Audio Engineering Society

Convention Paper 5801

Presented at the 114th Convention
2003 March 22–25 Amsterdam, The Netherlands

Application of Wave Field Synthesis in life-size videoconferencing

Werner P.J. de Bruijn[1] and Marinus M. Boone[1]

[1]1Laboratory of Acoustical Imaging and Sound Control, Delft University of Technology,
Lorentzweg 1, 2628 CJ, Delft, The Netherlands.

ABSTRACT

Spatial reproduction of the voices of conference participants can greatly enhance the performance of a life-size videoconferencing system in terms of qualities such as speech intelligibility, speaker identification and more generally the naturalness of a conference. A very suitable technique to implement accurate spatial sound reproduction including depth is Wave Field Synthesis (WFS). This paper presents results of research that has been carried out to investigate the combination of WFS with 2D video projection, including subjective experiments on sound localization, correspondence of perceived auditory and visual source directions and speaker identification in situations with multiple speakers, as well as speech intelligibility tests, investigations on the applicability of Distributed Mode Loudspeakers in WFS and coloration artifacts due to discretization of the loudspeaker array.

0. INTRODUCTION

In today's globalized world, efficient communication systems are a necessity. In an ideal situation, we would be able to have a virtual meeting with people anywhere in the world in a completely natural way, as if we were all present in the same room. This would require a system that has the capability to provide us with both a realistic visual and acoustical representation of 'the other side' and vice-versa. Many different systems for videoconferencing exist that aim to achieve this, all with different target audiences, ranging from small PC based desktop systems for one-to-one or one-to-many applications to systems with large video walls for many-to-many applications on which a complete meeting room is projected life-sized, so that the video screen appears to be a virtual window to the other side. Such systems are usually permanently installed, for instance in boardrooms of large multinational companies.

The performance of videoconferencing systems, especially these large, life-sized, systems, can be enhanced greatly by the addition of spatialized audio reproduction. Spatially separating the voices of remote conference participants facilitates the identification of individual speakers and it is a well known fact that this also improves speech intelligibility. Additionally, including a proper reproduction of auditory 'depth' and room acoustics can increase the overall 'naturalness' of a virtual meeting.

The audio reproduction parts of most current videoconferencing systems fail to provide this natural spatial sound reproduction in a large listening area, using conventional techniques such as stereophony or a small number of discrete loudspeakers.

A sound reproduction technique that seems very suitable for providing a natural spatialized sound reproduction in videoconferencing is Wave Field Synthesis (WFS). Using arrays of many small loudspeakers, this technique is able to synthesize the sound field of any sound source in a highly natural way, including proper reproduction of depth, in a large listening area. Therefore, a research project was started to investigate the requirements, benefits and possible limitations of using this technique in a life-size videoconferencing system. This paper gives an overview of the results from this research.

First, a short review of the WFS technique is given. Then, in section 2, several audio-visual experiments are described that address the need for vertical localization and the problems that occur when spatialized, true-perspective audio is combined with 2D video projection. Also, in section 2.5, we propose a way to avoid these problems.

Section 3 describes an experiment in which the improvement of speech intelligibility by using WFS is investigated.

A subject that was also studied in this project is the application of so-called Distributed Mode Loudspeakers in WFS. This is the subject of section 4.

Finally, section 5 deals with coloration artifacts that are possibly introduced by using discrete, instead of continuous, loudspeaker arrays.

1. WFS

In the late eighties a fundamentally new concept for sound reproduction was proposed by Berkhout; see e.g. Berkhout [1] and Berkhout et al. [2]. In this new concept, wave theory plays an essential role and individual loudspeakers are replaced by loudspeaker arrays (or 'loudspeaker-strips') that generate wave fronts from true or notional sources. Unlike all existing methods, the wave front solution is a so-called volume solution that generates an accurate representation of the original wave field in the entire listening space (and not at one or a few listening spots).

In the ideal situation the listening area is surrounded by planes of loudspeakers, which are fed with signals so that they produce a volume flux proportional to the normal component of the particle velocity of the original sound field at the corresponding position.

For practical purposes, this method has been adapted to make use of linear loudspeaker arrays surrounding the listening area, rather than planes of loudspeakers. It can be shown [3] that for linear arrays the input signals of the loudspeakers are given by:

$$E(\omega) = K\sqrt{jk}V_n(\vec{r},\omega), \qquad (1)$$

where $V_n(\vec{r},\omega)$ equals the normal component of the particle velocity, virtually at the loudspeaker position $\vec{r}$, k is the wave number and K is a constant depending on the loudspeaker sensitivity, the distance between the loudspeakers and the desired sound pressure of the reproduction. In case of loudspeakers with a flat frequency response, K is frequency independent.

The WFS concept can be applied for the reproduction of virtual sources which can be behind or in front of the arrays, because WFS can simulate any wave field shape, with convex or concave wave fronts.

2. AUDIO-VISUAL EXPERIMENTS

2.1. Vertical Localization

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 subsection 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.

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

- **Single-speaker reproduction**: The source is positioned at a specific vertical position by simply sending the source signal to the single speaker, out of a vertical array of speakers, that is closest to the desired source position.
- **Wave Field Synthesis reproduction**: The 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.
- **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.

2.1.1 Source Material

As source material for the experiment an audio-visual recording was made of the head of a person speaking 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.

