Proceedings of the Institute of Acoustics

VIRTUAL SOURCE IMAGING OVER LOUDSPEAKERS

O. Kirkeby (1), P.A. Nelson (1), and H. Hamada (2)

(1) Institute of Sound and Vibration Research, University of Southampton, Highfield, SO17 1BJ, UK

(2) Department of Electrical and Communications Engineering, Tokyo Denki University, Tokyo 101, Japan

1. INTRODUCTION

At the Institute of Sound and Vibration Research at the University of Southampton, we have for more than five years been working in collaboration with Tokyo Denki University in Japan on using digital signal processing to improve the quality of sound reproduction systems. The ultimate goal is to be able to produce the illusion in a listener of being in a "virtual" acoustic environment which is entirely different from that of the space in which the listener is actually located. Sound systems designed for this purpose are often referred to as "surround sound systems", or "3D-sound systems"; we prefer to label such systems virtual source imaging systems.

A virtual source imaging system can convey the sound to the listener either via headphones or over loudspeakers. We use loudspeakers for the reproduction. Even though the multiple transmission paths from the loudspeakers to the listener's ears make it difficult to control the sound field directly, we are capable of controlling the interference of the sound waves at any number of points using any number of loudspeakers. In this paper, we outline the progress that has been made to date, and we also point out some of the limitations of this new technology.

2. BINAURAL TECHNOLOGY

The overwhelming part of current research into virtual source imaging relies heavily on binaural technology [1], [2], [3] (a notable exception is when large arrays of loudspeakers are used for the reproduction. In that case it is possible to synthesize the entire sound field under certain conditions [4]). Binaural technology is based on the sensible engineering principle that if a sound reproduction system can generate the same sound pressures at the listener's eardrums as would have been produced there by a real sound source, then the listener should not be able to tell the difference between the virtual image and the real sound source. In order to know these binaural signals, or "target" signals, it is necessary to know how the listener's torso, head, and pinnae (outer ears) modify incoming sound waves as a function of the position of the sound source. This information can be obtained by making measurements on "dummy-heads" or human subjects [5], [6]. The results of such measurements are usually called head-related transfer functions, or HRTFs.

In practice, HRTFs vary enormously between listeners, particularly at high frequencies [7]. Consequently, it does not make much sense to talk about the detailed characteristics of the HRTFs of an "average" listener (incidentally, the dummy-head systems currently available are not considered to be representative of a significant number of human listeners anyway [6]). The large statistical variation in HRTFs between listeners is one of the main problems with virtual source imaging over headphones [8], [9].

Headphones offer good control over the reproduced sound. There is no "cross-talk" (the sound does not run round the head to the opposite ear), and the acoustical environment does not modify the reproduced sound (room reflections do not interfere with the direct sound). Unfortunately, though, when headphones are used for the reproduction, the virtual image is often perceived as being too close to the head, and sometimes even inside the head. This phenomenon is particularly difficult to avoid when one attempts to place the virtual image directly in front of the listener. It appears to be necessary to compensate not only for the listener's own HRTFs, but also for the response of the headphones used for the reproduction [10], [11]. In addition, the whole sound stage moves with the listener's head (unless head-tracking is used, and this requires a lot of extra processing power [3]). Loudspeaker reproduction, on the other hand, provides natural listening conditions but makes it necessary to consider the effect of cross-talk and the reflections from the acoustical environment [12], [13], [14].

3. CROSS-TALK CANCELLATION SYSTEMS

The cross-talk cancellation problem is in a sense the ultimate sound reproduction problem since an efficient cross-talk canceller gives one complete control over the sound field at a number of "target" positions. The objective of the cross-talk canceller is to be able to reproduce a desired signal at any single target position while cancelling out the sound perfectly at all remaining target positions.

3.1 Cross-talk cancellation using two *widely spaced* loudspeakers

The basic principle of cross-talk cancellation using only two loudspeakers and two target positions has been known for more than 30 years. In 1966, Atal and Schroeder used physical reasoning to determine how a cross-talk canceller comprising only two loudspeakers placed symmetrically in front of a single listener could work. In order to reproduce a short pulse at the left ear only, the left loudspeaker first emits a positive pulse. This pulse must be cancelled out at the right ear by a slightly weaker negative pulse emitted by the right loudspeaker. This negative pulse must then be cancelled out at the left ear by another even weaker positive pulse emitted by the left loudspeaker, and so on. Atal and Schroeder's cross-talk canceller is both elegant and simple, and it is straightforward to implement using analogue electronics. However, it has never gained widespread use, and this is probably because of the following two problems with its performance.

