one-way ANOVA of the results of the SRT experiment for the four combinations of source configuration and listening position, showing the statistical significance of the difference in SRT means for the WFS and discrete loudspeaker set-ups in each configuration.

Next we look at the results at listening position 2 and compare them to the results at listening position 1. For both source configurations a highly significant SRT difference between WFS and discrete loudspeakers is found at listening position 2, of 3.4 dB (p=5.5*10^{-8}, table 4.7) for source configuration 1 and 1.7 dB (p=8.1*10^{-5}, table 4.7) for source configuration 2. However, an ANOVA comparing the results for WFS for source configuration 1 at listening position 1 to those for listening position 2 shows that the decrease of the SRT between these two conditions

(0.6 dB) is not significant (p=0.23). Combined with the unexpected SRT difference between WFS and the discrete loudspeaker system that was found at listening position 1 for source configuration 1, this seems to indicate that for this source configuration, the found SRT difference between WFS and the discrete loudspeaker set-up is not due to the spatial source separation that occurs for WFS when the subject moves from listening position 1 to position 2, but is due to an SRT bias that was present between WFS and the discrete loudspeaker system for source configuration 1. The origin of this bias is not clear. One possible explanation might be that since we are using a discrete array with a spacing between the loudspeakers of 12.7 cm, spatial aliasing is present in the reproduced sound field above approximately 1.5 kHz. This may result in colour differences between the sound fields of virtual sources that are located at different positions, which might provide subjects with a cue to separate the signals, not spatially but spectrally. Possibly, the specific positions of the target- and noise source in source configuration 1 were such that they resulted in a significant colour difference between the target- and noise source signals for the WFS reproduction, making it easier for the subjects to separate them, resulting in a lower SRT for WFS than for the the discrete loudspeaker set-up. To check this hypothesis the experiment could be repeated with a source signal that does not contain any frequencies above the Nyquist frequency, although it might be questioned how meaningful it is do a speech intelligibility test using signals that contain no frequencies above 1.5 kHz while normal speech contains relevant energy for frequencies up to about 5 kHz.
For source configuration 2, the results are as expected. As mentioned above, no significant SRT difference is found between WFS and the discrete set-up at listening position 1, while at listening position 2 a highly significant SRT difference of 1.7 dB (p=8.1*10^{-5}, table 4.7) is found between WFS and the discrete set-up and for WFS a highly significant SRT decrease of 1.5 dB (p=4.6*10^{-4}) is found between the results at listening positions 1 and 2. These values are in agreement with the predicted value of about 2 dB that was obtained from extrapolation of the results reported in [Plom81].

The observed SRT differences between WFS and discrete loudspeakers expressed in decibel might not seem very large at first inspection. This, however, is not the case, as rather small differences in SRT can have a very strong impact on the actual speech intelligibility in terms of ‘intelligibility score’, which is defined as the percentage of sentences understood correctly. It can be argued that this is the measure that actually matters when we are looking at the speech

intelligibility of a system.

The approximate shape of the “intelligibility score versus SNR relative to the SRT” curve is shown in figure 4.19 (solid line) in which 0 dB on the horizontal axis corresponds to the SRT and thus by definition corresponds to 50% of the sentences being understood correctly. It is seen that the curve is very steep at signal-to-noise ratios close to the SRT, so that small changes in SNR lead to large changes in intelligibility score. The actual steepness (and thus sensitivity) of the intelligibility-score-versus-SNR curve that results from a specific speech intelligibility experiment depends on the experimental method used. Basically it is the cumulative probability function of the results from the experiment, which are assumed to have a normal distribution. For the test designed by Plomp & Mimpen that was used in this experiment the sensitivity at SRT is about 20%/dB when the individual SRT differences among subjects are taken into account, reducing to about 15%/dB when they are not.

To illustrate the consequences for the results from our experiment let us assume that the solid curve that is shown in figure 4.19 is the speech intelligibility curve that resulted from the experiment for the discrete loudspeaker set-up, source configuration 2, listening position 2. From the overall results of the experiment we have seen that the mean SRT that was found for the same situation but using WFS instead was on average 1.7 dB lower (table 4.7). In effect, this means that for the WFS situation the curve is shifted 1.7 dB to the left (assuming the shape of the curve is the same for both reproduction methods, which is an acceptable assumption). This curve is drawn dashed in figure 4.19. Also shown are two vertical dotted lines: one at the SRT level of the WFS curve and one at the SRT level of the discrete loudspeakers curve. The interpretation of this is as follows: the vertical line through the SRT level for WFS intersects the discrete loudspeakers curve at a point where the vertical axis reads 20%. This means that at an SNR at which subjects understand a sentence 50% of the time for WFS, they only understand it 20% of the time when discrete loudspeakers would be used with this same SNR. This, obviously, is a very large difference. Likewise, the vertical line through the SRT for the discrete loudspeakers intersects the WFS curve at 80%, meaning that at an SNR at which subjects understand the sentences correctly 50% of the time for discrete loudspeakers, they understand the same sentence with the same SNR 80% of the time for WFS. Seen in this light it becomes clear that the seemingly modest differences in SRT, expressed in dB, between WFS and discrete loudspeakers result in a significant improvement of speech intelligibility, especially at signal-to-noise ratios around the SRT.

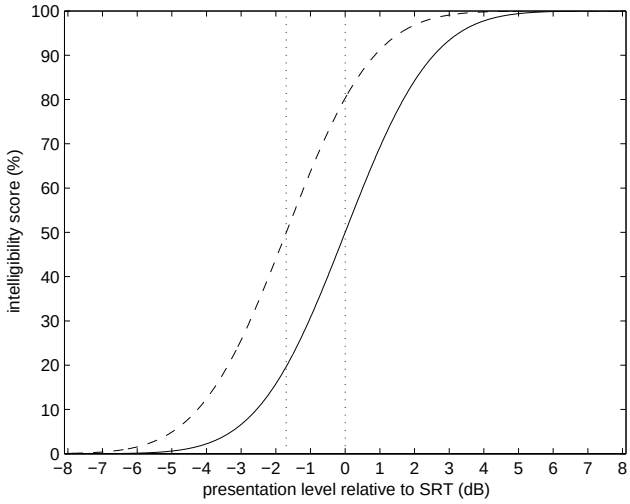


FIGURE 4.19. **Solid line: Intelligibility score (% of sentences understood correctly) versus SNR relative to SRT. The solid line represents the results for an arbitrary condition 'A'. The dashed line is the same curve shifted 1.7 dB to the left, representing the results for a condition 'B' that has a mean SRT that is 1.7 dB lower than that of condition 'A'. Curves have a slope of 20%/dB at the 50% level (after Plomp & Mimpen ([Plom79])).**

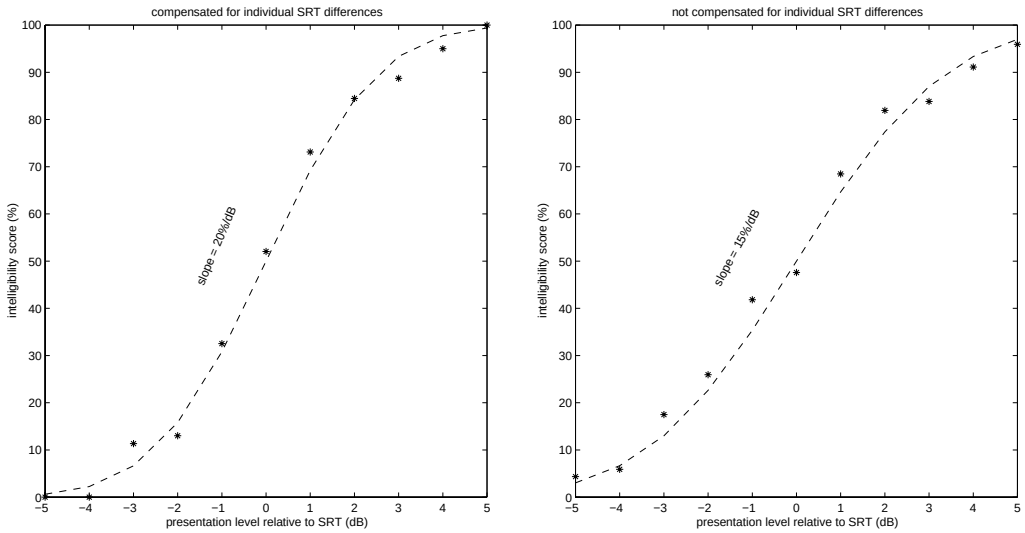


FIGURE 4.20. **Intelligibility score (% sentences understood correctly) versus presentation level relative to SRT (dB). Left plot: compensated for SRT differences among subjects (slope at SRT: 20%/dB), right plot: not compensated for SRT differences among subjects (slope at SRT: 15%/dB). The asterisks are the intelligibility scores extracted from the experimental data, the dashed curves are the expected curves.**

Finally, to check that the experimental data obtained in our speech intelligibility test is actually in close enough agreement with the intelligibility score vs. presentation level curves shown above to justify the above conclusions, the intelligibility score vs. presentation level curve is extracted from the actual data from the experiment.

This is done as follows: for each individual series of 14 sentences the presentation level of each sentence relative to the SRT of that particular series is calculated by subtracting that SRT (calculated as described in section 4.5.1) from each sentence's SNR at which it was presented to the subject. Since the SRT's are not integer dB values and vary from series to series, the resulting relative presentation levels can take any value within a certain interval. Since we want to calculate the percentage of sentences understood correctly at different presentation levels relative to the SRT from a finite number of observations, a discrete set of relative presentation levels for which we want to determine the intelligibility score has to be chosen and the individual observations are binned together to contribute to the intelligibility score at the discrete presentation level they are closest to[2]. Then, for each observation it is determined whether the sentence was understood correctly by the subject or not. This can be determined by checking whether the SNR at which the next sentence was presented is lower or higher than the SNR of the sentence under consideration. The above procedure is carried out separately for all series of 14 sentences (or actually 13, since the 14th sentence was not presented in the experiment as explained in section 4.5.1), resulting in *16* x *8* x *13* (*subjects* x *conditions* x *sentences per series*) transformed observations of the type (*relative presentation level*, *correct/incorrect*). Finally, from these transformed observations the intelligibility score as percentage of sentences understood correctly can be calculated for each discrete presentation level.

The result is shown in the left plot of figure 4.20, in which a 1 dB interval between discrete presentation levels has been chosen. Also shown (dashed) is the expected curve. It is seen that the results match the expected curve very well. Especially important is that the slope of the experimental curve at the SRT (0 dB in the plot) is exactly 20%/dB, the same value that was found by Plomp and Mimpen.

As mentioned before, the slope of the curve at SRT is reduced to 15%/dB when the results are

2. The size of the intervals between discrete levels is a compromise: if the intervals are chosen too small, then the number of observations that fall within each interval and are available to calculate the percentage of correctly understood sentences is too small to obtain a reliable result, if they are chosen too large, the number of levels for which the percentage is calculated is too small to make a meaningful curve-fit.

not compensated for individual SRT differences among subjects, which was achieved in the process of generating the left plot in figure 4.20 by subtracting each subject's individual SRT from his/her personal results. The analysis has also been carried out without compensating for these inter-subject differences by not subtracting the individual SRT's but the mean, averaged over subjects, SRT's as given in table 4.7. This results in the right plot of figure 4.20. Also in this case the results from the experiment are in close agreement with the expected curve.
This confirms that the results that have been obtained are reliable and justifies the conclusions that have been drawn from the found SRT differences between WFS and discrete loudspeaker reproduction.

4.5.4 Conclusions From The Speech Intelligibility Experiment

It has been shown that in situations in which individual conference participants' voices would be reproduced by the same individual loudspeaker in a discrete loudspeaker set-up, reproduction of the voices by WFS leads to a significant improvement of speech intelligibility for non-viewpoint observers, both in terms of SRT and in terms of 'intelligibility score'.
For one source configuration, a lower SRT for WFS than for the discrete loudspeaker set-up was found also with the subjects sitting at the viewpoint, which was not as expected as the target- and noise source were located in the same direction for this listening position. The reason for this result is not clear, but a possible explanation may be that for that specific source configuration, a colour difference was present between the sound fields of the target- and noise source for WFS reproduction, due to spatial aliasing.
Note that in this experiment only the case was considered in which the two sources are located on the same line through the viewpoint and that the experiment was carried out for only two source configurations and one specific discrete loudspeaker set-up. Depending on the configuration of the discrete loudspeaker set-up and the source positions, the speech intelligibility improvement by using WFS could be more prominent or less prominent than was found in this experiment. For instance, when the number of discrete loudspeakers is only small (as is the case in most current systems), so that individual sources are more likely to be assigned to the same individual loudspeaker, or when the differences in distance between sources as seen from the viewpoint are larger, the benefit of using WFS will be larger. On the other hand, in cases where the number of discrete loudspeakers is large, so that sources are more likely to be assigned to different individual loudspeakers, or when the differences in distance between

sources as seen from the viewpoint are smaller, then the benefit of using WFS will be smaller than was found in this experiment.

4.6 Conclusions

In this chapter several audio-visual perception experiments were discussed that were carried out to investigate the perception of spatial audio that includes a realistic reproduction of auditory depth in an audio-visual environment, in particular a life-size videoconferencing system.
First of all, it has been shown that human vertical sound localization, which is already quite poor in an audio-only situation, is even poorer when the sound is accompanied by a matching visual image, as the visual image tends to ‘capture’ the localization of the sound source. A separation between vertical source positions of 22 degrees was found to be acceptable. It can be concluded that in the life-size videoconferencing application it is sufficient to use a set-up that has only a few horizontal array bars that are positioned above each other, in which each source is assigned to the horizontal array closest to that source’s position. Additionally it was shown that stereophonic ‘phantom source imaging’ techniques do not work in the vertical direction.
Then a series of experiments was discussed that were carried out to investigate the possible problems that might occur when true perspective spatial audio reproduction is combined with conventional two-dimensional video projection for observers located at positions other than the unique viewpoint of the 2D video projection. At several observation positions, subjects carried out experiments with a single sound source, in which they had to position the sound source at the lateral position that corresponded to their perception of the visual image on the screen when the source was reproduced at the same depth as the ‘true’ source and in which they had to grade the discrepancies they experienced when the sound source was placed at the ‘correct’ position. It was concluded that in situations that can be expected to occur in practice, positioning the sound sources at their true positions leads to discrepancies between the perceived directions of the auditory and corresponding visual sources that are both noticeable and even annoying for observers located at positions other than the viewpoint of the video projection. Especially lateral displacements relative to the viewpoint are likely to cause problems, even if the displacements are only moderate. Observing the audio-visual scene from positions too close or too far from the screen should be of less concern.
Similar results were obtained in experiments with multiple sound sources, in which subjects had to identify a target speaker out of three visual alternatives and in which subjects had to

grade the realism of the reproduced audio-visual scene at several observation postions. Addionally it was shown that stereophonic reproduction is not suitable for application in life-size videoconferencing.

Given the results of these experiments, a method was proposed to avoid or reduce these problems. Compression of the reproduced depth of the auditory scene reduces the discrepancy between the perceived auditory and visual source positions for non-viewpoint observers. Given the geometry of the system at both the local and the remote side, i.e., the areas in which observers and sources should be allowed to be located, it is possible to determine for each source position which observer position at the other side will exhibit the largest discrepancy and what compression factor is needed to reduce this discrepancy to a value that is within acceptable limits.

It is important to note that although the experiments described above have been carried out within the specific context of a life-size videoconferencing system that combines WFS sound reproduction with 2D video projection, the results and conclusions are of relevance to all audio-visual systems that combine true perspective audio reproduction (not necessarily WFS) and 2D images.

Finally, a speech intelligibility experiment was described that showed that the natural spatial source separation that is achieved by WFS reproduction can lead to a significant improvement of speech intelligibility compared to non-spatialized audio reproduction, especially at signal-to-noise ratios around the Speech Reception Threshold. This is a strong argument to apply WFS sound reproduction in life-size videoconferencing, provided that the audio perspective is properly adjusted as described above. In the experiment, two particular source configurations were investigated in which a target-speech source and a speech-noise source were located on the same line through the viewpoint, so that they were reproduced by the same loudspeaker in the discrete loudspeaker system. For subjects that were located at positions other than the viewpoint, a significant SRT improvement was found for WFS reproduction compared to the discrete system for one of the two source configurations. The results for this source configuration were as expected. For the other source configuration, an SRT benefit for WFS compared to the discrete system was also found for observers at the viewpoint, contrary to expectation. The reason for this was not clear.

CHAPTER 5 *Coloration*

The use of a discrete, finite-length loudspeaker array results, in general, in a frequency response that is not flat across the whole reproduced spectrum, but has some spectral distortions. When these spectral distortions are large enough, coloration artifacts are perceived in the reproduced sound field. How strong the perceived coloration is, depends on several factors, such as the loudspeaker configuration, the source position, the listening position and the source signal.

Although some of this coloration is caused by diffraction at the edges of the array (finiteness), the main contribution to the perceived coloration is due to the spatial aliasing introduced by the finite distance between the individual loudspeakers of the array (discretization).

In a practical application of WFS, there is usually a desire to limit the number of required loudspeakers and reproduction channels as much as possible, for economical reasons, reasons of required processing power and sometimes also reasons of limited transmission bandwidth. In general however, the coloration artifacts that are introduced increase when the distance between neighboring loudspeakers is increased. Therefore, for the practical design of a WFS system it is important to know what distance between the loudspeakers can be allowed, such that no annoying coloration is introduced. Investigating the coloration that occurs in WFS systems with various specifications is the main objective of this chapter.

First, in section 5.1 the definitions of colour and coloration are given. Then, in section 5.2 an overview is given of the psycho-acoustic mechanism of perception of coloration. The mechanism of auditory filtering which is responsible for the limited frequency selectivity of the human hearing system is discussed (section 5.2.1) and simple models are presented to relate the physical spectra of signals to the perceived coloration of those signals, both in terms of the coloration threshold of individual signals (section 5.2.2) and the coloration differences between differently coloured signals (section 5.2.3). The section closes with a short discussion of the phenomenon of 'binaural decoloration' (section 5.2.4).
Then, in section 5.3, a closer look is taken at the coloration that occurs in a practical Wave Field Synthesis system as a result of the spatial aliasing that occurs because of the discretization of the loudspeaker array. First the origin of the coloration will be discussed, after which simulations will be shown of the response of several configurations, which will then be analyzed using the model for the auditory filtering of section 5.2.1. Of particular interest in the present application is the variation of colour as function of the position within the listening area. These spatial colour variations are discussed in section 5.3.2 and the simulated responses of section 5.3.1 are analyzed regarding spatial colour variations, using the model of section 5.2.3.
To check the validity of the described model, which aims to link the physical properties of the sound field to its perceived coloration, and to obtain an idea about the maximum distance between the loudspeakers of the array that can be allowed in the particular application that is the subject of this thesis, such that the perceived coloration stays within acceptable limits, a perception experiment was done in which subjects compared the responses of different configurations to each other regarding coloration. This experiment is described in section 5.4.
Finally, section 5.5 summarizes the conclusions regarding coloration.

5.1 Definitions Of Colour And Coloration

The *colour* of a sound signal is a perceptual attribute for which it is somewhat harder to give a commonly accepted qualitative definition than for several other perceptual attributes, such as 'loudness' and 'pitch'. 'Loudness', for instance, is defined in an American National Standards Institute definition as "that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from soft to loud" ([ANSI60]). Likewise, 'pitch' is defined as "that attribute of auditory sensation in terms of which sounds may be ordered on a scale

extending from low to high" ([ANSI60]). Physically, the pitch of a signal is mainly determined by distinct peaks in the spectrum of the signal, but it is important to note that despite (or, one might argue, because of) the very simple qualitative definition, the pitch of a signal is actually a quite complex property that depends also on several other physical properties of the signal, such as the sound pressure level. The same holds for loudness, as is noted also in the ANSI definition, which continues to state that "the loudness depends primarily upon the sound pressure although it also depends upon the frequency, waveform and duration of the sound.". Furthermore, to determine the pitch or loudness of a given signal, the complex properties of the human hearing system have to be taken into account.

For 'colour' there is no such strict definition as for 'pitch' and 'loudness'. In practice, the term 'colour' is often used as being synonymous to 'timbre', which is defined by the American National Standards Institute as "that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar" ([ANSI60]). 'Timbre' is a term that is commonly used in musical acoustics. For instance, the fact that two instruments sound different when they both play the same musical note with the same loudness can be attributed to the fact that the timbre of the acoustical signal that is produced by one instrument is different from that of the other. Physically, the timbre of a signal is mainly determined by the shape of the spectrum of the signal.

Also slightly different definitions of 'colour' can be found in literature. Salomons ([Salo95]), for instance, defines colour as "that attribute of cochlear sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness are dissimilar". As can be seen, this definition is slightly broader than the former one, as, besides timbre, it also includes pitch. This broader definition of 'colour' is the definition that is used in this thesis, since in analyzing the perceptual performance of a WFS system and in comparing different loudspeaker configurations to each other, we are interested in all aspects of spectral distortion that cause audible differences between the reproduced sound field and the original source field or between different positions within the reproduced sound field, including both timbre and pitch effects.

The audible change, by whichever cause, of the colour of a signal is called *coloration* and the signal is said to be *coloured* compared to the original signal. 'Coloration' is therefore a relative term, whereas 'colour' is not.

At this point, it will be clear that the use of the words 'colour' and 'coloration' to indicate

those perceptual sensations that are associated with the spectral shape of sounds and spectral differences between sounds, respectively, stems from the analogous everyday use of the same words to indicate the optical sensations that are determined by the spectral properties of light, i.e., ‘colour’ for the optical sensation determined by the shape of the spectrum of light and ‘coloration’ to indicate the sensation associated with a change of the spectrum.
The coloration of an audio signal is determined by comparing the signal to a reference signal, preferably the original, uncoloured, source signal. An example is comparing the colour of the output signal of an audio mixer to the colour of the input signal. Sometimes the original signal is not available as a reference or not practical to use, in which case another reference may be used. For example, when judging the coloration introduced by a loudspeaker, it is not possible to directly compare the original electrical input signal to the acoustical output signal, even though the original signal is available. It is possible, however, to compare the colour of the acoustical signal coming out of the loudspeaker to the signal coming out of a loudspeaker that one is very familiar with. In this case, the familiar speaker serves as a reference. Sometimes, the main point of interest is the colour difference between two coloured signals, rather than the coloration of each individual signal relative to the original signal. In that case it is also more convenient to directly compare the coloured signals to each other. This is the case in the perception experiment that is described later on in this chapter in section 5.4.

In many cases, the cause for coloration lies in the fact that, somewhere along the signal path, time-shifted copies of the original signal are generated and added, possibly with modified amplitude and phase, to the original signal. In a concert hall, these time-shifted copies are the reflections of the direct sound at the room boundaries, which all add together with the direct sound signal to make up the impulse response of the hall at the position of the listener. In figure 5.1 two examples are shown of the effect of adding one or multiple reflections to a signal. As can be seen, adding a single reflection leads to a harmonic modulation of the flat spectrum of the original signal (‘cosine noise’), while adding multiple reflections that are irregularly spaced in time leads to an irregularly distorted spectrum.
In a Wave Field Synthesis system, also time-shifted copies of an input signal are generated and reproduced by individual loudspeakers. For band-limited signals that contain only frequencies below the spatial Nyquist frequency, these contributions from individual loudspeakers merge together into a single event which has the same spectrum as the original signal, but for signals that contain also frequencies above the Nyquist frequency this is not the case anymore. In that

case, the contributions from the individual loudspeakers arrive at the listener as individual events, separated in time, causing spectral distortions that are highly place dependent. This will be elaborated further in section 5.3.

It should be noted, by the way, that coloration is not necessarily a negative phenomenon. It can be both pleasant or highly undesirable, depending on the situation and the nature of the coloration. For instance, in a good concert hall some mild coloration that is introduced by the reflections of the hall can add a pleasant ‘warmth’ or ‘brightness’ to the music being played, while in another hall strong individual reflections can cause a highly undesirable ‘comb filter’-like coloration.

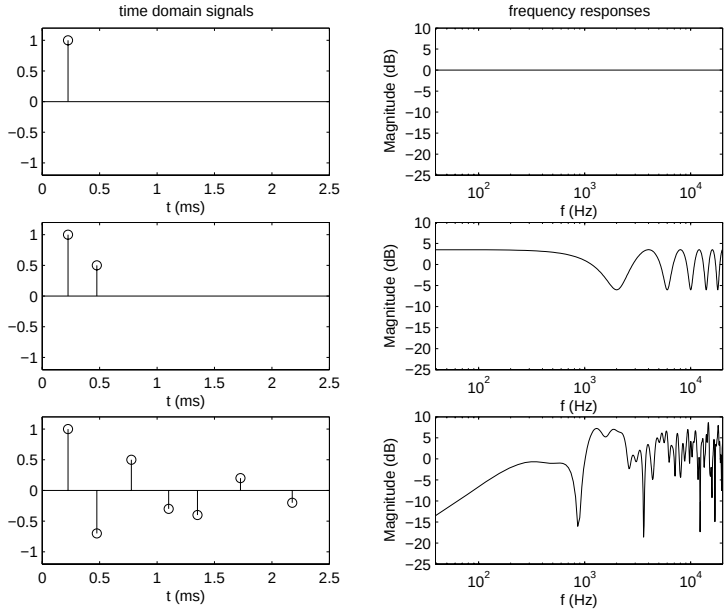


FIGURE 5.1. Examples of how the spectrum of a signal is changed by the addition of reflections. Shown on the left are the time domain signals, on the right the magnitudes of the corresponding frequency responses. Top: original signal. Center: original signal plus single reflection. Bottom: original signal plus multiple reflections.

5.2 Perception Of Coloration

To be able to determine whether or not distortion of the spectrum of a signal will result in audible coloration and, if that is the case, how strong the perceived coloration will be, knowledge is required about how the human hearing system processes and interprets a physical signal entering the ear. A first important stage in this process is the filtering of the physical spectrum by

the so-called *auditory filters*, which transform the spectrum of the acoustical signal into the so-called *internal spectrum* that is the starting point for subsequent neural processing.

5.2.1 Auditory Filters

It is well-known that the human hearing system has only a limited frequency resolution, with the resolution decreasing (getting worse) for higher frequencies. This phenomenon has a physiological origin. In the cochlea, the acoustical signal is spatially decomposed in its frequency components along the length of the basilar membrane, which contains inner hair cells. When stimulated, the inner hair cells trigger the fibres of the auditory nerve that are connected to them. Each inner hair cell, however, is not just stimulated by a single frequency, but by all frequencies within a certain range. Consequently, the nerve fibres that are connected to the hair cell are also triggered by the same range of frequencies, rather than just by a single frequency, which explains the limited frequency resolution of the hearing system. From a signal-processing point of view, one might say that the acoustical signal that enters the cochlea is filtered by a set of auditory filters, each of which integrates all physical signal components within a certain range of frequencies. Zwicker et al. ([Zwic57]) investigated the bandwidth of the spectral integration in the hearing system, which led to the introduction of the 'Bark' measure for the so-called 'critical bandwidth'. They found that the bandwidth of the spectral integration increases with frequency, in other words: the higher the frequency, the worse the frequency resolution.

The shape of the auditory filters has been studied extensively. Patterson ([Patt86]) proposed the following shape for the auditory filter with center-frequency f_c as function of frequency f:

$$W(f, f_c) = \left(1 + \frac{4|f - f_c|}{ERB(f_c)}\right) \cdot \exp\left(-\frac{4|f - f_c|}{ERB(f_c)}\right), \tag{5.1}$$

in which $ERB(f_c)$ is the *Equivalent Rectangular Bandwidth* at frequency f_c: the bandwidth of a filter with unit magnitude which has the same area as the auditory filter with center-frequency f_c, so:

$$ERB(f_c) = \int_{-\infty}^{\infty} W(f, f_c)\,df. \tag{5.2}$$

The values of the ERB for different center-frequencies that were found by several investigators led Moore and Glasberg ([Moor83]) to the following empirical expression for the equivalent rectangular bandwidth of the auditory filters (with f_c in kHz):

$$ERB(f_c) = (6.23 f_c^2 + 93.39 f_c + 28.52) \cdot 10^{-3} \ \ (kHz). \tag{5.3}$$

In figure 5.2 the shape of the auditory filters according to (eq. 5.1) and (eq. 5.3) is shown for various center-frequencies. It is seen that indeed the width of the filters increases for increasing f_c.

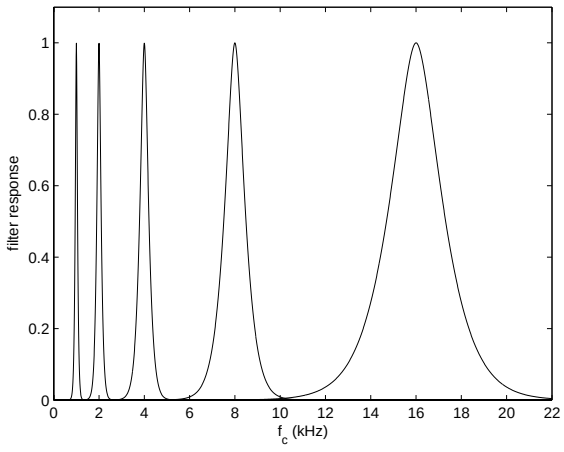


FIGURE 5.2. Shape of the auditory filters according to (eq. 5.1) and (eq. 5.3) for center-frequencies of 1, 2, 4, 8 and 16 kHz.

The magnitude at frequency f_c of the internal spectrum S_{int} of a signal with spectrum S is now obtained by integration of the spectrum S of the incoming physical signal, filtered by the auditory filter with center-frequency f_c:

$$\left|S_{int}(f_c)\right|^2 = \frac{\int_{-\infty}^{\infty} W(f, f_c) \left|S(f)\right|^2 df}{\int_{-\infty}^{\infty} W(f, f_c) df}, \tag{5.4}$$

in which the denominator is a normalization factor that corresponds to the internal spectrum of white noise, so that a white noise input signal leads to a flat internal spectrum. Note that this

normalization factor equals the ERB of the auditory filter with center-frequency f_c ((eq. 5.2)). The effect this filtering has on the spectrum of an incoming signal is that the spectrum is smoothed, as single frequency components are smeared out over a range of frequencies, especially for higher frequencies. This is demonstrated in figure 5.3 for both a very simple and a more complex signal.

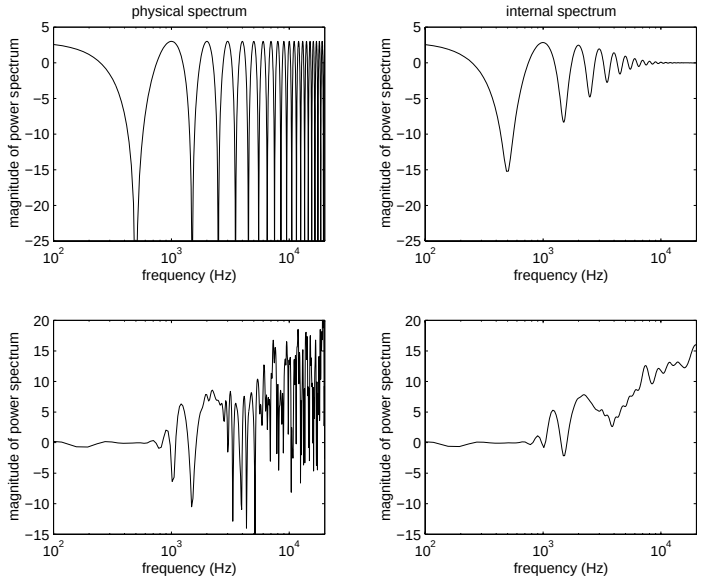


FIGURE 5.3. Examples of the effect of the auditory filtering process on the spectrum of two signals.
Top: Physical (left) and internal (right) spectrum of a pulse with a single repetition with unit gain and 1 ms delay ('cosine noise').
Bottom: Typical simulated physical (left) and internal (right) spectrum from the sound field reproduced by a WFS array (loudspeaker spacing 25 cm).

5.2.2 Criterion For Coloration Threshold

It would be convenient to have a criterion that enables to predict whether or not spectral coloration will be perceived when comparing a signal with given spectral distortion to the original undistorted signal. From the discussion of the auditory filtering process above, it seems plausible that such a criterion should be based on the internal, auditory filtered, spectrum of the signal, rather than on the physical spectrum of the incoming acoustical signal itself. Salomons ([Salo95]) developed such a criterion based on the internal spectrum of the signal, intended to predict coloration thresholds[1] of signals compared to white noise: the 'A_0' criterion. This A_0 criterion is actually a modified version of a criterion by the same name that was developed ear-

lier by Atal et al. ([Atal62])[2] and reads: “coloration is perceptible if the maximum level difference between maxima and minima of the internal spectrum exceeds a certain threshold A_0”, or in equation, if:

$$A_0 \leq 20 \log \left(\frac{|S_{\text{int}}|_{\max}}{|S_{\text{int}}|_{\min}} \right). \tag{5.5}$$

From coloration threshold data for fully correlated- and fully uncorrelated[3] harmonic cosine noise, Salomons derived an average value for A_0 of about 1.5 dB and it was shown that the revised criterion was indeed more successful in predicting coloration thresholds than the original one, also for more complex signals than harmonic cosine noise, including signals with multiple repetitions. Unfortunately though, it appears to be very difficult to include a satisfactory model of the binaural processing of dichotic signals in the criterion to accurately predict the increase of coloration threshold due to binaural decoloration (section 5.2.4).

It is important to note that the A_0 criterion does not tell us anything about the *strength* of the perceived coloration, nor does it give any information about the colour *difference* between two individual coloured signals. As was shown already by Salomons herself, the correlation between the coloration thresholds that were found by applying the A_0 criterion to signals recorded in various concert halls and subjective coloration scale values obtained from a paired-comparison test between the same concert hall signals, was very poor. To be able to make predictions about the perceived colour difference between two signals that are both coloured above threshold, the A_0 criterion is therefore not suitable and it is necessary to use a different analysis, such as the one discussed in the next subsection.

1. The ‘coloration threshold’ of a signal is defined as the gain difference between the signal and white noise for which a colour difference is just noticeable between the signal-plus-noise stimulus and white noise.
2. Salomons’ modification of the A_0 criterion of Atal et al. is mainly that she uses the internal spectrum in the analysis, whereas Atal used the short-time averaged power spectrum, which only resulted in a satisfactory prediction of the coloration threshold for a very limited range of signals.
3. ‘Fully correlated’ refers to the diotic condition of presenting the same signal to both ears, while ‘fully uncorrelated’ refers to the dichotic condition of presenting a different signal to each ear, with the two signals being uncorrelated.