2.1.2 Experiment Set-Up

The experiment set-up is illustrated in figure 1. A 15-element electro-dynamic loudspeaker array with a speaker distance of 12.7 cm was placed vertically at the side of a projection screen. All 3 reproduction methods could be handled by this single array by appropriately changing the driving signals of the loudspeakers. 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 centre speaker located at the desired source position.
- **Phantom source imaging**: The two loudspeakers directly above and below the screen were used (see figure 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. 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 reproduction level, comparable to a 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 1).

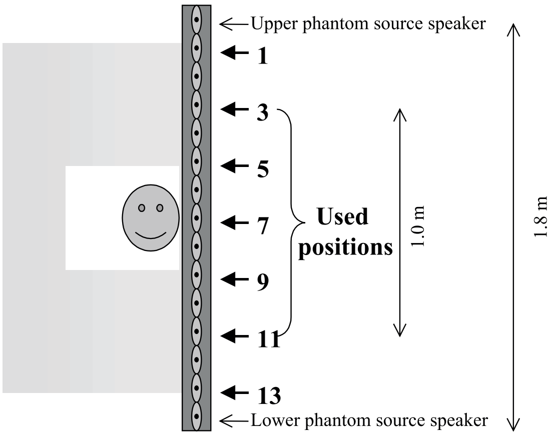

Figure 1. Schematic drawing of the loudspeaker set-up used in the vertical localization experiment.

Because it was expected that vertical localization becomes more critical when videoconference participants approach the screen, due to the larger effective angles between neighboring auditory source positions and between auditory and visual source positions, the experiment included two listening positions:

- **Position 1**: 1.5 m in front of the array.
- **Position 2**: 3.0 m in front of the array.

2.1.3 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 videoconferencing 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?

It seemed unnecessary to have a separation smaller than about 10 degrees between neighboring source positions at the listening positions, because this is about the localization accuracy for a familiar voice in front of the listener. Therefore it was decided to use 7 source positions separated by two loudspeaker distances (.25 m) with the centre position in the middle 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 centre position) and 8.2 degrees for the outer positions.

To obtain an unbiased response from the 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 centre. 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 1).

To investigate the audio-visual interaction, the experiment was carried out both with and without the video image present.

The hypothesis that we wanted to test was that the effect 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 centre position (which already is located at the position of the video image), while for the centre position itself the mean would be more firmly anchored to the centre, as reflected by the mean bias and standard deviation of the observations for the centre position.

12 subjects participated in the experiment. Subjects were free to move their head. Numbers on the array indicated the 13 positions from which subjects could choose.

The stimuli used were fragments with duration of 4 seconds of the source A/V material that played in a loop.

For each of the 4 combinations of listening position and video/no-video, three sequences of 20 stimuli each, in which all of the 5 used source positions were presented 4 times in random order, were presented. 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.

2.1.4 Results

Table 1 shows a summary of the results of the vertical localization experiment. For the ‘No Video’ condition three columns are shown: the ‘mean absolute bias’ *<bias>* (defined as the average over source positions and subjects of the absolute value of the difference between the mean of the responses and the true source positions), the sample standard deviation of the responses <s> and the ‘mean signed error’ <ε>, which will be explained later on when the effect of the audio-visual interaction is quantified.

	No Video			Video	
	<bias> (m)	**<ε> (m)**	**<s> (m)**	**<ε> (m)**	**<s> (m)**
Single Speaker Position 1	0.06	-0.07	0.17	+0.05	0.20
Single Speaker Position 2	0.03	+0.01	0.22	+0.14	0.22
WFS Position 1	0.05	-0.04	0.22	+0.08	0.21
WFS Position 2	0.08	+0.01	0.31	+0.10	0.22
Phantom Sources Position 1	0.11	-0.12	0.40	+0.02	0.30
Phantom Sources Position 2	0.06	-0.06	0.34	+0.10	0.28

Table 1. Summary of the results of the Vertical Localization experiment.

Looking at the results for ‘No Video’ we observe that the results for ‘single speaker’ and ‘WFS’ are similar: the mean absolute biases are small and the standard deviations are as could be expected for real sources. For ‘phantom sources’ we see that at position 1 the mean bias is significantly larger and the standard deviation is about twice as large as for the other two methods. This can easily be explained by the fact that typically subjects did not perceive a well-defined phantom image between the two loudspeakers but either heard the sound coming from one of the two loudspeakers or perceived an unclear image. This is very clear when the results are studied in detail. The interested reader is referred to a paper presented at a previous AES Convention, which deals with the vertical localization experiment in detail [4].

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 has to reflect the direction of the error relative to the image position (towards or away from the centre) 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. This 'mean signed error' <ε> is given in table 1 for both the 'no video' and 'video' condition. It is seen that in all cases a mean shift towards the image of the order of 0.1 m was observed, the average shift being 0.12 m.

Also, an analysis of variance (ANOVA) was done to investigate the statistical significance of the found differences between the mean errors in the situations with video and without video. This was done for all six combinations of reproduction method and listening distance. It was found that in all cases the effect of adding video was highly significant ($p<0.01$).

2.1.5 Conclusion

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). 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.

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.

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 decimeters, in which sources are assigned to the array closest to the true position.

2.2. Combining 2D video and true-perspective audio

Ideally, an audio-visual system should render a reproduction that is perceived as being a completely natural representation of the reproduced scene, both visually and acoustically, including a realistic impression of depth in both modalities.

For the visual part this would require a system that is able to create a 3D image of the scene that is to be reproduced. Although much research is being done in the development of such systems, it is at the moment not yet feasible to have a system that is able to provide a stable, high quality, multi-viewpoint 3D image, especially not in applications where several people have to be able to observe the scene simultaneously and where the image has to be recorded, transmitted and reconstructed in real-time, as is the case in for example life-size video conferencing or live television applications. Therefore, at least in these types of applications, the only feasible option is still a 2D projection of the real 3D visual scene.

