Audio Engineering Society

Convention Paper 5800

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

Improving speech intelligibility in teleconferencing by using Wave Field Synthesis

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

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

ABSTRACT

Large screen teleconferencing can be enhanced considerably with the application of spatial sound recording, transmission and reproduction. True spatial sound reproduction can be obtained with Wave Field Synthesis (WFS) which gives a sound reproduction that is independent of the listener position. Our research has shown that a significant improvement of speech intelligibility can be obtained with WFS as compared with a single loudspeaker reproduction, when there are several interfering speech signals. The improvement in Speech Reception Threshold (SRT) can be more than 2 dB, making a change in speech intelligibility from 50% to more than 85%.

1. INTRODUCTION

Wave Field Synthesis is a high quality spatial sound reproduction and sound synthesis system, developed since 1988 by the Laboratory of Acoustical Imaging and Sound Control of TU Delft [1, 2].

The method makes use of arrays of closely spaced loudspeakers which are fed with audio signals in such a way that a highly natural sound field is produced with wave front curvatures as would be obtained from real sources. In such a way an arbitrary number of so-called virtual sources can be reproduced simultaneously. Each virtual source is stored or transmitted as a separate sound channel, accompanied with meta-data describing the intended position of the source. Also information can be supplied to place the source in a virtual environment, based on reflection and reverberation parameters. The reflections and reverberation can be generated at the reproduction site at will. Alternatively, reflection and reverberation signals can be stored or transmitted as well. In such a way, WFS can very well be combined with the MPEG-4 standard as well as with standard 2/0 stereophonic or 3/2 surround recordings.
An interesting application of WFS is the combination with large screen video projection in for instance video-conferencing. An overview of this application is presented at the same Convention [3. It was shown in an earlier paper [4] that there may be a discrepancy between the observed visual direction and the perceived acoustical direction because of the two-dimensional visual projection that is normally used in

video systems. For ideal audio-visual reproduction the listener/observer should therefore be at the correct viewpoint. From that point of view it would be better for off viewpoint listeners to use a discrete loudspeaker reproduction for distinct screen areas instead of making the effort with WFS to include acoustic source depth information.

However, we did have the feeling that the inclusion of the depth information would not only make the audio experience more natural, but also give a better speech intelligibility. To test the improvement of the speech intelligibility, we carried out subjective tests by measuring the Speech Reception Threshold with discrete loudspeaker reproduction and with WFS. With the true acoustic perspective of WFS a spatial separation of different speech signals is possible, depending on the listener position, whereas this is not possible with discrete loudspeaker reproduction, because then the spatial setup remains constant for different listening positions.

2. The SRT method

A well known method to measure the speech reception threshold (SRT) is the standardized method developed by Plomp & Mimpen [5]. It is a subjective test to find the signal to noise ratio (SRT) at which the averaged understanding of speech sentences is 50%. Test persons listen to recordings of sentences spoken by well trained people and continuous noise with the same spectral content as averaged speech is included at a certain level. The test person has to repeat the spoken sentence.

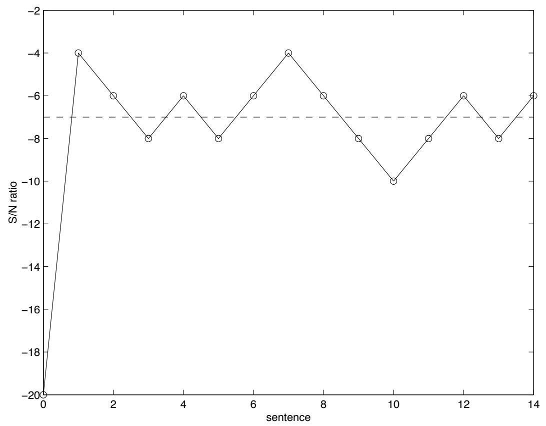

Figure 1: Example of the results of one series of 13 sentences for one subject. The horizontal axis denotes the number of the sentence in the list, the vertical axis is the SNR at which a sentence was reproduced. The SRT for this series is calculated as the mean of the SNR of the last 10 sentences, which in this case is -7 dB (dashed horizontal line).

