Hierarchical System of Surround Sound Transmission for HDTV

[3TM2.04]
Preprint 3339

GERZON, Michael A.;
Technical Consultant, Oxford, United Kingdom

**Presented at
the 92nd Convention
1992 March 24–27
Vienna**

AN AUDIO ENGINEERING SOCIETY PREPRINT

Hierarchical System of Surround Sound Transmission for HDTV

Michael A. Gerzon

Technical Consultant, 57 Juxon Street, Oxford OX2 6DJ, U.K.

Abstract

A new five-channel transmission hierarchy is described for HDTV surround sound, incorporating previously-described frontal stage multispeaker stereo hierarchies and a full 360^{o} ambisonic surround-sound stage. The structure of the hierarchy is different from previously proposed 3:2 hierarchies, but supports a much larger number of source and reproduction modes, including those allowing direct sounds from the sides, front and rear stages with optimised psychoacoustic quality. The hierarchy supports full up-, down- and sideways compatibility of all supported reproduction modes, and allows indefinite cascading down long broadcast production chains without increasing losses of directional reproduction quality.

1. INTRODUCTION

The design of hierarchies of transmission systems for surround sound for HDTV is one of the most complicated system design problems so far encountered in high-quality audio, both because of the large number of source directional encoding modes that need to be considered, and the even larger number of possible reproduction modes via many different loudspeaker layouts. The problem lies in supporting such a large number of modes while ensuring that sound mixed for every one mode is compatible with reproduction via every other mode.

This paper describes a 5-channel transmission hierarchy that directly supports an astonishingly large number (11) of source directional encoding modes and an even larger number of directional reproduction modes, as well as allowing future extensions to many more modes including those involving four or more frontal-stage stereo speakers and those supporting with-height reproduction modes, whether just across the width of a TV screen or around a full periphonic [1] sphere of directions.

The basic directional source encoding modes considered in this paper for use with the five-channel transmission hierarchy consist of the following eleven modes:
(i) so-called m:n stereo systems, assigning m channels to encoding speaker feed signals for a frontal stereo stage and n channels to encoding speaker feed signals for a rear stereo stage. (We include the cases m or n = 1 for mono feeds to a stage and n = 0 when no rear stage signals are

included). The following systems of front/rear stage stereo sound source coding are considered for use in the 5-channel hierarchy: 1:0 (i.e. mono), 2:0 (conventional 2-channel stereo), 3:0 (stereo encoded or intended to provide 3-speaker stereo feeds), 2:1 (using a conventional 2-channel frontal stereo stage encoding plus a monophonic rear stage "surround" signal), 2:2 ("quadraphonic" coding using both a front and a rear 2-channel stereo encoding), 3:1 (using a 3-channel frontal stereo encoding plus a monophonic "surround" encoding), and 3:2 (using a 3-channel frontal stereo stage and a 2-channel rear stereo stage).

Besides modes of encoding associated with using two stereo stages, there are also surround-sound encoding modes which encode the azimuthal direction of sounds within a 360° horizontal directional stage, in which sound encoding is specified not in terms of speaker feed signals (as is the case with "stereo" systems), but in terms of gains, one per signal channel, with which a sound from a given direction is assigned to each encoded signal channel. Such directional encoding modes do not directly convey reproduction speaker feed signals, but rely on the use of "psychoacoustic" decoding algorithms, designed to produce, for each specified speaker layout a consumer or end-user may wish to use, appropriate speaker feed signals adapted to that layout to produce an optimimum subjective illusion of the intended encoded directional effect. This philosophy is described in some detail in the author's reference [2], and an earlier generation Ambisonic technology to achieve this aim (not, however, optimised for HDTV use) is described in ref. [3].

In this paper, we describe four azimuthal directional encoding modes (ii) using five encoding channel signals denoted W, X, Y, E and F. The 3-channel mode, termed "B-format", has been used previously in earlier Ambisonic encoding systems, and consists of the signals W, X, Y with respective azimuthal gains 1, $2^{\frac{1}{2}}\cos\Theta$ and $2^{\frac{1}{2}}\sin\Theta$ as a function of azimuth Θ, measured anticlockwise from the due-front direction. The two additional channels E and F, supplementing the basic B-format signals, encode, in a manner described later in the paper, respectively aspects of azimuthal directionality that improve the directional stability of frontal stage sounds with change in listener position (a weakness with 2-speaker stereo and earlier Ambisonic reproduction methods) and aspects relating to the separation between frontal and rear sound stages. This yields another three azimuthal directional encoding formats, respectively termed BE-format (consisting of the signals W,X,Y,E), BF-format (consisting of the signals W,X,Y,F) and BEF-format (consisting of the signals W,X,Y,E,F).