For the acoustic part the situation is different. With Wave Field Synthesis (WFS) it is possible to achieve a natural reproduction of sound sources in which the true perspective is maintained.

It is attractive to use the best possible reproduction available for both modalities, which in the case of life-size video conferencing would be a combination of large-screen (scale 1:1) 2D video projection and Wave Field Synthesis sound reproduction. However, the benefit of combining a very natural audio reproduction with a less than perfect video projection is not as trivial as it might seem, because of the interaction that occurs between the visual and auditory modalities.

Basically, there are two possible problems:

- **Depth mismatch**: because several important visual cues for distance perception are lost in 2D projection there will be different, possibly disturbing, perceptions of depth for the auditory and visual modalities.

- **Source direction mismatch:** because there is only one unique position for which the perspective of the 2D video projection is correct ('the viewpoint') there will be a discrepancy between the perceived auditory and visual source directions for observers not located at the viewpoint.

Both phenomena are covered in more detail in a paper presented at a previous AES Convention [5]. A visual summary is shown in figure 2.

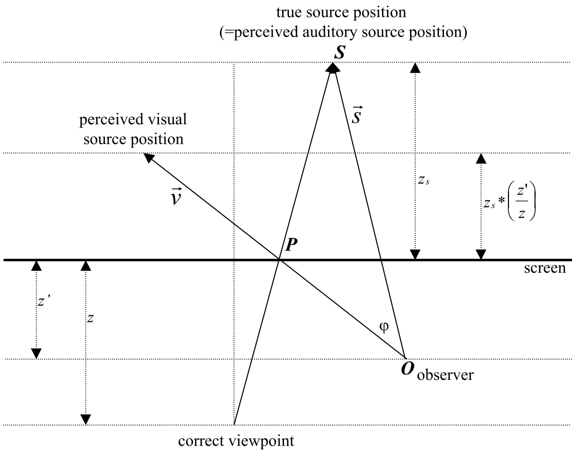

Figure 2. Mismatch of auditory and visual source direction and distance for an observer not located at the correct viewpoint.

To get more insight into these issues several subjective audio-visual experiments have been carried out, which are described in subsection 2.3 and 2.4.

2.3. Single Source A/V Experiments

2.3.1 Audio-visual source material and set-up

For the experiments that are described in this and the next subsection a visual projection was needed that had a true perspective (scale 1:1 in all dimensions, including depth) when viewed from the viewpoint. For this purpose a visual scene was constructed with one person standing at three different positions in a room. The visual scene is shown in figure 3. The source coordinates are given in table 2.

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

For all experiments that are described in this section the same loudspeaker set-up was used, consisting of a horizontal array with a total of 32 small loudspeakers with a spacing of 12.7cm located behind the (acoustically transparent) screen.

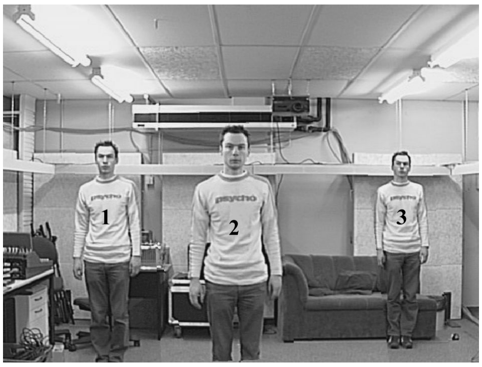

Figure 3. The perspective image that was used in the audio-visual experiments.

	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 2. Source coordinates *(x,z)* in meter, relative to both viewpoint and projection screen.

2.3.2 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 variable in this experiment was therefore the lateral position of the source.

The experiment was set up in the following way: the perspective visual scene of figure 3 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 centre 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 position having the depth level corresponding to the true source position (z-coordinates in table 2). The initial lateral position was chosen at random.

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.

The subjects received no feedback about the position at which the 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.

Using this procedure we get an estimation of the lateral sound source position interval that can be considered to correspond 'naturally' to each combination of source- and observer position. If for a certain source the obtained intervals for different observation positions 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 being 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:

- **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 1m 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).

6 normal hearing subjects participated in the experiment.

The SPL at listening position 1 (the viewpoint) corresponded to a typical SPL for speech of a person located at the virtual source position.

The results for observers at the viewpoint were as expected: the means of the subjects’ responses closely matched the true positions.

Figure 4 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 does not occur for source position 3 probably arises from the fact that in this particular geometry 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', which was reported by the subjects to be highly unnatural. In this case the subjects preferred to position the sound source completely at the position of the visual image.

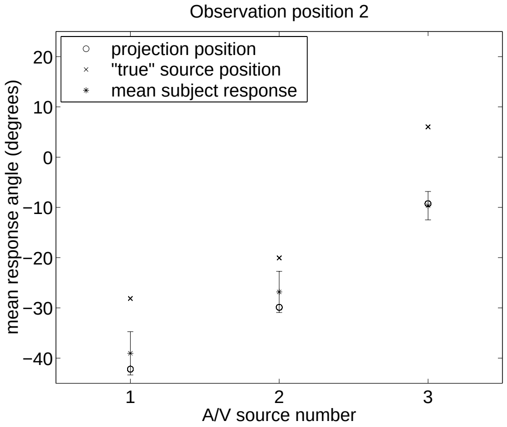

Figure 4. Localization responses of all the subjects for observation position 2. The asterisks are the means of the subjects' responses as seen from this position. Also the standard deviation of the responses is shown. The open circles indicate the direction of the visual image as seen from this observation position and the crosses represent the "true" source positions.