First, the widely spaced loudspeaker setup they used (the loudspeakers span 60 degrees as seen by the listener) makes the system very sensitive to head movement and head rotation [16], and it also gives the reproduced sound an unpleasant character since the frequency responses of the processed signals have sharp peaks at harmonics of approximately 2kHz [17]. Secondly, the HRTFs that the system is based on are very inaccurate since the influence of the listener's head on the incoming sound field is ignored. A free-field model does not account for the extra distance the sound has to travel around the circumference of the head at low frequencies, the attenuation caused by the shadowing of the head at high frequencies, or the shaping of the spectrum caused by the pinnae at very high frequencies [18].

3.2 Cross-talk cancellation using two *closely spaced* loudspeakers

Atal and Schroeder's system works a lot better when the two loudspeakers are close together than when they are far apart [17]. When the loudspeakers span only 10 degrees as seen by the listener, the robustness of the system with respect to head movement and head rotation is increased significantly. Furthermore, the sharp peaks in the frequency response of the reproduced sound are shifted up to harmonics of approximately 10kHz which leaves the main part of the audio frequency range unaffected. Such a system can still be implemented with analogue electronics as long as one does not attempt to include the listener's head in the model. Even though it is unlikely that anyone would be interested in producing silence at one of the listener's ears, a cross-talk cancellation network must be implemented if one has access to the binaural signals only (such as those recorded by a dummy-head, or synthesised from measured HRTFs). In addition, some naturally occuring signals, such as a whispering voice close to the head, are much louder at one ear than the other.

3.3 Cross-talk cancellation using four loudspeakers

Systems comprising two loudspeakers and two target positions are historically the most important, but the cross-talk cancellation principle is applicable to systems comprising any number of loudspeakers and any number of target points [19]. Imagine a dummy-head with four "ears", two on the left side of the head, and two on the right side of the head. When a recording made with such a dummy-head is played back over four loudspeakers after being filtered by a 4-by-4 cross-talk cancellation network, the original sound field is reproduced accurately at four points in the vicinity of the listener's two ears.

4. VIRTUAL SOURCE IMAGING SYSTEMS

A real sound source produces interaural time- and level differences that are used by the auditory system to localise the sound source [18]. For example, sound waves approaching the listener from the left will be louder, and arrive earlier, at the left ear than at the right. A virtual source imaging system works by accurately reproducing these cues.

4.1 Virtual source imaging using the "stereo dipole"

We use the term "stereo dipole" to describe a virtual source imaging system that comprises two closely spaced loudspeakers [17], [20], [21]. The two loudspeakers ideally span 10 degrees as seen by the listener, and so their centres need to be only about 20cm-30cm apart under normal listening conditions. They can therefore both be contained in the same cabinet. Although closely spaced loudspeaker setups have not received much attention from researchers in the past [22], [23], it turns out that there are some striking similarities between creating a virtual source close to the centre between two widely spaced loudspeakers and creating a virtual source well outside the angle spanned by two closely spaced loudspeakers [24].

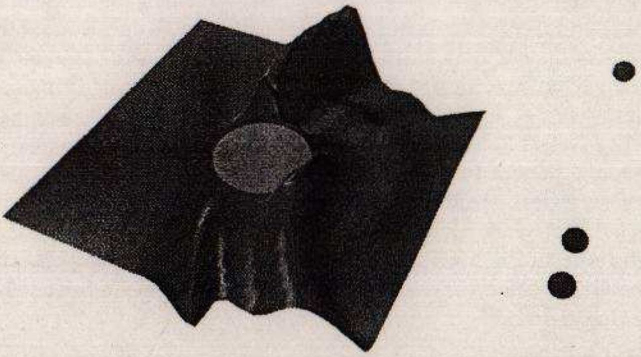

Figure 1. The sound field generated by two closely spaced monopole sources whose inputs have been adjusted to create a virtual source at 45 degrees to the left relative to straight front