5.2.3 Colour Differences Between Signals

To predict the perceived colour difference between differently coloured signals, Plomp ([Plom70], [Plom73]) proposed to quantify the colour difference (or actually: timbre difference, since pitch is not taken into account in the analysis) between two signals *i* and *j* with spectra S_i and S_j by calculating the sum of the squared differences between the respective auditory filtered power spectra:

$$D_{ij} = \sqrt{\sum_{n=1}^{N} \left(20\log\left|S_{\text{int}}\right|_{i,n} - 20\log\left|S_{\text{int}}\right|_{j,n} \right)^2}, \tag{5.6}$$

in which the summation is over N frequency bands. Plomp, in his first proposal for the model, used 1/3-octave bands for both the filtering of the original spectra and in the summation, since this roughly corresponds to the width of the auditory filters, but he already suggested that it would probably be better to use the actual shape of the auditory filters. This suggestion was adopted by Bladon and Lindblom ([Blad81]), who transformed the frequency axis of the filtered spectra to an axis corresponding to Zwicker's Bark scale, in which equal distance on the scale corresponds to an equal number of critical bandwidths. Also, they integrated the differences over the Bark scale, rather than just summing over individual critical bands. This approach will be used in this thesis also, with the difference that rather than transforming the frequency axis of the spectra to the Bark scale, it will be transformed to the so-called 'ERB-rate' (ERBR) scale, on which equal distance corresponds to an equal number of Equivalent Rectangular Bandwidths, as modeled by (eq. 5.3), which more accurately models the frequency selectivity of the auditory system than the Bark scale. This transformation of the frequency scale to the ERBR scale can be accomplished by means of the expression for calculating the ERBR, the number of ERB's below frequency *f*, for frequency *f*, which is given by ([Moor83]):

$$ERBR(f) = 11.17 \ln \left| \frac{f + 0.312}{f + 14.675} \right| + 43, \tag{5.7}$$

with *f* in kHz.

The expression for the colour difference between two stimuli *i* and *j* now follows from:

$$D_{ij} = \sqrt{\int_{ERBR_{\min}}^{ERBR_{\max}} \left(20\log\left|\tilde{S}_{\text{int}}\right|_i - 20\log\left|\tilde{S}_{\text{int}}\right|_j \right)^2 d\left(ERBR\right)}, \tag{5.8}$$

in which $\tilde{S}_{\text{int}}$ are the internal spectra, transformed to the ERBR scale. The integration boundaries $ERBR_{min}$ and $ERBR_{max}$ are determined by the effective bandwidth of the signals.

5.2.4 Binaural Decoloration

When listening to a dichotic signal, so different signals for the left- and right ear, the perceived amount of coloration can increase significantly when changing to listening with only one ear (monotic listening condition) or listening to either the left- or right ear signal with both ears (diotic listening condition). Apparently, the binaural processing by the auditory system of the signals received at both ears can to some extent result in less coloration being perceived than the amount of coloration that is perceived for each ear's signal individually. This is a phenomenon which is referred to as *binaural decoloration*. It is difficult to accurately model binaural decoloration, as the mechanism that the auditory system uses to achieve this seems to be very complicated (see [Salo95] for a detailed study of binaural decoloration), but as a very simplistic qualitative explanation it is sufficient to imagine that dips in the internal spectrum of one ear can be compensated by peaks at corresponding frequencies in the internal spectrum of the other ear. Although difficult to model, the effect is well-known and the amount of decoloration, in terms of the increase of the coloration threshold of the dichotic signal relative to that of the diotic signal, can be very significant, up to about 10 dB, depending on the type of signal ([Salo95], [Zure79]). Therefore, when doing experiments to investigate the perceived amount of coloration in realistic situations (for example in a concert hall or when listening to a multichannel sound reproduction system), it is important to use binaural signals in order to obtain meaningful results.

5.3 Coloration In A Wave Field Synthesis System

In section 2.3.1 it was shown that when a sound field is reproduced by a linear WFS array with spacing Δx, then the general formulation of the spatial Nyquist frequency above which spatial aliasing will be present in the reproduced field is given by:

$$f_{Nyq} = \frac{c}{\Delta x(\sin\theta_{\max,source} + \sin\theta_{\max,LS})}, \tag{5.9}$$

in which $\theta_{max,source}$ is the maximum angle present in the recorded source field and $\theta_{max,LS}$ is the maximum angle that is present in the sound field radiated by each individual loudspeaker. In the most unfavorable situation, i.e., a source field containing components from all spatial directions between -π/2 and +π/2 radians (with 0 being the direction perpendicular to the array), for example a diffuse sound field, and monopole array loudspeakers, $\theta_{max,LS}$ and $\theta_{max,source}$ are both equal to π/2 and the 'worst case' limit of (eq. 5.9) applies:

$$f_{Nyq} = \frac{c}{2\Delta x}. \tag{5.10}$$

For a monopole source very close to the array, $\theta_{max,source}$ also approaches π/2, so this is an unfavorable situation as well. On the other extreme end, the most favorable situation is when the source to be reproduced is a plane wave travelling in the direction perpendicular to the array ($\theta_{max,source}$=0). When the array consists of monopole loudspeakers, the spatial Nyquist frequency in this case is twice as high as for a diffuse source field or a monopole source very close to the array.

Looking at the operation of a Wave Field Synthesis system in the time-domain, we see that copies of the source signal with different time delays and gains are reproduced by the individual array loudspeakers (see (eq. 2.22) in section 2.2.3). Only for a band-limited input signal with no energy above the spatial Nyquist frequency, the contributions of the individual loudspeakers merge into a single event with the same spectrum as the source signal, while for signals containing frequencies above the Nyquist frequency, the individual loudspeaker contributions arrive at the listener position as individual events, separated in time, resulting in spectral distortions similar to those in figure 5.1 (bottom graphs). This is illustrated in figure 5.4, where the top graphs show the time domain response along a line parallel to a 9-element WFS array for three different input signals. The bottom graphs show the contributions from the nine individual loudspeakers and the total response, which is the sum of the nine individual contributions, at a single position (indicated in the top graphs by the dashed vertical line). The left graphs are for a signal that only contains frequencies below the spatial Nyquist

frequency. Here it can be seen how the individual responses merge together into a single event. The center graphs are for a signal that contains frequencies up to just above the spatial Nyquist frequency. In these graphs it can be observed how for this signal the individual contributions are just starting to be visible in the total response. The right graphs, finally, are for a signal that contains frequencies far above the spatial Nyquist frequency. In this case, serious aliasing is visible in the response along the line, while in the response at the single position the individual contributions are clearly visible as individual events, arriving separated in time. Because of this, the spectrum of the source signal is distorted at frequencies above the spatial Nyquist frequency.

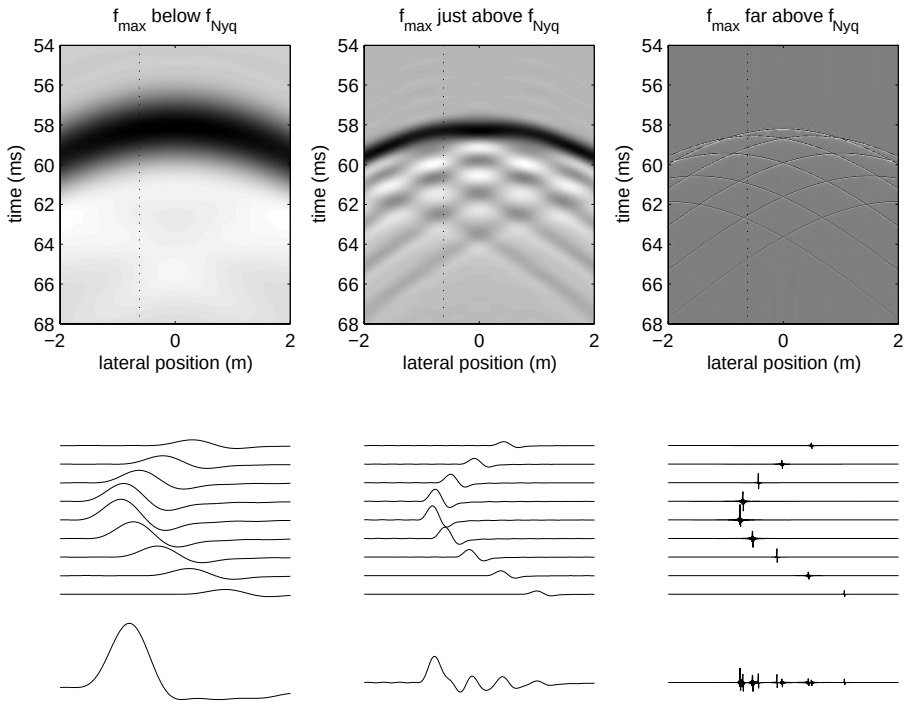


FIGURE 5.4. Illustration of the effect of spatial aliasing on the time-domain response along a line parallel to the array and at a single position for a WFS array consisting of 9 secondary sources, synthesizing a point source located behind the array at a central position.

The upper graphs show the time-response along a line for source signals containing only frequencies below (top-left), up to just above (top-center) and up to far above (top-right) the spatial Nyquist frequency.

The bottom graphs show the individual contributions of the 9 secondary sources and the total response at a single lateral position, indicated in the top graphs by the dashed line.

5.3.1 Simulations Of Physical And Internal Spectra

We will now present simulations of the physical and internal spectra of the sound fields reproduced by WFS arrays for various inter-loudspeaker spacings and for several source positions. The simulations were done for 1025 equally-spaced frequencies in the range 0-22.05 kHz, so the resolution was 21.5 Hz, which is smaller than the Equivalent Rectangular Bandwidth of the auditory filters, even at the lowest frequencies (see (eq. 5.3)). All simulations were done assuming monopole directivity for the secondary sources and a flat spectrum for the source signal and with amplitude tapering applied to the driving signals of the loudspeakers to reduce diffraction effects from the edges of the array.

loudspeaker array configurations. Simulations were done for five linear arrays with different loudspeaker spacings, each having a total length of 4 m:

- spacing 12.5 cm (33 loudspeakers)
- spacing 16.7 cm (25 loudspeakers)
- spacing 25.0 cm (17 loudspeakers)
- spacing 33.3 cm (13 loudspeakers)
- spacing 50.0 cm (9 loudspeakers)

source positions. Simulations were done for three sources:

- monopole point source at (0, -1) m: (center, 1 meter behind the array)
- monopole point source at (0, -5) m: (center, 5 meter behind the array)
- plane wave, 0 degrees (traveling perpendicular to the array)

simulation grid. Simulations were carried out on a simulation grid with these specifications:

- spacing in x direction (parallel to the array): 0.1 m, range: -2 to 2 m.
- spacing in z direction (perpendicular to the array): 0.1 m, range: 1 to 5 m.

In figure 5.5 and figure 5.6 the simulations are shown for the loudspeaker spacings 12.5, 25.0 and 50.0 cm, for the source position (0, -1) and for the plane wave, respectively. Each figure shows the magnitude of the response as function of frequency along the line z=3 m. The physical spectra are shown in the left- and the auditory filtered spectra in the right graphs.

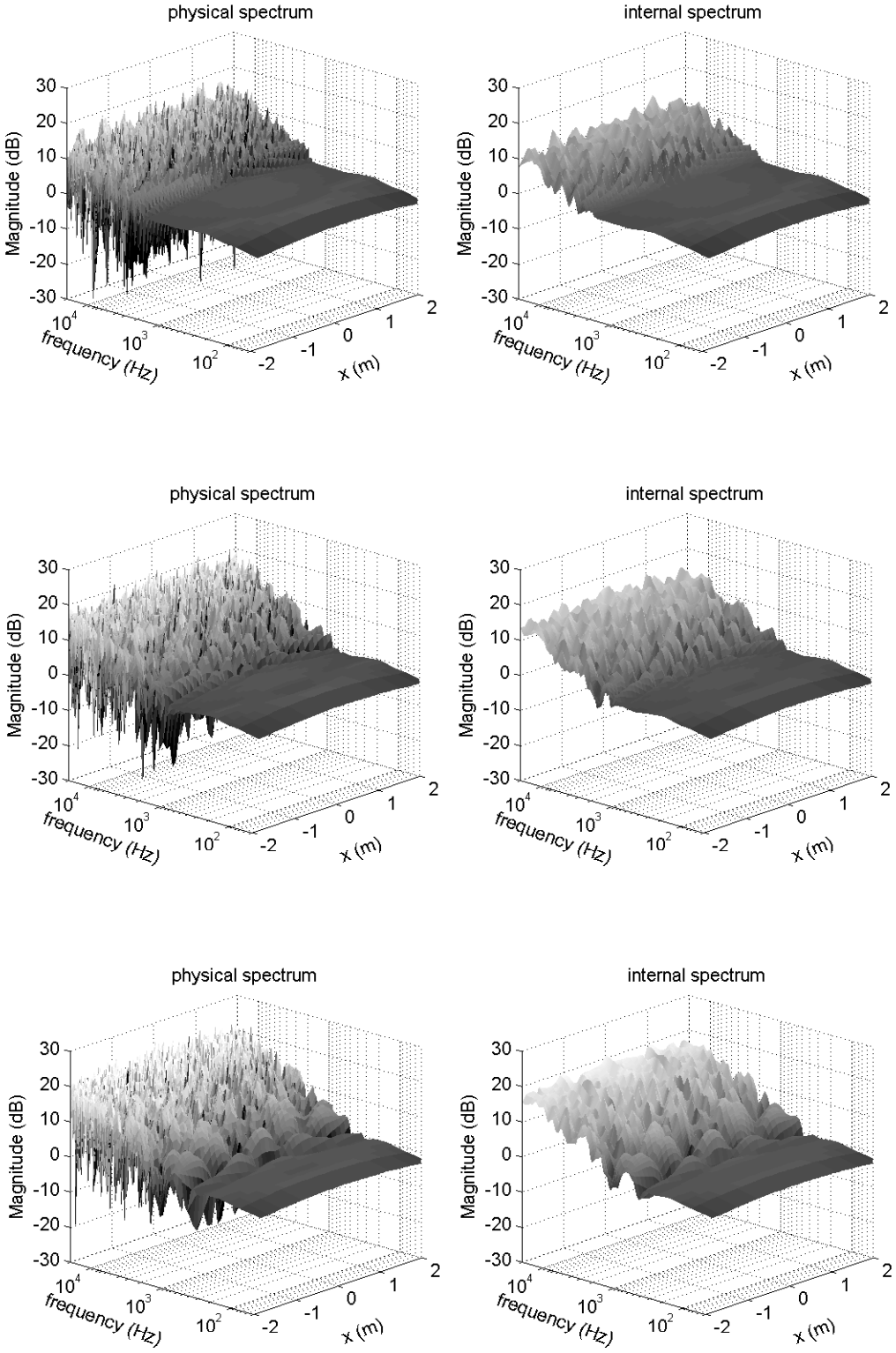


FIGURE 5.5. **Simulated response along the line z=3 m as function of frequency for a loudspeaker spacing of 12.5 cm (top), 25.0 cm (center) and 50.0 cm (bottom) for a virtual point source at (0, -1) m. Left: Physical spectrum. Right: Auditory spectrum.**

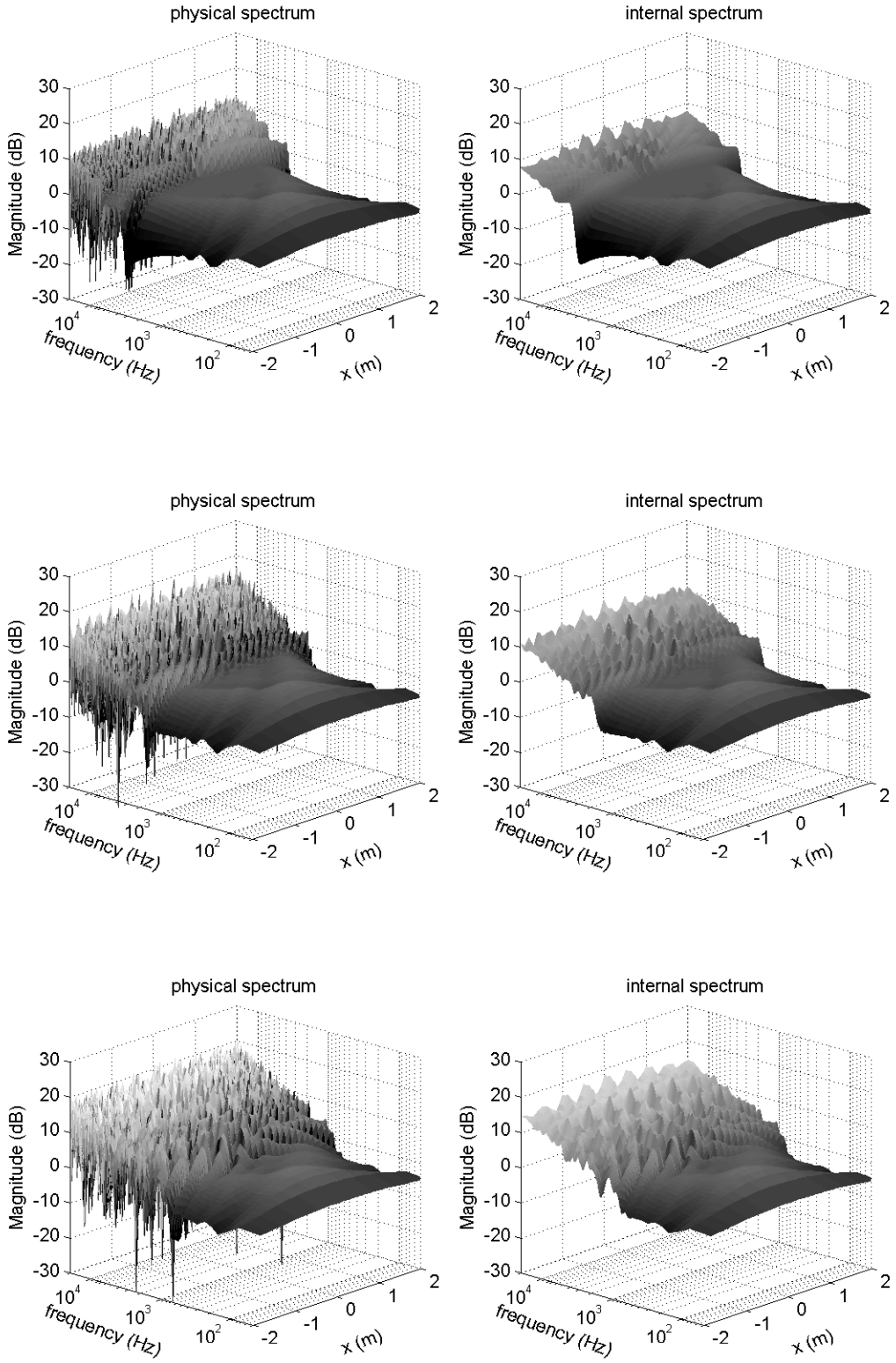


FIGURE 5.6. Simulated response along the line z=3 m as function of frequency for a loudspeaker spacing of 12.5 cm (top), 25.0 cm (center) and 50 cm (bottom) for a plane wave traveling perpendicular to the array. Left: Physical spectrum. Right: Auditory spectrum.

We can conclude from these simulations that, even though the processing through the auditory filters smooths out the extreme fluctuations, the resulting internal spectra still have quite large variations. Specifically, when we apply the A_0 criterion of section 5.2.2, which stated that coloration will be audible in a signal, compared to white noise, if the difference between the overall minimum and maximum of the internal spectrum exceeds about 1.5 dB, it is obvious that this will be the case even for the simulated system with the smallest loudspeaker spacing (12.5 cm, top-right graphs of figure 5.5 and figure 5.6). This is especially clear when we look at figure 5.7, in which the auditory filtered spectra at four different lateral positions are shown for the simulation of the top graph of figure 5.5.
Clearly visible in figure 5.7 is also the inappropriate systematic increase of the response for frequencies above the spatial Nyquist frequency, caused by the fact that for a discrete array, the $\sqrt{jk}$ filter that is present in the formula for the driving signals of a continuous linear WFS array ((eq. 2.22)) is not valid for this frequency range, as was discussed in section 2.3.1. Comparing the spectra for the different loudspeaker spacings in figure 5.5 and figure 5.6 it is also clear that the error that is introduced because of this, increases for increasing loudspeaker spacing. However, because this trend of increasing frequency response is systematic, i.e., the same in the whole listening area, it is very well possible to compensate for it by applying a simple pre-filter to the input signal, as will be shown later (section 5.4.2). It will be clear, however, from figure 5.7, that even then quite substantial fluctuations in the internal spectrum will remain, which are well above the threshold for coloration perception and which are place-dependent and therefore difficult to compensate for.

In conclusion, we should expect that a colour difference between the original source signal and the sound field reproduced by the WFS system will indeed be present. As such, however, this is usually not a real problem in practice, because of the fact that the original source signal is in general not available to the listener for comparison. So even though the reproduced signal may sound coloured when compared to the original signal, as long as there are no changes of the signal spectrum of a more systematic nature (for example a strong boost or attenuation on certain frequency regions, rather than peaks and dips at more or less isolated frequencies, resulting in the reproduced signal having a clearly ‘unnatural’ colour), listeners will not notice that the reproduced signal is coloured.
What should be of more concern are *spatial variations* of the colour of the reproduced sound

field, i.e., the change of the sound colour that the listener perceives when he moves his head or walks around in the room. These spatial variations of the colour will be discussed now.

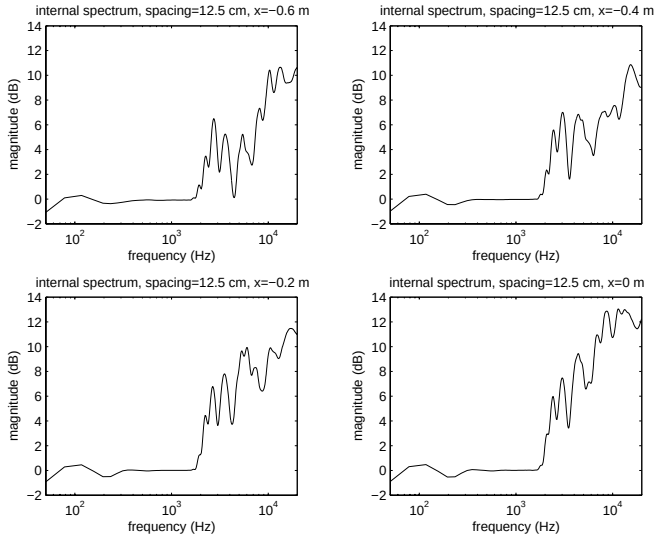


FIGURE 5.7. Auditory filtered spectra at positions x= -0.6 m, -0.4 m, -0.2 m and 0 m for the simulation of the top graph of figure 5.5 (loudspeaker spacing 12.5 cm, source at (0, -1) m, listening distance 3 m).

5.3.2 Spatial Colour Variations

When a listener perceives clearly audible colour changes when he changes his listening position within the sound field of an audio reproduction system, this can be very annoying and decrease the listener's impression of the quality of the system. In addition to the fact that it is simply unpleasant to listen to a sound of which the colour is constantly changing, it makes the listener aware that he is listening to a reproduction system and not to the original source. In the case of a life-size videoconferencing system this is highly undesirable, because the aim is to have a reproduction that is as natural as possible. In other words, clearly perceptible spatial colour variations would decrease the illusion of telepresence (see section 1.1.2) of the conference.

In the previous subsection it was shown by simulations that spatial colour variations are introduced in a WFS system for frequencies above the spatial Nyquist frequency. Because of the fact that, as can be observed in figure 5.5 and figure 5.6, these variations are highly place-dependent, it is not possible to compensate for them by simple filtering, contrary to the system-

atic trend of increasing response for frequencies above the Nyquist frequency. In this subsection these spatial colour variations that occur in a WFS system are investigated and quantified. The aim is to obtain an objective measure for the spatial colour variations that occur for a given loudspeaker configuration. Combined with the results of a perception experiment in which spatial colour variations are compared for WFS systems with different loudspeaker spacings, this should ultimately enable us to predict from simulations whether clear spatial colour variations will be perceived for a given system and what is the maximum distance between the loudspeakers that can be tolerated if no clearly perceptible spatial colour variations are allowed. The perception experiment will be described in section 5.4.
The basis for the analysis will be the internal spectra that were obtained from the simulations described in the previous subsection. In section 5.2.3 a model was described to quantify the colour difference between two signals that are both coloured above threshold. This model will now be used to analyze and quantify the spatial colour variations that occur in a WFS system, by applying it to the internal spectra at different positions, as obtained from the simulations.

From the right-hand graphs of figure 5.5 and figure 5.6 we can already get an indication of the magnitudes of the spectral variations that occur along a line parallel to the array, for different loudspeaker spacings. To get more insight into the spatial spectral variations in the whole 2D listening space, we can look at the magnitude of the internal spectrum for a single frequency at all points of the 2D grid for which the simulations of section 5.3.1 were done. Examples are shown in figure 5.8 and figure 5.9 for the same configurations as those of figure 5.5 and figure 5.6 and for two frequencies: f=2 kHz (left) and f=8 kHz (right). The range of the gray-scale is 25 dB in each figure.
It can be observed from figure 5.8 and figure 5.9 that the spectral variations are in general the strongest along the lateral (x) direction, so for a listener who moves parallel to the array. The variations along the perpendicular (z) direction are proportional to those along the lateral direction, so it is sufficient to do the analysis of spatial variations along the lateral direction.

As said, the starting point are the simulations of the internal spectra of section 5.3.1. To quantify the colour difference between the internal spectra at any two points of the simulation grid, we can apply (eq. 5.8). To obtain a measure for the colour variations along a line $z=z_0$ of the simulation grid, we can calculate (eq. 5.8) for the pair of internal spectra at positions (x_i, z_0) and (x_i+*offset*, z_0) for each x_i of the simulation grid and fixed value of *offset*. The value of the

offset, the distance between the lateral positions for which the colour difference is calculated, is somewhat arbitrary. If a small value is chosen (in the order of several centimeters), this gives an indication of the variations that will be perceived when a listener moves his head while seated or standing at a fixed position, whereas larger values (in the order of several decimeters) are indicative for variations that are perceived when a listener walks around in the room. Larger values (in the order of meters) are considered to be less important, since a listener does not jump instantaneously from a position at one side of the room to a position at the other side of the room, so due to the limited memory capacity of the human auditory system, an observer will not be able to directly compare the colour of the sound field at those two positions. Furthermore, from looking at the simulations in figure 5.8 and figure 5.9 we observe that, in the lateral direction, the spectral variations are of a periodic nature. For the frequencies of interest, the frequencies above the spatial Nyquist frequency for which spatial aliasing is present, a range of 0.5 m to each side of a given lateral position includes essentially the whole range of spectral variations that occur along the lateral direction. Likewise, it can be seen that significant variations occur in the decimeters rather than the centimeters range, so we will focus our analysis on variations in the range from 0.1-0.5 m.

For each of the simulated configurations described in section 5.3.1, the value of D_{ij} is calculated for x= -2...2 m (resolution: 0.1 m) and z=1, 3 and 5 m and offsets of 0.1, 0.2, 0.3, 0.4 and 0.5 m, so at each position (x, z) the colour difference between the internal spectrum at that position and the internal spectrum at position (x+*offset*, z) is calculated from (eq. 5.8). In this calculation, the whole audio range is taken into account by using $ERBR_{min}$=1 and $ERBR_{max}$=37, corresponding to 30 Hz and 20 kHz respectively ((eq. 5.7)), as the lower and upper boundaries for the integration.

Figure 5.10 shows the value of the integrand of (eq. 5.8) for the configurations with loudspeaker spacing 12.5 cm (left) and 50 cm (right) for a virtual point source at (0, -1) m, along the line z=3 m and for an offset of 0.1 m. It is clearly seen that spatial colour variations only occur for frequencies above the spatial Nyquist frequency of each configuration and that the magnitudes of the variations are larger for the configuration with the larger loudspeaker spacing. The value of D_{ij} now follows from integrating these graphs over the frequency dimension and taking the square root of the result.

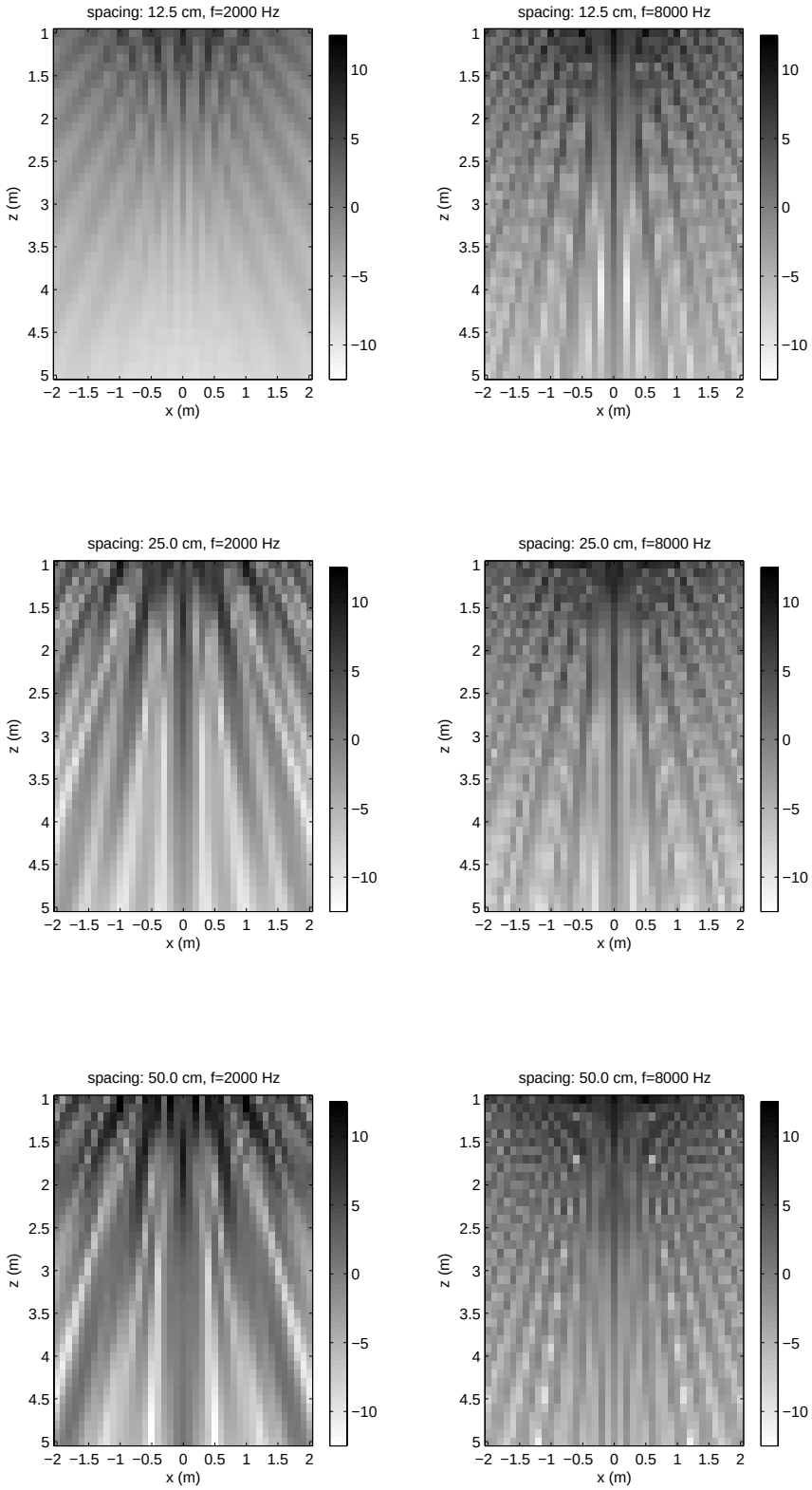


FIGURE 5.8. Simulation of the auditory filtered spectrum for loudspeaker spacing 12.5 cm (top), 25.0 cm (center) and 50.0 cm (bottom) for a virtual point source at (0, -1) m, for single frequencies of 2 kHz (left) and 8 kHz (right).

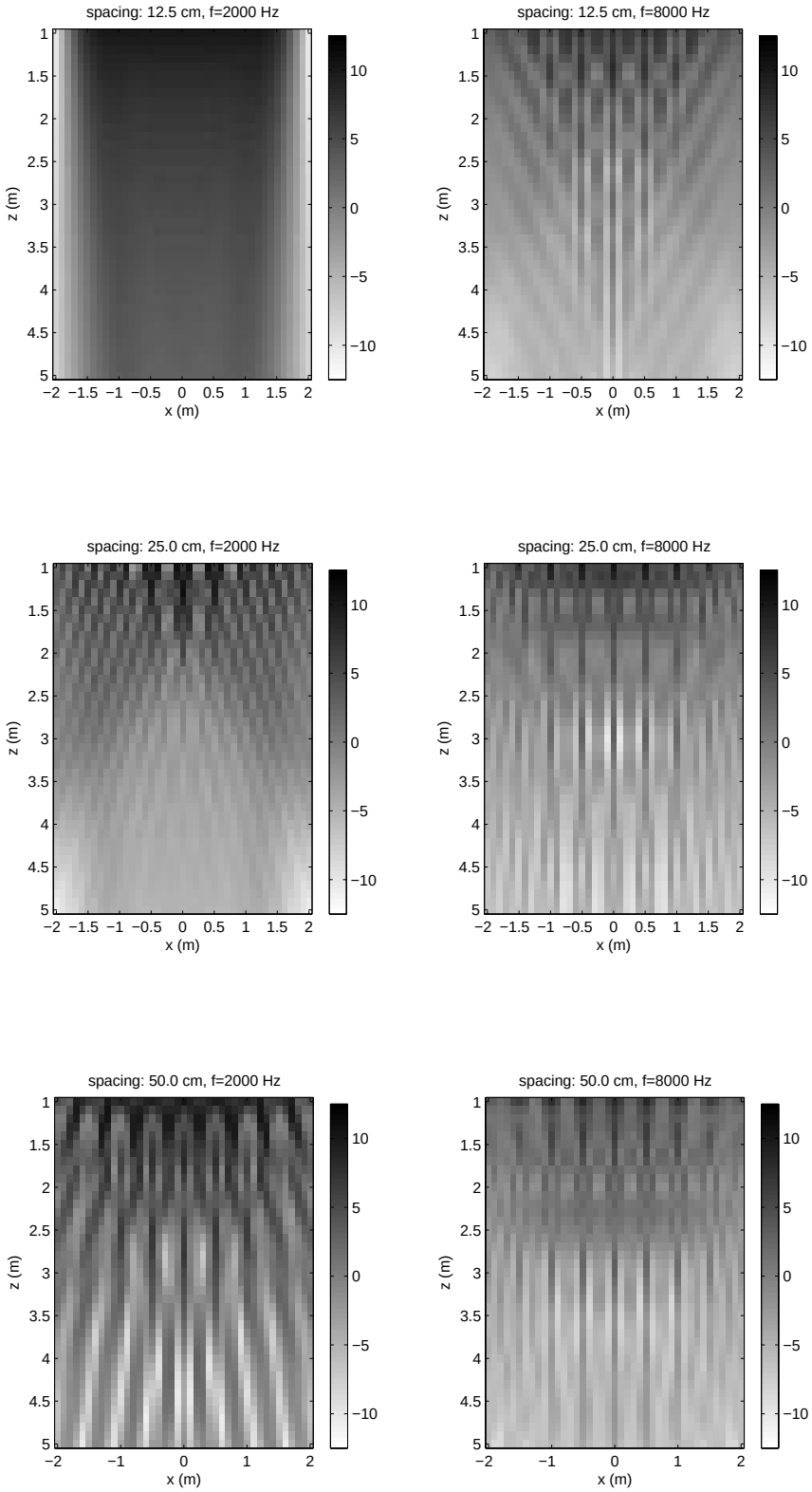


FIGURE 5.9. Simulation of the auditory filtered spectrum for loudspeaker spacing 12.5 cm (top), 25.0 cm (center) and 50.0 cm (bottom) for a plane wave traveling perpendicular to the array, for single frequencies of 2 kHz (left) and 8 kHz (right).