The results for observation position 3 (on axis, too close to the screen) were as expected: the mean of the subjects' responses matched the true position for source 2 (which is on an on-axis position) and was shifted slightly towards the visual position for both position 1 and 3.

Figure 5 gives an overview of the results. 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 centre of the screen, intervals contain 75% of the subjects' responses) for the three different observation positions (three bars for each source number, marked with a different symbol for each of the three observation positions).

As can be seen, for sources 1 and 3 there is no region where the preferred intervals of all three observation positions overlap, so that for those sources the reproduction will not be perceived as being natural by all observers in the room when the sources are reproduced at their 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.

For a more detailed description of this and the following experiments the reader is referred to [5].

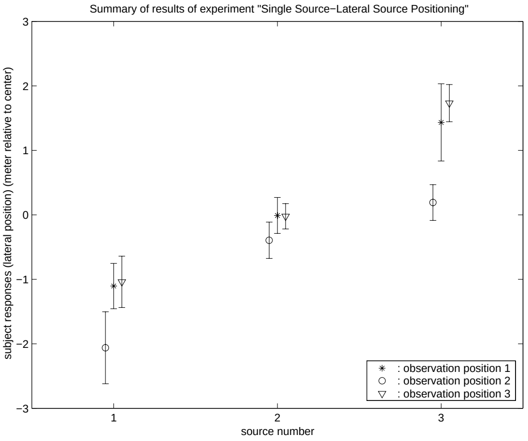

Figure 5. Summary of the results of the single source ‘Lateral Source Positioning’ experiment. Horizontal axis: source number, vertical axis: lateral source position intervals preferred by subjects (bars contain 75% of subjects' responses).

2.3.3 Discrepancy grading

Although the results of the ‘lateral source positioning’ experiment described above give a good indication of the source position ranges that are perceived as most natural for several observation positions, they do not give an actual indication of the 'degree of annoyance' that subjects feel when the source is located at a position that they perceive as not being the optimum position. Therefore, an experiment was carried out to investigate the subjective discrepancy between perceived audio and video source positions at different observation positions when the sound source is positioned at the 'true' position.

For this a 5-point impairment scale according to ITU standards 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’

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. 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 same 6 subjects participated as in the previous experiment.

In table 3 the discrepancy grading results are given. Shown are the means of all the subjects' grades, the standard deviation of the responses and the length of the 95% confidence interval of the mean.

As expected, at observation position 1 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.

		Source 1	**Source 2**	**Source 3**
Observation Position 1	mean grade	1.4	1.7	1.6
	stand. dev.	0.7	0.8	0.7
	95% conf. int.	±0.3	±0.3	±0.3
Observation Position 2	mean grade	3.5	2.3	3.4
	stand. dev.	1.1	1.1	0.9
	95% conf. int.	±0.4	±0.4	±0.3
Observation Position 3	mean grade	1.6	1.5	1.6
	stand. dev.	0.7	0.7	0.6
	95% conf. int.	±0.3	±0.3	±0.2

Table 3. Results of the single source ‘Discrepancy 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 source positions.

For observation position 2 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 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.

The results for observation position 3 are comparable to those for 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.

In conclusion, comparing the grading results for observation positions 1 and 3 to those of position 2, we see that a rather serious degradation in 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.

2.4. Multiple Source A/V Experiments

As explained in the introduction, 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 2.3 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 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.

2.4.1 Speaker identification

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-voice situation, as compared to stereophonic reproduction using two loudspeakers at the sides of the screen and reproduction with a configuration of discrete loudspeakers.

The visual set-up for this experiment was the same as in the experiments described in section 2.3.

The procedure was as follows: the computer chose one of the 3 projected persons as the ‘target speaker’. The ‘target’ speech signal was then routed 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 as closely as possible (so they corresponded to the projected images on the screen when viewed from this position).

The target speech signal 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 methods were as follows:

- **WFS:** the three sources were reproduced using the 32-element loudspeaker array.
- **Stereo:** reproduction was done with one array loudspeaker to each side of the screen. The balance of the levels of left and right loudspeaker signals was calculated according to the ‘law of sines’ for stereophonic reproduction.
- **Discrete:** 5 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.

The experimental procedure was as follows: the three signals (target speech- 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.

Subjects performed the task at three observation positions: positions 1 and 2 as defined in section 2.3 and a new position 4. Position 3 was not used in this experiment since it was found from the experiments described in section 2.3 that for this source position the results are similar to those in the viewpoint. Therefore this position was replaced by observation position 4, which is symmetrical to position 2 with

respect to the line through the viewpoint, but for which the perspective of the audio-visual scene is different, due to the non-symmetrical audio-visual source configuration.
11 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.

Table 4 gives an overview of the results of the experiment. For detailed results the reader is again referred to [5].
For observation position 1 it is clear from the results that the subjects had no problems whatsoever to correctly identify the target speaker.
For observation position 2 we see that several things are different from the results for position 1. A detailed look at the results for WFS shows a clear shift in identification to the right, especially for a source at position 1, which was in many cases perceived as corresponding to position 2. 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 up completely and the sound is heard as coming almost entirely from the loudspeaker that is closest to the subject. In the case of discrete reproduction the identification is only slightly less perfect as for the viewpoint.
For observation position 4 we observe the same phenomenon for stereo 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, subjects reported that they found the task easier at this position than at position 2.