An example of the sound field produced by the stereo dipole is illustrated in Figure 1. This figure shows a "snapshot" in time of the sound pressure produced by two closely spaced monopole sources. The listener's head is modelled as a rigid sphere whose diameter is 18cm [25]. The objective is to give the listener the impression that a relatively short pulse (most energy concentrated below 3kHz) is emitted from a virtual source positioned at 45 degrees to the left relative to straight front. It is seen that two wavefronts are radiating away from the two sources. The first wavefront is intended for the left ear. The second wavefront, whose amplitude is slightly smaller than the first, is intended for the right ear. The resulting interaural time- and level differences are identical to the ones generated by a sound source at the position of the virtual source.

It is important to realise that the performance of the stereo dipole deteriorates dramatically if the two loudspeaker inputs are not matched accurately. Even very small deviations from the optimal input signals upset the careful phase match that is required to reproduce the binaural signals at the listener's ears. Consequently, an implementation based on analogue electronics is bound to be extremely difficult because the components have to be within a fraction of a percent of their nominal values (and it is not possible to compensate for non-minimum-phase components in the plant). For the same reason, it is also very important that the two loudspeakers have very similar frequency responses. If the two loudspeakers are not well matched, the sound stage tends to appear asymmetric, or "lean to one side". Good quality full-range units usually work well together, but this is not the case for all two-way loudspeakers. Three-way loudspeakers rarely work well together. Finally, it must be mentioned that the two loudspeakers generally have to work quite hard at low frequencies. For virtual images well outside the angle spanned by the loudspeakers, frequencies below 500Hz are boosted by approximately 10dB.

Listening experiments have demonstrated that the stereo dipole can create very convincing virtual sources in front of the listener, and even sometimes behind the listener as well [26]. However, it is possible to create the illusion of a rear virtual source only when the listener's head is fixed. Any head movement will make the listener aware that the sound waves are generated by sound sources in the front hemisphere. In order to be able to create stable rear images, it is necessary to add a pair of loudspeakers behind the listener.

4.2 Virtual source imaging using four loudspeakers

By using four loudspeakers around the listener, rather than only two in front of the listener, it is possible to eliminate the front-back confusion problem. A "four-ear dummy head" can supply the necessary four tracks of "target" material. The four target signals can be synthesized from a database of HRTFs. For example, four "binaural" signals can be synthesized by combining the two signals obtained by rotating the head ten degrees clockwise and the two signals obtained by rotating the head ten degrees anti-clockwise [19]. Alternatively, a recording can be made with four microphones mounted in a sphere [27]. Such a "sphere dummy-head" includes the shadowing effect of the head but not the effect of the pinnae. Listening experiments have demonstrated that in addition to overcoming the front-back confusion problem, 4-by-4 systems offer good robustness with respect to head rotation [27].

The four loudspeakers can also be used to cater for two listeners rather than one [28]. For example, this could be useful in a car since the driver and the front seat passenger inevitably sit relatively close to each other. Indeed, there is nothing that prevents one from also including the backseat passengers in the model, but in that case even more loudspeakers are needed to control the sound field at the extra points.

5. DIGITAL FILTER DESIGN

Whether the objective is to design a cross-talk canceller or a virtual source imaging system, the fundamental problem to be tackled is one of multi-channel inversion [29]. Since the inversion techniques that are usually applied to common engineering problems [30] are not entirely appropriate for audio purposes, we have developed our own filter design methods [19], [31], [32], [33], [34]. These methods can determine a matrix of digital finite impulse response (FIR) filters that are optimal in a quantifiable sense.

The idea central to our filter design algorithms is to minimise, in the least squares sense, a cost function of the type

$$J = E + \beta V. \tag{1}$$

The cost function is a sum of two terms: a performance error E, which measures how well the desired signals are reproduced at the target points, and an effort penalty βV which is a quantity proportional to the total power that is input to all the loudspeakers. The positive real number β is a regularisation parameter that determines how much weight to assign to the effort term. By varying β from zero to infinity, the solution changes gradually from minimizing the performance error only to minimizing the effort cost only [35]. In practice, regularisation works by limiting the power output from the loudspeakers at frequencies at which the inversion problem is ill-conditioned. This is achieved without affecting the performance of the system at frequencies at which the inversion problem is well-conditioned. In this way, it is possible to prevent sharp peaks in the spectrum of the reproduced sound. If necessary, a frequency dependent regularisation parameter can be used to attenuate peaks selectively.