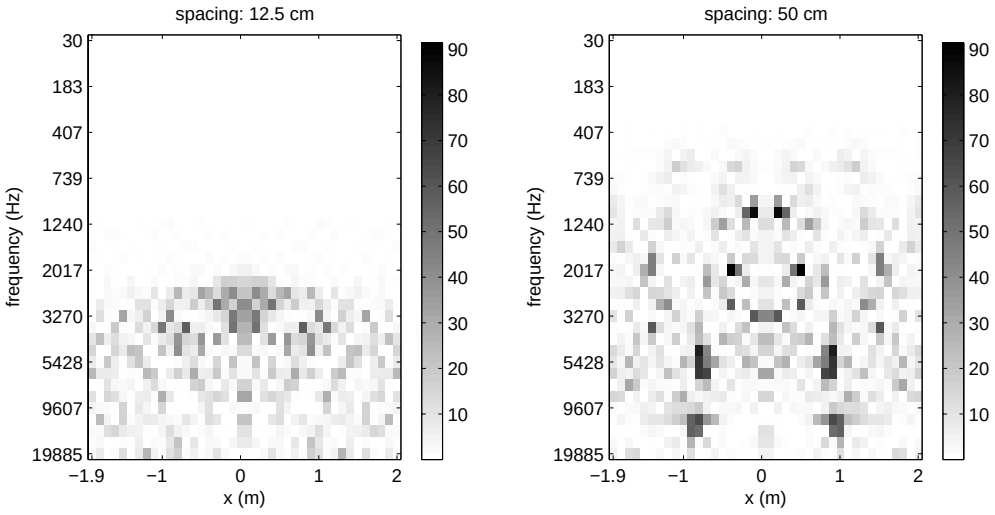


FIGURE 5.10. Value of the integrand of (eq. 5.8) (in (dB)^2) for the configurations with a virtual point source at (0, -1) m and loudspeaker spacing 12.5 cm (left) and 50 cm (right) along the line *z*=3 m, calculated for an offset of 0.1 m.

The top-left graph of figure 5.11 shows the value of D_{ij} for the configuration with loudspeaker spacing 12.5 cm and a virtual point source at (0, -1) m and an offset of 0.1 m. In the top-right graph the value of D_{ij} is shown for the same configuration but using an offset of 0.3 m in the analysis. The center- and bottom graphs of figure 5.11 show the values of D_{ij} for the loudspeaker spacings of 25 and 50 cm. The corresponding results for the plane wave traveling perpendicular to the array are shown in figure 5.12.

We can make several general observations from looking at figure 5.11 and figure 5.12. First of all, comparing the graphs of configurations with different loudspeaker spacings for the same offset, we observe that, as expected, the magnitudes of the colour differences that occur increase for configurations with increasing loudspeaker spacing.

Secondly, from comparing the results for a point source at (0, -1) m (figure 5.11) to those for a plane wave (figure 5.12), we see that the colour variations are larger for the point source close to the array than for the plane wave, which is in agreement with the discussion of spatial aliasing following (eq. 5.9).

Furthermore, we observe from comparing the results for different values of the distance *z* in each individual graph that, generally speaking, the colour variations along the lateral direction decrease for increasing distance from the array, as expected. In the remainder of this section we will continue to analyze in more detail the colour variations along the line *z*= 3 m.

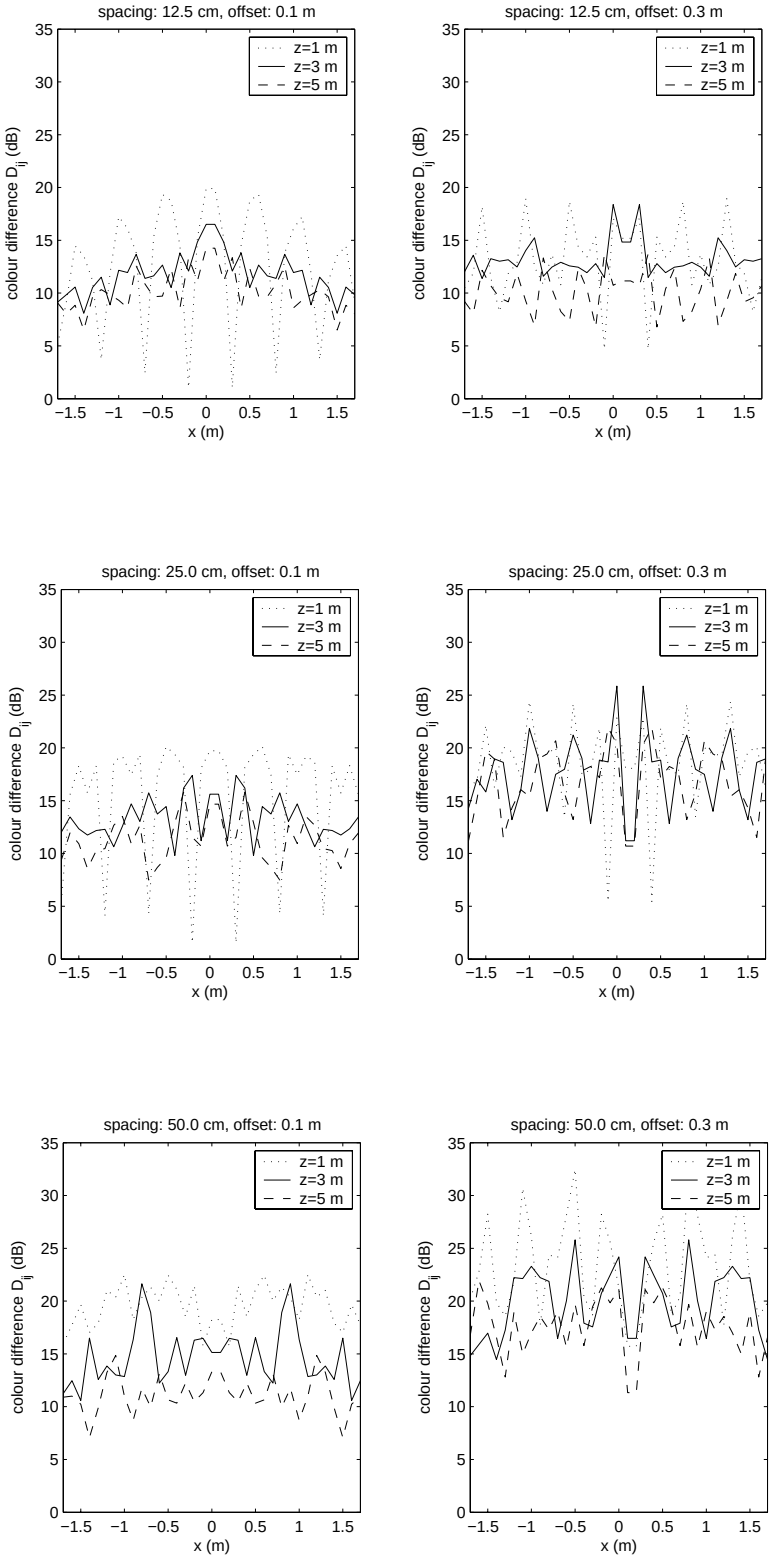


FIGURE 5.11. Value of the colour difference D_{ij} (calculated from (eq. 5.8)) along the lines z=1 m (dotted), z=3 m (solid line) and z=5 m (dashed line) for loudspeaker spacing 12.5 cm (top), 25.0 cm (center) and 50.0 cm (bottom) for a virtual point source at (0, -1) m, for two offsets: 0.1 m (left) and 0.3 m (right).

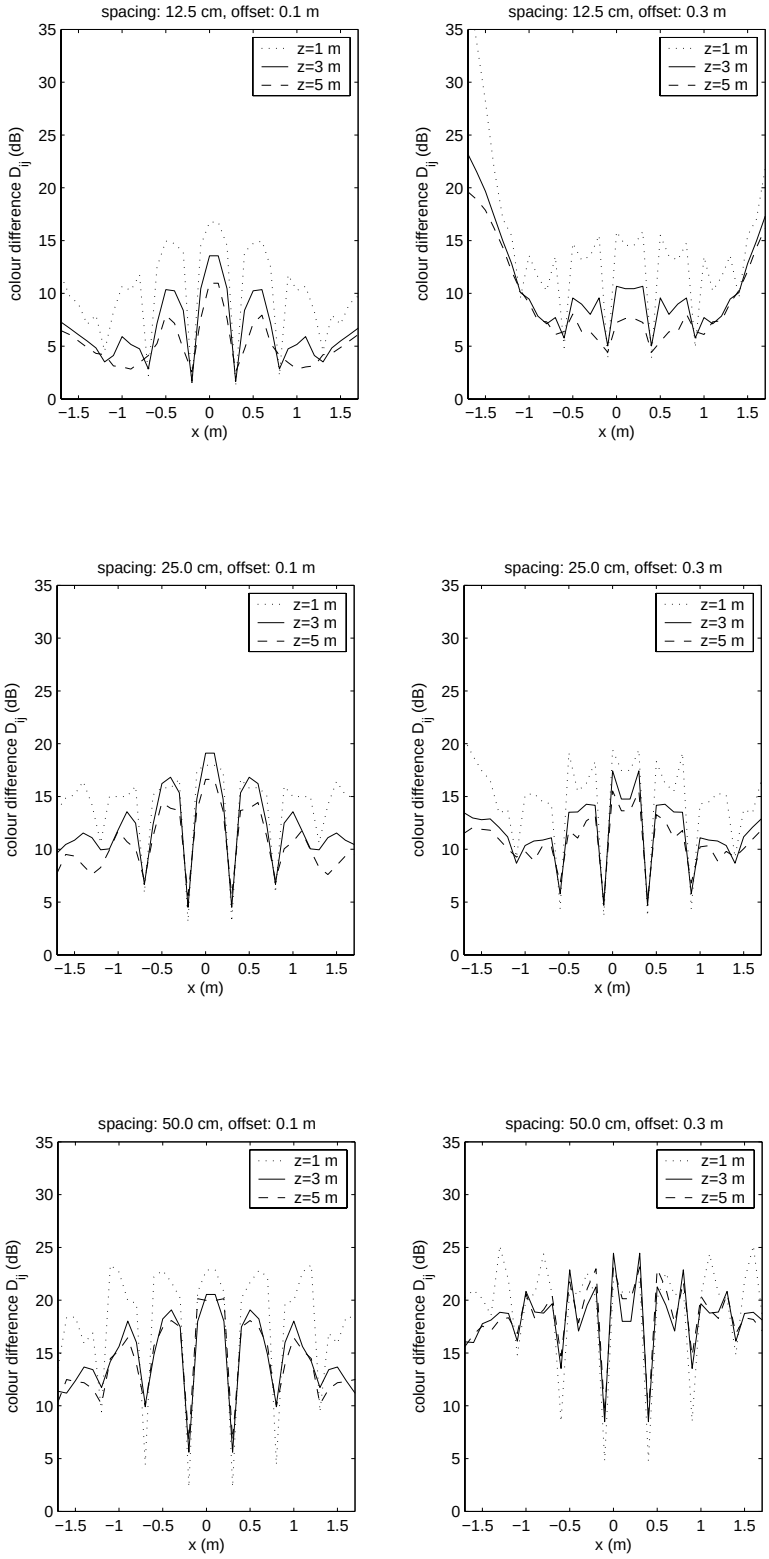


FIGURE 5.12. **Value of the colour difference D_{ij} (calculated from (eq. 5.8)) along the lines z=1 m (dotted), z=3 m (solid line) and z=5 m (dashed line) for loudspeaker spacing 12.5 cm (top), 25.0 cm (center) and 50.0 cm (bottom) for a plane wave traveling perpendicular to the array, for two offsets: 0.1 m (left) and 0.3 m (right).**

Finally, if we compare the graphs for 0.1 m offset and those for 0.3 m offset for individual configurations, it is clear that, although the variations are of the same order of magnitude and exhibit the same general patterns, the results can be significantly different. We observe that for large loudspeaker spacings, the colour differences found for the 0.3 m offset are overall a bit larger than those for the 0.1 m offset. The results for 0.2, 0.4 and 0.5 m offset (not shown) exhibit a similar resemblance in overall magnitude and pattern, with similar differences between individual situations. In some cases, the differences in the results can be very significant, for instance, when, for a certain combination of configuration and offset, the offset coincides with the period (along the lateral direction) of the largest spectral variations. In that case, the colour differences that are found between pairs of internal spectra with that offset appear to be smaller than those found along a line at another distance from the array or with another offset. Similarly, when the offset equals half the period of the largest fluctuations, the colour differences that are found can be significantly larger than average. To avoid that the results of the analysis are influenced too much by the exact choice of the offset and because there are no obvious reasons to attribute more importance to one of the offsets from the 0.1-0.5 m range than to the others, it is suggested to look at the average value of the spatial colour variations averaged over the five offsets as a measure of the degree of spatial colour variation at each position, rather than looking at values for any single offset.

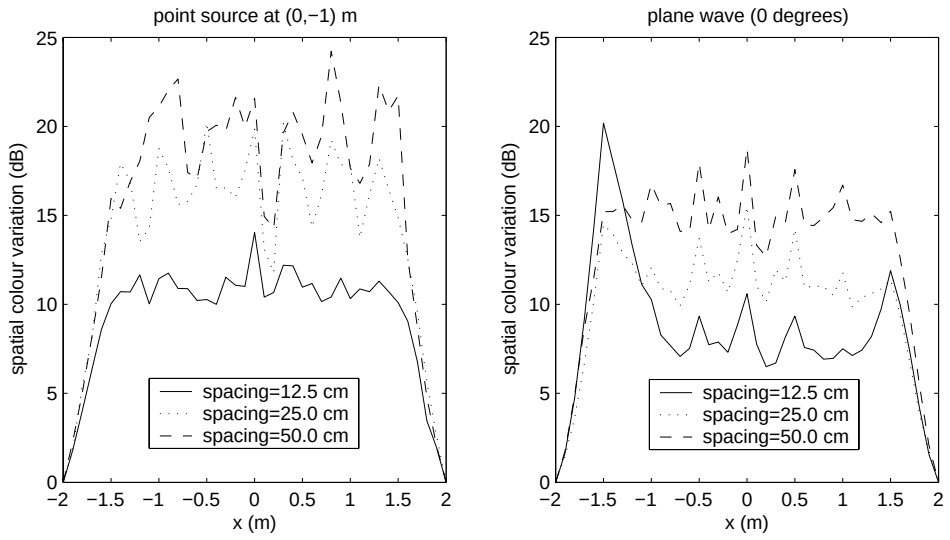


FIGURE 5.13. Spatial colour variation (in dB) averaged over five offsets from 0.1 to 0.5 m along the line *z*=3 m for a source at (0, -1) m (left) and a plane wave traveling perpendicular to the array (right) and for three different spacings (12.5 cm: solid line, 25.0 cm: dotted line, 50.0 cm: dashed line).

In figure 5.13 the value of the spatial colour variation along the line z=3 m, averaged over the five offsets from 0.1-0.5 m, is shown for the various configurations. The left graph shows the results for the source at (0, -1) m and spacings of 12.5, 25.0 and 50.0 cm and the right graph shows the results for a plane wave. From these graphs, the increase of spatial colour variations with increasing loudspeaker spacing is very clear. The peak at x= -1.5 m for the plane wave and spacing 12.5 cm is an effect of the tapering of the outer loudspeakers of the array, which, in case of a plane wave, causes the reproduced sound pressure level to drop very quickly when moving out of the reconstruction area of the array (cf. figure 5.6 and figure 5.9 (top graphs)).

Finally, to arrive at a single-number quantitative measure for the spatial colour variations of a specific configuration, we take the spatial average of the averaged-over-offsets spatial colour variations over the range of positions for which the results are not influenced by the tapering of the array. We will refer to this quantity as the 'Spatial Colour Variation Index' (or 'SCV Index'), expressed in decibels:

$$SCV\ Index = \frac{1}{N_x N_{offset}} \sum_{i=(N_{offset}+1)}^{N_x+N_{offset}} \sum_{j=1}^{N_{offset}} \sqrt{\int_{(ERBR)_{\min}}^{(ERBR)_{\max}} \left(20\log\left|S_{\text{int}}\right|_{x_i} - 20\log\left|S_{\text{int}}\right|_{x_{i-j}}\right)^2 d\left(ERBR\right)}, \qquad (5.11)$$

in which N_{offset} is the number of offsets over which is being averaged and N_x is the number of equidistant lateral positions over which is being averaged. The minimum and maximum value of index i, $N_{offset+1}$ and N_x+N_{offset}, indicate the boundaries of the range of lateral positions that is taken into account in the averaging. Note that simulation results have to be available for $x_1 \ldots x_{Noffset}$ as well.

loudspeaker spacing	source at (0, -1) m	source at (0, -5) m	plane wave
12.5 cm	11.1	5.1	8.0
16.7 cm	14.3	11.1	9.8
25.0 cm	17.0	13.1	11.6
33.3 cm	18.4	16.0	15.4
50.0 cm	19.7	19.2	15.3

TABLE 5.1. Spatial Colour Variation Index (in dB) for all the simulated configurations.

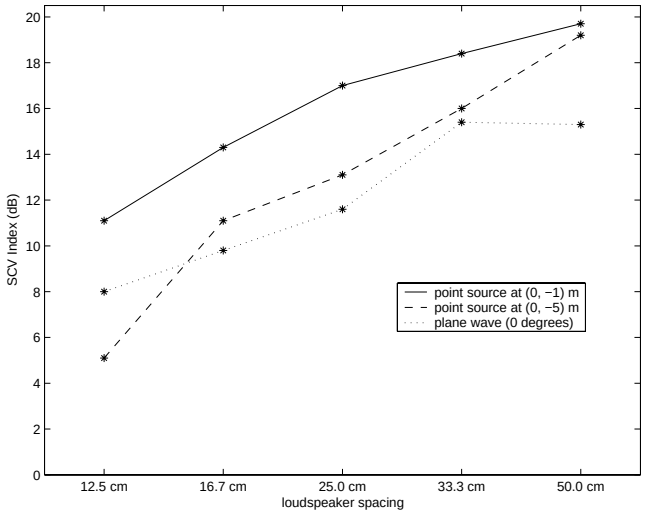


FIGURE 5.14. Spatial Colour Variation Index (in dB) for all the simulated configurations. The solid line is for the monopole source at (0, -1) m, the dashed line for the monopole source at (0, -5) m and the dotted line for the plane wave traveling perpendicular to the array.

In table 5.1 the value of the SCV Index is given for all the simulated configurations. The range of lateral positions that was taken into account in the spatial averaging was x= -1....+1 m. Figure 5.14 shows the same data in graphical format. It is seen that also this single-number representation clearly shows the increase of spatial colour variations for increasing loudspeaker distance and the decrease of spatial colour variations for increasing source distance (the atypical result that the SCV Index for the plane wave and loudspeaker spacing 12.5 cm is larger than that for a source at (0, -5) m is due to the effect of tapering that was mentioned in the discussion of figure 5.13).

Up to this point, the analysis of the spatial colour variations was done assuming signals with a flat power spectrum covering the whole audio bandwidth. For a videoconferencing system, however, the most important signals to reproduce are speech signals, which have a more narrow bandwidth. In figure 5.15 the spectrum of both male and female speech noise, i.e., noise with the same spectrum as the long-term averaged spectrum of speech, is shown. It can be seen that there is almost no significant energy present at frequencies below about 75 Hz and 150 Hz for male and female speech noise, respectively, and almost no energy above about 6 kHz and 8 kHz, respectively. Higher frequencies do occur in the short-term spectrum of speech, mostly because of short transient-like sounds from consonants.

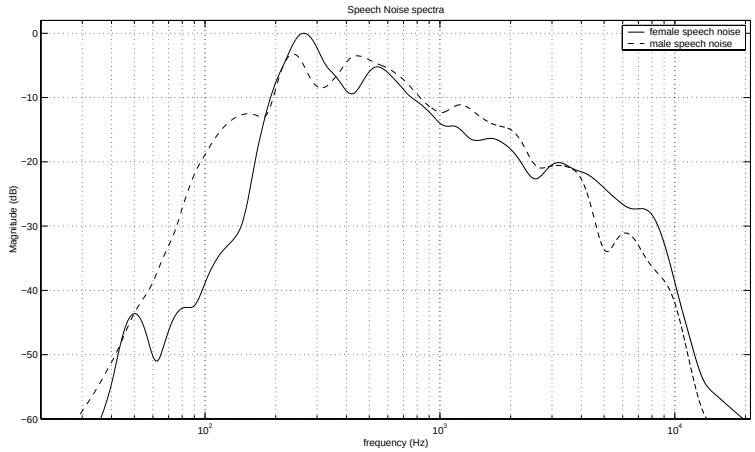


FIGURE 5.15. Long-term averaged spectrum ($^1/_3$-octave filtered, A-weighted) of male and female speech (source: male and female speech noise recordings from [TNO88]).

The main difference in the analysis of spatial colour variations when using speech or speech noise as source signal rather than full-bandwidth audio is that now in (eq. 5.11) different lower- and upper limits for the integration have to be used, according to the bandwidth of the used signal. In the case of the male and female speech noise from figure 5.15 it seems appropriate to use values of $ERBR_{min}$=2 and $ERBR_{min}$=4, corresponding to 63 Hz and 139 Hz, for male and female speech noise, respectively, and $ERBR_{max}$=30 and $ERBR_{max}$=32, corresponding to 6.2 kHz and 8.3 kHz, respectively.

An additional difference between the outcome of (eq. 5.11) for a signal with a flat power spectrum and a signal with a non-flat spectrum, like speech, is that the effect of the filtering process by the auditory filters depends on the spectrum of the input signal to the filters, as can be seen from (eq. 5.4). In other words, the difference (in dB) between the internal spectra of two different signals will in general not remain the same after multiplying the physical spectra of both signals by the same modulator spectrum. In the particular case of the present analysis of spatial colour variations, this means that the value of the difference between the spectra of the reproduced sound field at two positions x and $x+offset$ will be different when speech noise is used as input signal instead of white noise.

Figure 5.16 shows the result of the analysis for female speech noise. Comparing figure 5.16 to figure 5.13, it is clear that the overall magnitudes of the spatial colour variations are somewhat lower in the case of speech noise for each configuration. Table 5.2 shows the value of the SCV

Index for all configurations for female speech noise, figure 5.17 shows the same data in graphical format. Figure 5.18, figure 5.19 and table 5.3 show the results for male speech noise. The slight difference between the results for male and female noise result mainly from the inclusion of the frequencies above 6.2 kHz for female speech noise. The fact that the frequency range of the male noise extends to lower frequencies than that of the female noise does not influence the results significantly, since at those frequencies there is no spatial aliasing present in the reproduced sound field.

In the next subsection an attempt will be made to link the objective SCV Index that was developed in this subsection to the results of a perception experiment on spatial colour variations, in order to determine its effectiveness in predicting their perceived strength.

Two final remarks about the analysis of spatial colour variations:
First, in this analysis no binaural decoloration has been taken into account, due to the complexity of modelling this phenomenon (see section 5.2.4). However, since the intention of the proposed analysis of spatial colour variations is not to predict coloration thresholds or to derive an absolute measure of coloration strength of a specific signal, but rather to be able to make a prediction of the colour differences between signals of a similar nature, this does not seem to be a real problem.
Second, the analysis also does not take into account the spatial colour variations that are possibly introduced by the reflections from the reproduction room. In a reverberant room, the strength of the coloration introduced by the WFS array is masked to a certain extent by the spatial colour variations introduced by the reflections from the room. For strong early reflections, these spatial room colour variations can have a distinct spatial pattern, as for the spatial colour variations introduced by the WFS array, while for the reverberant part of the room response these variations are of a statistical nature. In [Plom73] it is shown that for a diffuse field, the SPL at a single frequency measured at various spatial locations has a theoretical standard deviation of 5.57 dB. On the other hand, a too reverberant reproduction room is undesirable for use as a videoconferencing room, because of the increase of acoustical feedback problems (see section 1.3.2) and the negative effect on speech intelligibility (although some properly generated early reflections can actually enhance the intelligibility).

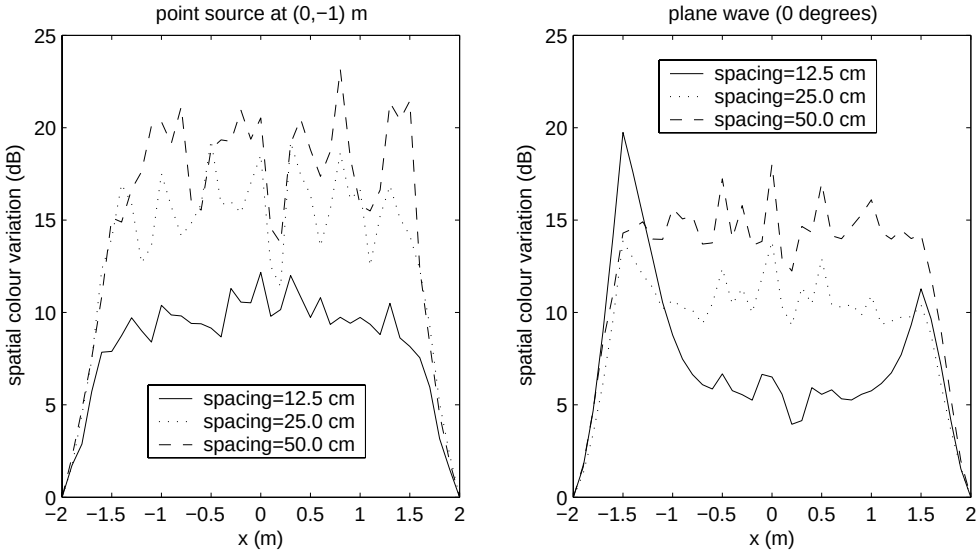


FIGURE 5.16. Spatial colour variation (in dB) for female speech noise (frequency range used in the analysis: 139 Hz - 8.3 kHz), averaged over five offsets from 0.1 to 0.5 m along the line *z*=3 m for a source at (0, -1) m (left) and a plane wave traveling perpendicular to the array (right) and for three different spacings (12.5 cm: solid line, 25.0 cm: dotted line, 50.0 cm: dashed line).

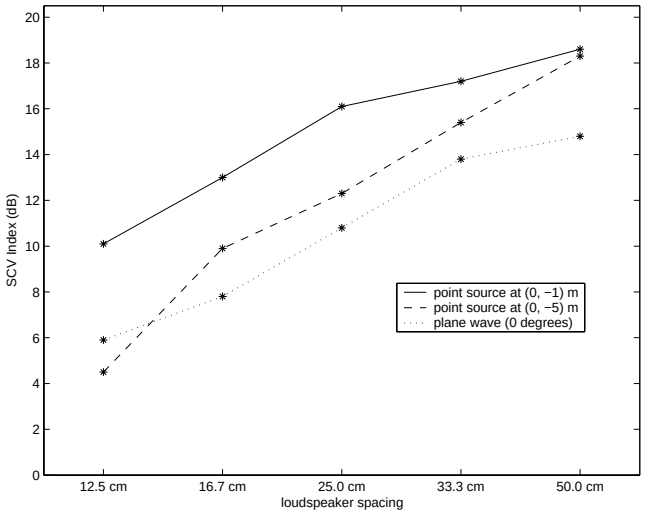


FIGURE 5.17. Spatial Colour Variation Index (in dB) for all the simulated configurations for female speech noise (frequency range used in the analysis: 139 Hz - 8.3 kHz). The solid line is for the monopole source at (0, -1) m, the dashed line for the monopole source at (0, -5) m and the dotted line for the plane wave traveling perpendicular to the array.

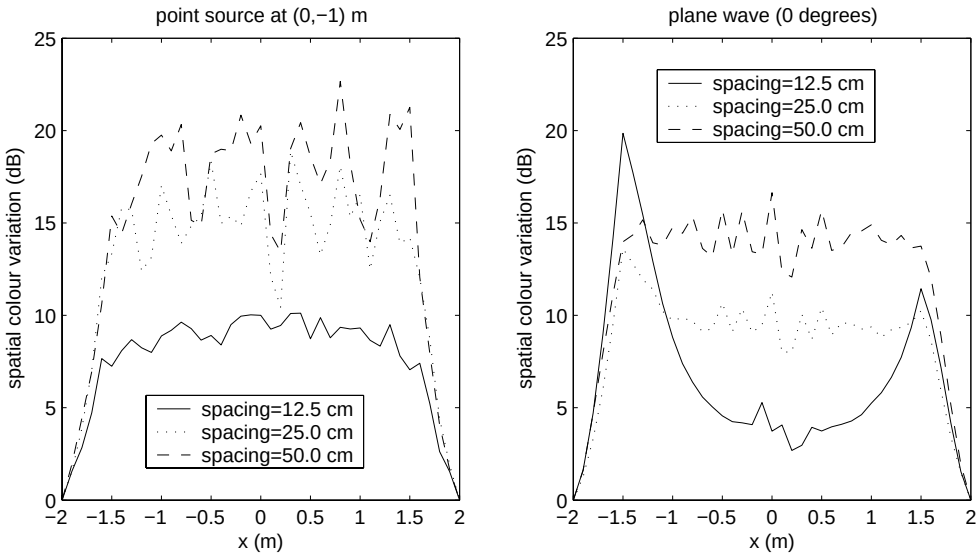


FIGURE 5.18. Spatial colour variation (in dB) for male speech noise (frequency range used in the analysis: 63 Hz - 6.2 kHz), averaged over five offsets from 0.1 to 0.5 m along the line *z*=3 m for a source at (0, -1) m (left) and a plane wave traveling perpendicular to the array (right) and for three different spacings (12.5 cm: solid line, 25.0 cm: dotted line, 50.0 cm: dashed line).

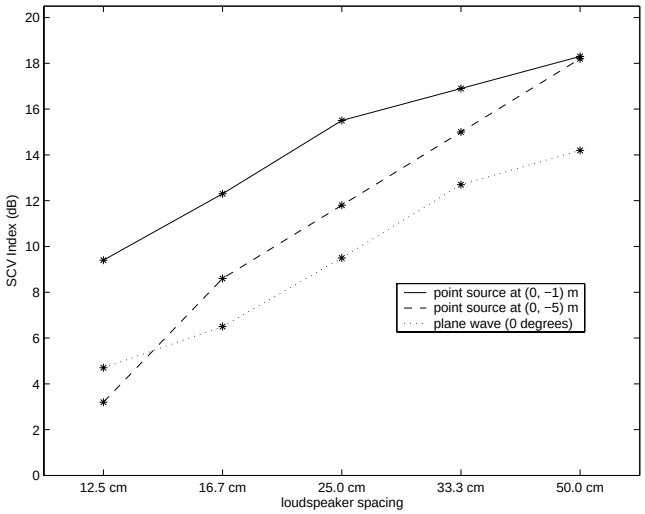


FIGURE 5.19. Spatial Colour Variation Index (in dB) for all the simulated configurations for male speech noise (frequency range used in the analysis: 139 Hz - 6.2 kHz). The solid line is for the monopole source at (0, -1) m, the dashed line for the monopole source at (0, -5) m and the dotted line for the plane wave traveling perpendicular to the array.

loudspeaker spacing	source at (0, -1) m	source at (0, -5) m	plane wave
12.5 cm	10.1	4.5	5.9
16.7 cm	13.0	9.9	7.8
25.0 cm	16.1	12.3	10.8
33.3 cm	17.2	15.4	13.8
50.0 cm	18.6	18.3	14.8

TABLE 5.2. Spatial Colour Variation Index (in dB) for all the simulated configurations for female speech noise (frequency range used in the analysis: 139 Hz - 8.3 kHz).

loudspeaker spacing	source at (0, -1) m	source at (0, -5) m	plane wave
12.5 cm	9.4	3.2	4.7
16.7 cm	12.3	8.6	6.5
25.0 cm	15.5	11.8	9.5
33.3 cm	16.9	15.0	12.7
50.0 cm	18.3	18.2	14.2

TABLE 5.3. Spatial Colour Variation Index (in dB) for all the simulated configurations for male speech noise (frequency range used in the analysis: 63 Hz - 6.2 kHz).

5.4 Coloration Perception Experiment

In this section a perception experiment is described in which subjects compared the spatial colour variations they perceived for various WFS configurations. The objective of the experiment was two-fold:

- To obtain an indication of the maximum distance between array loudspeakers that can be allowed in the specific case of the videoconferencing application of WFS that is the subject of this thesis, such that no excessive amount of spatial colour variations will be perceived by users of the system.
- To link the objective measure for spatial colour variations that was developed in the previous section, the SCV Index, to the actual perceived amount of spatial colour variations, thereby checking the validity of the underlying model, and checking its useful-

ness in predicting the relative amount of perceived spatial colour variations of given WFS configurations.

In accordance with the discussion at the end of section 5.3.1, the spatial variations of colour were investigated in the experiment, rather than the absolute coloration (i.e., the difference between the reproduced signal and the original source signal) because:

- In practice the original source signal is never available to the user of the videoconferencing system, so the user is unable to compare the reproduced signal to it.
- Contrary to the spatial variations of colour, it is very well possible to compensate for systematic coloration of the reproduced signals, by pre-filtering of the source signal, as will be shown later on in section 5.4.2

As was discussed in section 5.3.2, colour variations are in general the strongest when an observer moves in a direction parallel to the loudspeaker array (rather than perpendicular to it), so in the experiment spatial variations were investigated on a line parallel to the array.

First, in section 5.4.1, the general design of the experiment will be discussed. Section 5.4.2 gives the details about how the stimuli that were used in the experiment were generated, starting from the frequency domain simulations of section 5.3.1. Section 5.4.3 gives the details about how the experiment was set up. Finally, the results are presented and discussed in section 5.4.4.