Thus in this paper, we show how all of eleven encoding formats, seven of them being m:n stereo formats, and four being B-, BE-, BF- and BEF- azimuthal directional encoding formats, can be incorporated into just five transmission signals.

The problem of compatibility of these different encoding modes is to ensure that the result of encoding into any one of the eleven encoding modes, and of decoding, via reception of the transmission signal via any of the other encoding modes into a reproduction speaker layout, will result in acceptable conversion from every mode to every other mode. With 11 supported encoding modes, there are 10 × 11 = 110 possible conversion matrices converting any one mode into any other - hence the comment at the start of the paper about this being a complicated system design problem.

In fact, the compatibility problem is even worse than so far indicated, since so far it has only described what may be termed "one-step" conversion matrices between encoding modes. In broadcast and other professional applications, one wishes to be able to take material intended for any mode and convert it into any other mode, as best as is possible, whether or not the source material includes the results of any earlier mode conversions, i.e one requires that the results of cascading mode conversion does not result in any further uneccesary degradation of the resulting directional effect. This "cascadability" requirement has already been considered by the author in detail for frontal stage stereo systems in refs. [4] to [6], but complicated though the design algorithms neccesary to ensure cascadability in the front-stage case may be, the abstract formulation of cascadability in the surround case is even more mathematically complicated. Fortunately, despite this, we are able in this paper to present a relatively simple approach, using notional intermediate transmission channel signals (which may or may not actually be used), that allows the design of a practical cascadable hierarchy of conversion matrices between different coding systems.

Such an approach yields an hierarchy of transmission and conversion matrices between all directional encoding modes that not only gives "compatibility" after one step of conversion, but that allows indefinite cascading of conversion processes that give results that are never worse than the results of "downconverting" to the the "bottleneck" simplest encoding system associated with the intermediate stages and then "upconverting" back to the output encoding system.

Such a cascadability of conversions ensures that several successive conversions do not cause any unexpected degradations, and that, provided that the original balance engineer for a sound source has checked that the "downward" compatibility of the mix is acceptable, then at later stages in the signal chain, the results should remain acceptable without further checking.

Because the problem we are tackling is inherently a complex one, this paper is conceptually quite complicated. We recommend that the reader might consult earlier papers on the frontal-stage multispeaker stereo case first as an introduction [4-7]; of these, the simplest papers are [5] and [6], but the one that introduces the basic technical ideas in detail is [4]. However, this paper is intended to be independent of these, and contains a brief summary of the results of [4].

To deal with the description and design of a fully compatible cascadable hierarchy requires many steps: the description of the directional encoding systems themselves, the presentation of methods of defining and constructing cascadable hierarchies, and not least, descriptions of the methods of reproducing sounds from any directional encoding system into any desired loudspeaker layout.

This last problem of describing methods of reproduction is a vast topic, far too large to be covered in this paper. Some of the required methods have been described in previous papers. For example, ref. [3] has described conventional Ambisonic decoding of B-format signals, and ref. [7] has described methods of reproducing n_1-speaker stereo signals via larger numbers n_2 of frontal stage stereo speakers. Better methods of decoding surround sound from B-, BE-, BF- and BEF-format encodings, suitable for use with HDTV, will be the subject of another paper [8] of considerable technical complexity.

In this paper, we only go into some detail on one method of reproduction that is not dealt with elsewhere, namely a method of reproducing B-format 360°-encoded sounds via a frontal stereo stage using three (or more) loudspeakers. This optimal frontal-stage reproduction from B-format means that there is a 3-channel surround-sound format capable of reception not only in at least two true surround-sound modes, but also as true frontal-stage 3-speaker stereo - an essential requirement for HDTV. This is particularly valuable for situations where there is considerable pressure on channel capacity in the transmission medium, such as in DAB or DCC applications, or in the case when broadcasts must be multilingual, where a 3-channel surround broadcasting mode may well make the difference between being able to broadcast in surround and stereo-only transmission.