We always include a modelling delay in order to allow the optimal filters to compensate for non-minimum phase components in the plant [36]. We do not favour the use of minimum phase approximations [22] since these can alter the time structure of the original waveform.

6. PRACTICAL CONSIDERATIONS

The use of a modelling delay causes the reproduced sound to be delayed by a few milliseconds (typically 1ms-2ms, rarely more than 10ms). This can potentially be a problem in applications where perfect synchronization with, for example, images is crucial.

It does not seem to be of crucial importance to use individualised HRTFs in order to create a convincing virtual image, and at the moment we tend to rely on a single database of HRTFs measured on a KEMAR dummy-head at the MIT media lab (this database is free to download from the internet site http://sound.media.mit.edu/~kdm/hrtf.html). Nevertheless, not all listeners report the same perception of the reproduced sound. The perceived height varies between listeners, and it is quite common that a listener will consistently judge virtual images to the left to be higher, or lower, than virtual images to the right. This is probably because the listener's HRTFs are not symmetric with respect to the median plane (the left and right half of the listener's head are not exactly the same acoustically). Preliminary experiments suggest that the performance is improved when individualised HRTFs are used. In practice, it might be sufficient to give a listener the option of choosing between, say, ten different sets of HRTFs [37].

Room reflections will generally make the performance of the system worse than if the listening environment is anechoic [38]. It is not a problem in principle to compensate for the room response, but it does add colouration to the reproduced sound, and it also makes the system less robust with respect to head movement. The stereo dipole usually performs well in an ordinary listening environment, such as an office or a living room, as long as the listener sits in the direct field. If the listener, or the loudspeakers, are positioned close to a wall, or in the corner of a room, the performance of the system will inevitably suffer unless the early reflections are taken out by a suitably designed filter matrix. Interiors of cars are particularly "hostile" acoustic environments and it is unlikely that one can get away with ignoring their influence on the reproduced sound [39].

At low frequencies, the cross-talk cancellation problem is almost always ill-conditioned. Consequently, the digital filters that make up the cross-talk canceller are likely to boost low frequencies by 30dB or more. When the outputs from these filters are added together, most of the low-frequency energy ought to cancel out and form a set of loudspeaker input signals with relatively well-behaved frequency responses (dynamic range less than 15dB). This can only happen if the digital signal processing does not introduce any rounding or truncation errors. Unfortunately, the low-frequency boost makes the system very sensitive to uncorrelated low-frequency noise which can enter the recorded signals through, for example, inaccurate representation of low-frequency sound sources such as air-conditioning systems.

When an HRTF is measured digitally, an analogue low-pass filter is always used to prevent aliasing [40]. This causes the spectrum of the measured transfer function to contain only very little energy at frequencies just below the Nyquist frequency (half the sampling frequency). If one attempts to invert such a transfer function, the solution will inevitably boost high frequencies. Although this high frequency boost is inaudible, it must be taken out by using regularisation. This helps to protect the loudspeakers from overloading, and it also ensures a large dynamic range of the audible part of the processed signals.

It ought to be possible to improve the quality of the perceived sound immensely by developing loudspeakers particularly for the purpose of virtual source imaging. A key requirement to such loudspeakers is that they must have almost identical frequency responses (not just their amplitude responses, but also their phase responses must be the same). They should ideally be able to cope with an unusually large amount of low frequency energy. They need to work well on the axis only. The off-axis response is not important as long as the loudspeaker is not too omni-directional; in that case an unnecessary large amount of room reflections will interfere with the direct sound.

7. CONCLUSIONS

We believe that digital signal processing for multi-channel sound reproduction has got great potential. The "stereo dipole" technology, which uses only two closely spaced loudspeakers to generate virtual images in front of a single listener, has applications to multi-media computers, car audio, home entertainment systems, and video games. Systems comprising four loudspeakers have many applications. Since the auditory system is a very subjective judge of the quality of the reproduced sound, it ought to be possible to compromise the hard engineering criteria that we have largely relied on this far. This leaves plenty of scope for the developement of even more sophisticated filter design algorithms.