5.4.1 Experiment Design

Presentation Method. For reasons of practicality and control, the stimuli in this experiment were simulated responses of various WFS configurations, presented to subjects dichotically over headphones. Since the objective was to have subjects compare the colour variations between different listening positions, reproduction by a WFS array would mean that subjects would have to change listening position all the time between stimuli, which is not practical. Alternatively, subjects could remain at one position and the source could be moved instead, but, due to the discretization and finiteness of the array, this situation is not fully equivalent. Furthermore, since colour variations can occur over rather small distances, it would be necessary to fix the subjects' listening position quite accurately, to make sure they are actually lis-

tening to the signals that we want them to listen to. With headphone reproduction these issues are of no relevance. Finally, using simulated signals, rather than, for instance, signals recorded with an artificial head, gives more flexibility regarding making changes in loudspeaker configurations, source positions and listening positions.

Experimental Method. The experimental method that was used is the well-known psychometric method of '*paired-comparisons*'. Generally, in this method subjects compare pairs of signals in terms of a certain quality, with the aim of being able to order the different stimuli on a one-dimensional subjective scale that relates to that quality. The essentials of the method of paired-comparisons are explained in Appendix C.1.
The basic concept of the experiment was as follows. Each trial consisted of presenting a subject with two pairs of dichotic signals:

- **Pair 1**: two dichotic signals, simulated for a certain source position and loudspeaker spacing *A*: one for position (*x, z*), (*x* being the lateral coordinate, parallel to the array and *z* being the coordinate perpendicular to the array) and one for position (*x+offset, z*). The left-ear signal of each dichotic signal was a simulation of the response at 0.1 m to the left and the right-ear signal at 0.1 m to the right of the lateral position *x* (or *x+offset)*, so the position *x* (or *x+offset*) can be thought to be the position of the center of the subject's head, while a between-ears distance of 0.2 m is assumed.
- **Pair 2**: two dichotic signals simulated for the same source position and for the same positions (*x, z*) and (*x+offset, z*) as used in Pair 1, but now for loudspeaker spacing *B*, so the only difference between Pairs 1 and 2 is the loudspeaker spacing.

The details of how the stimuli were generated will be given in the next subsection.
It was the subject's task to indicate in which of the two pairs, Pair 1 or Pair 2, the two dichotic signals were *the most dissimilar* regarding colour. So in this variation of the 'paired-comparison' method, the subjects actually made 'inter-pair' comparisons of 'intra-pair' colour differences, thus comparing *pairs of pairs* of signals rather than just pairs of signals. These inter-pair comparisons were made for different lateral positions *x* for all possible combinations of loudspeakers spacings *A* and *B*, with the remark that the case of comparing *B* to *A* was assumed to be equivalent to comparing *A* to *B*, so a comparison of *B* to *A* was regarded as a replication of the comparison of *A* to *B* and the value in the preference matrix for comparing *B* to *A* was assumed to be 1 minus the value for comparing *A* to *B*. Also, no comparisons of *A* to *A* were

done and a value of 0.5 was assumed in the resulting preference matrix (see Appendix C.1).
In the first preparations of the experiment it was considered to present diotic signals, i.e., identical signals for the left- and right ear, simulated for position (x, z), rather than dichotic ones, in order to avoid localization cues which might distract from the cue of interest: colour. However, after listening to the signals both diotically and dichotically, it was clear that the decoloration effect of presenting the signals dichotically (see section 5.2.4) was so significant that it would be unrealistic to use a diotic presentation. Besides, since in the experimental method that has just been described the subjects compare within each trial two pairs of signals that only differ regarding loudspeaker configuration, the localization cues that are present in the signals are the same for Pair 1 and Pair 2, so their influence is minimized. Additionally, also other spatial cues, such as source broadening, are introduced by using dichotic signals, which can be different for different loudspeaker configurations. Subjects were instructed to try to ignore these cues.

Processing of Results. The experimental procedure described above results in a so-called ‘preference matrix’ which shows for each pair of pairs of signals how often, in terms of the fraction of the total number of times the two pairs were compared, one was judged to have more intra-pair colour difference than the other pair. By assuming a normal distribution for the variations in the choice between the two stimuli in each pair, the fractions can be converted to scale values on a one-dimensional relative ‘colour difference’ scale that ranges from ‘minimum colour difference’ to ‘maximum colour difference’ on which the stimuli can be ordered. Details of the statistical background of the method can be found in Appendix C.1.
Ideally, given our goal of determining the maximum allowable loudspeaker spacing such that no excessive spatial colour variations are perceived, the resulting scale values would show a relatively constant amount of perceived spatial colour variations up to a certain loudspeaker spacing, with a steep increase of perceived colour differences for larger spacings, so that a clear ‘breaking point’ can be identified.

5.4.2 Generating The Stimuli

5.4.2.1 Simulation of frequency-domain responses

The starting point for generating the stimuli were the frequency-domain simulations of section 5.3.1.

5.4.2.2 Compensation for systematic differences between configurations

Because of the different spacings between loudspeakers and the different number of loudspeakers in the different configurations, there are systematic differences between the simulated responses of the different configurations:

Systematic spectral differences. As mentioned earlier, the $\sqrt{jk}$ filter that is present in the driving signals of the secondary sources of a linear WFS is only valid for a continuous array. For a discrete array, spatial aliasing occurs above the Nyquist frequency, the value of which depends on the loudspeaker spacing, the source position and the listening position. Applying the $\sqrt{jk}$ filter above that frequency results in a 3 dB/octave increase of the response for these frequencies, as illustrated in figure 5.7. Applying the $\sqrt{jk}$ filter to the different speaker configurations having different loudspeaker spacings therefore results in systematic colour differences between configurations. This is undesirable when we want to compare the spatial variations of colour between configurations.

Systematic level differences. Because of the different number of loudspeakers of the different configurations, systematic level differences result between simulated responses of different configurations. This is also undesirable when we want to make fair comparisons of spatial colour variations between the responses of the different configurations.

Fortunately, both categories of systematic differences can be compensated for quite easily and simultaneously by designing a compensation filter for each configuration that is the inverse of the spatially averaged response of the configuration. By spatially averaging the responses, the spatial spectral variations are ‘averaged out’ and only the systematic spectral characteristics of the reproduced sound field remain. By filtering the driving signals for each configuration with the inverse of its averaged response, the systematic deviations from a flat spectrum are removed, the level is normalized and the spectral variations that remain are due to spatial aliasing only.

For each combination of loudspeaker configuration and source position, the average of the magnitude of the responses at the simulation points from x= -1.2 to x=1.2 m along the line z=3 m was calculated. Not the full simulated range of x positions was used to avoid that the average would be influenced by the effects of the amplitude tapering of the edges of the array. This average response was inverted to obtain the desired compensation filter and a 3rd-order IIR fil-

ter was designed to implement an approximation of this filter. Then, all simulated responses for this combination of loudspeaker configuration and source position were filtered by this filter to obtain the compensated responses. Figure 5.20 shows an example of the efficiency of this procedure.

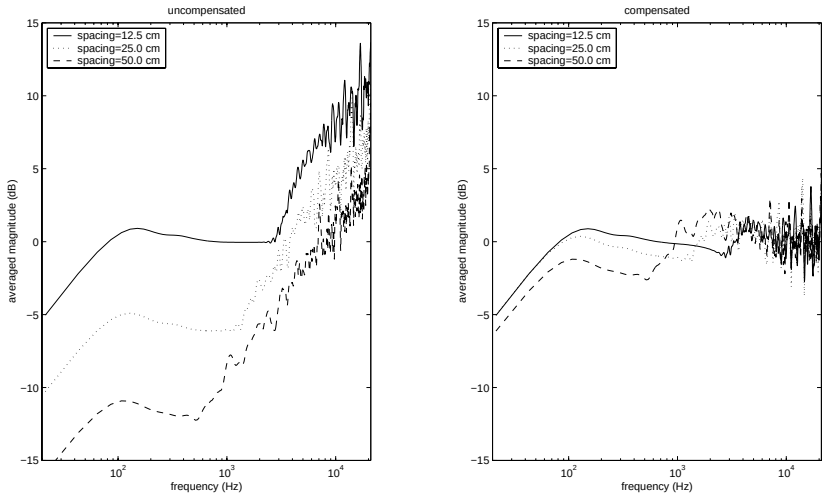


FIGURE 5.20. Spatially averaged system responses for loudspeaker spacings 12.5, 25.0 and 50.0 cm and source position (0, -5) m. The left plot shows the average response, averaged over the range *x*= -1.2 to 1.2 m along the line *z*=3 m, before compensation, the right plot shows the compensated responses.

5.4.2.3 Generating the time-domain signals

To finally arrive at the dichotic time-domain signals for the experiment, the following steps were carried out:

- The compensated frequency-domain responses were inverse-Fourier transformed to obtain the corresponding time-domain impulse responses. Figure 5.21 shows an example of the responses along a line parallel to the array for one configuration, figure 5.22 shows the impulse response at one position on this line.
- Because of the specific application of videoconferencing investigated in this thesis, the obtained impulse responses were convolved with a 2 second sample of either male- or female speech noise: noise having the same amplitude spectrum as the long-term spectrum of male or female speech. The advantage of using speech noise instead of real speech recordings is that it is continuous, so that a shorter sample is sufficient to judge

the colour, and its characteristics are constant, contrary to real speech, of which the characteristics depend much on the choice of a specific speech sample. The speech noise samples used were extracted from [TNO88]. Figure 5.15 shows the spectra of both types of noise.

The time-domain data sets that were generated in Matlab were exported to hard disk as `.wav` files (mono, 16 bit, 44.1kHz) of 2 seconds duration.

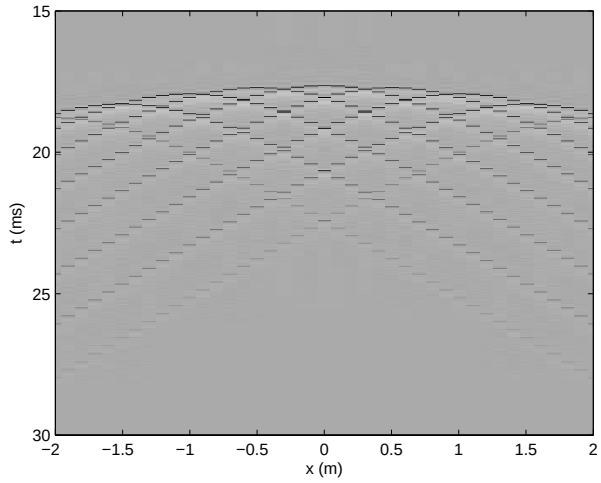


FIGURE 5.21. Time-domain responses along the line *z*=1 m for loudspeaker spacing 50.0 cm and source position (0, -5) m.

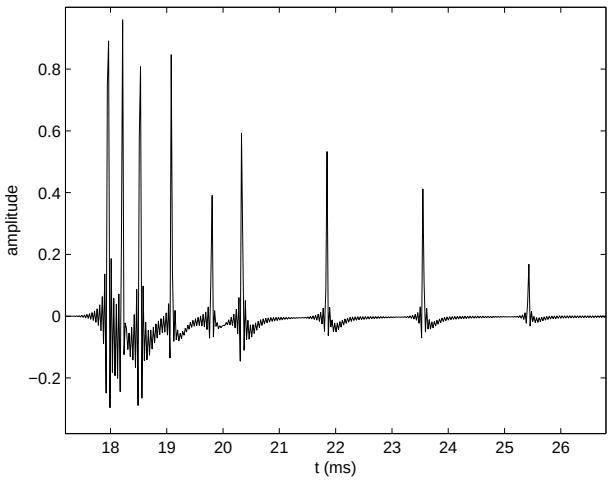


FIGURE 5.22. Time-domain response at *x*= -1.1 m on the line *z*=1 m for loudspeaker spacing 50.0 cm and source position (0, -5) m.

5.4.3 Experimental Set-Up

Five different values of lateral position *x* were used as 'position of the center of the head' in the experiment: *x*= -1.4 m, -0.7 m, 0 m, 0.7 m and 1.4 m. The value of distance *z* was fixed at *z*=3 m throughout the experiment. The value of the offset was chosen randomly from the range -0.5 m to +0.5 m in steps of 0.1 m for each trial individually. This range of offsets corresponds to the range of the offsets used in calculating the SCV Index, as discussed in section 5.3.2.
A single run of the experiment consisted of paired-comparisons of all combinations of speaker coonfigurations (10) for each value of *x* (5), so a single run consisted of 50 trials in total. The presentation order of the 50 trials was randomized, as was the presentation order of configuration *A* and *B* within each pair. A single run of the experiment took about 20-30 minutes per subject to complete.
A set of Matlab scripts was written that controlled the whole experiment automatically and included a graphical user interface for the subject, which is shown in Appendix A.4.

The procedure was as follows: the subject was presented with dichotic stimulus *A1* (loudspeaker configuration *A*, position *x*) and immediately after that dichotic stimulus *A2* was played (loudspeaker configuration *A*, position *x+offset*). Then, after a 1 second pause the second pair was presented: first dichotic stimulus *B1* (loudspeaker configuration *B*, position *x*) and immediately after that dichotic stimulus *B2* (loudspeaker configuration *B*, position *x+offset*). The subject then had to decide in which of the two pairs, *A*1/*A2* or *B1/B2* the two signals were the most dissimilar regarding colour. He/she then simply pressed the button on the user interface labeled 'Pair 1' or the button 'Pair 2'. It was also possible to press 'Repeat' to hear the same two pairs again. After a complete run had been finished, the results were saved for further processing.
Within a single run of the experiment, the parameters *source position* and *source signal* were kept fixed. Separate experiments were executed for three different combinations of values of these parameters:

- **Experiment 1**: source position (0, -5) m, female speech noise.
- **Experiment 2**: source position (0, -5) m, male speech noise.
- **Experiment 3**: source position (0, -1) m, female speech noise.

17 Normal hearing subjects of varied age and sex participated in Experiment 1, 11 in Experiment 2 and 5 in Experiment 3.
The actual instructions that were given to the subjects can be found in Appendix A.4.

5.4.4 Results And Discussion

For each experiment (Experiment 1, 2 and 3) all responses of trials with the same combination of loudspeaker spacings *A* and *B*, source position and source signal were grouped together, so the results for individual subjects, all lateral positions *x* and all values of the offset were regarded as samples from the same population of '*responses for comparing loudspeaker configuration A to loudspeaker configuration B for source position C and source signal D'*.
The results for the three experiments are shown in figure 5.23. The "perceived spatial colour variation" scale values of the individual loudspeaker spacings of each experiment were obtained in the following way (for a more detailed description, see Appendix C.1). For each combination of loudspeaker configurations *A* and *B*, the fraction p_{AB} of the total number of trials in which configuration *A* was compared to configuration *B* and in which the stimulus pair of configuration *A* was judged to have a larger colour difference than the stimulus pair of configuration *B* was calculated. These fractions p_{AB}, numbers between 0 and 1, were collected in a (*5* x *5*) preference matrix. Then, each fraction p_{AB} was transformed to the scale value difference z_{AB} between *A* and *B*, by determining the value *z* for which $F(z)=p_{AB}$, in which *F* is the cumulative distribution function of the standard normal distribution. Finally, the scale value for configuration *A* was calculated by taking the average of all five z_A**.** values. For convenience, the minimum of the five scale values for each experiment was subtracted from the scale values, so that for each of the three experiments the resulting minimum scale value is zero. The preference matrices of the three experiments are given in Appendix C.2.

Looking first at the results for Experiment 1 (figure 5.23, left: source at (0, -5) m, female speech noise) we see an interesting phenomenon. The scale values for the loudspeaker spacings of 12.5, 16.7 and 25.0 cm are relatively low and almost constant, indicating that no clear differences in the amount of spatial colour variations were perceived when pairs of signals of these loudspeaker spacings were compared to each other. This is also clear when we look at the corresponding *p* values for trials that compared these three configurations to each other in the preference matrix of Experiment 1 (table C.1 in Appendix C.2), which are all between .41 and

.59, very close to chance level, indicating that subjects did not notice a clear systematic difference between them, regarding spatial colour variations. When the loudspeaker spacing is increased from 25.0 to 33.3 cm, we see a sudden increase of the scale value, indicating that this loudspeaker spacing was consistently judged to have more colour variations than the three smaller spacings. When the spacing is increased further to 50.0 cm, the colour variations increase even more. This also is reflected in the *p* values in the preference matrix.

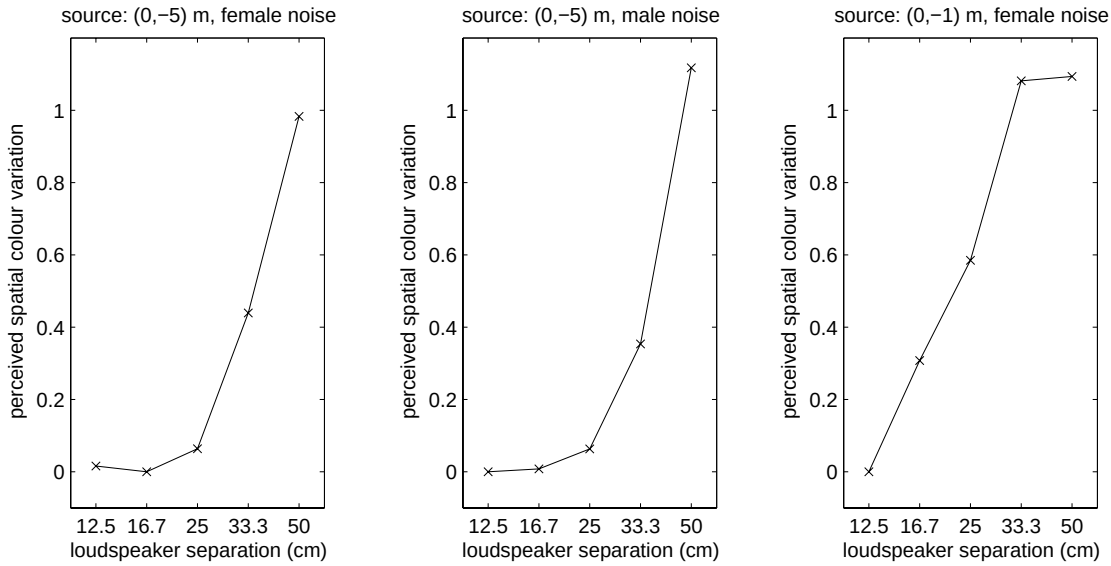


FIGURE 5.23. Results of the three paired-comparison experiments on perceived spatial colour variations. Scale values (vertical axis) are relative, with a scale value of 0 corresponding to the loudspeaker spacing for which the minimum amount of spatial colour variations was perceived. Left: Experiment 1 (source at (0, -5) m, female speech noise). Center: Experiment 2 (source at (0, -5) m, male speech noise). Right: Experiment 3 (source at (0, -1) m, female speech noise).

When we look at the values of the SCV Index for this same situation (table 5.2, center column) we see a value of the SCV Index of 12.3 dB for the spacing of 25.0 cm and a value of 15.4 dB for the spacing of 33.3 cm, so the results of Experiment 1 seem to indicate a critical value of the SCV Index, between these two values, below which an increase of the SCV Index does not result in a noticeable increase in spatial colour variations, while an increase of the SCV Index to a value just above this critical value results in a sudden increase in perceived spatial colour variations. The increase of the perceived colour variations that were found in Experiment 1 when going from a spacing of 33.3 cm to 50.0 cm is also consistent with the corresponding increase of the SCV Index in table 5.2 and indicates that above the critical value of the SCV

Index, the perceived colour variations increase for increasing SCV Index.

Now let us look at the results of Experiment 2 (figure 5.23, center: source at (0, -5) m, male speech noise). The results are very similar to those of Experiment 1, with similar low and almost constant scale values for spacings up to 25.0 cm, a sudden increase for 33.3 cm and a further increase for 50.0 cm. In section 5.3.2 we found that due to the fact that the male speech noise contains almost no energy above 6 kHz, while this upper limit is about 8 kHz for female speech noise, the values of the SCV Index for male speech were lower than those for female speech noise. Looking at those values of the SCV Index for this situation (table 5.3, center column), we see a value of 11.8 dB and 15.0 dB for spacings of 25.0 cm and 33.3 cm, respectively. This is consistent with the critical value between 12.3 dB and 15.4 dB that was suggested by the results of Experiment 1. Also, the slightly larger increase of perceived colour variations when going from 33.3 cm to 50.0 cm that was found in Experiment 2 compared to Experiment 1 is consistent with the slightly larger increase in SCV Index that was found for male speech noise compared to female speech noise.

Finally, we look at the results of Experiment 3 (figure 5.23, right: source at (0, -1) m, female speech noise). Here we do not see a ‘knee’ as in Experiments 1 and 2, but a steady increase of perceived colour variations with increasing loudspeaker spacing, except when going from 33.3 cm to 50.0 cm, in which case the perceived colour variations increase only slightly. When we look at the values of the SCV Index for this situation (table 5.2, left column), these results can be interpreted as being consistent with the results of Experiment 1 and 2. From Experiments 1 and 2 we concluded that a sudden increase in colour variations was perceived when the SCV Index exceeded a critical value, which, according to the combined results of Experiment 1 and 2, should be located between 12.3 dB and 15.0 dB. From table 5.2 we see that the SCV Index for the spacing of 12.5 cm in Experiment 3 is 10.1 dB, which is below the range of the critical value that was found in the first two experiments. For the spacing of 16.7 cm the value has already increased to 13.0 dB, which is within the range for the critical value that was found in the first two experiments. Seeing that there is a significant increase in the amount of perceived colour variations, this seems to suggest that this value of 13.0 dB exceeds the actual critical value of the SVI Index, so that the range within which the critical value is located can now be narrowed down to between 12.3 dB and 13.0 dB. Finally, note that the only small increase of perceived colour variations when going from 33.3 to 50.0 cm agrees with the also only small increase of the SCV Index in table 5.2 (first column).

Note that the results of the paired-comparison test do not suggest that no spatial colour variations were perceived for situations with an SCV Index below the critical value, as there was no absolute reference in the experiment. They merely suggest that the strength of the perceived colour variations is more or less constant below this value and increases noticeably above it. However, from extensive practical experience with reproduction of numerous types of sound sources, including speech, with a WFS array with a loudspeaker spacing of 12.5 cm[4], it is safe to state that, in general, no clear spatial colour variations are perceived with that spacing. This notion can serve as a more or less absolute reference, indicating that the spacings that in the perception experiment resulted in similar scale values as the 12.5 cm spacing, do not exhibit clear spatial colour variations either. Combined with the fact that a critical value of the SCV Index has been found that is consistent with the results of all three experiments, this implies that as long as the SCV Index of a configuration is below this critical value, no significant perception of spatial colour variations is to be expected.

In conclusion, the results of the three perception experiments described in this section seem to indicate that the objective analysis of spatial colour variations of section 5.3.2 and the SCV Index that was proposed as an objective measure are indeed suitable to predict the amount of spatial colour variations that is perceived when listening to the sound field reproduced by a given WFS configuration. Also, there seems to be a critical value of the SCV Index of about 12-13 dB, below which no clear spatial colour variations are perceived, while a sudden increase of colour variations is perceived when the critical value is exceeded and a monotonous increase of perceived colour variations is found for increasing values of the SCV Index above the critical value. With this, the second objective of the experiment, as stated at the start of this section, has been reached.

With regard to the first objective, finding the maximum loudspeaker spacing that can be allowed in the particular application of videoconferencing, we can conclude from the experiments that 25 cm can be taken as a reasonable limit. The results of Experiment 3 suggest a smaller upper limit, but the small distance of only 1 m from the source to the array can be considered to be a somewhat extreme case in the life-size videoconferencing application.

4. This is the default set-up of the WFS demonstration system at the Laboratory of Acoustic Imaging and Sound Control at Delft University of Technology

5.5 Conclusions

In this chapter an analysis was made of the coloration artifacts that can arise in the reproduced sound field of a WFS array, due to spatial aliasing. First the definitions of ‘colour’ and ‘coloration’ were given in section 5.1, after which the perception of coloration was discussed in section 5.2. The essential first processing step in the perception of coloration is the filtering of the incoming acoustical signal by the auditory filters, resulting in the internal or auditory spectrum, which is the basis for all subsequent processing (section 5.2.1). A criterion for determining the threshold of coloration of signals was discussed (section 5.2.2) and a model, based on the internal spectrum, for determining the colour difference between two coloured signals was presented (section 5.2.3).

Section 5.3 specifically addressed the coloration that occurs in WFS reproduction. From simulations of various configurations of loudspeaker arrays, source- and listening positions, it was shown that coloration of the original source signal should be expected to occur in practical applications (section 5.3.1). However, it was argued that the main concern should not be the absolute colour difference between the reproduced signal and the original signal, but rather the spatial variations of colour that a listener experiences when moving around in the listening area. These spatial colour variations were investigated in section 5.3.2. An objective model was developed and a single-number measure for the spatial colour variations of a given configuration, the SCV Index, was proposed. The model was applied to the simulations of the configurations of section 5.3.1.

Section 5.4, finally, described a paired-comparisons perception experiment in which subjects compared the spatial colour variations of the various configurations. The results were consistent with the proposed objective model and single-number measure for spatial colour variations and indicated the existence of a critical value of the SCV Index of about 12-13 dB, below which no strong spatial colour variations are perceived, while when the SCV Index is increased above the critical value there is a sudden increase in the perceived amount of spatial colour variations. If the SCV Index is increased further, the perceived amount of spatial colour variations increases as well.

For the specific application of life-size videoconferencing, a maximum inter-loudspeaker distance of 25 cm seems to be reasonable.

CHAPTER 6 *Multi-Actuator Panel (MAP) Arrays*

One of the main factors that can be considered as still limiting the widespread application of Wave Field Synthesis is the fact that a large number of loudspeakers is required, taking up a significant amount of space in the listening environment, which in some applications is undesirable for aesthetic or practical reasons.
A solution might be found in looking for different types of acoustical transducers for WFS reproduction instead of conventional electro-dynamic cone loudspeakers. In previous research at TU Delft, the possibility of building arrays consisting of electrostatic transducers was considered and tried (although this was more because of the expected better control of directivity compared to electro-dynamic speakers) but this appeared to be impractical, mainly because of the low efficiency ([Verh97]).

The past couple of years, a new type of loudspeaker has made its appearance in the audio world, called the ‘Distributed Mode Loudspeaker’ (DML), first introduced by British company NXT ([NXT], [Azim97], [Harr97]). A DML consists of a flat, thin panel made of some stiff material, which is forced to vibrate by an electro-dynamic transducer, the *exciter*, attached to the back of the panel. The vibrations of the exciter give rise to bending waves in the panel, which as a result starts to radiate acoustical waves into the surrounding air, thus acting as a loudspeaker.

A big advantage of DML loudspeakers is that they can be extremely thin and lightweight. Furthermore, they do not look like typical loudspeakers and can more easily be integrated in the room interior. Both points are especially advantageous in situations where many loudspeakers are necessary, as is the case with WFS. Using DML panels for WFS reproduction, it would be possible to mount the loudspeakers directly on or even enclosed into the walls, which would eliminate some of the practical and aesthetic problems mentioned before.
An additional advantage in the specific application of videoconferencing is that the panels, due to their flatness and due to the fact that, contrary to a cone loudspeaker, they are not moving as a whole, can be used as a video projection screen at the same time as well, thus eliminating some of the practical problems that occur with conventional loudspeakers, which have to be placed either behind the screen, creating the need for an acoustically transparent screen, which in general is not beneficial for the visual quality, or below, above or beside the screen, which possibly introduces sound localization problems.
However, because of the completely different working mechanism of DML panels compared to conventional loudspeakers, the DML panels behave differently than those conventional loudspeakers in several aspects. In particular, their radiation behavior is often characterized as being “diffuse”, both in a spatial and temporal sense, by which is meant that both the spatial and temporal coherence of their radiated sound field is low ([Azim97]). Therefore, their suitability for application in WFS is not trivial and should be investigated first. A first exploratory study of the applicability of DML’s for WFS reproduction is the subject of this chapter.

In section 6.1 we will first look at the basic theoretical aspects of Distributed Mode Loudspeakers. From this, some of the characteristics of DML panels will become clear that are important to take into account when we want to apply them in a WFS system.
Then, in section 6.2, measurements on small DML panels and on an array of small DML panels will be analyzed that were meant as a pilot study to get more insight into the characteristics of individual panels and to prove experimentally that it is indeed possible to do Wave Field Synthesis with DML panels by constructing an array of individual small DML panels.
We will see that one of the problems with DML loudspeakers is the low-frequency response, which is limited mainly by the physical size of the panels. Therefore, the use of larger panels is preferred. However, for WFS reproduction a small spacing between the individual secondary sources is required in order to avoid artifacts caused by spatial aliasing. Therefore, an interesting question is: is it possible to construct a properly working WFS array by attaching several

exciters to a single large panel of DML material? If this is indeed possible, then the practical problem of constructing an array of many individual loudspeakers is reduced to attaching an array-like arrangement of exciters to the back of a single large DML panel, while at the same time this is expected to improve the low-frequency response, because of the larger size of the panel. This concept will be referred to as a *Multi-Actuator Panel* (MAP). Additionally, in the videoconferencing application, using a MAP as both loudspeaker array and projection screen would eliminate the seams between individual panels that might interfere with the requirements for a high-quality projection screen. The first investigations on Multi-Actuator Panels are the subject of section 6.3.

Most of the experimental work presented in this chapter was carried out in the context of the Master's project of Wilfred van Rooijen. More details about various subjects described in this chapter can be found in Van Rooijen's Master's thesis ([Rooij01]) and reference will be made to his thesis at several points.

6.1 Distributed Mode Loudspeaker Theory

The basic concept of a Distributed Mode Loudspeaker (DML) is nothing new; in fact it is a phenomenon that is experienced in everyday life. Whenever a thin plate is caused to vibrate by an external source, the vibrations in the plate in turn cause acoustical waves to be radiated from its surface into free air. Examples are someone knocking on the door or window, motor noise of a car or airplane being passed on from the exterior enclosure to the interior, etcetera.
In all these examples, the vibrations of a mechanical source are transferred to a plate of solid, stiff material with which the source is in direct contact and in which, as a result, bending waves are induced. These bending waves propagate through the material and cause the surface of the plate to vibrate. At the interface of the plate and the surrounding air, part of the energy of the bending waves is converted into acoustical waves through the acoustic radiation impedance of the system, finally resulting in audible sound waves in the surrounding air.
A DML works according to the same principle. It consists of a thin piece of stiff material that may or may not be built into an enclosure. To the back of the material an electro-dynamic exciter is connected, which converts an electrical input signal into a mechanical movement that is passed on to the DML material, thus generating bending waves that are then converted into acoustical waves at the free surface of the material, so that the DML starts radiating sound.

Although the electro-dynamic exciter of a DML is very similar to the magnet/voice-coil excitation system of a conventional cone loudspeaker, the way in which the mechanical vibration of the exciter is converted to acoustic waves is very different. In a cone loudspeaker, the cone ideally moves back and forth as a whole, thus acting like a piston. In a DML however, as explained above, the vibrations are converted to bending waves, causing the material to vibrate more or less randomly instead of moving as a whole in a piston-like fashion.

The mathematical description of bending waves is more complicated than that of acoustical waves, because they involve transverse as well as rotational movement and a detailed description of the their theory is beyond the scope of this thesis. For a detailed description of bending wave theory, the reader if referred to the vast amount of literature on the subject (for instance, [Crem88], to which reference will be made several times in this section).

6.1.1 Panel Vibrations

We will discuss the behavior of the DML according to the geometry shown in figure 6.1. The DML is a thin, flat plate, with thickness *h*, positioned in the *x*-*y* plane. The excursion $\zeta(x,y,t)$ of the plate in the *z* direction normal to its surface is then governed by the two-dimensional bending wave equation ([Crem88]):

$$B'\left(\frac{\partial^4\zeta}{\partial x^4}+2\frac{\partial^4\zeta}{\partial x^2\partial y^2}+\frac{\partial^4\zeta}{\partial y^4}\right)+\rho h\frac{\partial^2\zeta}{\partial t^2}=q(x,y,t), \tag{6.1}$$

in which ρ is the density of mass of the plate material, $q(x,y,t)$ is the external pressure applied to the plate in the normal direction *z* and B' is the bending stiffness of the plate, which is given by:

$$B'=\frac{Eh^3}{12(1-\nu^2)}, \tag{6.2}$$

with *E* Young's modulus (in general frequency-dependent) and ν Poisson's ratio, which are both properties of the plate material.

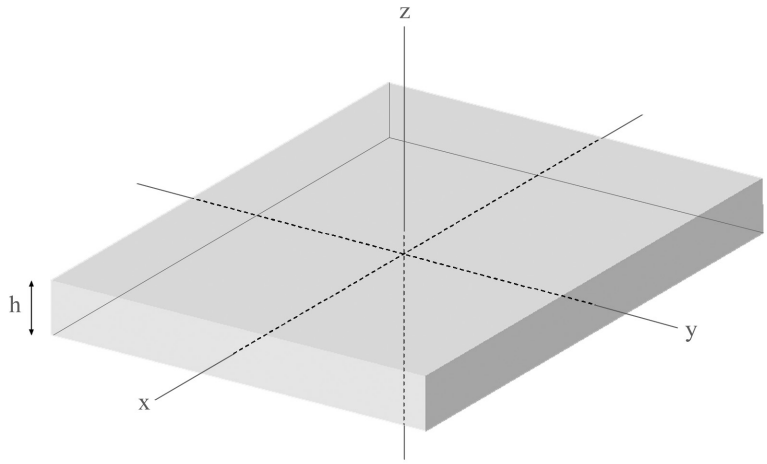


FIGURE 6.1. The geometry for the analysis of the vibrations of a DML panel.

By taking the Fourier transform of (eq. 6.1) we obtain the two-dimensional bending wave equation in the space-frequency domain:

$$\left(\frac{\partial^4 \tilde{\zeta}}{\partial x^4} + 2 \frac{\partial^4 \tilde{\zeta}}{\partial x^2 \partial y^2} + \frac{\partial^4 \tilde{\zeta}}{\partial y^4} \right) - k_B^4 \tilde{\zeta} = \frac{Q(x, y, \omega)}{B'}, \tag{6.3}$$

with $\tilde{\zeta}(x, y, \omega)$ the Fourier transform of excursion $\zeta(x,y,t)$, $Q(x,y,\omega)$ the Fourier transform of external pressure function $q(x,y,t)$ and k_B the *bending wave number*:

$$k_B = \sqrt[4]{\frac{\omega^2 \rho h}{B'}}. \tag{6.4}$$

From (eq. 6.4) follows the bending wave phase velocity:

$$c = \frac{\omega}{k_B} = \sqrt[4]{\frac{\omega^2 B'}{\rho h}} \tag{6.5}$$

and the bending wave group velocity:

$$c_g = \left[\frac{dk_B}{d\omega} \right]^{-1} = 2 \sqrt[4]{\frac{\omega^2 B'}{\rho h}} = 2c. \tag{6.6}$$

From (eq. 6.5) it is seen that the propagation of the bending waves is dispersive.
Using the relation that the speed of the plate in the direction normal to the plate surface is the time derivative of normal excursion ζ, an expression similar to (eq. 6.3) can be derived for the normal velocity $V(x,y,\omega)$ of the plate:

$$\left(\frac{\partial^4 V}{\partial x^4} + 2\frac{\partial^4 V}{\partial x^2 \partial y^2} + \frac{\partial^4 V}{\partial y^4} \right) - k_B^4 V = \frac{j\omega Q}{B'}. \tag{6.7}$$

For arbitrary boundary conditions of the plate and source function $Q(x,y,\omega)$, it is very difficult to obtain an analytical solution of (eq. 6.7) and numerical methods, such as finite element methods, have to be used to obtain the normal velocity profile.