Besides 3-speaker presentation from B-format encoding, this paper will also make use of the results of ref. [7] on "upconversion" of stereo signals in m:n stereo systems. In particular, an m:n stereo signal can be reproduced via any speaker layout with m_2 frontal stage stereo speakers and n_2 rear stage stereo speakers whenever $m_2 \geqslant m$ and $n_2 \geqslant n$, using the stereo upconversion methods given in ref. [7]. This is already familiar in 3:1 stereo systems where, in cinema film sound applications, the 1-channel "surround" signal is usually fed to many speakers spread across the rear and sides of the listening area in order to delocalise it. In a similar way, the frontal stage signals of a 2:1 stereo transmission need not be fed to only two speakers, but can be fed to a frontal 3-speaker stereo system using the optimised 3 × 2 decoders described in refs. [7] and [4] to give enhanced frontal stage image quality and stability.

It is important not to confuse the numbers m and n of stereo channels assigned to representing the front and rear stages of an m:n stereo signal with the number of loudspeakers used for reproduction in each of the two stages; there is no reason why the signals should not be matrixed or otherwise processed for reproduction via more loudspeakers across each of these stages.

Additional complexities arise from the existence of several methods of encoding a cascadable hierarchy of conversion matrices between directional encoding modes into actual transmission channels, and the main thrust of this paper is to concentrate on the "compatibility matrixing" approach rather than the "downmixing" approach, without wholly excluding the latter.

We assume that the reader is familiar with other earlier approaches to HDTV sound, as surveyed in the two excellent papers by Meares [9] and Theile [10], in which the philosophy of the "compatibility matrixing" and "downmixing" approaches is discussed.

Beyond the technical complexities of this paper, it is hoped that the reader will see that our aim is to ensure that systems of sound, not just for HDTV, but also for cinema and for audio-only uses, should all prove operationally and technically compatible with one another. It is important that programmes intended for one medium should always be usable via any other medium. With the importance of HDTV broadcast/cinema co-productions, there is already an obvious need to ensure that the sound systems for HDTV and cinema should easily be convertible for use in the other medium. This is not to say that technical standards should be identical - indeed the home environment is capable of subleties of directional effect impossible in a cinema auditorium environment due to the latter's longer time delays, differing acoustics and larger audience area, so that HDTV sound systems should be designed to take advantage of this potential for improved domestic directional effect.

Nevertheless, the structure of directional sound systems in the different media should be such that conversion is always possible. The methods in this paper are based on detailed studies of how such conversion can be best effected so as to avoid operational problems and to retain acceptable results when inter-medium conversion is used. There are inevitably some remaining compromises (after all even conventional mono and 2-speaker stereo are not perfectly compatible), and it is expected that experience will suggest minor alterations of the matrix coefficients suggested in this paper for the conversion matrices between directional encoding systems.

However, it is believed that the structure of the cascadable hierarchy of systems that use five transmission channels is the best that can be found to maximise operational flexibility and to minimise operational problems. The structure of interconversions should permit all present and future audio media to use common methods of signal handling without the operational problems associated with earlier proposals that are not fully cascadable and which exclude many of the possible directional encoding options.

The numerical values of the matrix coefficients in this paper are intended as a starting point for more detailed studies, but are believed to be quite close to the operationally best values. Since they involve 132 conversion matrices from 11 encoding modes to 12 reproduction modes, the choices are constrained, and a complete experimental check of all possible conversion's compatibility is probably impractical. Nevertheless, we feel that practical trials and optimisation of such a cascadable hierarchy are feasible, by allowing theoretical methods to take some of the design burden, and by doing careful experimental checks of the main upconversion and downconversion modes.

2. DIRECTIONAL SOUND ENCODING FORMATS

We shall start off by listing and describing the various methods of encoding directional sound that will be considered in this paper, including three formats that require more channels than are used in the five-channel transmission hierarchy considered in this paper.