Depending on the correct or incorrect answer, the level of the noise is changed up or down, giving a response curve as shown in figure 1. The test results are averaged to obtain the SRT under the given conditions. Usually the sentences are in the native language of the listener and are also spoken by a native speaker of the same language. In our experiments we used prerecorded speech material that was kindly supplied by Plomp's institute. These speech sentences were recorded with special level adjustments to balance the SRTs of different sentences during playback.

3. Experimental set-up

We tested combinations of two source configurations and two listening positions for both the WFS and discrete loudspeaker system, resulting in eight individual conditions. In each source configuration there was a target source (speech signal) and a noise source. Figure 2 shows the lay-out of the source and listener positions for both tested source configurations. The position of the noise source was chosen such that for an observer at listening position 1 it was located on the same straight line as the target source, while for an observer at listening position 2 the speech and noise source were spatially separated by an angle of 10 degrees. This applies for the WFS reproduction. With the discrete loudspeaker reproduction both signals were reproduced by one loudspeaker.

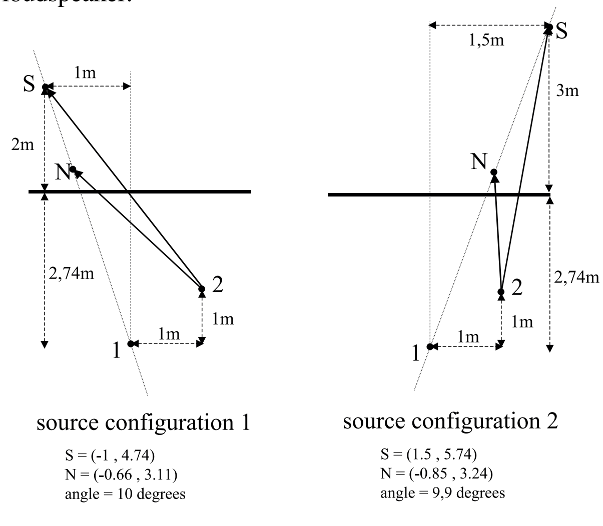

Figure 2: Source and listening positions used in the speech intelligibility experiment. On the left are the target (S) and noise (N) source positions for source configuration 1 and on the right those for source configuration 2.

The target source and noise source levels were calibrated to a value of 65 dB(A), corresponding to a normal speech level, for each of the 8 conditions. Thereby all parameters were kept as constant as possible and only the spatial separation remained as a distinct parameter.

The experimental procedure was manually controlled by an operator, who changed the noise level depending on the correct reproduction of the spoken sentence. The set-up with the two listening positions is shown in figure 3.

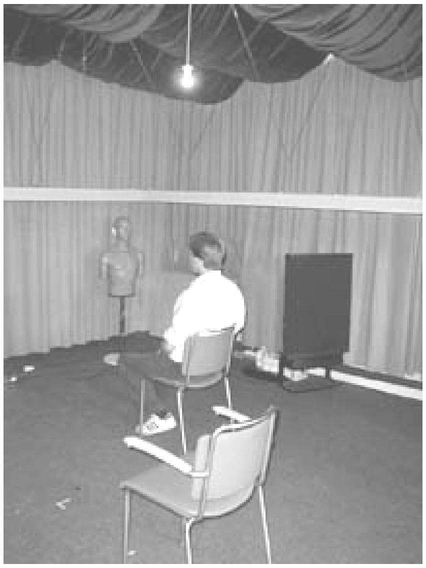

Figure 3: Experiment set-up with the two listening positions (chairs). The subject is sitting at position 2, the empty chair is listening position 1.

4. Results

There were 16 test subjects who participated in the experiments. The results are summarized in figure 4, showing the means and the 95% confidence intervals of the means for the different tested conditions.

The SRT differences between WFS and discrete loudspeaker reproduction are given in table 1, together with the p-statistic of a one-way ANOVA of the results (this number shows the probability that the differences between the means of the tested conditions are only by chance).

The results clearly show that there is a significant improvement of the SRT for WFS reproduction as compared to discrete loudspeaker reproduction at listening positions 2, confirming our expectation that the angular separation between target and noise source will help to discriminate between both sources. It even seems to be that there is a small increase in intelligibility with WFS at listening position 1, where this would not be expected. Until now we have not found a good explanation for that.