Infinite Plate with Point Excitation. For the particular case of an infinite plate that is excited by an external point force F_0, the normal velocity V_0 at the excitation point follows from the mechanical point impedance $Z_{m,0}$ through the relation ([Crem88]):

$$Z_{m,0} = \frac{F_0}{V_0} = 8\sqrt{B'\rho h}, \tag{6.8}$$

and the resulting normal velocity field $V(x,y,\omega)$ is given as a function of $k_B r$ by:

$$V(k_B r) = V_0 \left[H_0^{(2)}(k_B r) - H_0^{(2)}(-jk_B r) \right], \tag{6.9}$$

with r the distance from the excitation point and $H_0^{(2)}$ the zeroth-order Hankel function of the second kind ([Crem88]). The normalized absolute value of (eq. 6.9) is shown in figure 6.2 for values of $k_B r$ up to 25. As an indication, a typical value of k_B for a plate of polymer material, such as PVC, of 5 mm thickness at 1 kHz is k_B=25 m^{-1}, so that the range of $k_B r$ in figure 6.2 corresponds to an area of about 1 m around the excitation point in this case. It is seen that the excitation results in a high normal velocity only in a small area around the excitation point and decreases rapidly at larger distances. This will prove to be of high importance later on in this chapter when we will discuss Multi-Actuator Panels (section 6.3). For a metal plate with a thickness of 5 mm, k_B is typically about 10 m^{-1} at 1 kHz.

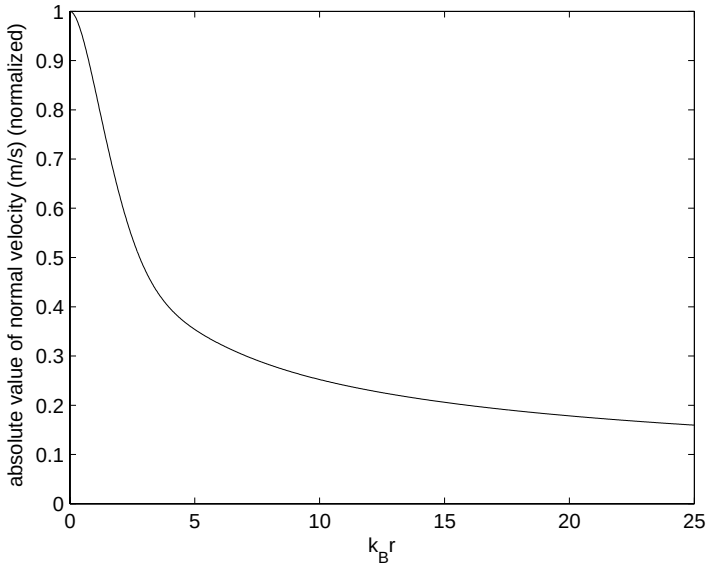


FIGURE 6.2. Absolute value of the normal velocity of an infinite plate due to point excitation at r=0 as function of $k_B r$ according to (eq. 6.9). The vertical axis has been normalized.

Finite Plates. For the more general case of a finite plate, the solution of (eq. 6.7) depends on the boundary conditions and the external excitation Q and is in general very hard to obtain in analytical form. However, it can be shown ([Crem88]) that it can be expressed as an infinite summation of all the normal velocity eigenfunctions of the plate, which are monochromatic orthogonal functions that are solutions of the homogeneous version (Q=0) of (eq. 6.7) and satisfy the boundary conditions, so that we can write:

$$V(x,y,\omega) = \sum_{n=0}^{\infty} a_n(\omega)\varphi_n(x,y), \tag{6.10}$$

in which φ_n is the eigenfunction corresponding to eigenfrequency ω_n and $a_n(\omega)$ is the weighting coefficient that determines the contribution of φ_n to the total normal velocity field at frequency ω

Since, according to (eq. 6.10), the velocity field of a finite plate is built up from a discrete set of eigenfunctions corresponding to a discrete set of eigenfrequencies ω_n, the frequency response of a vibrating DML panel will in general not be flat. It can be shown ([Crem88]) that the modal density (the number of eigenfunctions per unit frequency) is proportional to the size of the panel, so it will be higher for larger panels. In general, the frequency response will there-

fore be flatter for large panels than for small panels. Furthermore, the size of the panel determines the lowest value of ω_n for which an eigenmode of the panel exists. This means that the size of the panel sets a lower limit on its frequency range and the frequency range of larger panels extends down to lower frequencies than that of smaller panels.

Exciter Position. The eigenfunction weighting factor a_n of eigenfunction φ_n in (eq. 6.10) can be shown to be proportional to the integral of the product of external driving function Q and eigenfunction φ_n over the plate surface S ([Crem88]):

$$a_n \propto \int_S Q(x,y)\varphi_n(x,y)dxdy. \tag{6.11}$$

In the case of point excitation, or when the area over which $Q \neq 0$ is small compared to the spatial period of eigenfunction φ_n, so that φ_n can be regarded as being constant over this area, φ_n can be taken out of the integral in (eq. 6.11) and a_n is proportional to the value of φ_n at the center of the excitation area (x_0,y_0):

$$a_{n,\text{point excitation}} \propto \varphi_n(x_0, y_0). \tag{6.12}$$

From (eq. 6.12) it is clear that if the point excitation is applied at a position at which φ_n is close to zero, this eigenfunction will not contribute significantly to the total response of the plate. Likewise, if the excitation is applied at a position at which φ_n has a maximum, it will contribute maximally. This means that the frequency response of the plate depends on the excitation position.

In the case of DML panels, the common exciter types are attached to the back of the panel by a thin metal ring with a typical diameter of several centimeters. As an example, consider a polymer plate with thickness 5 mm to which an exciter is attached having a diameter of 5 cm. Using typical values of the relevant material properties, it follows from (eq. 6.4) that the exciter diameter equals one-tenth of the bending wavelength $\lambda_B = 2\pi/k_B$ at 250 Hz and $0.03\lambda_B$ at 20 Hz, so for the lowest audio frequencies the exciter can indeed be regarded as a point excitation, so that (eq. 6.12) applies and the response is strongly dependent on the exciter position. At higher frequencies the exciter can no longer be regarded as a point excitation. Because of

the increasingly complex spatial shapes of the eigenfunctions for increasing eigenfrequencies, it is difficult to make general statements about the result of the integration in (eq. 6.11) as a function of exciter position. It can however be stated that the range of variation of a_n with varying exciter position will not exceed the range of variation for the point excitation case (assuming Q is constant over the area where it is unequal to zero and equal total force applied to the panel in both cases) and therefore the total frequency response will depend less on the exact exciter location than for the low frequencies.

Damping. Until now, internal damping in the plate was not taken into account in the analysis of DML panel vibrations. Internal damping of the material can be accounted for by introducing a complex Young's modulus E in (eq. 6.2) ([Crem88]):

$$E = E_0(1 + j\eta), \tag{6.13}$$

with η the *damping factor* (or: *loss factor*), which is frequency dependent. This leads to a modified version of (eq. 6.3), with now a complex bending wave number k_B, which follows from (eq. 6.13), (eq. 6.2) and (eq. 6.4) and which, from a first-order approximation of (eq. 6.4), is given by:

$$k_B = k_{B,0}(1 - j\frac{\eta}{4}), \tag{6.14}$$

with $k_{B,0}$ the bending wave number without damping. Note that from (eq. 6.5) it follows that also the velocity c becomes complex when damping is present.

The main effect of damping is a decrease of the amplitude of the bending waves over distance and hence a reduction of the total vibrational energy of the plate. The energy corresponding to this reduction of amplitude is converted to heat by the damping mechanism. As an indication of order of magnitude, damping factor η is in the range 10^{-4}~10^{-2} for various metals ([Crem88]) and in the range 10^{-2}~10^{-1} for a PVC foam panel laminated with cardboard paper on both sides ([Boon04b]).

6.1.2 Acoustic Panel Radiation

What we are primarily interested in is not so much the vibrational behavior of a DML panel per se, which was analyzed in the previous subsection, but the sound field that the DML panel radiates into the surrounding space due to these vibrations. Because of the complicated bending wave patterns that are responsible for the sound radiation of a DML panel, their radiation characteristics are not as easy to describe as those of cone loudspeakers, which essentially act like a moving piston for sufficiently low frequencies. In [Azim97] an attempt is made to model DML panel radiation by means of a rather arbitrary distribution of so called 'elementary sources' of alternating signs on a virtual planar 'control surface' in the air just in front of the DML panel. In this model, no relation is present between the vibrations of the panel's surface and the distribution of elementary sources and some assumptions are made that seem unjustified. Since also no theoretical justification of the model is offered, this model will not be discussed further here.

A more elegant and theoretically justifiable way to describe the sound field radiated by a DML panel is to use the Rayleigh I integral (eq. 2.13), which is repeated here for convenience:

$$P(\vec{r}_R,\omega) = \frac{j\omega\rho_0}{2\pi}\int_S V_n(\vec{r},\omega)\frac{e^{-jk|\vec{r}-\vec{r}_R|}}{|\vec{r}-\vec{r}_R|}dS, \tag{6.15}$$

which describes the sound pressure at receiver position r_R by a continuous distribution of monopole sources on surface *S*, the monopole source at position *r* having an amplitude equal to the value of the normal component of the particle velocity V_n at that same position *r*. Realizing that at the surface of the DML panel the air has to move with the same normal velocity as the panel and assuming that V_n is zero at all other positions in the plane in which the panel is located, equivalent to assuming that the panel is mounted in an infinite baffle, we see that we can describe the sound field radiated by the DML panel into the air by taking the surface of the DML panel for *S* and the normal bending wave velocity profile of the panel surface, given by (eq. 6.10), for V_n in the Rayleigh I integral (eq. 6.15).

Unfortunately, as explained in the previous subsection, it is very difficult to obtain analytical solutions for the eigenfunctions that are needed to calculate the normal bending wave velocity of a practical panel of finite size, so that exact evaluation of (eq. 6.15) to obtain the radiated sound field is in general not possible. However, from looking at the relatively simple case of an

infinite point-excited plate, for which the normal velocity is given in (eq. 6.9), we can explain one of the most well-known properties of DML panels, namely that their directivity pattern exhibits less beamforming at high frequencies than conventional cone loudspeakers ([Azim97], [Harr97]). Since the normal velocity of the plate according to (eq. 6.9) is a function of the product $k_B r$, the velocity profile as depicted in figure 6.2 corresponds to a decreasing range of the distance r from the excitation point for increasing frequency. In other words, the higher the frequency, the narrower the area around the excitation point becomes in which the normal velocity has a significant value. This effectively means that if we insert the normal velocity profile of (eq. 6.9) for V_n in the Rayleigh I integral (eq. 6.15) to calculate the sound field radiated by the panel, we introduce a sort of aperture function in the integral, the size of which decreases for increasing frequency. This counteracts the effect of increasing beamforming for increasing frequency that is observed for constant-size apertures, such as a cone loudspeaker, resulting in less beamforming at high frequencies for a DML panel than for a conventional cone loudspeaker.

This is illustrated in figure 6.3, which shows the far-field directivity plot of an infinite point-excited PVC plate of 5 mm thickness (drawn in the right half of each directivity plot) versus the directivity of a cone loudspeaker (drawn in the left half of each directivity plot) for six different frequencies. The response of the infinite plate was calculated by numerical evaluation of the Rayleigh I integral (eq. 6.15) using normal velocity profile (eq. 6.9) with $k_B = 0.32\sqrt{\omega}$ (which follows from inserting the values of the material properties of PVC in (eq. 6.4)) . The integration was carried out over a plate area of 1 x 1 m centered around the excitation point, with a spacing between grid points of 1 cm in both the x and the y direction. For the lowest frequency that was considered, 200 Hz, $k_B r$=11 at the edges of the integration area. Looking at figure 6.2, we see that this means that at the edges the amplitude of the normal velocity is already 12 dB below the amplitude at the excitation point, so limitation of the integration to these boundaries seems acceptable for the purpose of a qualitative illustration. The directivity plot of the conventional cone loudspeaker was calculated using the well know directivity function for a rigid piston:

$$P_{cone} \propto \frac{J_1(kd\sin\theta)}{kd\sin\theta}, \tag{6.16}$$

with J_1 the Bessel function of the first kind, d the radius of the piston and θ the angle relative to the on-axis direction. The radius d was chosen to be 10 cm. All responses in figure 6.3 were normalized to an on-axis response of 80 dB.

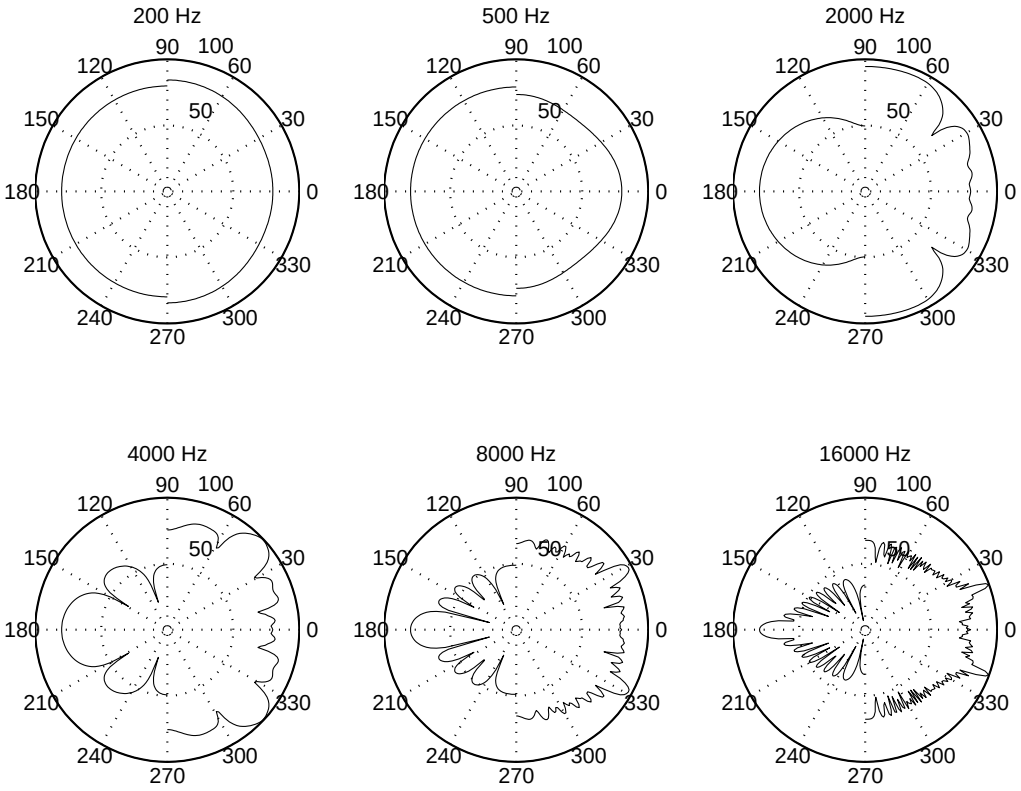


FIGURE 6.3. Far-field directivity plots of an infinite point excited plate (shown in the right half of each plot) and a piston loudspeaker with radius 10 cm (shown in the left half of each plot) for six frequencies. All responses were normalized to an on-axis response of 80 dB.

In figure 6.3 we clearly see that the simulated radiation of the DML panel indeed exhibits much less beamforming at high frequencies than a conventional cone loudspeaker and that the radiation remains quite broad up to high frequencies. What is also interesting to note is the occurrence at high frequencies of two off-axis directions with increased response. These can be shown to occur at angles θ for which the bending wavenumber in the panel k_B equals the spatial wavenumber of the radiated sound field $k_x=k\sin\theta$, so when $\sin\theta=k_B/k$. This is an effect known as ‘coincidence’ ([Crem88]). For frequencies for which $k_B>k$ (in the example of figure 6.3: for frequencies below 1.9 kHz) the effect does not occur.

Although in practice we are dealing with finite-sized DML panels, the reasoning for infinite

panels given above seems to be valid in a qualitative sense also in the finite panel case, since we have seen from (eq. 6.9) and figure 6.2 that the amplitude of the normal velocity decreases very rapidly with increasing distance from the point of excitation, so that the influence of the edges of the panel will not be very strong, provided that the panel is not too small and the exciter is not positioned too close to an edge of the panel. In this last statement, “too small” is to be interpreted as relative to the size of the ‘high velocity area’ of the velocity response of the infinite panel (figure 6.2), which is determined by the value of k_B, which in turn is determined by the material properties and the thickness of the plate. Boone ([Boon04b]) describes measurements which indeed confirm the validity of the above reasoning for finite plates. He measured the impulse response of a 1.3 x 0.75 m panel of PVC foam of 5 mm thickness laminated at both sides with cardboard paper, on a grid of measurement points at 0.5 cm distance from the surface of the panel. From these measurements, the normal velocity at the panel surface was calculated by means of inverse wave field extrapolation based on the inverse of the discretized version of Rayleigh I integral (eq. 6.15). The resulting calculated normal velocity profile corresponded quite well to that of an infinite plate as given by (eq. 6.9).

Another characteristic of the sound field of a DML that has been reported in literature is that the radiated sound field is *diffuse*, both in a temporal and in a spatial sense, by which is meant that both the temporal and the spatial coherence of the sound field is low ([Azim97]). This diffuseness can be understood as follows: after the initial excitation of the panel by the exciter, first circular bending waves will start to propagate from the exciter position. When they reach the edges of the panel, they are reflected and as the number of reflections increases, the bending wave pattern of the panel surface becomes more and more complicated, until it eventually becomes a two-dimensional diffuse field, with both virtually random local temporal velocity variations and virtually random spatial velocity variations. This is similar to a reverberant field that is built up in a room from an increasing number of reflections from the room boundaries, until the sound field in the room becomes fully diffuse. Since the normal movements of the panel surface are transferred to the air at the panel surface, the radiated sound field is expected to show a similar temporal and spatial behavior. Temporally, this means that first a relatively sharp peak is expected, resulting from the time interval between excitation and the moment the bending waves reach the first edge, after which the time response becomes increasingly dense, until it finally becomes virtually stochastic. Spatially, it means that at first the spatial sound field will resemble that of a point source at the exciter position, after which it will become less

and less spatially localized, while the response will also become increasingly dissimilar for different receiver positions.

To be able to apply DML panels in Wave Field Synthesis, it is very important that the initial part of the panel response indeed behaves in the deterministic way described above, since for WFS it is essential that the phase relation between individual secondary sources can be fully controlled. In the next section experiments are described that were carried out to determine if this is the case.

6.2 Experiments On Individual DML Panels And A Multi-Panel Array

For our first experiments on the applicability of DML panels for WFS, we were supplied with a small number of prototype DML panels by New Transducers Ltd., UK. They were small panels mounted in a shallow enclosure, intended for automotive applications. The specifications were as follows:

- Panel material: pressed paper.
- Panel size: 17.6 x 12.7 x 0.3 cm (w x h x d).
- Panel mass: 15 g.
- Low-frequency limit: 300 Hz.
- Exciters: 2 (to increase radiated power), diameter 2.5 cm, mass: 63 g each.
- Enclosure size: 21 x 17 cm size, filled with damping material. Panel attached to enclosure by foam strip suspension.

First, in the next subsection, measurements of the time-, frequency- and directional response of individual panels will be presented. Then, in section 6.2.2, the measurements on a multi-panel array consisting of 9 individual panels are presented.

6.2.1 Measurements On Individual DML Panels

The impulse response was measured for several individual DML panels. The individual panels were placed on an electronic turntable in an anechoic room and the impulse response was measured at 3 m distance for the full 360 degrees of the horizontal plane in steps of 5 degrees, using a MLSSA measurement system and a B&K measurement microphone. Also the impulse response of a small conventional cone loudspeaker (Vifa M110) in a small box was measured as a reference.

Figure 6.4 shows the on-axis impulse response for two individual panels and the reference loudspeaker. It is seen that, as expected, the impulse responses of the DML panels have a longer ‘tail’ after the initial peak than the reference loudspeaker, due to the complicated bending wave vibrations of the panels after excitation. It is also interesting to compare the impulse responses of the two individual panels. We see that the first parts of the responses, containing the initial pulse and the first millisecond or so after it, are very similar to each other. After this first part, the similarity becomes smaller. This seems to indicate that the response of a DML panel can indeed be divided into a deterministic initial response that is similar for different panels, followed by a more or less diffuse part. As mentioned before, the existence of a deterministic initial response is essential for application in WFS.

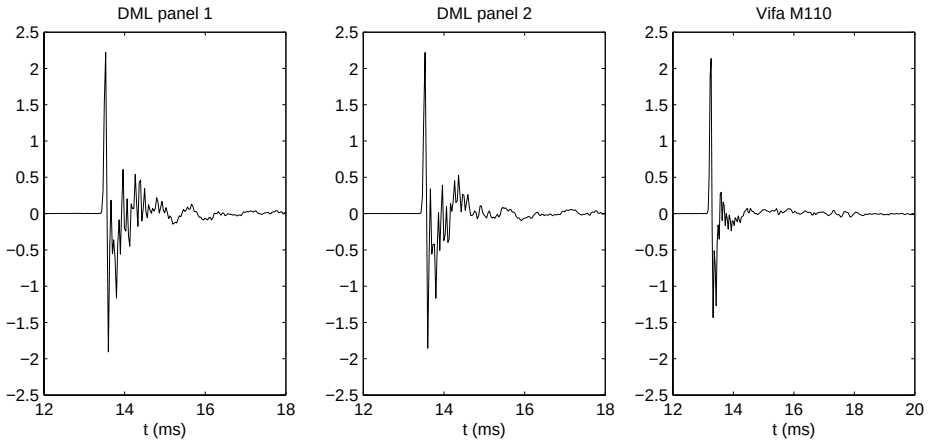


FIGURE 6.4. On-axis impulse response of two individual DML panels and the reference loudspeaker.

Figure 6.5 shows the frequency response of the same two panels and the reference loudspeaker. The left graph shows the on-axis responses of DML panel 1 (top curve), DML panel 2 (center curve) and the reference loudspeaker (bottom curve). For visualization purposes, a 25 dB offset was introduced between the three responses. It is seen that the frequency response of both DML panels is quite irregular, with local variations that are much stronger than in the response of the reference loudspeaker. These variations are mainly due to the small size of the panels, which, as explained in section 6.1.1, results in a relatively low density of eigenmodes. Also, although the responses of the two DML panels are similar, there are also clear differences. These are probably due to the fact that the positions of the exciters were not identical on both panels. As was explained in the discussion of (eq. 6.10), the values of the weighting factors of the individual eigenmodes depend on the position of the exciter, so that different exciter

positions result in different relative contributions of each eigenmode to the total response.
The right graph of figure 6.5 shows the responses measured at an angle of 45 degrees. Several observations can be made. First of all, the frequency responses of both DML panels are significantly different from their on-axis responses. This is a result of the complicated pattern of bending waves of the panel, as explained in section 6.1.2. What is also clear is that as predicted by the radiation theory of section 6.1.2, the DML high-frequency response at an angle of 45 degrees has not dropped significantly compared to the on-axis response, whereas the effect of beamforming at higher frequencies is very clear in the 45 degree response of the reference loudspeaker.

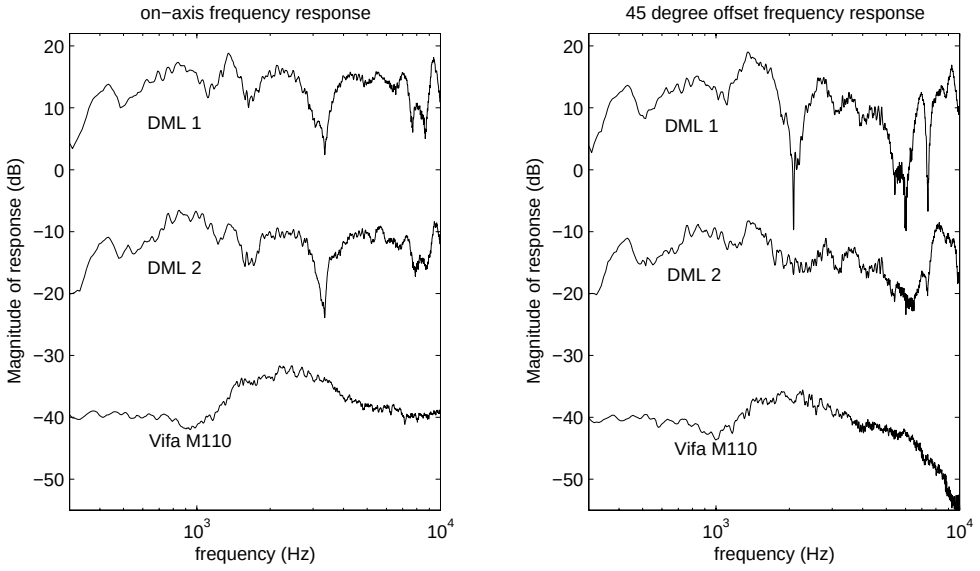


FIGURE 6.5. Frequency responses on-axis (left plot) and at 45 degrees (right plot) for two individual DML panels (top and middle curve) and the reference loudspeaker (bottom curve). In both plots, the three individual curves have been shifted 25 dB relative to each other for visualization purposes.

Figure 6.6 and figure 6.7 show the directivity patterns of the two DML panels (plotted in the same figure as a solid line and a dashed line) and the reference loudspeaker, respectively, at six frequencies. All responses were normalized such that for each curve the maximum response corresponds to 50 dB. Again, several observations can be made. Comparing figure 6.6 and figure 6.7, we notice that the response as a function of angle is more irregular and less symmetrical around 0 degrees for the DML panels than for the conventional loudspeaker. Again, this can be attributed to the complicated bending wave vibrations of the DML panels.

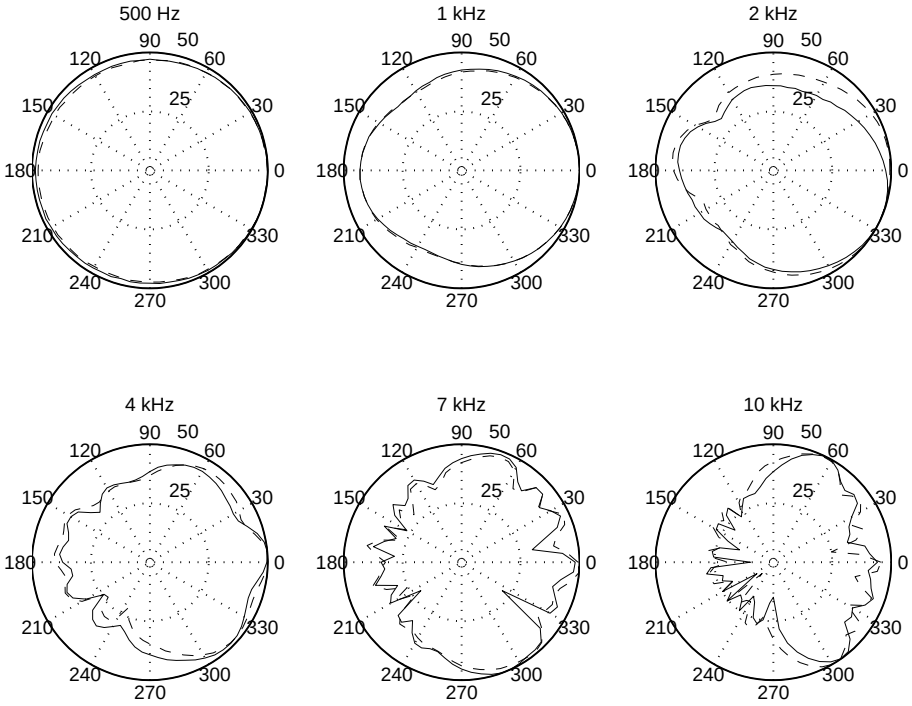


FIGURE 6.6. Directivity pattern for two different DML panels (solid line and dashed line) at six frequencies. All responses are normalized to a maximum response of 50 dB.

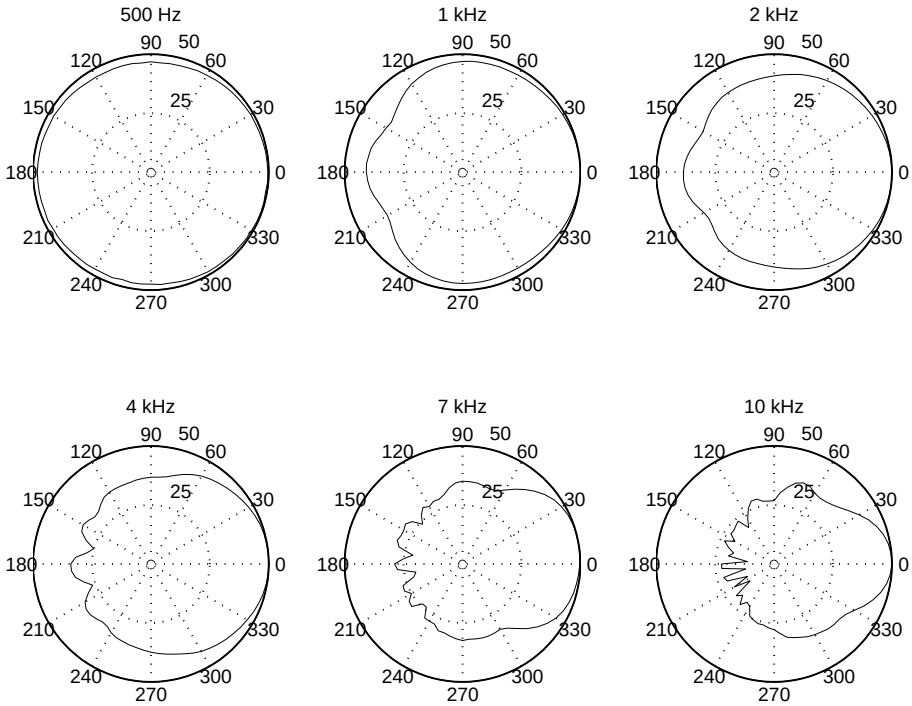


FIGURE 6.7. Directivity pattern for Vifa M110 loudspeaker. at six frequencies. All responses are normalized to a maximum response of 50 dB.

Furthermore, we see that while both types of loudspeakers are more or less omnidirectional at low frequencies, their behavior at higher frequencies is very dissimilar. In figure 6.7, we clearly see that as the frequency increases, the beamforming effect becomes more and more prominent for the conventional loudspeaker. The radiation of the DML panel, on the other hand, remains quite broad even for very high frequencies, in accordance with the theory of section 6.1.2. At 10 kHz, we also see the occurrence of 'coincidence peaks' in the DML directivity pattern, as predicted by the radiation theory in section 6.1.2, at approximately +60 and -60 degrees. It can be checked that from this it follows that the effect does not occur for frequencies below 7.5 kHz for these panels.

Finally, it can be stated that, although there are differences at higher frequencies, the directional responses of both DML panels are quite similar, even though their frequency responses differ significantly due to non-identical exciter positioning on both panels.

6.2.2 Measurements On A Multi-Panel Array

From nine of the individual DML panels, a short WFS array was constructed by attaching them in a frame with the panels placed next to each other in landscape orientation, so that the spacing was 22 cm and the total length of the array was 2.0 m. The array was placed in the anechoic room and the individual panels were connected to individual outputs of a DSP system, which was able to generate WFS driving signals according to (eq. 2.22) for any virtual source position, including sources in front of the array[1].

For several virtual source positions, the impulse response of the array was measured along a line parallel to the array at a distance of 3 m, with a step-size of 5 cm. This line was also the reference line for the WFS reproduction. A taper was applied to the gain of the outer two panels of both sides of the array to reduce diffraction artifacts. As a reference, for the same virtual source positions also the response was measured of a WFS array consisting of 16 conventional loudspeakers with a spacing of 12.7 cm, giving an equal total length of 2.0 m. Since the spacing was larger for the DML array, more spatial aliasing artifacts were expected for the DML array.

Figure 6.8 shows the measured response of the DML array (top-left graph) and the conven-

1. For synthesis of these so-called *focused sources*, a modification of the driving signals (eq. 2.22) is required. For details the reader is referred to [Verh97].

tional array (top-right) for a virtual point source at position (0,-1) m (center, 1 m behind the array). Clearly, the initial wave fronts synthesized by both arrays are almost identical, indicating that the DML array is able to synthesize a correct first wave front. What is also clear is that in the case of the DML array, the wave front is followed by stronger aliasing artifacts than in the case of the conventional array. This was expected, since the array spacing was larger for the DML array. Also, it can be observed that for the DML array the response after the initial wave front, excluding the aliasing artifacts, is more 'noisy' than for the conventional array. This is caused by the diffuse tail of the impulse response of the individual panels (see figure 6.4).
The center graphs of figure 6.8 show the measurements for a plane wave travelling in a direction perpendicular to the array. Also in this case, the DML array is able to synthesize a correct wave front, with the same remarks as above for the virtual source 1 m behind the array.
Finally, in the bottom graphs of figure 6.8 the responses are shown for a focused source 1 m in front of the array. The synthesis of focused sources is a critical test to check the phase coherence of the individual secondary sources, since only when the sources are sufficiently coherent, addition of their individual sound waves will result in a very narrow focus point. The measurements were carried out on a line through the focus point, parallel to the array, to determine how effective the focusing was and therefore how coherent the responses of the individual panels were. As can be seen, also in this case the DML array performs well. The width of the focus point is a bit larger than in the case of the conventional array, but this was to be expected, again because of the larger array spacing.
Measurement results for additional source positions can be found in the M.Sc. thesis of W. van Rooijen ([Rooij01]). These measurements confirm the findings for the source positions discussed above, so they are not included here.

In addition to the measurements, also some informal listening tests were carried out to evaluate the quality of the virtual sources synthesized by the 9-panel DML array. These confirmed that the DML array indeed was able to generate well-localized, stable, virtual sources, including focused sources. The sound quality, however, was quite poor, due to the lack of low-frequencies and the strongly irregular frequency response (figure 6.5) of the panels.