2.1 Frontal stereo formats

We consider 5 frontal-stage stereo formats intended to provide speaker feed signals for any number from one to 5 loudspeakers in a frontal stereo stage. (The 1-speaker mono case can be considered, by an abuse of terminology, as being "one-speaker stereo" for ease of description!)

The notations we shall use for the channel signals associated with loudspeaker feed signals are given here with reference to figures 1a to 1e.

1:0 stereo also known as mono! This uses a single signal denoted C_1 intended to feed a front-centre speaker as shown in fig. 1a.

2:0 stereo, i.e. conventional 2-channel stereo, which uses 2 speaker feed signals L_2 and R_2 intended for a respective left and right frontal loudspeaker as shown in fig. 1b.

3:0 stereo. This conveys frontal stage stereo via three signals L_3, C_3 and R_3 intended for respective reproduction via a left, centre and right loudspeaker in a frontal 3-speaker stereo system, as shown in fig. 1c.

4:0 stereo . This conveys frontal-stage stereo via 4 signals L_4, L_5, R_5 and R_4 intended respectively for an outer left, inner left, inner right and outer right speaker of a 4-speaker frontal stereo layout as shown in fig. 1d.

5:0 stereo . This conveys frontal-stage stereo via 5 signals L_6, L_7, C_5, R_7, R_6 intended for respective outer left, inner left, centre, inner right and outer right loudspeakers of a frontal-stage 5-speaker stereo layout as shown in fig. 1e.

Reference [7] describes in detail the methods used to convert speaker feed signals intended for one number of frontal stage stereo speakers via a matrix for reproduction via a larger number of frontal stage stereo speakers, and refs. [4-6] deal with associated transmission systems for frontal stage stereo, including "downconversion" matrices for reducing the number of speaker feeds.

2.2 m:n front/rear 2-stage stereo

Figure 2 shows a typical 5-speaker layout such as might be used with a front stage plus rear stage stereo system. The four m:n stereo systems we shall consider in detail in this paper use speaker feed signals for subsets of this layout, as follows:

2:1 stereo. This 3-channel method uses three signals, 2 being signals L_{2F} and R_{2F} being intended for reproduction from a frontal stereo stage, and the third signal B is intended as a monophonic signal allocated to the rear stage.

2:2 stereo This 4-channel system, in the past called "quadraphonic", allocates two signals L_{2F}' and R_{2F}' to a frontal stereo stage and another two signals L_{2B} and R_{2B} to a rear stereo stage.

3:1 stereo This 4-channel method allocates three signals L_{3F}, C_{3F} and R_{3F} to a frontal 3-channel stereo stage and a single channel B to a rear monophonic stage.

3:2 stereo This 5-channel system, based on 5-channel cinema standards, allocates 3 signals, respectively L_{3F}', C_{3F}', R_{3F}' from left to right, to a frontal stereo stage, and two signals L_{2B} and R_{2B} to left and right of a 2-channel rear stereo stage.

The reason for using different symbols for the 2 frontal channels for 2:0, 2:1 and 2:2 stereo is that the conversion matrices we shall consider will fold different amounts of rear-stage sounds into these frontal channels, and similarly for 3:0, 3:1 and 3:2 proposals. It is helpful, therefore, to use distinct symbols to avoid confusions.

2.3 B-format

Unlike the "stereo" systems considered above, the remaining directional encoding systems here do not use signals representing speaker feeds, but assign each direction in a surround-sound stage to a set of gains, one per channel, with which sounds assigned to that direction are mixed into the encoded signal. We first describe two versions of B-format, one for horizontal azimuthal sound, and the other for full-sphere sound.

(Azimuthal) B-format This encodes sounds in a horizontal 360° azimuthal sound stage into three signals W, X and Y with respective gains 1, $2^{\frac{1}{2}}\cos\theta$ and $2^{\frac{1}{2}}\sin\theta$ for sounds assigned to a directional azimuth θ, measured anticlockwise from due front. The polar gain patterns of these 3 signals is shown in figure 3.

(Periphonic) B-format This encodes sounds from any direction from the full sphere of directions in 3 dimensions, using 4 signals W, X, Y and Z with respective gains equal to 1, $2^{\frac{1}{2}}x$, $2^{\frac{1}{2}}y$ and $2^{\frac{1}{2}}z$ for sounds assigned to a direction with direction cosines (x,y,z) in respective front, left and up directions. Figure 4 shows the polar gain patterns of these 4 signals.