8. REFERENCES

[1] F.L. Wightman and D.J. Kistler, "Headphone simulation of free-field listening", J. Acoust. Soc. Am. 85, part I pp. 858-867 and part II pp. 868-878 (1989)

[2] H. Møller, "Fundamentals of binaural technology", Applied Acoustics 36, pp. 171-218 (1992)

[3] D.R. Begault, "3-D Sound for Virtual Reality and Multimedia". AP Professional, Cambridge MA, 1994

[4] A.J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis", J. Acoust. Soc. Am. 93, pp. 2764-2778 (1993)

[5] M. Kleiner, "Problems in the design and use of "dummy-heads" ", Acustica 41, pp. 183-193 (1978)

[6] H. Møller, C.B. Jensen, D. Hammershøi, and M.F. Sørensen, "Evaluation of artificial heads in listening tests", presented at the 102nd Audio Engineering Society Convention, 1997 March 22-25, Munich, Germany. AES preprint 4404 (A1)

[7] H. Møller, M.F Sørensen, D. Hammershøi, and C.B. Jensen, "Head-related transfer functions of human subjects", J. Audio Eng. Soc. 43 (4), pp. 203-217 (1995)

[8] H. Møller, M.F Sørensen, C.B. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?", J. Audio Eng. Soc. 44 (6), pp. 451-469 (1996)

[9] W.M Hartmann and A. Wittenberg, "On the externalization of sound images", J. Acoust. Soc. Am. 99, pp. 3678-3688 (1996)

[10] D. Pralong and S. Carlile, "The role of individualized headphone calibration for the generation of high fidelity virtual auditory space", J. Acoust. Soc. Am. 100, pp. 3785-3793 (1996)

[11] H. Møller, C.B. Jensen, D. Hammershøi, and M. Friis Sørensen, "Design criteria for headphones", J. Audio Eng. Soc. 43 (4), pp. 218-232 (1995)

[12] P. Damaske, "Head-related two-channel stereophony with loudspeaker reproduction", J. Acoust. Soc. Am. 50, pp. 1109-1115 (1971)

[13] D. Griesinger, "Equalization and spatial equalization of dummy-head recordings for loudspeaker reproduction", J. Audio Eng. Soc. 37 (1/2), pp. 20-29 (1989)

[14] D.H. Cooper and J.L. Bauck, "Prospects for transaural recording", J. Audio Eng. Soc. 37, pp. 3-19 (1989)

[15] B.S. Atal, M. Hill, and M.R. Schroeder, "Apparent Sound Source Translator", United States Patent Office, No. 3,236,949, February 22, 1966

[16] P.A. Nelson, F. Orduna-Bustamante and H. Hamada, "Inverse filter design and equalisation zones in multi-channel sound reproduction", IEEE Transactions on Speech and Audio Processing 3 (3), pp. 185-192 (1995)

[17] O. Kirkeby, P.A. Nelson, H. Hamada, "The "stereo dipole" - binaural sound reproduction using two closely spaced loudspeakers", presented at the 102nd Audio Engineering Society Convention, 1997 22-25 March, Munich, Germany. AES preprint 4463(I6). Also submitted to the Journal of the Audio Engineering Society

[18] Jens Blauert, "Spatial Hearing, The Physchophysics of Human Sound Localization", MIT Press, 1997

[19] O. Kirkeby, P.A. Nelson, H. Hamada, and F. Orduna-Bustamante, "Fast deconvolution of multi-channel systems using regularisation", ISVR Technical Report No. 255, University of Southampton, 1996. Less comprehensive version accepted for publication in IEEE Transactions on Speech and Audio Processing

[20] O. Kirkeby, P.A. Nelson, H. Hamada, "Stereo dipole", Patent Application, PCT/GB97/00415, 1997

[21] P.A. Nelson, O. Kirkeby, T. Takeuchi, and H. Hamada, "Sound fields for the production of virtual acoustic images", Letters to the editor, Journal of Sound and Vibration 204 (2), pp. 386-396 (1997)