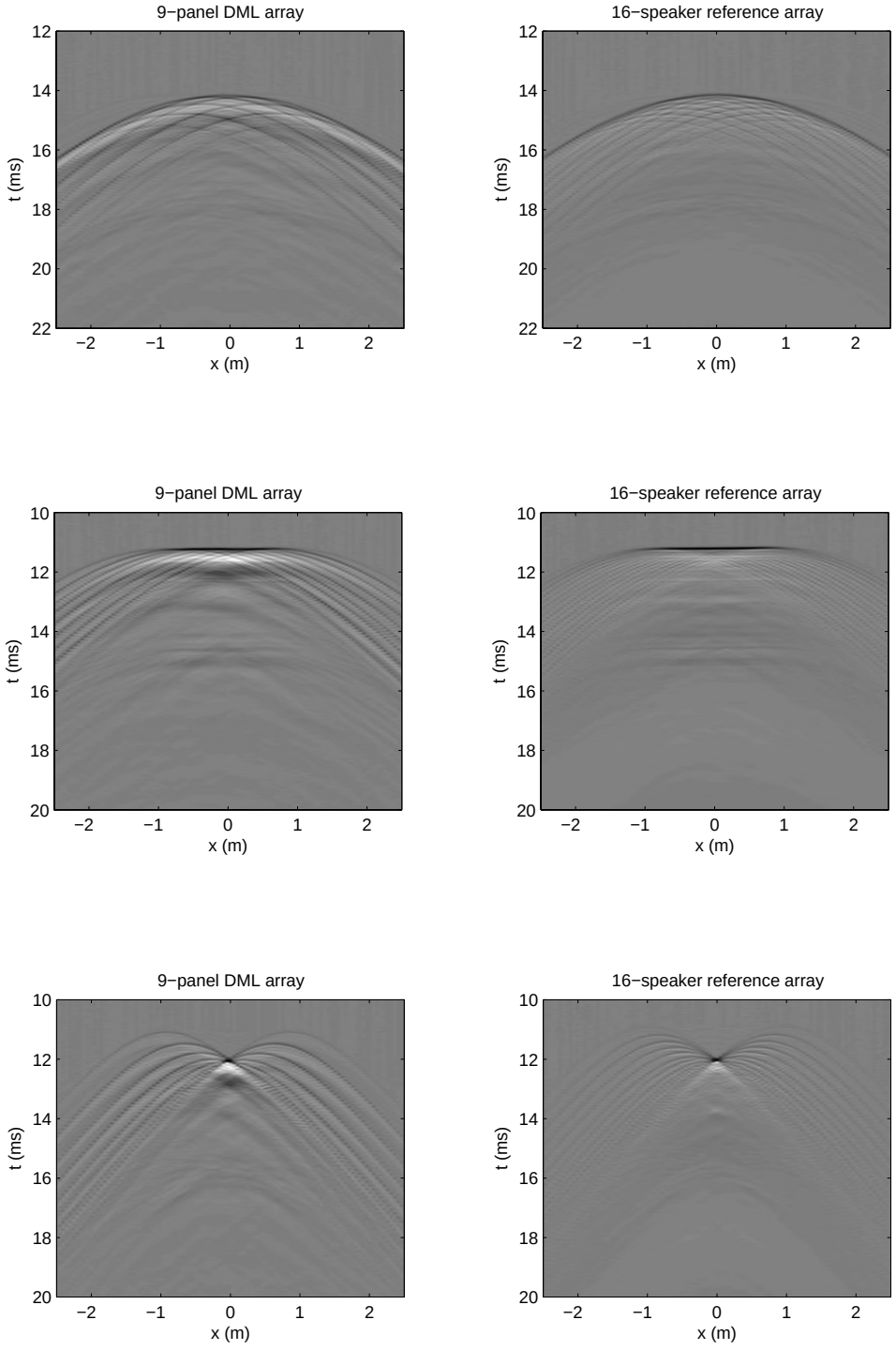


FIGURE 6.8. **Measured responses of the 9-panel DML array (left) and 16-speaker reference array (right) for a virtual point source 1 m behind the array (top), a plane wave traveling perpendicular to the array (center) and a focused source 1 m in front of the array (bottom). The responses for the source in front of the array and the plane wave were measured at a line parallel to the array at 3 m distance, the responses for the focused source were measured at 1 m distance (through the focus point).**

Horizontal axis: offset from the center of the array (m). Vertical axis: time (ms).

In conclusion, the measurements presented in this section have shown that it is possible in principle to construct a WFS array from small individual DML panels. Such an array is able to synthesize sound fields of virtual sound sources with similar spatial quality as an array of conventional loudspeakers. This indicates that the initial peaks of the impulse responses of the panels are indeed deterministic and sufficiently identical for different panels. The diffuse tail of the impulse response does not influence the spatial quality of the reproduced sound field significantly, but merely results in a higher loudness, due to the temporal integration mechanism of the human hearing system in the evaluation of loudness.
However, the sound quality of the small panels used in the array was insufficient for high-quality WFS reproduction, especially regarding reproduction of low frequencies and the non-flatness of the frequency response. As explained in section 6.1.1, these properties are, at least in part, inherent to the use of small panels, because of their eigenmode characteristics.

6.3 Experiments On Multi-Actuator Panels (MAP's)

Generally speaking, the sound quality of a large DML panel is better than that of a small panel. As explained in section 6.1.1, the lowest eigenfrequency of a large panel will be lower than for a smaller panel, so that the frequency response extends to lower frequencies and also the number of eigenmodes per unit frequency will be higher, resulting in a smoother frequency response. For WFS, however, a small spacing between individual secondary sources is required to avoid spatial aliasing artifacts, so building a WFS array from individual large panels in order to solve the problem of insufficient sound quality that was found for the array of small panels discussed in the previous subsection, is not an option.
It would be very interesting if it would be possible to construct a WFS array by attaching multiple, individually driven exciters to a single large panel, which effectively acts as an array of individual secondary sources with a spacing equal to the spacing between the exciters. This way, it would be possible to perform high-quality WFS with the benefit of enhanced sound quality of using a large panel. Additional advantages would be the easier construction compared to constructing an array out of individual panels and, for the specific application of videoconferencing in which we want to project a video image on the panels, the absence of seams between individual panels. This is the concept of a *Multi-Actuator Panel* (MAP) array.
For the idea of a MAP array to be feasible in practice, it is necessary that each exciter, effectively, only influences the movement of the panel in the area directly surrounding it. This is

necessary first of all because, as explained above, each exciter should effectively act as an individual secondary source, separated from its neighboring secondary sources by a distance equal to the distance between the exciters. Furthermore, it should be avoided that the movement of an exciter is influenced by the movements of the other exciters, because it is essential for proper WFS reproduction that it is possible to control the movements of the individual exciters independently from each other.

Fortunately, as we have seen in section 6.1, the normal velocity profile of a point-excited large plate is indeed such that the amplitude of the normal velocity is maximum at the point of excitation and then decreases very rapidly with increasing distance ((eq. 6.9), figure 6.2). As was explained, the spatial extent of the 'high velocity' area depends on several variables, most notably the amount of internal damping in the plate material and the nominal value of k_B of the plate. Therefore, in order to limit the 'active area' of the individual exciters, different strategies are possible:

Panel material with high internal damping. One way to limit the active area of the individual exciters is to choose a panel material with a high internal damping factor η. The disadvantage is that high internal damping results in a high loss of vibrational energy and thus in low sound radiation efficiency.

High nominal value of k_B. From figure 6.2 it can be seen that another way to limit the active area is to choose a panel material that has a high nominal value of the bending wave length k_B. Looking at (eq. 6.4) it is seen that this can be achieved by choosing the panel material and thickness such that the bending stiffness B' is low. From (eq. 6.8) it is seen that this also results in a low mechanical point impedance, so that the normal velocity at the excitation point is also high, which is advantageous for the efficiency. From (eq. 6.2) it is seen that a low bending stiffness can be achieved by choosing a panel material with a low Young's modulus E. However, a low value of E typically is associated with materials which also have a relatively low mass density ρ, which counteracts the influence of a reduced value of B' on the value of k_B ((eq. 6.4)). As can be seen from (eq. 6.2), another way to reduce B' and thus increase k_B, without the complicated influence of the material properties, is to decrease the thickness h. However, there is a lower limit to the value of the bending stiffness B' that can be used in practice. Since we are not dealing with a theoretical point-excited plate but with an exciter with

a relatively large mass that is attached to the panel, if B' becomes too small and the panel becomes too 'flexible', the mass of the exciter will significantly influence the movement of the panel. Especially at low frequencies, there is also the risk of excessive panel excursions when B' is too small, resulting in distortion. Furthermore, it follows from (eq. 6.5) that increasing k_B results in reduction of propagation speed c. This means that the diffuse tail of the panel impulse response will become longer, which, if it becomes too long, is also undesirable (think of a metal reverb plate).

From the above it should be clear that selecting the 'ideal' panel material is not trivial, because of the fact that many material properties are involved that often can not be chosen independently and sometimes have opposite effects on the MAP performance (E and ρ) or negative side effects (η: loss of efficiency). Therefore, a variety of different materials was tried. The next subsections discuss results of two different prototypes.

6.3.1 5-Exciter Polycarbonate MAP

A first prototype was built from a plate of polycarbonate material ('Lexan'), which has a high amount of internal damping, with thickness 0.75 mm and a size of 0.70 x 0.70 m, to which five exciters were attached with a spacing of 12.7 cm. A shallow enclosure was built around the panel. From preliminary tests it was clear that the sound quality and the efficiency of the panel were very poor, but this prototype was only built to test if it is possible, in principle, to perform WFS with an array built according to the MAP concept.

Figure 6.9 shows an example of the measurement results obtained with this panel. From this measurement it is clear that a correct first wave front is synthesized and thus that the individual exciters indeed act as individual secondary sources, which is also clear from the individual aliasing traces that can be seen. From this measurement it can be concluded that it is indeed possible to build a WFS array according to the MAP concept.

More measurements on the 5-exciter polycarbonate MAP can be found in [Rooij01]. We will not go into more detail here.

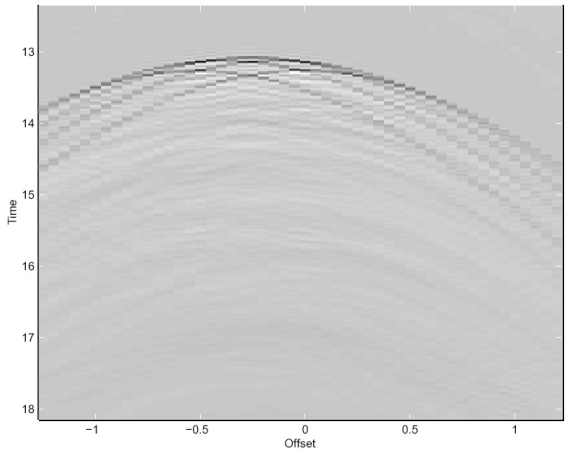


FIGURE 6.9. Measured responses of the 5-exciter polycarbonate MAP for a virtual point source behind the array, measured at a line parallel to the array.

6.3.2 Foamboard MAP Array

After it had been established by the experiments with the 5-exciter polycarbonate MAP described in the previous subsection that it is indeed possible to construct a properly working WFS array based on the MAP concept, a search was started for material that could result in higher efficiency and better sound quality. A promising candidate material that was found was so-called ‘foamboard’, a sandwich-type of material, consisting of PVC foam with a thickness of 5 mm, laminated with cardboard paper on both sides.

To check if the material was suitable to construct a MAP array from it, a single exciter was attached to a panel of size 100 x 70 x 0.5 cm and the response was measured along a line parallel to the panel at a close distance. Using inverse wave field extrapolation, it was determined from these measurements that the main part of the radiated acoustic energy originated from a region around the exciter with a diameter of about 10-15 cm ([Rooij01]). This means that if the exciters are separated by this distance, they will effectively act as individual sources. This indicates that the panel is indeed suitable to use as a MAP for WFS reproduction.

It was then decided to build a prototype MAP system using these foamboard panels. It consisted of three panels of the dimensions mentioned above, each one built into a separate enclosure. A total of 20 exciters (diameter: 6 cm) was available. These were distributed over the panels as shown in figure 6.10. The spacing of the exciters was 12.7 cm and they were attached

slightly below the middle of the panel, to avoid excitation at a point where many eigenfunctions of the panel are zero (see section 6.1.1).

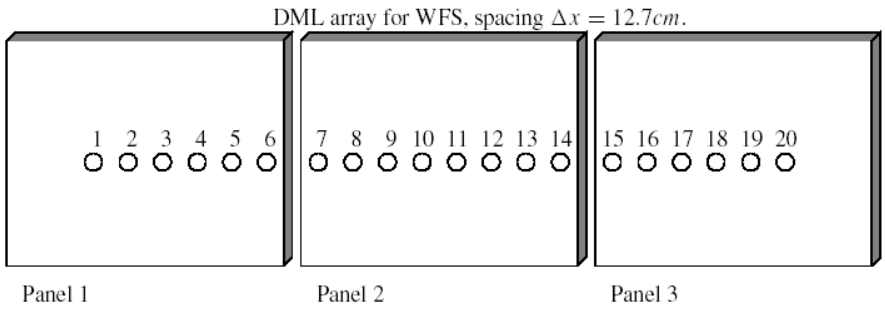


FIGURE 6.10. Distribution of the 20 exciters over the three foamboard MAP's. The dimensions of each panel are 100 x 70 x 0.5 cm. The exciter spacing is 12.7 cm. (from [Rooij01]).

Measurements on Individual Exciters and Filtering. Before evaluating the WFS performance of the panels, the response of the panel to excitation by individual exciters was measured, to determine both the overall response of the panel and the variation of the response for excitation by different exciters. This was done by positioning a measurement microphone close to the panel surface in front of the exciter to be measured.

Figure 6.11 shows the measured frequency responses for exciters 1, 7 and 10, so they represent responses for different panels and different exciter positions on the panels (figure 6.10). Looking at the overall responses, it is clear that the frequency response of the panel is far from flat[2]. Also, although the overall response is similar for all three exciters, there are clear differences between them, as could be expected. Listening to the panels, the non-flat frequency response was noticed as an obvious coloration of the reproduced sound. Therefore, it was decided to try to correct the frequency response by digital filtering. In order to keep the required processing as limited as possible, it was decided to first try to improve the frequency response by a single FIR filter at the input of the system, instead of applying an individual filter to each individual exciter. It was realized that this would not give a perfect correction of the frequency response, given the differences between the responses of the individual exciters and the dependence of the frequency response on angle (see section 6.1.2), but nevertheless it was considered to be interesting to see if a significant improvement could be made by using only a single filter.

2. The strong peak at 4 kHz might be a resonance from the enclosure, although it was filled with damping material.

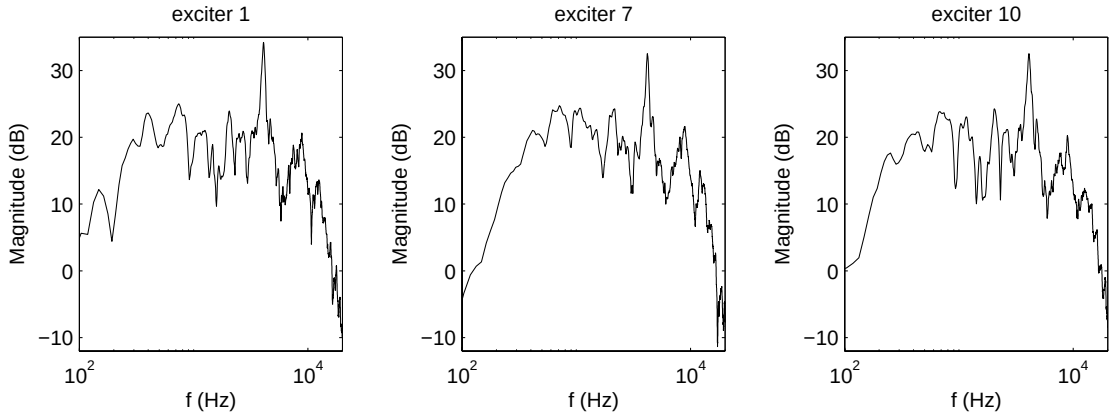


FIGURE 6.11. Frequency responses of the foamboard panels for excitation by three individual exciters. Responses were measured close to the panel, right in front of the active exciter. The frequency axis is logarithmic.

In figure 6.12, a detailed look is given at the impulse responses of exciters 1, 7 and 10 for the early part (top graphs) and the later part (bottom graphs) separately. It is seen that the first 2 ms of the impulse response is very similar for all three exciters. The later parts, corresponding to the diffuse tail of the panel response, are much less similar. Therefore, the FIR filter design should be based on the first part of the impulse response only. Figure 6.13 shows the result of filtering the three exciters by an FIR filter based on the first 2 ms of the response of exciter 1. Details of the filter design can be found in [Rooij01]. It is seen that the responses have improved considerably, although, as expected, the result is not perfect. In order to achieve an optimal compensation of the frequency response in the whole listening area, it is necessary to use advanced multi-channel filtering techniques. Because of the angle dependency of the frequency responses, it is not sufficient to design the filters of the individual exciters by simply inverting the individual responses measured in the way described above. Instead, it is necessary to measure the response of the panel at many points in the room for each individual exciter and to calculate all filters simultaneously in a multi-channel optimization procedure. In this way, in principle also the influence of the room on the reproduced sound field can be compensated. Because of the exploratory nature of the experiments described here, this was not done in this case, but research on this subject has been carried out by Corteel et al. ([Cort02]) in the context of the EC project CARROUSO ([CARROU]), in which the study of the application of MAP arrays in WFS was continued, based on the first results presented here.

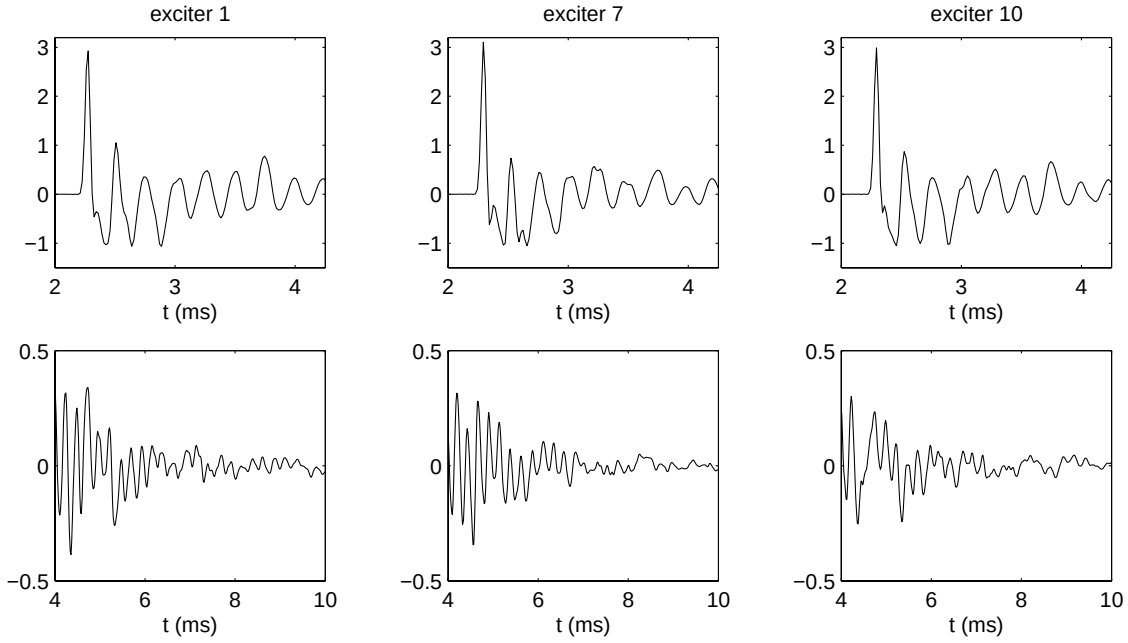


FIGURE 6.12. Impulse responses for exciters 1 (left), 7 (center) and 10 (right), separated into the first 2 ms (top) and the later part (bottom).

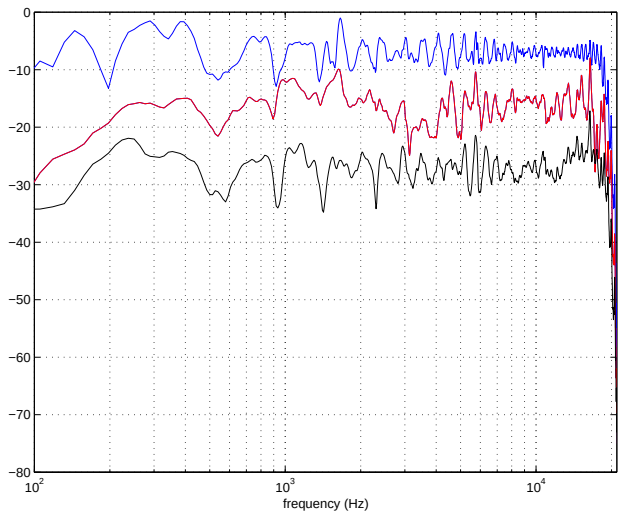


FIGURE 6.13. Frequency responses of exciters 1 (top), 7 (center) and 10 (bottom) after filtering with an FIR filter, based on the first 2 ms of the impulse response of exciter 1. An offset of 10 dB was introduced between the individual responses for visualization purposes (from: [Rooij01]).

Measurements of WFS performance. The sound field synthesized by the 3-panel foamboard MAP array was measured in the anechoic room for several virtual source positions. Measurements were carried out on a line parallel to the array at a distance of 3 m, with a step-size of 5 cm.

Figure 6.14 shows the measured responses for four source configurations: a virtual point source behind the array at (0,-1) m (top-left), a virtual point source behind the array at (-0.75,-3) m (top-right), a plane wave (bottom-left) and a focused source 1 m in front of the array (bottom-right). The measurement of the focused source was carried out on a line parallel to the array at 1 m distance, through the focus point. It is clear from all four measurements that the synthesis of the wave fronts is very good and comparable to the wave field synthesized by an array of conventional loudspeakers with the same spacing, as can be checked by comparing the top-left, bottom-left and bottom-right graphs to the corresponding right-hand graphs of figure 6.8. Especially the very narrow focus point that is seen in the measurement of the focused source proves that the individual exciters of the MAP array essentially behave as individual, phase coherent, secondary sources.

From the experiments presented in this section it can be concluded that it is indeed possible to construct a properly working WFS array by attaching multiple exciters to a large panel of a suitable material. For this so-called ‘Multi-Actuator Panel’ (MAP) concept, including the option of multi-channel digital filtering, an international patent application has been filed by Delft University of Technology together with Studer Professional Audio Equipment ([Boon01]).

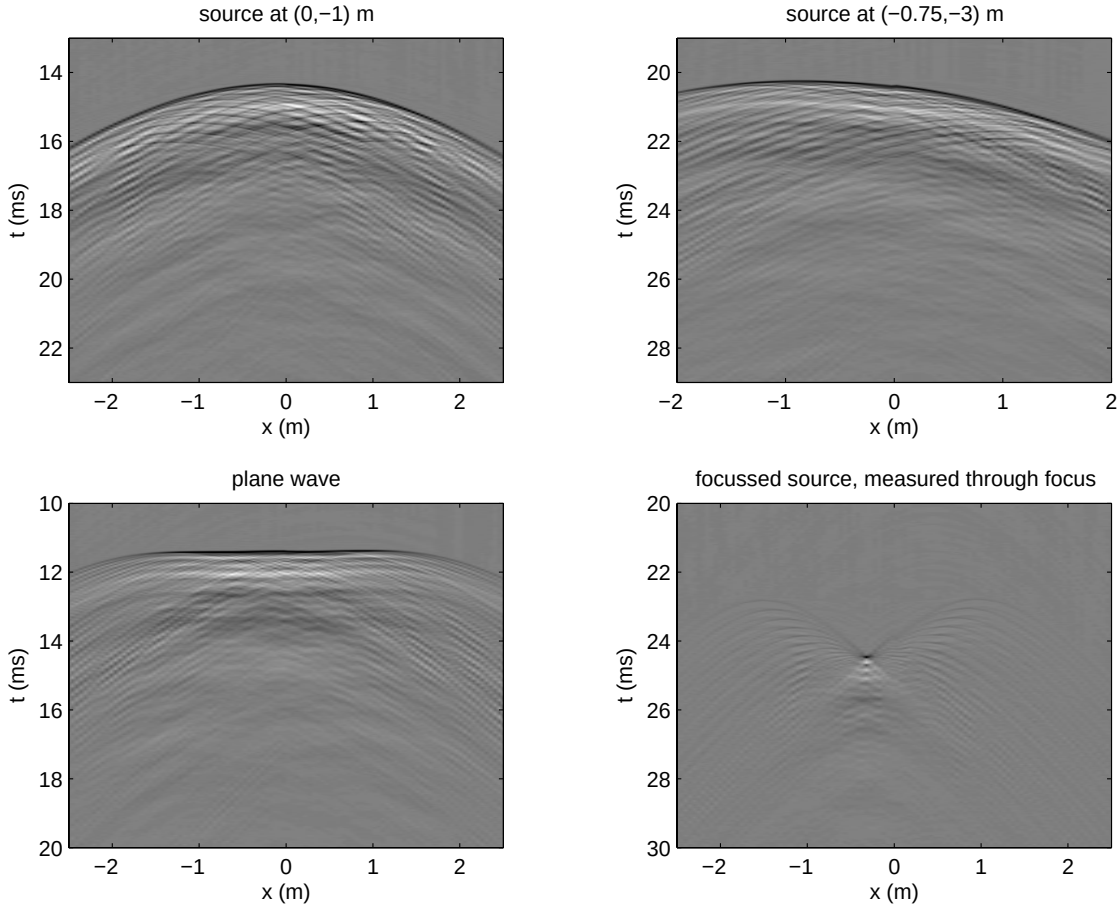


FIGURE 6.14. Measured sound field synthesized by the three-panel foamboard MAP array for a virtual point source at (0,-1) m (top-left), a virtual point source at (-0.75,-3) m (top-right), a plane wave (bottom-left) and a focused source 1 m in front of the array (bottom-right). Measurements were carried out on a line parallel to the array at 3 m distance, except the measurement for the focused source, which was carried out on a line parallel to the array at 1 m distance (through the focus point).

6.4 Conclusions

By the experiments that have been carried out it has been established that DML technology is suitable for WFS reproduction.

The experiments on the Multi-Panel array have shown that, despite the different working mechanism of DML loudspeakers compared to conventional loudspeakers, an array built up from individual small DML panels behaves in a way very similar to an array of conventional loudspeakers, as far as the spatial properties of the WFS-reproduced sound field are concerned. Virtual sources can be generated very effectively, as confirmed by measurements and informal listening tests and even focused sources, for which the phase relation between individual sec-

ondary sources is especially critical, can be generated. A disadvantage of using individual small panels is the frequency response, which is far from flat and lacks low frequencies.

The subsequent experiments on the Multi-Actuator Panel (MAP) arrays show that, additionally, it is possible to construct a properly working WFS array by attaching multiple exciters to a single panel of suitable material. In this case, the material properties become more critical, as on the one hand the performance of an individual exciter should be fairly uninfluenced by the other exciters, implying the need for a rather high internal damping in the material, while on the other hand the desire for high efficiency requires the opposite.

A clear disadvantage of all prototypes was the non-flat frequency response, causing the panels to have a rather low sound quality. This can be improved by means of digital filtering, but it was found that in the case of the MAP's, filtering all the individual exciter signals with a common FIR filter does not give the desired result, so more sophisticated filtering techniques are required.

It will be clear that the investigations presented in this chapter represent only the first steps in the development of MAP arrays for WFS and there is still much room for improvement regarding the panel material, the exciters, the panel enclosures and the digital filtering of the exciter signals. After the finishing of the investigations presented here, the concept has been adopted by the EC project CARROUSO ([CARROU]) in which the technology has been improved considerably, especially regarding the filtering ([Cort02]). Successful demonstrations of WFS systems using MAP's have been made at various major audio conferences by the CARROUSO consortium.

In conclusion, the MAP concept seems to be a promising way to make widespread implementation of WFS reproduction more feasible. In the particular case of audio-visual systems such as videoconferencing systems, the MAP concept enables the merging of the loudspeaker array with the projection screen.

CHAPTER 7 *Synthesis of Results*

At the end of this thesis, we collect the main results of the research that was carried out and relate them to the main objectives that were stated in section 1.4, which are repeated here for convenience.

The first objective of this thesis was:

> *To investigate whether application of Wave Field Synthesis results in an audio reproduction system for videoconferencing that enables a more effective virtual meeting than can be achieved when conventional audio reproduction systems are used, with emphasis on acoustical telepresence, speech intelligibility and speaker identification and within the restriction of using two-dimensional video projection for the visual part of the system.*

The questions that are asked in this first main objective were investigated in the perception experiments presented in chapter 4. It was found that combining audio reproduction that includes an accurate reproduction of distance, such as WFS, with conventional two-dimensional video projection introduces some problems for observers that are laterally displaced relative to the viewpoint of the video projection, because of a mismatch between the perceived directions of auditory sources and the corresponding video images on the screen. This mis-

match is caused by the fact that, whereas the auditory perspective is correct for all observers in the reproduction area, the visual perspective is only correct at the unique viewpoint of the video set-up. From the experiments in section 4.2 and section 4.3, it was found that in a typical life-size videoconferencing set-up, such discrepancies are likely to be noticeable and even annoying and were found to degrade visual-to-auditory source-matching performance and the overall perceived ‘naturalness’ of the reproduced audio-visual scene, compared to a conventional ‘discrete loudspeaker’ set-up, even for observer positions with only a modest lateral displacement from the viewpoint. The influence of watching the audio-visual scene from a too small or too large distance from the screen appeared to be much less of an issue.

This would seem to suggest that combining WFS with two-dimensional video projection is not beneficial for the overall performance of the system in terms of achieving ‘acoustical telepresence’, which was the main goal of the TU Delft/France Telecom project that the research presented in this thesis was a part of. However, it was found in the SRT test of section 4.5 that in specific situations, reproducing the voices of conference participants by WFS results in a significant increase of speech intelligibility, compared to conventional reproduction. Since high speech intelligibility is one of the most important characteristics of a good speech communication system such as a videoconferencing system, this is an argument in favour of using WFS.

Furthermore, it was shown in section 4.4 that it is possible to avoid or reduce the problems introduced by the mismatch between the auditory and visual perspective, by reducing the perspective of the reproduced auditory scene to some extent, by pulling all auditory sources somewhat closer to the screen in the direction of the visual viewpoint. By applying a proper amount of this so-called ‘auditory depth compression’, it should be possible to find a compromise between avoidance of mismatch problems and increased speech intelligibility.

It should be noted that the conclusions given above are not only relevant to the specific combination of WFS and two-dimensional video projection, but are of relevance to all audio-visual systems that combine spatialized audio reproduction that includes depth with two-dimensional video, for instance, systems that combine binaural (HRTF-based) sound reproduction with 2D video. Furthermore, the general conclusions are to a large extent even applicable also to audio-visual systems that provide three-dimensional video images with only a single viewpoint, since, just as is the case for a two-dimensional video projection system, for these systems the visual perspective is also only correct for an observer at the unique correct viewpoint.

Due to practical limitations, no investigations were carried out on the combination of WFS

sound reproduction with three-dimensional video. It may be expected, however, that when WFS sound reproduction is combined with multi-viewpoint three-dimensional video, this will result in a significant increase of the perceived degree of ‘overall telepresence’ (section 1.1.2), compared to a similar system in which conventional sound reproduction is used.

The second main objective of this thesis was:

> *To optimize the Wave Field Synthesis reproduction system in the context of the videoconferencing application, such that all participants of the local room have a natural sound perception of all the voices of the remote participants with correct localization.*

The most important aspect of this objective was to investigate what is the minimum number of array loudspeakers that is needed in a life-size videoconferencing system, since this will be one of the main factors that determine the overall cost of the total system, not only in terms of production- and signal-processing costs, but also in terms of required transmission bandwidth. This issue was separated into two main questions:

- What spatial resolution is needed in the reproduction of participants’ voices with regard to sound source localization, in both the horizontal and vertical direction, taking into account the sensory interaction with the visual information that is received from the video projection?
- What maximum distance can be allowed between the individual loudspeakers of the reproduction array, such that no annoying coloration of the reproduced sound field is perceived by users of the videoconferencing system?

The first question was investigated also in chapter 4. Regarding horizontal source localization, results from literature and previous studies were available on both the source localization accuracy of a WFS system and the required accuracy of the match between the directions from which an auditory source and a corresponding visual source are reproduced. In previous studies on WFS it was found that source localization is stable for quite a large range of loudspeaker spacings, even though the sources appear to broaden somewhat when the spacing is increased to values for which the spatial Nyquist frequency becomes significantly lower than 1.5 kHz. Combined with the strong audio-visual interaction effect of ‘visual capture’ that can be expected to occur in a videoconferencing system, it was expected that coloration due to spatial

aliasing would be the limiting factor for the maximum horizontal loudspeaker spacing, rather than the source localization accuracy.

Regarding the required accuracy of vertical source positioning, it was found from the experiment described in section 4.1 that this is not critical at all. The low ability of humans to localize sounds in the median plane, combined with the visual capture effect that also works in the vertical direction, resulted in the conclusion that a vertical source positioning accuracy of about 22 degrees, measured at the observation position, is sufficient to ensure that no discrepancy between the vertical position of a voice signal and the visual image of the corresponding moving lips on the screen is perceived. If the minimum distance to which participants approach the screen is assumed to be 1.5 m, then this leads to the requirement that the system should enable positioning sound sources with a vertical resolution of about 60 cm. From this, it is concluded that in this application it is not necessary to apply WFS reproduction in the vertical direction, but that it is sufficient to use a single horizontal array bar at the average height of the conference participants' mouths, or possibly two horizontal array bars, for instance, one at the average height of a standing conference participant and the other at the average height of seated participants. In the latter case, each voice signal is reproduced by either the upper or the lower array, whichever one is closest to the height of the virtual source.

The second question, concerning coloration caused by spatial aliasing due to the distance between the array loudspeakers, was investigated in chapter 5. It was shown that colour differences between the original sound field and the sound field reproduced by a WFS array should be expected to occur for arrays with a practical loudspeaker spacing, but it was explained that the main concern should be the spatial variations of the colour of the reproduced sound field that a listener perceives when walking around in the room, rather than the colour difference between the reproduced sound field and the original sound field, because the original sound field is not available for comparison to the user and because systematic colour differences between the original and reproduced field can easily be compensated by means of digital filtering. Using a model of the perception of colour differences between individual signals, an analysis was made of the spatial colour variations in the simulated sound fields reproduced by WFS arrays with different loudspeaker spacings and a single-value measure, the 'Spatial Colour Variation Index' (or 'SCV-Index'), was proposed to quantify the strength of the spatial colour variations that occur in the sound field reproduced by a specific WFS configuration.