Since this paper is concerned mainly with horizontal azimuthal sound reproduction, the term "B-format" in this paper without any qualification will refer to the azimuthal 3-channel case with signals W, X, and Y.

2.4 Enhanced B-formats

A problem with B-format using previous methods of Ambisonic reproduction

[3] is that, good though the reproduced surround-sound illusion is for listeners away from the centre of the listening area, the actual direction of central-stage frontal images does not remain aligned to that of a television screen forsuch listeners. While the improved surround-sound decoders of ref. [8] greatly reduce this problem, it is found useful to add further channels to B-format to provide further information to help stabilise sound images. We thus introduce 3 new formats for azimuthal horizontal sound as follows:

BE-format This 4-channel format uses 4 signals W,X,Y,E to provide improved image stability for front-centre sounds.

BF-format This 4-channel format uses 4 signals W, X, Y, F to provide better separation between frontal and rear sound stages.

BEF-format This 5-channel format uses 5 signals W, X, Y, E, F to provide more stable frontal stage images and a better front/rear stage separation.

The five signals W, X, Y, E, F are defined to have respective gains for sounds assigned to an azimuthal direction θ measured anticlockwise from due front as follows:

W has gain 1

X has gain $2^{\frac{1}{2}}\cos\theta$

Y has gain $2^{\frac{1}{2}}\sin\theta$

E has gain $k_e(1 - k_g(1-\cos\theta))$ for $|\theta| \leq \theta_S$
and gain 0 otherwise

F has gain $2^{\frac{1}{2}}k_f\sin\theta$ for $|\theta| \leq \theta_S$
and gain $-2^{\frac{1}{2}}k_b\sin\theta$ for $|180^\circ-\theta| \leq \theta_B$
and gain 0 otherwise,

where θ_S is the half-stage width of a frontal stage, typically between 60° and 70° (so that the total azimuthal frontal stage is between 120° and 140° wide) and θ_B is the half-stage width of a rear stage, typically 70° (so that the rear stage is typically 140° wide), and where k_g is a fixed gain preferably equal to 3.25, but which may be standardised elsewhere in the range 3 to $3\frac{1}{2}$, and where k_e, k_f and k_b are user-defined gains between 0 and 1 determining the degree of "enhancement" added to B-format across the front and rear stages.

Figures 5a and 5b show the gains as a function of azimuthal angle θ of the signals E and F for $k_g = 3.25$, $k_e = k_f = k_b = 1$, $\theta_S = 60^\circ$ and $\theta_B = 70^\circ$.

2.5 Encoding format labels

To simplify future descriptions, it is convenient to label the above directional encoding systems as systems A_j with j an integer, as follows: 1:0 stereo (A_1), 2:0 stereo (A_2), 3:0 stereo (A_3), 4:0 stereo (A_{12}), 5:0 stereo (A_{13}), 2:1 stereo (A_4), 2:2 stereo (A_{11}), 3:1 stereo (A_5), 3:2 stereo (A_6), B-format WXY (A_7), BE-format (A_8), BF-format (A_9), BEF-format (A_{10}) and periphonic B-format WXYZ (A_{14}).

3. NOTIONAL TRANSMISSION CHANNELS

The quickest way of describing the complicated hierarchy of systems based on the above directional encoding methods is to introduce what we term "notional transmission signals", which may be actual transmission signals for a "compatibility matrixing" system, or may merely play the role of a mathematical intermediary used to describe the structure of an actual transmission system without themselves physically existing anywhere within the system.

The basic hierarchy we consider is based on five notional transmission signals M_T, S_T, B_T, T_T and F_T, whose meaning can roughly be described as follows:

M_T is a monophonic transmission signal.

S_T is a stereophonic left-minus-right difference signal.

B_T is a rear-stage monophonic signal.

T_T is a third-channel stereo signal representing the difference in information between 2-speaker and 3-speaker stereo.

F_T is a front-stage difference minus rear-stage difference signal.