[22] J.L. Bauck and D.H. Cooper, "Generalized transaural stereo and applications", J. Audio Eng. Soc. 44 (9), pp. 683-705 (1996)

[23] Fr. Heegaard, "The reproduction of sound in auditory perspective and a compatible system of stereophony", EBU Rev., pt. A-Technical, no 50, pp.2-6 (1958 Dec.); reprinted in J. Audio Eng. Soc. 40, 802-808 (1992)

[24] O. Kirkeby, P.A. Nelson, "Virtual source imaging using the stereo dipole", to be presented at the 103nd AES Convention in New York, September 26-29, 1997

[25] O. Kirkeby, P.A. Nelson, T. Takeuchi, H. Hamada, "Acoustic fields generated by virtual source imaging systems", pp. 941-954, in Proceedings of the Active 97, The international symposium on active control of sound and vibration, Budapest, Hungary, August 21-23, 1997, OPAKFI

[26] T. Takeuchi, P.A. Nelson, O. Kirkeby and H. Hamada, "Robustness of the performance of the "Stereo Dipole" to misalignment of head position", Presented at the 102nd AES Convention, 22-25 March 1997, Munich, Germany. AES preprint 4464 (I7)

[27] Y. Kahana, P.A. Nelson, O. Kirkeby and H. Hamada, "Multi-channel sound reproduction using a four-ear dummy-head", presented at the 102nd AES Convention, 22-25 March, 1997, Munich, Germany. AES preprint 4465 (I8)

[28] Y. Kahana, P.A. Nelson, and O. Kirkeby, "Objective and Subjective assessments of the reproduction of virtual images for multiple listeners", to be presented at the 103nd AES Convention in New York, September 26-29, 1997

[29] O. Kirkeby and P.A. Nelson, "Properties of least squares inverse filters used for multi-channel sound reproduction", pp. 1259-1271, in Proceedings of Active 95, the 1995 International Symposium on Active Control of Sound and Vibration, S. Sommerfeldt and H. Hamada (editors), Technomic publishing co., 1995

[30] S. Haykin (editor), J.H. Justice, N.L. Owsley, J.L. Yen, and A.C. Kak, "Array Signal Processing", Prentice-Hall, 1985

[31] P.A. Nelson, H. Hamada, and S.J. Elliott, "Adaptive inverse filters for stereophonic sound reproduction", IEEE Transactions on Signal Processing 40 (7), pp. 1621-1632 (1992)

[32] P.A. Nelson and F. Orduna-Bustamante, "Multi-channel signal processing techniques in the reproduction of sound", J. Audio Eng. Soc. 44, pp. 973-989 (1996)

[33] O. Kirkeby, P.A. Nelson, F. Orduna-Bustamante, and Hareo Hamada, "Local sound field reproduction using digital signal processing", J. Acoust. Soc. Am. 100 (3), pp. 1584-1593 (1996)

[34] F. Orduna-Bustamante, P.A. Nelson and H. Hamada, "Sparse-update LMS and filtered-x LMS algorithms", submitted to the IEEE Transactions on Signal Processing, 1997

[35] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery, "Numerical Recipes in C", Second edition, Cambridge University Press, 1992

[36] Bernard Widrow and Samuel D. Stearns, "Adaptive Signal Processing", Prentice-Hall, 1985

[37] S. Shimada, N. Hayashi, and S. Hayashi, "A clustering method for sound localization transfer functions", J. Audio Eng. Soc. 42, pp. 577-584 (1994)

[38] T. Takeuchi, P.A. Nelson, O. Kirkeby, H. Hamada, "The effect of reflections on the performance of virtual acoustic imaging systems", pp. 927-940, in Proceedings of the Active 97, The international symposium on active control of sound and vibration, Budapest, Hungary, August 21-23, 1997, OPAKFI

[39] F. Orduna-Bustamante, P.A. Nelson and H. Hamada, "Stereophonic imaging for in-car entertainment systems", Proceedings of Autotech'93, Seminar 36; Recent Advances in NVH Technology, IMechE Paper C462/36/206 (1993)

[40] Alan V. Oppenheim and Ronald W. Schafer, "Digital Signal Processing", Prentice-Hall, 1975