In order to find the maximum loudspeaker spacing that can be allowed in the videoconferenc-

ing application such that no strong spatial colour variations are perceived by users of the system, a perception experiment was carried out. In this paired-comparison experiment, subjects compared the spatial colour differences that were present in the simulated sound fields of WFS loudspeaker arrays with various loudspeaker spacings, for various source positions and for both male and female speech noise signals. The results of this experiment were shown to be consistent with the proposed model for spatial colour variations and a critical value of the SCV-Index could be found, above which a clear increase in spatial colour variations was perceived, while for configurations with values of the SCV-Index below this critical value no strong spatial colour variations are to be expected. For the typical WFS configuration for videoconferencing that was investigated in the experiment, the critical value corresponded to a loudspeaker spacing of about 25 cm. Given the arguably rather critical nature of the signal that was used in the experiment (continuous speech noise samples, as opposed to real speech), this value should be regarded as a lower bound, so the spacing in a practical system can probably be larger before spatial colour variations are actually perceived.

Finally, in chapter 6 the application in WFS of Multi-Actuator Panel (MAP) arrays was investigated. MAP's are an extension of the concept of so-called distributed mode loudspeaker (DML) panels, in which multiple individually driven exciters are attached to a single large panel of suitable material in such a way that the whole panel acts like a WFS array. Since in the specific application of videoconferencing the possibility of using MAP arrays would enable a complete integration of the WFS array and the video projection screen, this concept can be considered to be another aspect of the second main objective, the optimization of the loudspeaker array for videoconferencing. First experiments with an array constructed from small individual single-exciter DML panels showed that DML panels are in principle suitable for WFS reproduction, but that the frequency response of small panels is too poor to be used in critical applications. From a theoretical discussion of the properties of DML panels, it was expected that using larger panels should result in a better sound quality. This triggered the investigation of the feasibility of constructing arrays according to the MAP concept. It was shown that for the MAP concept to be feasible, the panel material should have specific properties, resulting from the requirement that each individual exciter should only excite the region of the panel directly surrounding it, so that effectively, each exciter acts like an individual secondary source, while at the same time the efficiency of the system, in terms of radiated acoustic power, should be sufficiently high. Experiments were carried out with prototypes of

different materials. From measurements of the sound field reproduced by an array of MAP's made of foamboard, it could be concluded that it is indeed possible to construct a properly working WFS array according to the MAP concept that is able to reproduce sound fields that, regarding spatial quality, are comparable to the sound fields reproduced by an array of conventional loudspeakers.

CHAPTER 8 *Conclusions*

In the ideal videoconferencing system for group-to-group communication, participants at all conference sites should be able to communicate with each other as if they are all in the same room together. If people are able to have a completely natural conversation and forget that they are using an audio-visual system, this will enhance the effectiveness of the human communication during the virtual meeting.

Conceptually, in the case of two conference sites this situation is established by virtually connecting the two rooms in such a way that they form a new, virtual room in which all participants are present, with the videowall at both sites forming the interface between the two rooms. An important aspect in achieving the illusion of telepresence and in preventing fatigue and loss of concentration during long meetings, is that the voices of the remote participants are reproduced in a spatially correct way, without a mismatch between the perceived locations of the voices and their corresponding visual images on the screen. Additionally, in this ideal system, the audio recording- and reproduction techniques that are used should enable the participants at both sites to be located anywhere in the room and to walk around in it, while all of them perceive a correct spatial image of the sound field of the other site at all times. This means that the system should be a *double-dynamic system*.

Such an accurate spatial reproduction of the remote sound field that is correct within an exten-

sive listening area can be achieved by means of Wave Field Synthesis. At the recording site, a spatial recording is made of the sound field, either by recording the voice of each individual person by an individual microphone close to that person or by using a microphone array. The recorded audio signals are transmitted to the remote site, where they are processed to generate the driving signals for an array of closely-spaced loudspeakers that is installed behind the video screen. This loudspeaker array generates a copy of the sound field of the recording site that is correct within the whole listening area, ensuring a correct localization of the voices for all participants at the reproduction site, including a correct sense of distance.

The dual-site concept described above can be extended to the case in which more than two rooms are virtually connected to form one large meeting room. At each site, the remote rooms can be visualized life-sized on individual screens that are mounted on the walls side-by-side. Regarding this visualization, care will have to be taken that the visual orientation of all individual rooms relative to each other is consistent for the participants at all sites, so that a participant at one site who is addressing a participant at one of the other sites will be visually perceived as such by the participants at all sites. This will be a challenge, especially as far as the development of the visual part of the system is concerned. Once this spatial orientation of the different rooms relative to each other has been established, the extension of the dual-site concept to this multi-site concept will not introduce any additional problems regarding the WFS audio reproduction, apart from transmission issues.

An interesting issue arises when, due to practical limitations imposed by the current technical state of visualization systems, WFS sound reproduction is combined with conventional two-dimensional video projection. Because the visual perspective of the video is only correct at a single position (the viewpoint), while the auditory perspective of the WFS sound field is correct within the whole listening room, a discrepancy is perceived between the directions of the voices and their corresponding visual images on the screen by observers located at positions other than this viewpoint. This discrepancy is perceived as being annoying for even modest lateral displacements from the viewpoint and degrades the performance of the system in terms of speaker identification and naturalness. By reducing the perspective of the reproduced sound field to some extent, this problem can be avoided or reduced. It should be noted, however, that the problem actually arises from the fact that in the case of our system, the spatial audio reproduction by WFS is more realistic than the visual reproduction offered by the 2D visualization system. It is believed that the present study of this phenomenon is relevant to all audio-visual

systems in which the perspectives of the audio- and video reproduction are not identical and that the issue should therefore be taken into account in the design process of such systems.
Besides the aspect of ‘naturalness’, a very important property of a videoconferencing system is its performance in terms of speech intelligibility. If the system performs badly in this respect, this will hinder the communication between the local and remote participants and will eventually result in fatigue and loss of concentration. The fact that with WFS the individual voices can be reproduced from their correct spatial locations is very beneficial for the speech intelligibility. It was found that, compared to a system that provides no spatialized sound reproduction, the speech intelligibility is improved significantly when WFS is used.
One of the factors that is still limiting the widespread application of WFS is the fact that many loudspeakers are required. A possible solution has been found in an extension of the concept of so-called Distributed Mode Loudspeaker (DML) panels to the new concept of the Multi-Actuator Panel (MAP). A MAP consists of a large piece of light, stiff material, to which multiple individually-driven electro-dynamic transducers are attached. It has been shown that it is possible to construct a MAP in such a way that each exciter effectively acts as an individual source, so that MAP’s can very well be used for WFS reproduction. In the application of videoconferencing, an additional advantage is that the panels can be used as a projection screen as well, so that the loudspeaker array and the video screen can be integrated.

APPENDIX A *Subject Instructions and User Interfaces*

This appendix shows the actual written instructions for all the perception experiments described in this thesis that were given to subjects before they carried out each experiment.
Also included are screenshots of the user interfaces that were used in the experiments.
In all experiments (except the vertical localization experiment, in which only Dutch native speakers participated) both English and Dutch versions of the instructions were available. Dutch native speakers always received the Dutch instructions, Non-Dutch natives received English instructions.

A.1 Vertical Localization

These were the Dutch instructions that were given to the subjects before they participated in the experiment on vertical localization of section 4.1:

Dit experiment is bedoeld om de requirements voor de vertikale lokalisatie van een videoconferencing systeem, gebaseerd op golfveldsynthese, te onderzoeken.
Je krijgt zometeen 3 sequences van 20 korte (4 sec.) spraakfragmenten te horen die gereproduceerd worden van verschillende vertikale posities. De mogelijke posities zijn aangegeven op het array met de nummers 1 tm 13. Het is de bedoeling dat je na ieder fragment het nummer

intypt waar je het vandaan hoorde komen.
Let op: De vraag die gesteld wordt is dus nadrukkelijk: "Waar HOOR je het vandaan komen?", in tegenstelling tot: "Waar VERMOED je dat het vandaan komt.". Belangrijk is verder op te merken dat de posities volkomen random gekozen worden en dat per sequence bepaald wordt welke posities wel en welke niet aan bod komen. Het is dus niet zo dat elke positie even vaak aan bod moet komen in het experiment.
Elk van de 3 sequences wordt voorafgegaan door een testsequence, waarbij de geluidsbron achtereenvolgens van tien willekeurig gekozen posities komt. Je wordt hierbij al gevraagd een nummer in te voeren, maar dit wordt niet meegenomen in de resultaten. Dit is enkel bedoeld om te wennen aan de stimuli, je krijgt geen feedback over de werkelijke posities.
Het is de bedoeling dat dit experiment gedaan wordt op twee luisterposities. Totaal duurt dit ongeveer 20 minuten.
Vervolgens wordt het experiment herhaald, maar nu in combinatie met het beeld van de spreker, geprojecteerd op het scherm. Hierbij is het de bedoeling je aandacht zoveel mogelijk op het visuele beeld van de spreker gericht te houden. Deze 2e run kan eventueel ook op een ander tijdstip of na een korte pauze plaatsvinden.

Concentratie, WE GAAN BEGINNEN!

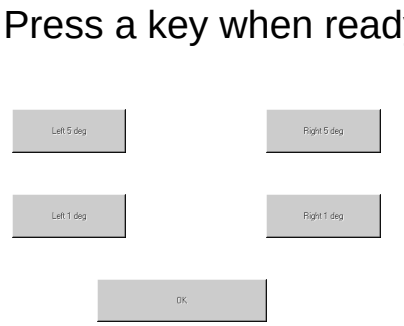
A.2 Correspondence Of Perceived Source Positions In Auditory And Visual Modalities: Single Source Experiments

Below are the instructions that were given to subjects before they took part in the ‘single source’ experiments (‘Experiment A’ and ‘Experiment B’) described in section 4.2. These experiments were carried out in one session, so the instructions were combined.

EXPERIMENT DESCRIPTION

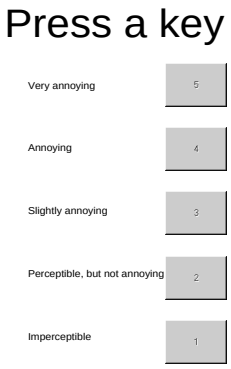
In this experiment you are going to match auditory space to visual space.
You will be seated at a specific position relative to the projection screen. On the screen an image is projected of a room with three people standing at different locations in the room. The voice of one of the persons is reproduced by the system, initially from a random position.
It is your task to move the auditory source so that its position appears to correspond naturally to the position of the person on the screen. There is no right or wrong answer, the task is simply

to configure the audio-visual scene such that it appears (as) natural (as possible) to you.
The way this works is as follows. On the computer screen you see a picture as shown below:



After you have pressed a button, one of the 3 visual source positions is chosen (the number will be indicated on the screen and they correspond to the persons on the screen from left to right) and the sound is reproduced from a random position.
You can now move the source left and right by pressing the corresponding buttons. There are 'coarse' buttons that move the source 5 degrees and 'fine' buttons that move the source 1 degree. You can move the source left and right as much as you like. When you feel that the source is at the most natural position you press 'OK' and a new visual source is chosen. You will match each visual source 5 times.

When you have completed this part of the experiment a second experiment is started in which you will evaluate how well a specific auditory source position corresponds to a certain visual source position. The screen you will see looks like this:



After you have pressed a button, one of the three visual sources is chosen. The voice is reproduced from a certain position relative to this source. It is your task to indicate subjectively, by giving a grade, how you would qualify the discrepancy between the auditory and visual source positions or, in other words: how natural the reproduced audio-visual scene appears to you (in terms of spatial lay-out).
The grading is done on a 5-point scale, with descriptions as indicated in the screenshot above:

- *1 corresponds to 'no noticeable discrepancy', so 'a perfect match',*
- *5 corresponds to an obvious discrepancy that is so large that it makes it impossible for you to associate the reproduced sound with the person in the image.*

After you have pressed one of the buttons a new source is chosen. As in the first part, you will evaluate each of the three visual sources 5 times.
When you have completed this part of the experiment the whole experiment is repeated for a different evaluation position. This can be immediately after the previous session or after a break.

A.3 Correspondence Of Perceived Source Positions In Auditory And Visual Modalities: Multiple Source Experiments

Below are the instructions that were given to subjects before they took part in the ‘multiple source’ experiments (‘Experiment C’ and ‘Experiment D’) described in section 4.3. These experiments were carried out in one session, so the instructions were combined.

SOURCE IDENTIFICATION EXPERIMENT

Part 1:

In this part of the experiment you are going to try to identify a specific speaking person in a situation where multiple people are speaking at the same time, by means of the spatial position of the person's voice.
The way in which you are going to do this is fairly simple.On the projection screen you will see 3 people. After you have pressed a mouse button, the voices of 3 people are reproduced:

- *One male: english speaking, the speaker to be identified and*
- *Two female: dutch speaking, both identical but saying different things.*

Each of the 3 voices is reproduced from an individual position, where each voice, in principle, corresponds to an individual person on the screen. So one voice is connected to person 1, one voice to person 2 and one voice to person 3. This however, does not mean that during the experiment you will in all cases actually hear them coming from the intended positions or even from separate positions.
On your monitor you will see 3 buttons, labelled "1", "2" and "3" which correspond to the 3 persons on the screen, from left to right (see screenshot 1). It is your task to press the button corresponding to the person on the screen of which you think that the male voice is most likely to be coming from. Whether or not you feel that the position of the voice actually corresponds to the position of the person on the screen is irrelevant. Don't hesitate too long before making a choice.
After you have made your choice the next situation is reproduced. You will make 27 of such choices.

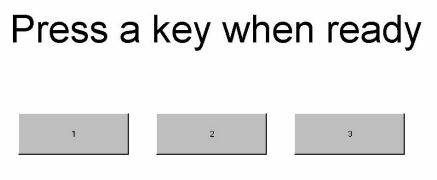
Part 2:

When you've finished part 1 the screen changes and 5 buttons appear (numbered "1" to "5") with labels varying from "completely realistic" (1) to "completely unrealistic" (5) (see screenshot 2). You will hear a series of audio fragments, comparable to the first series.
Now, it is your task to judge for each of the fragments to what extent you feel that the spatial layout of the 3 voices (so not just the male voice) appears as being realistic, given the visual presentation of the 3 persons in visual space.

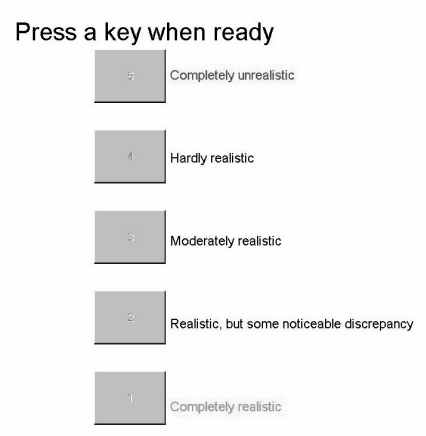
- *1:"completely realistic" in this context means that the spatial layout of the voices is in complete accordance with reality, while*
- *5:"completely unrealistic" means that the spatial layout of the voices doesn't correspond in any way to reality.*

Attention: so "5" means BAD, "1" means "GOOD"!!!
Also in this part you will make 27 choices. After this you are finished for this observation posi-

tion and the experiment will be repeated for a total of 3 observation positions (if you prefer in separate sessions).



Screenshot 1.



Screenshot 2.

A.4 Coloration Experiment

Below are the English instructions for the coloration experiment described in section 5.4.

'COLORATION' EXPERIMENT

In this experiment you will judge for noise samples to which extent the 'colour' of one sample is different from that of another. You will do this by repeatedly comparing two pairs of noise samples, presented over headphones.
Following first now is an explanation of what exactly we mean by the term 'colour'.

‘**Colour**’ *(or ‘klankkleur’ in Dutch) is a subjective property of sound which is a bit harder to describe as for example terms like ‘loudness’ or ‘pitch’. ‘Loudness’, for instance, is defined relatively simple as “that property of sound that enables ordering different sounds on a scale varying from ‘soft’ to ‘loud’”. Similarly, ‘pitch’ can be defined as “that property of sound that enables ordering different sounds on a scale varying from ‘low’ tot ‘high’. It is already a bit harder to give a definition of the term ‘timbre’. The formal definition of this is “that property of sound that makes it possible for a listener to judge that two sounds having the same loudness, pitch and spatial properties are different”. Think for example of two singers who are standing at the same position, singing the same note with the same loudness but yet their voices sound different.*
‘Colour’ is a term that is yet a bit more general than ‘timbre’ and is defined as:

“‘Colour’ is that property of sound that makes it possible for a listener to judge that two sounds having the same loudness and spatial properties are different.”.

This means that two sounds having the same loudness and spatial properties can be said to have a different colour if we can distinguish one from the other. This also means that the term ‘colour’ combines both the properties ‘pitch’ and ‘timbre’. It is important to fully understand this definition of ‘colour’ before you start the experiment.

The Experiment. *In a couple of minutes you will judge the colour differences of pairs of noise samples. You will see a screen that looks like figure 1. In every trial, you will hear two pairs of stereo samples of which each pair consists of a ‘stimulus A’ and a ‘stimulus B’ of 2 seconds each, with stimulus B presented right after stimulus A. So the sequence is:*

- *Pair 1, Stimulus A (2 seconds)*
- *Pair 1, Stimulus B (2 seconds)*
- *Pause of 0.5 second*
- *Pair 2, Stimulus A (2 seconds)*
- *Pair 2, Stimulus B (2 seconds)*

On the screen it is indicated which sample is playing at each moment.
After both pairs have finished playing you are asked to answer the question:

In which of the two pairs (‘Pair 1’ or ‘Pair 2’) was the difference between the colour of the ‘A’ and ‘B’ stimulus the largest?

In other words: you will compare the amount of colour difference within Pair 1 to the amount of colour difference within Pair 2.

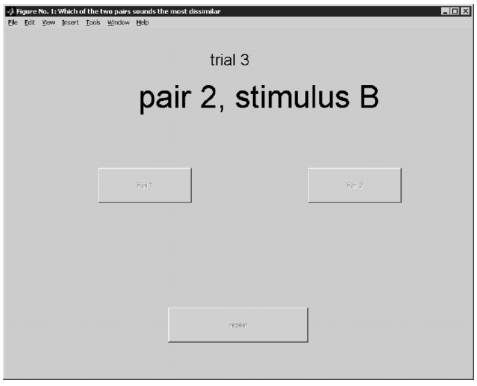
Attention! *There will be small differences in spatial properties (direction and/or ‘width’ of the signal) and/or loudness between the A and B stimuli of each pair and also between the two pairs. Given the definition of ‘colour’ that you have just read it is important that you pay* **as little attention as possible** *to this in making your decisions!*

You indicate your choice by pressing the button labelled ‘Pair 1’ or ‘Pair 2’. If after listening to a trial you are not sure yet, you can listen to the same two pairs again by pressing ‘Repeat’. You can listen to each pair as often as you like before you make a decision. After you have made your decision for Pair 1 or 2 you have finished one trial and the next trial of 2 pairs is presented. In total you will do 50 trials, which should take about 15 minutes.

To get used to the noise samples and the way in which they are presented you will start with listening to one or more test trials which are representative of the trials you will hear in the actual experiment. In these test trials you don’t have to choose for Pair 1 or 2 yet, they are intended only to get used to the stimuli. After each test trial of 2 pairs you can indicate that you want to listen to another test trial by pressing ‘test stimulus’. When you feel you have a good idea of what to expect you press ‘start experiment’, after which the screen of figure 1 appears. After pressing a mouse/keyboard button the first trial of the actual experiment starts.

Have fun!

FIGURE 1.



APPENDIX B *Analysis of Variance (ANOVA)*

Analysis of Variance (ANOVA) is a widely-used statistical method to investigate the probability that two or more groups of experimental data are samples from populations that have the same mean. Essentially, this is done by comparing the variance *among* the groups to the variance *within* the groups. It is a very useful tool to investigate the significance of the effect that a specific variable, that can take several discrete values, has on some quantity that is studied. This is done by dividing the experimental data into groups, or *categories*, according to the different values that the variable of interest can take and then performing an analysis of variance to find out if there are significant differences between the means of any subset of the individual groups, thus revealing the significance of the variable of interest. In section 4.1 of this thesis, ANOVA was used to investigate the significance on vertical localization accuracy of the presence of a matching video image, by separating the experimental data into ‘audio+video’ and ‘audio only’ categories and in section 4.5 the experimental data of the speech intelligibility experiment was separated into ‘WFS’ and ‘discrete loudspeakers’ categories to investigate the significance of the variable ‘reproduction system’ on the Speech Reception Threshold.

In this appendix the ANOVA method will be described in some more detail, following the explanation given in [Jobs91], Chapter 5: *‘Analysis of variance and experimental design’*.

Suppose we have done an experiment in which observations have been done on a continuous random variable Y with distribution $f(Y)$ for a total of g categories of some qualitative classification variable X and we want to investigate whether there are systematic differences between the means of Y for the g individual groups or not. For example: X could be 'brand of cars' and Y could be 'life expectancy' and we want to know whether there are systematic differences between the mean life expectancies $\mu_1....\mu_g$ of g different brands of cars.

First, several assumptions are made: the variances of the g groups are assumed to be homogeneous with magnitude σ^2 and the distribution of Y is assumed to be normal for each group, so the only possible difference between the distributions of Y between the g groups are the means. The objective of the Analysis Of Variance method is to test the null hypothesis:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_g . \tag{B.1}$$

If H_0 can be rejected, then we can conclude that there are significant differences between the means $\mu_1....\mu_g$ of the g groups.

Let y_{ij} be the i-th observation on Y in the j-th group, with j = 1, 2, ..., g and i = 1, 2, ..., n_j, (n_j is the number of samples in group j). Then the sample mean of the j-th group is:

$$\bar{y}_j = \frac{\sum_{i=1}^{n_j} y_{ij}}{n_j}, \tag{B.2}$$

and the sample mean of all n (which is the sum of all n_j's) observations is:

$$\bar{y} = \frac{\sum_{j=1}^{g}\sum_{i=1}^{n_j} y_{ij}}{n}. \tag{B.3}$$

The sample variance of the j-th group is:

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2}{(n_j - 1)}. \tag{B.4}$$

Now several ‘sums of squares’ can be calculated. First of all, there is the ‘total sum of squares’, *SST*, defined as:

$$SST = \sum_{j=1}^{g} \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2, \tag{B.5}$$

which measures the variation of all *n* samples around the overall mean.
Secondly, there is the ‘sum of squares among groups’, *SSA*, defined as:

$$SSA = \sum_{j=1}^{g} \sum_{i=1}^{n_j} (\bar{y}_j - \bar{y})^2 = \sum_{j=1}^{g} n_j (\bar{y}_j - \bar{y})^2, \tag{B.6}$$

which describes the variation among the *g* group means and finally, there is the ‘total sum of squares within groups’, *SSW*:

$$SSW = \sum_{j=1}^{g} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2, \tag{B.7}$$

describing the total variation *within* the *g* groups.
It can be shown that:

$$SST = SSA + SSW. \tag{B.8}$$

Under the null hypothesis H_0, both *SSA/*(*g-1*) and *SSW/*(*n-g*) provide estimates of the true variance σ^2 and, again: under H_0, should be equal.
In the case that H_0 is *not* true, *SSA/*(*g-1*) will be greater than *SSW/*(*n-g*), so their ratio, *F*:

$$F = \frac{SSA}{SSW} \frac{(n-g)}{(g-1)}, \tag{B.9}$$

can be used as a statistic to test H_0. It should be close to unity for H_0 to be true, while H_0 is rejected if *F* is sufficiently larger than unity.
It can be shown that *F* has a so-called F-distribution with (*g-1*) and (*n-g*) degrees of freedom and the critical value that *F* should exceed for H_0 to be rejected at a certain confidence level

can be looked up in a table of the F-distribution. If indeed H_0 is rejected, this means that it can be concluded, at the used level of confidence, that there are significant differences among the means of the g groups.

Alternatively, given the value of F obtained from the observations and the degrees of freedom in the experiment, the probability p that H_0 is true can be calculated. This is the most commonly used way to present the results of an ANOVA analysis, since it does not just prove that the means can or can not be said to be different at a specific confidence level, but it actually gives the probability that the means of all groups are equal, so that any found differences between them can be considered to be due to chance only. Critical values of this p-statistic that are commonly found in literature to prove the significance of a certain effect are $p<0.01$ or, as a less strong criterion, $p<0.05$.

APPENDIX C *Paired-Comparison Experiments*

This appendix describes how a set of stimuli can be ordered on a one-dimensional scale that corresponds to a certain property of the stimuli, by analyzing the results of a so-called ‘paired-comparison test’. In such a test, subjects make pair-wise comparisons of all stimuli to each other regarding the property under investigation and a judgement is made by the subjects which of the two stimuli in each pair has the highest ‘value’ of the property.

The nature of the property that is investigated can be very diverse. It can be a property that is highly correlated to one or more physical properties. For instance, the stimuli could be noise bursts and the property under investigation could be ‘loudness’, which would result in an ordering of the noise stimuli on a subjective ‘loudness’ scale, extending from ‘least loud’ to ‘loudest’. However, the property that is investigated can also be of a higly subjective nature. For instance, a paired-comparison test could be performed on a set of paintings regarding the property ‘beauty’, in which case the result would be an ordering of the paintings on a ‘beauty’ scale’, extending from ‘least beautiful’ to ‘most beautiful’.

In section C.1 a well-known method to extract the scale values from the results of a paired-comparison test is described.

In section 5.4 a paired-comparison test was used to compare the perceived strength of the spatial colour variations in the sound fields reproduced by WFS arrays with various loudspeaker

spacings, which resulted in an ordering of the different loudspeaker configurations on a ‘spatial colour variations’ scale. In section C.2, the preference matrices for this experiment are given.

C.1 Extracting Scale Values From Paired-Comparisons Experiments

Suppose we have *n* stimuli that we want to order on a scale that corresponds to a certain property of the stimuli. From the set of *n* stimuli that are investigated, all possible pairs are formed. There are:

$$\frac{1}{2}\frac{n!}{(n-2)!} = \frac{n(n-1)}{2} \tag{C.1}$$

pairs. The factor 1/2 is included so that pair *ij* and pair *ji* are treated as one and the same pair. Each of these pairs is presented to the subjects, who have to indicate for each pair which of the two stimuli in the pair has the highest ‘value’ for the property under investigation. In the simplest implementation the method is forced-choice, meaning that the subject has to choose one of the two stimuli, even if no difference is observed[1]. In total, each pair is evaluated *m* times by the subjects. This can be achieved by letting one subject compare all pairs *m* times, or by letting several subjects compare each pair m_k times, such that the sum of all m_k’s equals *m*.
Then, for each pair *ij*, the number of times that stimulus *i* was indicated to have a higher value for the quality under investigation than stimulus *j* is determined as a fraction p_{ij} of the total number of times *m* that this pair was evaluated:

$$p_{ij} = \frac{\text{number of times } i \text{ indicated to have higher value than } j}{m}. \tag{C.2}$$

These fractions p_{ij} are collected in a so-called ‘preference matrix’ of size (*n* x *n*). Since pairs *ij* and *ji* are not treated as separate pairs, p_{ji} is taken to be $1\text{-}p_{ij}$. For a similar reason p_{ii} is assumed to be 0.5. These fractions p_{ij} are now interpreted as the *probability* that stimulus *i* is judged to have a higher value than stimulus *j* in a direct comparison.
A very well-known method to convert the probabilities in the resulting preference matrix to the

1. It is possible to extend the method to include the possibility for ties or the possibility to quantify the difference between the two stimuli of a pair, usually on a 5- or 7-point scale. Here we will only discuss the forced-choice case.

desired scale values for the *n* stimuli was proposed by Thurstone ([Guil54]). The crucial step in this method is the assumption that the variations in the choice between two stimuli in a pair have a standard normal distribution. Furthermore it is assumed that the standard deviation of the choices is equal for all pairs and that the judgements of different pairs are independent[2]. Under these assumptions, for each pair *ij* the value z_{ij} can be calculated which satisfies the relation:

$$p(x \leq z_{ij}) = p_{ij}, \tag{C.3}$$

in which the operator *p(•)* denotes probability and *x* is a variable which has a standard normal distribution. The value of z_{ij} can be determined from a table of the cumulative distribution function *F* of the standard normal distribution *N(0,1)*:

$$z_{ij} = F^{-1}(p_{ij}). \tag{C.4}$$

Each z_{ij} can now be interpreted as an estimate of the difference between the scale values of stimulus *i* and *j* on the one-dimensional scale that corresponds to the property under investigation. All these z_{ij}'s are also collected in an (*n* x *n*) matrix. Finally, all the *n* elements of each row (all the $z_{i\bullet}$ values corresponding to stimulus *i*) are summed and divided by *n*, so that the resulting value z_i for row *i* is the average of the values $z_{i\bullet}$ of all the pairs in which stimulus *i* appeared. In [Guil54] it is stated that this averaging yields the best estimate for the scale values z_i of the stimuli in a least squares sense. These resulting z_i's are now interpreted as the scale values of the corresponding stimuli on a linear scale corresponding to the quality under investigation.

An important point to note about the method of paired-comparisons, is that only relative scale values are obtained, that give information about the *difference*s between the locations of the stimuli on the subjective scale. No information is available about the absolute locations on the scale. This means that in the example of the judgement of beauty of paintings, it would be possible that all paintings were considered to be beautiful, but some were just considered to be

2. To check if these assumptions are fulfilled to a satisfactory degree, a Chi-Square test can be done afterwards to check the internal consistency of the results ([Guil54]).

more beautiful than others. It would however be equally well possible that none of the paintings were considered to be beautiful, but some were just considered to be even less beautiful than others. In the interpretation of the results one therefore has to be careful not to translate high and low scale values into terms of ‘low’ and ‘high’, ‘bad’ and ‘good’ etcetera, but rather into terms of ‘least’ and ‘most’. Consequently, the location of zero of the scale is arbitrary and it is allowed to add or subtract any constant value to the scale values. If it is desired that the scale values also have some absolute meaning, then it is required to have one or more absolute references to which the stimuli can be compared. For instance, in the example of the paintings it would be possible to include a painting which is considered to be ‘very beautiful’ and one that is considered to be ‘not beautiful at all’.

C.2 Preference Matrices Of Experiments On Spatial Colour Variations

Below, the preference matrices resulting from the three paired-comparison experiments on spatial colour variations of section 5.4 are shown. Each cell contains the fraction of times p_{ij} that the configuration of the corresponding row i was judged to have more spatial colour variations than the configuration of the corresponding column j.

loudspeaker spacing	12.5 cm	16.7 cm	25.0 cm	33.3 cm	50.0 cm
12.5 cm	.50	.41	.44	.38	.24
16.7 cm	.59	.50	.47	.28	.14
25.0 cm	.56	.53	.50	.36	.14
33.3 cm	.62	.72	.64	.50	.29
50.0 cm	.76	.86	.86	.71	.50

TABLE C.1. Preference matrix for Experiment 1 (source at (0, -5) m, female speech noise).

loudspeaker spacing	12.5 cm	16.7 cm	25.0 cm	33.3 cm	50.0 cm
12.5 cm	.50	.47	.47	.36	.15
16.7 cm	.53	.50	.53	.25	.16
25.0 cm	.53	.47	.50	.33	.22
33.3 cm	.64	.75	.67	.50	.11
50.0 cm	.85	.84	.78	.89	.50

TABLE C.2. Preference matrix for Experiment 2 (source at (0, -5) m, male speech noise).

loudspeaker spacing	12.5 cm	16.7 cm	25.0 cm	33.3 cm	50.0 cm
12.5 cm	.50	.32	.28	.20	.12
16.7 cm	.68	.50	.28	.24	.24
25.0 cm	.72	.72	.50	.20	.32
33.3 cm	.80	.76	.80	.50	.48
50.0 cm	.88	.76	.68	.52	.50

TABLE C.3. Preference matrix for Experiment 3 (source at (0, -1) m, female speech noise).

References

[ANSI60] American National Standards Institute, '*USA Standard Acoustical Terminology (Including Mechanical Shock and Vibration) Sl.1-1960 (R1976)*', New York, 1960.

[Atal62] B.S. Atal, M.R. Schroeder and K.H. Kuttruff, '*Perception of coloration in filtered Gaussian noise - short-time spectral analysis by the ear*', Proceedings of the 4th Int. Congress on Acoustics, Copenhagen, 1962.

[Atma02] S.A. Atmadja, '*Implementation of a source tracking system using cross correlation of speech signals*', M.Sc. thesis, TU Delft, 2002.

[AT&T] `http://www.research.att.com/history/70picture.html`

[Azim97] H. Azima and N. Harris, 'Boundary interactions of diffuse field distributed-mode radiators', Proc. 103rd Convention of the Audio Eng. Soc. (preprint 4635), New York, USA, 1997.

[Bech95] S. Bech, V. Hansen and W. Woszczyk, '*Interaction between audio-visual factors in a home theater system: experimental results*', Proceedings of the 99th Convention of the Audio Eng. Soc., New York, USA, 1995.

[Beer99] J. Beerends and F. De Caluwe, '*The influence of video quality on perceived audio quality and vice versa*', J. Audio Eng. Soc. **47(5)**, pp355-362, 1999.

[Berk87] A.J. Berkhout, '*Applied seismic wave theory*', Elsevier Science Publishers, 1987.

[Berk88] A.J. Berkhout, '*A holographic approach to acoustic control*', J. Audio Eng. Soc. **36(12)**, pp977-995, 1988.

[Bert76] P. Bertelson and M. Radeau, '*Ventriloquism, sensory interaction, and response bias: remarks on the paper by Choe, Welch, Gilford, and Juola*', Perception & Psychophysics **19**, pp531-535, 1976.

[Blad81] R. Bladon and B. Lindblom, '*Modeling the judgment of vowel quality differences*', J. Acoust. Soc. Am. **(69)**, pp1414-1422, 1981.

[Blau69] J. Blauert, '*Sound localization in the median plane*', Acustica **22**, 1969

[Blau83] J. Blauert, '*Spatial Hearing*', MIT Press, 1983.

[Boon99] M. Boone, W. de Bruijn and U. Horbach, '*Virtual Surround Speakers with Wave Field Synthesis*', Proc. 106th Convention of the Audio Eng. Soc. (preprint 4928), Munich, Germany, 1999.

[Boon01] M. Boone, D. de Vries and U. Horbach, '*Wave Field Synthesis sound reproduction system using a distributed mode panel*', International Patent Application PCT/NL01/00843, filed 11/2001.

[Boon04a] M. Boone and H. Helleman, '*Audibility thresholds of spatial variations in a single acoustic reflection*', Proc. 116th Convention of the Audio Eng. Soc., Berlin, Germany, 2004.

[Boon04b] M. Boone, '*Multi Actuator Panels (MAPs) as loudspeaker arrays for Wave Field Synthesis*', J. Audio Eng. Soc. **52(7)**, 2004.

[Bron99] A. Bronkhorst and T. Houtgast, '*Auditory distance perception in rooms*', Nature **397**, pp517-520, 1999.

[Brui98a] W. de Bruijn, T. Piccolo and M. Boone, '*Sound recording techniques for Wave Field Synthesis and other multichannel sound systems*', Proc. 104th Convention of the Audio Eng. Soc. (preprint 4690), Amsterdam, 1998.