	Position 1	**Position 2**	**Position 4**
Discrete	100%	90%	98%
WFS	98%	71%	80%
Stereo	94%	59%	64%

Table 4. Overview of the results of the multiple source 'Source Identification' experiment. Shown are percentages of correct identifications for all combinations of reproduction method and observation position.

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 up completely for off-axis observers.

2.4.2 Realism grading

Although the results of the experiment described above give an indication how well subjects were able to identify a specific speaker out of several competing speakers, they say 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 actual 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 methods, source positions and observation positions that were used in the previous experiment. Therefore the following experiment was carried out.
After completion of the experiment of section 2.4.1 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. 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 table 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.

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.

		WFS	Stereo	Discrete
Observation Position 1	mean grade	2.5	2.4	1.8
	stand. dev.	1.2	1.1	1.1
	95% conf. int.	±0.2	±0.2	±0.2
Observation Position 2	mean grade	3.2	4.5	2.0
	stand. dev.	1.0	0.8	0.9
	95% conf. int.	±0.2	±0.2	±0.2
Observation Position 4	mean grade	2.8	4.7	1.7
	stand. dev.	1.1	0.6	0.7
	95% conf. int.	±0.2	±0.1	±0.2

Table 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.

2.5. Depth Compression

From the experiments described in section 2.3 and 2.4 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.

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 perceptible discrepancies occur. The principle is sketched in figure 6 for the source and observer configurations that were used in the experiments described in sections 2.3 and 2.4.

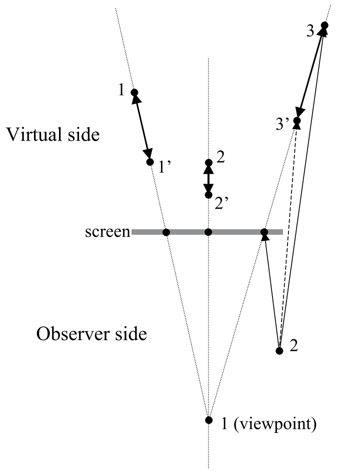

Figure 6: Principle of compression of the depth of the auditory scene to avoid discrepancies between perceived auditory and visual source directions. Sources 1, 2 and 3 are reproduced from positions 1', 2' and 3', respectively, closer to the viewpoint. For the combination of observation position 2 and source position 3 it is shown how applying a compression factor of 0.52 reduces the mismatch from 15 degrees (uncompressed) to the acceptable value of 11 degrees (compressed).

The optimum compression factor 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 (which will be discussed in section 3) are preserved. As a reference the maximum angle for which the 'ventriloquist' effect is effective can be taken, which is about 11 degrees. This is the maximum angle for which no discrepancy is perceived, so if instead the requirements are that no *annoying* discrepancies should occur, then the maximum mismatch angle that can be allowed is larger. With this criterion in mind the necessary compression factor can be determined, defined as the ratio between the distance between the 'compressed' source position and the projection on the screen and the distance between the 'true' source position and the projection on the screen (so 'no compression' corresponds to a factor of '1' and 'full compression' (sound source located at the position of the projection on the screen) corresponds to a factor of '0'). Preferably the compression factor 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' it is determined for which observation position on the other side the largest discrepancy is expected to occur, after which the compression factor that is needed to reduce this expected discrepancy to the maximum allowable value can be determined. This results in a 'map' of the source area that assigns a compression factor to each source position within that area. As an example, figure 6 shows how for an observer at position 2 the original discrepancy for source position 3 of 15 degrees is reduced to the acceptable value of 11 degrees by applying a compression factor of 0.52 (i.e.: the distance between the virtual sound source and its corresponding projection on the screen is reduced to 52% of its original value).

Since from the experiments described in section 2.3 and 2.4 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 an informal test was carried out which confirmed this assumption.

3. SPEECH INTELLIGIBILITY

The results of the audio-visual experiments described in section 2 indicate that the advantageous effects of using true-perspective spatialized audio reproduction in life-size videoconferencing are not as trivial as expected when the video part of the system consists of a conventional 2D projection, as will normally be the case. However, it was still expected that applying WFS can significantly improve the system performance, not only in terms of overall 'naturalness' of the reproduction, but also in terms of speech intelligibility, a very important quality for a speech communication system. This was expected because it is well known that spatially separating competing speech signals is highly beneficial for improving intelligibility. Systems applying conventional stereophonic or discrete multi-loudspeaker techniques are also able to reproduce individual voices in a spatially separated way, but since their reproduction does not include proper reproduction of the depth of the auditory scene, there is no separation of sound sources that are located more or less in the same direction when seen from the viewpoint of the video projection. In reality though, these sources, although located on the same straight line when seen from the viewpoint, should be perceived as being located in different directions when seen from other, non-viewpoint, positions.

Conventional systems are not able to reproduce this "listener position dependent" spatial separation of sound sources, while WFS is because it accurately reproduces the sound sources (virtually) at their true positions. For this reason it can be expected that in these situations the speech intelligibility will be better for WFS. This is illustrated in figure 7.

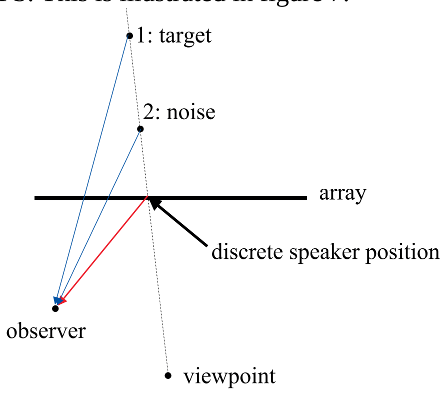

Figure 7: When voices are reproduced by WFS, the sources are spatially separated, while they are not when a discrete or stereophonic loudspeaker system is used in situations where the two sources are located on the same line through the viewpoint of the video projection.