3.1 Encoding Matrices

For every directional encoding system A_i considered above, using n_i signals, we shall define an $n_i \times n_i$ encoding matrix, which we shall denote by E_{ii}, which takes the signals of A_i and converts them into a subset Z_i of the notional transmission channel signals. The original system A_i signals can be recovered from the set Z_i of transmission signals by the inverse $n_i \times n_i$ decoding matrix

$$D_{ii} = E_{ii}^{-1} \quad . \qquad (3\text{-}1\text{-}1)$$

Figure 6 shows a schematic showing how the <u>transmission encoding matrix</u> E_{ii} encodes the n_i signals of the directional encoding system A_i into n_i notional transmission signals Z_i, and how its inverse <u>transmission decoding matrix</u> D_{ii} recovers the original n_i signals of the directional encoding system A_i from the set Z_i of n_i transmission channel signals.

Specifying such $n_i \times n_i$ encoding matrices E_{ii} for each directional encoding system A_i into a set Z_i of transmission channels, and hence also the decoding matrix D_{ii} via equ. (3-1-1), in turn specifies what we shall term a <u>conversion matrix</u> R_{ji} for converting the n_i signals of the system A_i into the n_j signals of the system A_j.

Figure 7 is a schematic showing how $n_j \times n_i$ conversion matrices R_{ji} from a system A_i to a system A_j are produced via an intermediate transmission encoding operation E_{ii} into the n_i notional transmission channels Z_i, followed by a transmission decoding operation D_{jj} from the n_j notional transmission channels Z_j into the system A_j. Since the set Z_j of transmission channels may not be a subset of the original set Z_i encoded from A_i, any missing transmission channels not in the set

Z_i may be represented by a zero-valued transmission signal, shown as entering the block I_{ji} in figure 7, and the subset Z_j of the possible set of transmission signals is then selected from the total set (comprising Z_i and the zero signals) of transmission signals before decoding into A_j.

The idea in designing a hierarchical system of conversion matrices between a number of directional encoding systems A_i is to select a set of notional transmission channels, and a set of transmission encoding matrices E_{ii} into these such that the resulting conversion matrices R_{ji} formed by the encoding from A_i and decoding into A_j gives good subjective (i.e. "compatible") reproduction of one system via the other in all cases.

This might seem an arbitrary complication in that it would seem easier at first sight simply to choose the conversion matrices between systems directly for best compatibility. However, the trouble with this is that, in general, the result of repeated conversions between systems, such as shown in figure 8 from a system A_i to A_j and then from A_j to A_k results in "cascade degradations", whereby even an upconversion to a more elaborate system followed by a downconversion back to the original system will not get back to the original signals. For example, in [6], we reported that the 4 × 3 upconversion suggested by Meares [9] followed by his 3 × 4 downconversion for 3- and 4-speaker stereo results in centre speaker levels being increased by 2.5 dB relative to outer speaker levels.

Apart from any once-and-for-all losses of information, conversion matrices constructed via the use of notional transmission channels via the construction of figure 7 do not result in such cascade losses, no matter how many intermediate stages of conversion there may be. Essentially, this is because the conversion process leaves the associated transmission signals in Z_i unchanged, unless they are discarded altogether.

In ref. [4], we showed that given desired "upconversion" matrices for converting n_1-speaker frontal stereo into n_2-speaker frontal stereo with $n_2 \geqslant n_1$, it was possible to construct encoding matrices into notional transmission channel signals such that the resulting conversion matrices for $n_2 \geqslant n_1$ were the desired upconversion matrices. It was further shown that there was some freedom of choice in this construction of transmission encoding matrices, so that a degree of optimisation for the "downconversion" conversion matrices from a larger to a smaller number of stereo speakers was possible.

The hierarchy proposed in this paper generalises the methods of ref. [4] by specifying for every one of 11 directional encoding systems A_1 to A_{11} given in section 2 above an associated subset of the set M_T, S_T, B_T, T_T and F_T of notional transmission channels, along with an encoding matrix E_{ii}. This has to be done in such a manner that the resulting conversion matrices R_{ji} between systems A_i and A_j gives acceptable subjective results in all cases.