[Brui98b] W. de Bruijn, '*Recording and reproduction of reflections and reverberation for Wave Field Synthesis*', M.Sc. thesis, Delft University of Technology, 1998.

[Buch02] H. Buchner, S. Spors, W. Kellermann, and R. Rabenstein, '*Full-Duplex Communication Systems with Loudspeaker Arrays and Microphone Arrays*', Proc. IEEE Int. Conference on Multimedia and Expo (ICME), Lausanne, Switzerland, 2002.

[Buse97] P. Buser and M. Imbert, '*Vision*', MIT Press, 1997.

[CARROU] http://www.emt.iis.fhg.de/projects/carrouso/

[CAVE] http://www.evl.uic.edu/pape/CAVE/

[Chat99] N. Chateau, '*Contribution of sound spatialization to audio-visual quality in videoconferencing*', in: *Collected papers from the joint meeting "Berlin 99"*, Acoust. Soc. of Am., 1999.

[Cole62] P. Coleman, '*Failure to localize the source distance of an unfamiliar sound*', J. Acoust. Soc. Am. **34**, pp345-346, 1962.

[Cort02] E. Corteel, U. Horbach and R. Pellegrini, '*Multichannel Inverse Filtering of multi-exciter distributed mode loudspeakers for wave field Synthesis*', Proc. 112th Convention of the Audio Eng. Soc., Munich, Germany, 2002.

[Cowa84] R. Cowan, 'Teleconferencing: maximizing human potential', Reston Publishing, 1984.

[Crem88] L. Cremer and M. Heckl, '*Structure-borne sound*', 2nd ed., Springer-Verlag, 1988.

[Dixo80] N. Dixon and L. Spitz, '*The detection of auditory visual desynchrony*', Perception **9**, pp719-721, 1980.

[Emer98] M. Emerit and A. Gilloire, '*Application of sound spatialization techniques to group telecommunications*', internal report of France Telecom R&D, 1998.

[Fisc91] S. Fisher, '*Virtual Environments, Personal Simulation, & Telepresence*', in: '*Virtual Reality: Theory, Practice and Promise*', S. Helsel and J.Roth (eds.), Meckler Publishing, 1991.

[Gard68] M. Gardner, '*Proximity image effect in sound localization*', J. Acoust. Soc. Am. **43**, p163, 1968.

[Gay00] S.L. Gay and J. Benesty (eds.), '*Acoustic signal processing for telecommunication*', Kluwer Academic, 2000.

[Gold91] E. Goldstein, '*Perceived orientation, spatial layout and the geometry of pictures*', in: S.R. Ellis (ed.), '*Pictorial communication in virtual and real environments*', Taylor and Francis, 1991.

[Guil54] J.P. Guilford, '*Psychometric methods*', McGraw-Hill, 1954.

[Harr97] N. Harris and M. Hawksford, '*The Distributed Mode Loudspeaker (DML) as a broad-band acoustic radiator*', Proc. 103rd Convention of the Audio Eng. Soc. (preprint 4526), New York, USA, 1997.

[Hebr74] Hebrank & Wright, '*Are two ears necessary for localization of sound sources on the median plane*', J. Acoust. Soc. Am. **56**, pp935-938, 1974.

[Huls02a] E. Hulsebos, D. de Vries and E. Bourdillat, '*Improved Microphone Array Configurations for Auralization of Sound Fields by Wave-Field Synthesis*', J. Audio. Eng. Soc. **50**, pp779, 2002.

[Huls02b] E. Hulsebos and D. de Vries, '*Parameterization and Reproduction of Concert Hall Acoustics Measured with a Circular Microphone Array*', Proc. 112th Convention of the Audio Eng. Soc. (preprint 5579), Munich, Germany, 2002.

[Huls03] E. Hulsebos, T. Schuurmans, D. de Vries and M. Boone, '*Circular Microphone Array for Discrete Multichannel Audio Recording*', Proc. 114th Convention of the Audio Eng. Soc. (preprint 5716), Amsterdam, 2003.

[Huls04] E. Hulsebos, '*Auralization using Wave Field Synthesis*', Ph.D. thesis, Delft University of Technology, 2004.

[Huyg90] C. Huygens, '*Traité de la lumière - ou sont expliquées les causes de ce qui arrive dans la reflexion, & dans la refraction. Et particulièrement dans l'étrange refraction du cristal d'Islande*', Pierre van der Aa, Leiden, 1690.

[IMAX] `http://www.imax.com`

[IOSONO] `http://www.iosono-sound.de/`

[ITU90] ITU-R Recommendation BS.562-3: *'Subjective assessment of sound quality'*, International Telecommunication Union, 1990.

[ITU97] ITU-R Recommendation BS.1286: '*Methods for the subjective assessment of audio systems with accompanying picture*', International Telecommunication Union, 1997.

[Jess73] M. Jessel, '*Acoustique théorique - propagation et holophonie*', Masson et Cie, 1973.

[Jobs91] J.D. Jobson, '*Applied multivariate data analysis, Vol. I: Regression and experimental design*', Springer Verlag, 1991.

[Kita02] N. Kitagawa and S. Ichihara, '*Hearing visual motion in depth*', Nature **416**, pp172-174, 2002.

[Komi89] S. Komiyama, '*Subjective evaluation of angular displacement between picture and sound directions for HDTV sound systems*', J. Aud. Eng. Soc. **37(4)**, pp210-214, 1989.

[Kutt73] H. Kuttruff, '*Room acoustics*', Applied Science, 1973.

[Loom98] J. Loomis, R. Klatzky, J. Philbeck and R. Golledge, '*Assessing auditory distance perception using perceptually directed action*', Perception & Psychophysics **60**, pp966-980, Psychonomic Society, 1998.

[Mass87] D. Massaro, '*Speech perception by ear and eye: a paradigm for psychological inquiry*', Erlbaum Associates, 1987.

[McGu76] H. McGurk and J. MacDonald, '*Hearing lips and seeing voices*', Nature **264**, pp746-748, 1976.

[Merk00] I. Merks, '*Binaural application of microphone arrays for improved speech intelligibility in a noisy environment*', Ph.D. thesis, Delft University of Technology, 2000.

[Moor83] B.C.J. Moore and B.R. Glasberg, '*Suggested formulae for calculating auditory-filter bandwidths and excitation patterns*', J. Acoust. Soc. Am. **(74)**, pp750-753, 1983.

[Nath97] C. Nathanail, C. Lavandier, J.D. Polack and O. Warusfel, '*Influence of sensory interactions between vision and audition on the perceptual characterisation of room acoustics*', Proc. Int. Computer Music Conf. (ICMC) 97, pp414-417, Thessaloniki, Greece, 1997.

[Nico98] R. Nicol, M. Emerit and A. Gilloire, '*Mur de telepresence pour la visioconference: une approche holophonique*', Proc. CORESA '98 (Compression et Réprésentation des Signaux Audiovisuels), Lannion, France.

[Nico99] R. Nicol, '*Restitution sonore spatialisée sur une zone étendue: application à la téléprésence*', Ph.D. thesis, Université du Maine, Le Mans, France, 1999.

[NXT] `http://www.nxtsound.com/`

[PANORA] `http://www.tnt.uni-hannover.de/project/eu/panorama/`

[Patt86] R.D. Patterson and B.C.J. Moore, '*Auditory filters and excitation patterns as representations of frequency resolution*', in: B.C.J. Moore (ed.), '*Frequency selectivity in hearing*', Academic Press, London, 1986.

[Plom70] R. Plomp, '*Timbre as a multidimensional attribute of complex tones*', in: '*Frequency analysis and periodicity detection in hearing*', R. Plomp and G.F. Smoorenburg (eds.), A.W. Sijthoff, 1970.

[Plom73] R. Plomp and H.J.M. Steeneken, '*Place dependence of timbre in reverberant sound fields*', Acustica **(28)**, pp50-59, 1973.

[Plom79] R. Plomp and A.M. Mimpen, *'Improving the reliability of testing the speech reception threshold for sentences'*, Audiology **18**, pp43-52, 1979.

[Plom81] R. Plomp and A.M. Mimpen, *'Effect of the orientation of the speaker's head and the azimuth of a noise source on the speech-reception threshold for sentences'*, Acustica **48**, pp325-328, 1981.

[Rede00] A. Redert, '*Multi-viewpoint systems for 3-D visual communication*', Ph.D. thesis, Delft University of Technology, 2000.

[Rime98] A. Rimell, M. Hollier and R. Voelcker, '*The influence of cross-modal interaction on audio-visual speech quality perception*', Proceedings of the 105th Convention of the Audio Eng. Soc. (preprint #4791), 1998.

[Roff67] S. Roffler and R. Butler, '*Factors that influence the localization of sound in the vertical plane*', J. Acoust. Soc. Am. **43(6)**, pp1255-1259, 1967.

[Rooij01] W. van Rooijen, '*Distributed Mode Loudspeakers for Wave Field Synthesis*', M.Sc. thesis, Delft University of Technology, 2001.

[Salo95] A.M. Salomons, '*Coloration and binaural decoloration of sound due to reflections*', Ph.D. thesis, Delft University of Technology, 1995.

[Schr00] O. Schreer and P. Sheppard, '*VIRTUE - The step towards immersive telepresence in virtual video-conference systems*', Proc. eWork and eBusiness 2000, Madrid, Spain, 2000.

[Sear75] C. Searle, L. Braida, D. Cuddy and M. Davis, '*Binaural pinna disparity: another auditory localization cue*', J. Acoust. Soc. Am. **57(2)**, pp448-455, 1975.

[Sedg91] H. Sedgwick, '*The effects of viewpoint on the virtual space of pictures*', in: S.R. Ellis (ed.), '*Pictorial communication in virtual and real environments*', Taylor and Francis, 1991.

[Sher92] T.B. Sheridan, '*Musings on Telepresence and Virtual Presence*', Presence: Teleoperators and Virtual Environments **1(1)**, pp120-125, 1992.

[Some66] Somerville, *'Recent work on the effects of reflectors in concert halls and music studios'*, J. Sound & Vib. **3**, pp127-134, 1966.

[Sonk00] J.-J. Sonke, '*Variable acoustics by Wave Field Synthesis'*, Ph.D. thesis, Delft University of Technology, 2000.

[Spei93] J. Speigle & J. Loomis, '*Auditory distance perception by translating observers*', Proceedings of IEEE Symposium on Research Frontiers in Virtual Reality, San Jose, USA, 1993.

[Star97] E. Start, '*Direct sound enhancement by Wave Field Synthesis*', Ph.D. thesis, Delft University of Technology, 1997.

[Ster3D] `http://www.stereo3d.com/hmd.htm`

[TNO88] Audio Compact Disc: '*Spraakmateriaal behorende bij de test voor het meten van de spraakverstaanbaarheidsdrempel voor korte zinnen in stilte en in stationaire of fluctuerende ruis*', Instituut voor Zintuigfysiologie TNO Soesterberg, The Netherlands, 1988.

[Verh97] E. Verheijen, '*Sound reproduction by Wave Field Synthesis*', Ph.D. thesis, Delft University of Technology, 1997.

[ViewMa] `http://www.fisher-price.com/us/view-master/default_html.asp`

[VIRTUE] `http://www.virtue.eu.com/`

[Voge93] P. Vogel, '*Application of Wave Field Synthesis in room acoustics*', Ph.D. thesis, Delft University of Technology, 1993.

[Vrie94] D. de Vries, E. Start and V. Valstar, '*The Wave Field Synthesis concept applied to sound reinforcement: restrictions and solutions*', Proceedings of the 96th Convention of the Audio Eng. Soc. (preprint 3812), Amsterdam, 1994.

[VRphob] `http://graphics.tudelft.nl/~vrphobia/`

[Warr83] D. Warren, T. McCarthy and R. Welch, '*Discrepancy and nondiscrepancy methods of assessing visual-auditory interaction*', Perception & Psychophysics **33(5)**, pp413-419, 1983.

[Watk78] Watkins, '*Psychoacoustical aspects of synthesized vertical locale cues*', J. Acoust. Soc. Am. **63**, pp1152-1165, 1978.

[Wigh92] F. Wightman and D. Kistler, '*The dominant role of low-frequency interaural time differences in sound localization*', J. Acoust. Soc. Am. **91(3)**, pp1648-1661, 1992.

[Wigh97] F. Wightman and D. Kistler, '*Monaural sound localization revisited*', J. Acoust. Soc. Am. **101**, pp1050-1062, 1997.

[Wosz95] W. Woszczyk, S. Bech and V. Hansen, '*Interactions between audio-visual factors in a home theater system: definition of subjective attributes*', Proceedings of the 99th Convention of the Audio Eng. Soc., New York, USA, 1995.

[Xu02] L-Q Xu, B Lei and E. Hendriks, '*Computer vision for a 3-D visualisation and telepresence collaborative working environment*', BT Technology Journal **20**(**1)**, pp64-74, 2002.

[Zure79] P.M. Zurek, '*Measurements of binaural echo suppression*', J. Acoust. Soc. Am. **(66)**, pp1750-1757, 1979.

[Zwic57] E. Zwicker, G. Flottorp and S.S. Stevens, '*Critical bandwidth in loudness summation*', J. Acoust. Soc. Am. **(29)**, pp548-557, 1957.

Summary

Application of Wave Field Synthesis in Videoconferencing

This thesis investigates the application of Wave Field Synthesis (WFS) in life-size videoconferencing systems. In the ideal videoconferencing system, the users have a perfect illusion of being together in the same room with their remote conference partners, a concept known as *telepresence*. To achieve acoustical telepresence, the conferencing system should be able to provide an accurate reproduction of the sound field of the remote site in both a spectro-temporal and a spatial sense. Artifacts in the reproduced sound, such as strong coloration or a mismatch between the perceived positions of the voices of remote participants and their corresponding images on the screen, can lead to fatigue and loss of concentration, resulting in a less effective virtual meeting. Existing systems are unable to provide a sufficiently accurate spatial reproduction, since they offer no or only very simple sound spatialization. Therefore, a research project was started with the aim to develop the audio part of a videoconferencing system providing a sufficiently accurate spatial reproduction to achieve acoustical telepresence.

The technique that is used in this project is *Wave Field Synthesis* (WFS), an audio reproduction concept developed at TU Delft that uses arrays of closely-spaced loudspeakers to synthesize the sound field of a desired sound source. This reproduction is correct within an extensive listening area, rather than only at a single ‘sweet spot’, so it is ideal for applications in which

multiple people should be able to experience a realistic spatial sound reproduction, even when they walk around in the room.

An audio system for videoconferencing can be divided into several subsystems that take care of the recording, the transmission and the reproduction of the sound, respectively. This thesis focuses on the reproduction part and has two main objectives: *To investigate whether application of Wave Field Synthesis results in an audio reproduction system for videoconferencing that enables a more effective virtual meeting than can be achieved when conventional audio reproduction systems are used,* and, *to optimize the WFS reproduction system for the application of videoconferencing*. The most important aspect of this optimization is to determine the minimum number of loudspeakers that is required to achieve a natural reproduction with accurate localization and without disturbing coloration of the sound.

In the study of audio-visual systems, it is important to have sufficient knowledge of auditory- and visual perception and also of the audio-visual interaction effects that occur when audio and video are combined. An overview is given of the most important aspects of auditory and visual perception of space and of spatial audio-visual interaction effects, the most well-known one being the effect of ‘visual capture’, which in the videoconferencing application can be expected to loosen the requirements for the resolution of auditory source positioning.

A perception experiment was carried out in which the required vertical resolution for sound source positioning was investigated. Results show that due to the low human ability to localize sound sources in the vertical plane, combined with the visual capture effect, vertical sound source positioning is not critical at all in the videoconferencing application and one, or at most two, horizontal WFS loudspeaker arrays are sufficient.

For reasons of practicality, the video system in this project was restricted to conventional two-dimensional (2D) video projection. An important property of such a system is that there is only a single viewpoint for which the visual perspective is identical to the perspective of the original real-life scene, while the perspective is distorted for all other viewing positions. Furthermore, several important cues for depth estimation are absent in a 2D projection, so that in general the perception of depth is not the same as in the original scene. This means that if 2D video is combined with WFS, the audio- and video perspective will not be the same. It was expected that this might introduce a discrepancy between the perceived positions of the auditory- and corresponding visual sources, which might degrade the performance of the system in some aspects. This issue was investigated in a series of audio-visual perception experiments. One experiment

with a single sound source positioned at its original distance, investigated for various observation positions what was the range of lateral positions that, according to the subjects, corresponded naturally to the corresponding visual source on the screen. It was found that for some visual sources it was impossible to position the sound source such that it corresponded naturally to the visual source for all observers in the room. In a next experiment, subjects graded the annoyance of the discrepancy between the perceived position of a sound- and visual source when the sound source was positioned at its original position. It was found that for observers with a lateral displacement relative to the viewpoint, resulting in a discrepancy between perceived auditory- and visual source directions, the discrepancy was judged as being annoying for a modest lateral displacement of 1 m. Viewing from a distance too close to the screen did not result in a significant annoyance. In a multiple-source experiment, audio-visual source identification performance was tested. Out of three simultaneous speech sound sources, randomly distributed over three visual sources on the screen and positioned at the corresponding original source positions, subjects had to identify a target source. It was found that observing the audio-visual scene from a position with a lateral displacement relative to the viewpoint resulted in a degradation of identification performance. In a final experiment, subjects graded the overall naturalness of the audio-visual scene when three speech sound sources were positioned at the original source positions corresponding to visual images on the screen. The results indicated a degradation of perceived naturalness for observation positions with a lateral displacement relative to the viewpoint. From this series of experiments, it is concluded that combining 2D video projection with WFS can indeed be expected to result in problems that are caused by the different perspectives of the audio and video.

A simple method is presented to avoid or at least reduce the problems discussed above. This method, *auditory depth compression*, consists of reducing the perspective of the auditory scene by pulling all sound sources somewhat closer to the screen in the direction of the viewpoint, to such an extent that the discrepancies that occur are reduced to acceptable values.

A very important characteristic of a speech communication system is its *speech intelligibility.* Using a standardized test, the speech reception threshold (SRT) of a system using WFS was compared to a system using conventional reproduction in a situation with a target speech source and one interfering speech noise source. It was found that WFS reproduction can result in a significant increase of the speech intelligibility. This is a strong argument in favour of applying WFS in a videoconferencing system. Using the auditory depth compression method

discussed above, it should be possible to avoid the problems that are introduced by combining WFS with 2D video, while retaining the benefit of increased speech intelligibility.

Discretization of the WFS array causes *spatial aliasing*, resulting in a distortion of the spectrum of the reproduced sound field above the spatial Nyquist frequency, which is inversely proportional to the loudspeaker spacing. This spectral distortion might be perceived as a place-dependent *coloration* of the sound field. It was investigated what is the maximum loudspeaker spacing that can be allowed so that no coloration artifacts are perceived. Based on a model of human perception of colour differences between individual signals, an analysis is made of the spatial colour variations that occur in the reproduced sound field of a WFS system with a given loudspeaker spacing, and a measure is proposed to quantify the strength of the perceived spatial colour variations, the *Spatial Colour Variation (SCV) Index*. In a perception experiment, subjects compared the spatial colour variations in the simulated sound fields of WFS systems with various loudspeaker spacings. The results were consistent with the model and a critical value of the SCV-Index was found, above which noticeable spatial colour variations can be expected to occur, corresponding to a loudspeaker spacing of about 25 cm for the videoconferencing application.

In the context of the objective of optimizing the WFS loudspeaker array, the application of so-called *Multi-Actuator Panel* (MAP) arrays in WFS was investigated. MAP's are an extension of the well-known concept of distributed mode loudspeaker (DML) panels to the case in which multiple individually-driven exciters are attached to a large panel of suitable material, such that each exciter acts as an individual secondary source. First, by experiments with an array of small single-exciter DML panels, it was established that in principle it is possible to use DML panels for WFS reproduction. A disadvantage of small panels is their low sound quality, especially for low frequencies. Therefore, it was investigated if it is possible to build a WFS array according to the MAP concept. It is shown that this requires panel material with specific properties. From experiments with various prototypes it can be concluded that MAP arrays are indeed capable of synthesizing sound fields with a spatial quality comparable to that of sound fields reproduced by an array of conventional loudspeakers. In the videoconferencing application, this means that the WFS loudspeaker array and the projection screen can be integrated.

Werner de Bruijn, Delft University of Technology, 2004.

Samenvatting

Toepassing van Wave Field Synthesis (‘Golfveldsynthese’) in Videoconferencing

In dit proefschrift wordt de toepassing van Wave Field Synthesis (WFS) in life-size videoconferentiesystemen onderzocht. In het perfecte videoconferentiesysteem hebben de gebruikers de illusie dat zij zich met hun gesprekspartners in dezelfde kamer bevinden, een concept dat bekend staat als *telepresence*. Om akoestische telepresence te bereiken moet het systeem in staat zijn een nauwkeurige weergave te maken van het geluidsveld van ‘de andere kant’, zowel in temporele als in ruimtelijke zin. Artefacten in de geluidsreproductie, zoals sterke kleuring of een discrepantie tussen de waargenomen posities van stemmen en de corresponderende beelden op het scherm, kunnen leiden tot vermoeidheid en concentratieverlies, resulterend in een minder effectieve vergadering. Bestaande systemen zijn niet in staat een voldoende nauwkeurige ruimtelijke weergave te maken, daar zij geen of slechts een zeer simpele ruimtelijke weergave bieden. Om deze reden werd een onderzoeksproject gestart met als doel het ontwikkelen van het audiodeel van een videoconferentiesysteem dat een voldoende nauwkeurige ruimtelijke weergave biedt om akoestische telepresence te bereiken.

In dit project is gebruik gemaakt van *Wave Field Synthesis* (WFS, Nederlands: *Golfveldsynthese*), een geluidsreproductietechniek ontwikkeld aan de TU Delft die gebruik maakt van arrays van luidsprekers om het geluidsveld van een geluidsbron te synthetiseren. Deze repro-

ductie is correct in een uitgebreid luistergebied, i.p.v. slechts op één ‘sweet spot’, zodat WFS ideaal is voor toepassingen waarin meerdere mensen een realistische ruimtelijke geluidsweergave moeten ervaren, zelfs als zij rondlopen in de kamer.

Een audiosysteem voor videoconferencing kan onderverdeeld worden in subsystemen voor de opname, transmissie en reproductie van het geluid. Dit proefschrift richt zich op de reproductie en heeft twee doelstellingen: *Onderzoeken of toepassing van Wave Field Synthesis leidt tot een geluidsreproductiesysteem voor videoconferencing dat een effectievere virtuele vergadering mogelijk maakt dan mogelijk is met conventionele technieken,* en, *het optimaliseren van het WFS reproductiesysteem voor toepassing in videoconferencing*. Het belangrijkste aspect van deze optimalisatie is het bepalen van het minimale aantal luidsprekers dat nodig is voor een natuurlijke reproductie met nauwkeurige lokalisatie, zonder storende kleuring van het geluid.

In een onderzoek als dit spelen akoestische en visuele perceptie alsmede audio-visuele interactie effecten die optreden wanneer audio en video worden gecombineerd, een grote rol. Er wordt daarom een overzicht gegeven van de belangrijkste aspecten van de akoestische en visuele perceptie van ruimte en van ruimtelijke audio-visuele interactie effecten.

In een perceptie experiment werd de vereiste verticale resolutie voor geluidsbronpositionering onderzocht, met als conclusie dat door het beperkte menselijke vermogen tot bronlokalisatie in het verticale vlak, gecombineerd met het ‘visual capture’ effect, verticale geluidsbronpositionering in het geheel niet kritisch is in de videoconferencing toepassing en dat één, of ten hoogste twee, horizontale WFS luidsprekerarrays volstaan.

Om praktische redenen bestond het videosysteem in dit project uit conventionele twee-dimensionale (2D) videoprojectie. Een eigenschap hiervan is dat er slechts één observatiepunt is, het ‘viewpoint’, waar het visuele perspectief gelijk is aan dat van de oorspronkelijke scène, terwijl het perspectief vervormd is op alle andere observatiepunten. Bovendien ontbreken in een 2D projectie een aantal belangrijke cues voor dieptewaarneming, zodat in het algemeen de perceptie van diepte niet hetzelfde is als in de oorspronkelijke scène. Dit betekent dat als 2D video gecombineerd wordt met WFS, het audio en video perspectief verschillend zullen zijn. Verwacht werd dat dit een discrepantie zou kunnen veroorzaken tussen de waargenomen posities van de akoestische en de corresponderende visuele bronnen, hetgeen de prestaties van het systeem negatief zou kunnen beïnvloeden. Dit is onderzocht in een serie audio-visuele perceptie experimenten. In een experiment met één geluidsbron, gepositioneerd op zijn originele afstand, werd voor verschillende observatiepunten onderzocht binnen welk bereik van laterale

posities de bron mocht staan zodat, volgens de proefpersonen, de positie van de geluidsbron op natuurlijke wijze overeenkwam met de corresponderende visuele bron op het scherm. Het bleek voor sommige visuele bronnen niet mogelijk te zijn de geluidsbron dusdanig te plaatsen dat hij voor alle waarnemers in de kamer op natuurlijke wijze correspondeerde met de visuele bron. In een tweede experiment beoordeelden proefpersonen hoe storend de discrepantie tussen waargenomen posities van een geluids- en visuele bron was wanneer de geluidsbron op zijn originele positie stond. Voor waarnemers met een laterale verplaatsing t.o.v. het viewpoint, resulterend in een discrepantie tussen de waargenomen richting van de geluids- en visuele bron, bleek dat de discrepantie al als storend werd ervaren bij een verplaatsing van 1 m. In een experiment met meerdere bronnen werd audio-visuele bronidentificatie getest. Uit drie gelijktijdige spraakbronnen, willekeurig verdeeld over drie visuele bronnen op het scherm en gepositioneerd op de corresponderende oorspronkelijke bronposities, moesten proefpersonen één specifieke bron identificeren. Het bleek dat wanneer de audio-visuele scène geobserveerd werd van een positie met een laterale verplaatsing t.o.v. het viewpoint, dit een negatieve invloed had op de identificatieperformance. In een laatste experiment beoordeelden proefpersonen de natuurlijkheid van de audio-visuele scène wanneer drie spraakbronnen waren gepositioneerd op de oorspronkelijke posities die correspondeerden met de visuele bronnen op het scherm. De resultaten toonden een vermindering van de beoordeelde natuurlijkheid voor observatieposities met een laterale verplaatsing. Uit deze serie experimenten wordt geconcludeerd dat inderdaad verwacht kan worden dat het combineren van 2D video met WFS leidt tot problemen die veroorzaakt worden door de verschillende perspectieven van audio en video. Er wordt een simpele methode voorgesteld om de hierboven besproken problemen te vermijden of verminderen. Deze methode, *auditory depth compression*, bestaat uit het reduceren van het akoestische perspectief door alle geluidsbronnen wat dichter naar het scherm te halen, in die mate dat de discrepanties die optreden gereduceerd worden tot acceptabele waarden.

Een belangrijke eigenschap van een communicatiesysteem is de *spraakverstaanbaarheid*. Door middel van een gestandaardiseerde test werd de spraakverstaanbaarheidsdrempel (SRT) van een WFS systeem vergeleken met die van een conventioneel reproductiesysteem in een situatie met één spraakbron en één stoorbron. Het bleek dat WFS reproductie kan resulteren in een significante verbetering van de spraakverstaanbaarheid, hetgeen een sterk argument is voor het toepassen van WFS in videoconferentiesystemen. Met behulp van de hierboven besproken ‘auditory depth compression’ methode moet het mogelijk zijn de problemen die veroorzaakt

worden door het combineren van WFS met 2D video te vermijden, terwijl het voordeel van verbeterde spraakverstaanbaarheid behouden blijft.

Discretisatie van het WFS array veroorzaakt *spatiële aliasing*, hetgeen resulteert in vervorming van het spectrum van het gereproduceerde geluidsveld boven de Nyquist frequentie, die omgekeerd evenredig is met de afstand tussen de luidsprekers. Deze spectrale vervorming kan mogelijk waargenomen worden als plaatsafhankelijke *kleuring* van het geluidsveld. Er is onderzocht wat de maximaal toegestane afstand tussen de luidsprekers is zodat er geen kleuring wordt waargenomen. Na een analyse van de spatiële klankkleurvariaties die optreden in het gereproduceerde geluidsveld van een WFS systeem, gebaseerd op een model van de waarneming van klankkleurverschillen tussen signalen, wordt een maat voorgesteld om de sterkte van de waargenomen klankkleurvariaties te quantificeren, de *Spatial Colour Variation (SCV) Index*. In een experiment vergeleken proefpersonen de spatiële klankkleurvariaties in gesimuleerde geluidsvelden van WFS systemen met verschillende luidsprekerafstanden. De resultaten zijn consistent met het model en er wordt een kritische waarde gevonden voor de SCV-Index waarboven waarneembare klankkleurvariaties verwacht kunnen worden, in de videoconferencing toepassing corresponderend met een luidsprekerafstand van 25 cm.

In het kader van het optimaliseren van het luidsprekerarray werd de toepassing van zogenaamde *Multi-Actuator Panel* (MAP) arrays in WFS onderzocht. MAPs zijn een uitbreiding van het concept van de distributed mode loudspeaker (DML) panelen tot het geval waarin meerdere individueel aan te sturen actuatoren bevestigd zijn op een groot paneel, zodanig dat elke actuator zich gedraagt als een individuele bron. Eerst werd d.m.v. experimenten met een array van kleine panelen met elk één actuator aangetoond dat het in principe mogelijk is DML panelen voor WFS reproductie te gebruiken. Een nadeel van kleine panelen is hun slechte geluidskwaliteit, vooral bij lage frequenties. Daarom is onderzocht of het mogelijk is om een WFS array te bouwen volgens het MAP concept. Dit blijkt specifieke eisen te stellen aan de materiaaleigenschappen van het paneel. Uit experimenten met verschillende prototypes kan worden geconcludeerd dat het mogelijk is m.b.v. MAP arrays geluidsvelden te synthetiseren die vergelijkbaar zijn met geluidsvelden van een array van conventionele luidsprekers. Voor de videoconferencing toepassing betekent dit dat het WFS luidsprekerarray en het projectiescherm geïntegreerd kunnen worden.

Werner de Bruijn, Technische Universiteit Delft, 2004.

Dankwoord

Het is al weer een jaar of zes geleden (plus of min een paar maanden) dat ik begon met mijn promotieonderzoek in de vakgroep akoestiek. En op het moment dat ik deze woorden schrijf, in de trein, op weg om te gaan genieten van een zonnig weekend reggae muziek, kan ik eindelijk zeggen dat mijn proefschrift af is. Een goed gevoel.
Dit is dan ook een mooi moment om mijn waardering uit te spreken voor een heleboel mensen die allemaal, op de één of andere manier, hebben bijgedragen aan de succesvolle afronding van deze zes jaar durende periode van mijn leven.

Allereerst wil ik mijn begeleider Rinus Boone bedanken, die een essentiële bijdrage heeft geleverd aan de totstandkoming van dit proefschrift. Ik bedank hem voor de prettige samenwerking en de vele goede discussies die we gehad hebben over mijn werk, voor zijn kritisch lezen van en vele nuttige commentaar op de eerste versies van alle individuele hoofdstukken van dit proefschrift en ook voor zijn geduld en vertrouwen in het feit dat, uiteindelijk, dit proefschrift af zou komen, nadat ik ruim een jaar geleden met mijn nieuwe baan bij Philips was begonnen, hetgeen betekende dat ik er voornamelijk in de avonduren en weekenden aan moest werken en het dus allemaal wat minder snel vorderde.
Ik wil mijn promotor professor Berkhout bedanken voor zijn stimulerende visie, die ondermeer heeft geleid tot de ontwikkeling van de Wave Field Synthesis techniek die de basis vormt voor

het werk in dit proefschrift en voor zijn inspirerende commentaar en suggesties op het manuscript.

Then a few “thank you”s in English. First of all, many thanks to the people at France Télécom R&D, Lannion, France, who were the sponsors of this project. In particular I thank Rozenn Nicol and Marc Emerit, for the pleasant cooperation and interesting discussions we’ve had.
I also want to thank the members of my committee for doing me the honor of being in the committee and for evaluating the manuscript of this thesis. In particular I would like to thank Armin Kohlrausch for his thorough evaluation of the manuscript and many useful suggestions.

Terug naar het Nederlands nu. Ik wil graag alle stafleden van de vakgroep akoestiek bedanken. De wetenschappelijke staf, in het bijzonder Diemer de Vries die ook nauw betrokken was bij dit project en de technische staf: Henry, Edo, Leen en Paul voor de technische- en computerondersteuning.
Ik bedank natuurlijk mijn kamergenoten gedurende mijn vijf jaar als promovendus op de vakgroep: Ivo, Jan-Jakob, Jan, Edo en Javier, voor de gezellige en stimulerende werkomgeving (en voor het tolereren van mijn reggae verslaving).
Verder bedank ik mijn collega-promovendi van het seismische kamp voor het bijdragen aan de gezellige sfeer in de vakgroep, evenals de vele studenten van de vakgroep gedurende de vijf jaar. In het bijzonder wil ik Wilfred van Rooijen bedanken, wiens uitstekende afstudeerwerk een belangrijk onderdeel van hoofdstuk 6 van dit proefschrift vormt (ik heb een paar plaatjes van je gejat, dacht dat je dat wel ok zou vinden...)

Buiten de academische muren bedank ik mijn vrienden, die er gelukkig voor zorgden dat er ook wat te beleven viel, daar buiten die muren.

Tenslotte wil ik nog de belangrijkste mensen in mijn leven bedanken. Pap en Mam, ontzettend bedankt voor alles, dit boekje is voor jullie. Jeroen, het is fijn om jou als broer te hebben. En tenslotte bedank ik jou, Carolien, simpelweg omdat je mijn súpermeisje bent!

Werner de Bruijn, 14 augustus 2004.

Curriculum Vitae

Werner Paulus Josephus de Bruijn was born in Tilburg, The Netherlands, on June 15, 1973.
In 1991, after receiving his “Gymnasium β” diploma from the Pauluslyceum in Tilburg (1985-1991), he started his studies of applied physics at Delft University of Technology. He chose to do his graduation project at the Laboratory of Acoustical Imaging and Sound Control, where he did an investigation of techniques for the recording and reproduction of reverberation for application in Wave Field Synthesis and other multi-channel sound reproduction systems. He received his M.Sc. degree in 1998.
Having developed an enthusiasm for doing scientific research and for the work on Wave Field Synthesis that was being carried out at the Lab. of Acoustical Imaging and Sound Control, he then started his Ph.D. research on the application of WFS in videoconferencing, in a project sponsored by France Télécom R&D. This research resulted in the present thesis.
In May 2003, after his contract at the university had ended, he started his new job as senior scientist in the Acoustics & Sound Reproduction cluster of the Digital Signal Processing group at Philips Research Laboratories in Eindhoven, The Netherlands, where he is currently working on the development of new audio features for consumer products.