For this reason an experiment was carried out in which the so-called ‘Speech Reception Threshold’ (SRT), the signal-to-noise ratio at which subjects have a 50% chance of repeating a standardized speech sentence correctly, was determined for both WFS reproduction and a discrete loudspeaker reproduction. This was done for two source configurations and two listening positions: the viewpoint of the video projection and a position 1 m off-center and 1 m too close to the screen. The two source configurations (each consisting of a target speech source and an interfering speech noise source) were chosen such that when seen from the viewpoint the target and noise sources were located on the same straight line, while they were separated by an angle of 10 degrees when perceived from listening position 2. The two configurations are shown in figure 8.

The results of the experiment confirmed the hypothesis: at the viewpoint, where the noise and target source are perceived from the same direction for both WFS and discrete loudspeaker reproduction, no significant difference in SRT was found, while for listening position 2 a highly significant SRT difference of up to 2.25 dB was found, implying an improvement of speech intelligibility in terms of ‘percentage of sentences repeated correctly’ of up to 38% when WFS is used.

The speech intelligibility experiment is the subject of a separate paper presented at this AES Convention, in which the experiment is described in more detail [6].

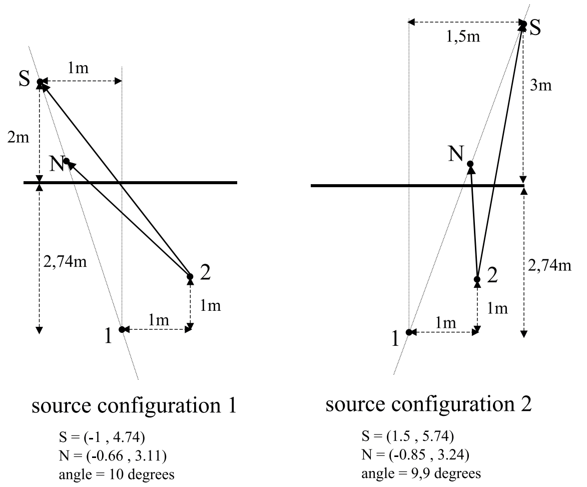

Figure 8: The two source configurations and the listening positions for which the speech intelligibility was investigated.

4. APPLICATION OF DML PANELS IN WFS

In the past couple of years a new loudspeaker technology has been introduced in the audio world called 'Distributed Mode Loudspeaker' (DML). Basically they consist of a flat panel of some light, stiff material to which one or more electro-dynamic exciters, comparable to the magnet/voice coil system of a conventional electro-dynamic loudspeaker, are attached. The audio signal is sent to the exciter which converts the electrical signal into a mechanical vibration, thereby forcing the panel material to vibrate, which in turn leads to the panel radiating acoustic energy into the surrounding air. The working mechanism of such a DML panel is principally different from a conventional loudspeaker in the sense that whereas a conventional cone loudspeaker ideally moves as a whole piston-like surface, the radiation mechanism of DML panels is through bending waves that travel across the surface of the panel.

For the wide-spread application of WFS it could be very attractive to use arrays of DML panels instead of conventional loudspeakers, for several reasons. First of all, the material that is used for the panels is very light and flat, making it possible to mount them on the wall for instance, which would take away some of the potential practical objections against installing a WFS system. In the particular case of videoconferencing or audio-visual applications in general, there is the additional advantage that the DML panels can be used as a projection screen at the same time, so that the screen and the WFS loudspeaker array are fully integrated.

Although the properties of DML panels described above make them attractive for WFS applications it is not trivial that WFS reproduction is actually possible with them. In several publications about DML panels it is said that the sound field they radiate is much more diffuse than that of conventional loudspeakers, while for proper Wave Field Synthesis it is required that the individual secondary sources can be controlled in a deterministic way. For this reason a series of pilot experiments was carried out, to investigate whether in principle DML's can be used for WFS reproduction.

The starting point was an array of nine individual small DML panels connected side-to-side, each one driven by an individual signal. This is a concept we refer to as a "multi-panel, single-exciter" array. Figure 9 shows an example of the 9-panel "multi-panel, single exciter" array synthesizing a virtual point source 1 m behind the centre of the array. It is seen that indeed a wave front is built up, comparable to what would be expected for an array of conventional loudspeakers. This can be explained by the fact that, although the impulse response of a DML panel has a 'stochastic' tail, the initial pulse is still very deterministic and strong, as is required for use in WFS. Based on the results obtained with this 9-panel array it could be concluded that it is indeed possible in principle to do WFS with DML panels.

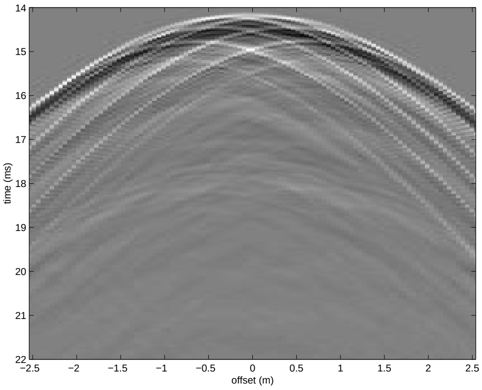

Figure 9: Measured sound field of a 9-panel DML array (spacing: 22 cm) synthesizing a virtual point source 1 m behind the centre of the array (measured at a line parallel to the array at a distance of 3 m).