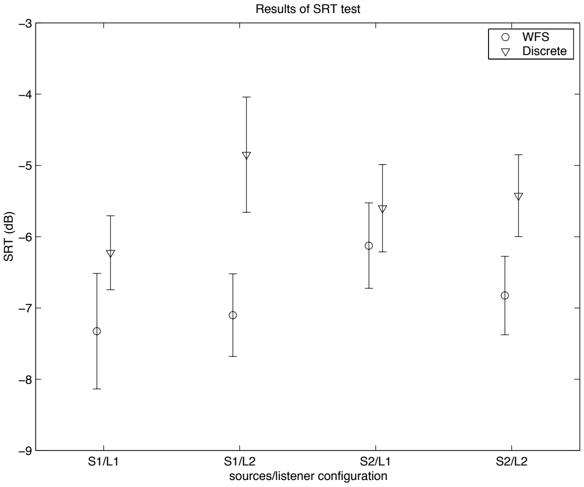

Figure 4: Results of the SRT experiment. The horizontal axis denotes the configuration of source and listener positions, 'S1/L2' meaning 'source configuration 1/listener position 2' (see figure 2). The vertical axis denotes the SRT (in dB) in terms of SNR. The open circles and triangles indicate the 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)	1.1	2.25	0.53	1.4
p=statistic	0.021	39×10^{-6}	0.20	750×10^{-6}

Table 1: 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 source/listening position configurations, showing the statistical significance of the difference in SRT means for the WFS and discrete loudspeaker set-ups in each configuration.

5. Discussion and conclusion

The results show that there is a significant difference in SRT for WFS and discrete loudspeaker reproduction, depending on the test conditions. Although the differences in dB are only small, the effects are significant. This is because the psychometric curve of the intelligibility score as a function of the signal to noise ratio is very steep, as shown in figure 5. Around the 50% intelligibility a change of 1 dB in signal to noise ratio has an effect of 20% on the intelligibility score. For the S1/L2 condition where the SRT difference is 2.25 dB this means that when the intelligibility score with single loudspeaker reproduction is 50%, this will improve with WFS reproduction to 88%, which is very significant. Hence, it can be concluded that the application of WFS in teleconferencing with many participants can be very advantageous, not only for the naturalness of the sound reproduction, but also for the understanding of speech under "cocktail party" conditions.

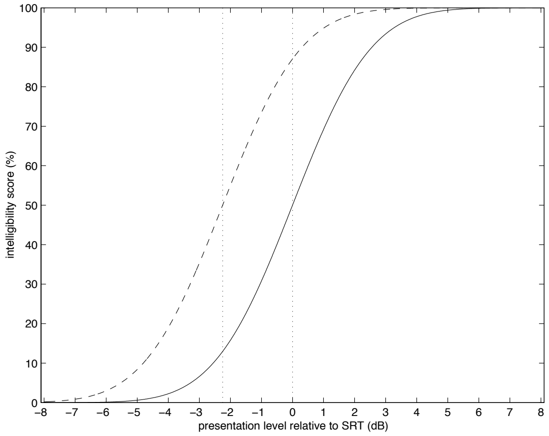

Figure 5: Solid line: Intelligibility score (% of sentences understood correctly) versus SNR relative to SRT. The solid line represents the results for an arbitrary condition 'A'. The dashed line is the same curve shifted 2.25 dB to the left, representing the results for a condition 'B' that has a mean SRT that is 2.25 dB lower than that of condition 'A'. Curves have a slope of 20%/dB at the 50% level (after Plomp & Mimpen [5]).

6. ACKNOWLEDGMENT

This work was supported by a project on teleconferencing, financed by France Telecom R&D, Lannion, France.

7. 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 . W.P.J. de Bruijn and M.M. Boone, "Application of Wave Field Synthesis in life-size videoconferencing", 114th Convention of the AES, 22 – 25 March 2003, Amsterdam,

4 . W.P.J. de Bruijn and M.M. Boone, "Subjective experiments on the effects of combining spatialized audio and 2D video projection in audio-visual systems", 112th Convention of the AES, 10 –13 May 2002, Munich, Convention paper 5582.

5 . R. Plomp and A.M. Mimpen, "Improving the reliability of testing the speech reception threshold of sentences", Audiology, 18 (1), p. 43-52, 1979.