A disadvantage of making an array out of small individual panels is that the frequency response is limited on the low end of the spectrum by the physical dimensions of the panels. This led us to the idea of attaching multiple individually driven exciters to a single larger piece of material, so this single large panel effectively behaves like an array of loudspeakers. This concept, which we refer to as a "single-panel, multi-exciter" array, should improve low-frequency response and additionally should make construction of WFS arrays easier and cheaper than when individual panels are used. Again however, it was not trivial this would actually work, because since the exciters are now attached to the same piece of material there is the risk of neighboring exciters influencing each other too much. Therefore, material is required that has sufficient internal damping to ensure that each exciter effectively acts like an individual loudspeaker. On the other hand: too high internal damping generally results in low efficiency and more harmonic distortion, so a compromise has to be found by choosing the right material.

Similar experiments were carried out as for the "multi-panel, single-exciter" array on several prototypes of multi-exciter panels. The first results indicated that in principle there are no problems, since the dominant initial response of each exciter appears to be highly localized at the position of the exciter on the panel. Figure 10 shows an example of the measured sound field of a 20-exciter DML array, consisting of three multi-exciter panels attached side-to-side, with 6, 8 and 6 exciters respectively, synthesizing a point source 1 m behind the centre of the array.

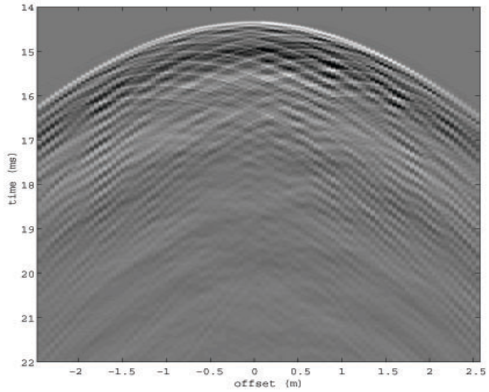

Figure 10: Measured sound field of a 20-exciter DML array consisting of three multi-exciter panels attached side-to-side, with 6, 8 and 6 exciters respectively (spacing: 12.7 cm), synthesizing a virtual point source 1 m behind the centre of the array (measured at a line parallel to the array at a distance of 3 m).

Details of the pilot study described above can be found in several papers presented at previous AES Conventions ([7], [8]).

After it had been established that WFS reproduction using DML panels according to the "single-panel, multi-exciter" concept is possible, the next step was to optimize the sound quality of the panels. Much can already be gained by choosing the right panel material. However, the frequency response will still be less flat than is required for high quality sound reproduction, so active filtering is necessary. The first simple approach to this was to apply a single FIR filter, that is the inverse of the spatially averaged frequency response of the array, to the source signal. This already improves the frequency response of the DML array in the sense that the overall response is flatter, but quite large fluctuations as function of frequency still remain because of differences in the response of individual exciters. To overcome this, the signals that are fed to the exciters have to be filtered individually. This further optimization of the DML panels for WFS reproduction is one of the main areas of research of the EC project 'Carrouso' [9] that TU Delft is involved in. Papers of Carrouso partners have been presented at previous AES conventions in which it is shown how this multi-channel filtering can be done efficiently, incorporating a compensation for the response of the listening environment in the same filtering process [10].

5. COLORATION

A consequence of using a discrete loudspeaker array instead of a continuous one is the introduction of spatial aliasing above the spatial Nyquist frequency which depends on the speaker distance:

$$f_{Nyq} = \frac{c}{\Delta x(\sin\theta_{max,source} + \sin\theta_{max,LS})}, \qquad (2)$$

in which c is the speed of sound in air, Δx is the distance between the loudspeakers, $\theta_{max,source}$ is the maximum spatial component present in the source field and $\theta_{max,LS}$ is the maximum spatial component radiated by the loudspeakers. Below this frequency the reconstructed wave field is identical to the desired wave field. Above this frequency however, the contributions of the individual loudspeakers do no longer interfere in a constructive way and they can be identified as individual contributions in the impulse response. The effects are clearly visible in the measurements shown in figures 9 and 10.

In the frequency domain this means that there will be a distortion of the frequency response above this frequency, which is position dependent. This is illustrated in figure 11.

The perceptual consequence of this is that the reproduced wave field exhibits spatial color fluctuations. Whether these color fluctuations can actually be perceived or even be annoying depends on several parameters, such as the distance between the loudspeakers, the frequency content of the source signal, the source position and the listener position. In general it can be said that the larger the spacing between the loudspeakers is, the larger the effect of the coloration will be. On the other hand: a larger distance between the loudspeakers (and thus smaller number of individual channels) is advantageous from a computational and system cost point-of-view. Therefore we want to investigate how strong these spatial color fluctuations are for WFS arrays with different loudspeaker spacings, in order to get an idea what the maximum acceptable spacing is (with regard to coloration).

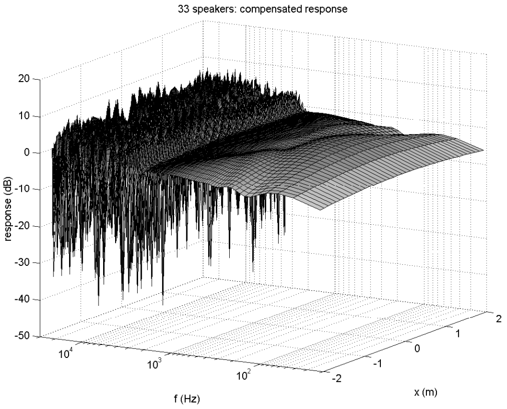

Figure 11: Frequency-domain example of the effect of spatial aliasing due to discretization of the loudspeaker array. The response of an array with spacing 12.5 cm is shown along a line parallel to the array. For frequencies above the Nyquist frequency the response is distorted.

To investigate this, a perceptual headphones experiment has been set up in which subjects compare simulated signals of different speaker configurations. Simulations have been made of the response of WFS systems with different loudspeaker spacings on a closely spaced grid of listening points. The loudspeaker spacings included in the experiment are 12.5 cm, 16 cm, 25 cm, 33 cm and 50 cm. Subjects carry out a 'paired comparison' test in which each trial consists of two pairs of signals:

- **Pair 1**: the binaural impulse response of loudspeaker system *A* at listening position *(x,z)*, convolved with a sample of speech noise (noise with the long-term averaged spectrum of either male or female speech) and the response of the same system at position *(x+offset,z)*, in which *offset* is chosen at random from the range -0.5 to 0.5 m in steps of 0.1 m.

- **Pair 2**: the same two binaural responses as pair 1, but now simulated for loudspeaker system *B*.

It is the subjects' task to indicate in which of the two pairs the two signals were the most different regarding color.

The main variable of interest in this experiment is the loudspeaker distance, so in one series of the experiment a subject compares the responses of all the different loudspeaker configurations to each other at different lateral positions *x* for one fixed value of the variables *source position, listening distance* and *male/female speech noise*. Then, in a next series, one or more of the fixed variables are changed. The results of all paired comparisons can then be analyzed to obtain a scale value on a one-dimensional 'coloration' scale for each of the configurations.

At the moment of finishing this paper the coloration experiment is being carried out and the results will be presented at the 114th AES convention.

6. CONCLUSIONS

This paper described research that has been carried out to investigate the application of Wave Field Synthesis in life-size videoconferencing. Although the focus was on WFS reproduction, it should be noted that several of the issues that were addressed are of relevance to audio-visual systems in general.

From the vertical localization experiment described in section 2.1 it could be concluded that vertical source resolution, which is already very non-critical in audio-only applications, is even less so in an audio-visual system such as a videoconferencing system.

From the experiments described in sections 2.3 and 2.4 it has to be concluded that adding true-perspective (in this case: WFS) audio reproduction to a 2-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.

A way to reduce the observed negative effects, compression of the reproduced depth of the auditory scene, was described in section 2.5.

In section 3 it was shown that using WFS sound reproduction in a speech communication system can greatly improve speech intelligibility, as compared to conventional systems.

Section 4 showed that Distributed Mode Loudspeakers are a technology that can very well be applied in WFS. Not only is it possible to build arrays of small individual panels, but it is also possible to construct WFS arrays by attaching multiple, individually driven, exciters to larger panels, which is especially attractive in audio-visual applications such as videoconferencing.

Finally, in section 5 the possible problem of coloration due to spatial aliasing caused by the use of discrete loudspeakers arrays was addressed and a subjective experiment was described, which at the time of writing is being carried out.

7. ACKNOWLEDGMENT

This work was supported by France Telecom R&D, Lannion, France.

8. REFERENCES

1. A. J. Berkhout, *'A Holographic Approach to Acoustic Control'*, J. Audio Eng. Soc., vol. 36, pp. 977-995, 1988.
2. A. J. Berkhout, D. de Vries, and P. Vogel, *'Acoustic Control by Wave Field Synthesis'*, J. Acoust. Soc. Am., vol. 93, pp. 2764–2778, 1993.
3. M. M. Boone and E. N. G. Verheijen, *'Multi-channel sound reproduction based on wave field synthesis'*, paper presented at the 95th AES Convention (paper 3719), New York, 1993.
4. Werner P.J. de Bruijn, Marinus M. Boone and Diemer de Vries, *'Sound localisation in a video conferencing system based on Wave Field Synthesis'*, paper presented at the 108th AES Convention (paper 5144), Paris, France, 2000.
5. Werner P.J. de Bruijn and Marinus M. Boone, *'Subjective Experiments on the Effects of Combining Spatialized Audio and 2D Video Projection in Audio-Visual Systems'*, paper presented at the 112th AES Convention (paper 5582), Munich, Germany, 2002.
6. Marinus M. Boone and Werner P.J. de Bruijn, *'Improving speech intelligibility in teleconferencing by using Wave Field Synthesis'*, paper presented at the 114th AES Convention, March 22-25 2003, Amsterdam, The Netherlands.
7. Marinus M. Boone and Werner P.J. de Bruijn, *‘On the applicability of Distributed Mode Loudspeaker panels for Wave Field Synthesis based sound reproduction’*, paper presented at the 108th AES Convention (paper 5165), Paris, France, 2000.
8. Marinus M. Boone, Werner P.J. de Bruijn and Wilfred van Rooijen, ‘*Recent developments on*

WFS for high quality spatial sound reproduction', paper presented at the 110th AES Convention (paper 5370), Amsterdam, The Netherlands, 2001.

9. http://emt.emt.iis.fhg.de/projects/carrouso/
10. E. Corteel, U. Horbach and R. Pellegrini, *'Multichannel Inverse Filtering of Distributed Mode Loudspeakers for Wavefield Synthesis'*, paper presented at the 112th AES Convention (paper 5611), Munich, Germany, 2002.