3.2 Some general results

From figure 7, we see that the conversion matrix R_{ji} from system A_i to system A_j is given by the formula

$$R_{ji} = D_{jj} I_{ji} E_{ii} , \qquad (3\text{-}2\text{-}1)$$

where I_{ji} is the $n_j \times n_i$ matrix that selects the channels of Z_j from those of Z_i in figure 7. I_{ji} has all entries equal to zero, except for entries of 1 for those "diagonal" entries corresponding to a notional transmission channel being present both in the set Z_i and the set Z_j of channels.

Using equ. (3-1-1), we get from (3-2-1) :

$$D_{jj} I_{ji} = R_{ji} D_{ii} , \qquad (3\text{-}2\text{-}2)$$

which asserts that the columns of D_{jj} corresponding to channels present in the set Z_i of channels are obtained by multiplying the transmission decoding matrix D_{ii} by the conversion matrix R_{ji} from A_i to A_j. In ref. [4], the design procedure for the transmission hierarchy was based on this result.

If systems A_i and A_j are such that Z_i is a subset of the transmission channels Z_j, then we shall term A_j an "upconversion" of the system A_i, and term R_{ji} an <u>upconversion matrix</u>, and conversely shall term the converse conversion matrix R_{ij} from system A_j to A_i a <u>downconversion matrix</u>. In particular, we have from (3-2-1) and (3-1-1) the result that, if A_j is an upconversion of A_i, then

$$R_{ij} R_{ji} = I_{ii} , \qquad (3\text{-}2\text{-}3)$$

where I_{ii} is the identity matrix on A_i, i.e. the result of following an upconversion matrix by a downconversion back to the original system is to leave the signals unaltered - as we would hope!

More generally, if three systems A_i, A_j and A_k are such that the three sets Z_i, Z_j and Z_k of associated notional transmission signals satisfy

$$Z_j \subseteq Z_i \cap Z_k , \qquad (3\text{-}2\text{-}4)$$

then it can be proved from (3-1-1) and (3-2-1) that the corresponding conversion matrices satisfy the general cascade relationship (see fig. 8):

$$R_{ki} = R_{kj} R_{ji} . \qquad (3\text{-}2\text{-}5)$$

Moreover, if we have an arbitrary cascade of systems A_{i_1}, $A_{i_2}, \ldots, A_{i_n}$, with successive conversion matrices $R_{i_{k+1} i_k}$ for $k = 1, 2, \ldots, n-1$, and if A_j is a system such that

$$Z_j = Z_{i_1} \cap Z_{i_2} \cap \cdots \cap Z_{i_n} , \qquad (3\text{-}2\text{-}6)$$

then

$$R_{i_n i_{n-1}} \cdots R_{i_3 i_2} R_{i_2 i_1} = R_{i_n j} R_{j i_1} , \qquad (3\text{-}2\text{-}7)$$

which asserts that the result of the cascaded conversion matrices is equivalent to just one downconversion to a "greatest common downconversion" of the systems in the chain from the original system, and one upconversion back up to the final system.

The above general results ensure that any collection of conversion matrices R_{ji} between pairs A_i and A_j of directional coding systems will be indefinitely cascadable down a long production chain involving repeated conversions between systems. (In particular, the trivial conversion matrix R_{ii} from a system to itself is, by (3-1-1) just the identity matrix I_{ii}).

3.3 Structure of the 5-channel hierarchy

The above general results on "cascadable hierarchies" of directional encoding systems will now be made more concrete. Figure 9 shows the structure of how the eleven directional coding systems A_1 to A_{11} considered in section 2 are encoded into the five notional transmission signals M_T, S_T, B_T, T_T and F_T. To avoid clutter on the diagram, the subscripts "T" are omitted from the transmission signals, and the encoding matrix process E_{ii} is indicated only by a rightward-pointing (near) horizontal arrow from a source/reproduction directional coding system to the corresponding set of notional transmission channel signals. The leftward-pointing (near-)horizontal arrows correspond to the inverse decoding operations D_{ii}. The near vertical lines correspond to upconversion and downconversion inclusion maps I_{ji} for sets Z_i, Z_j of transmission channel signals such that Z_j includes Z_i.

This is all more complicated to describe than to understand by direct inspection of figure 9. For example, mono signals are encoded just into the M_T transmission channel, whereas 2-channel stereo is encoded into a mono M_T and stereo difference S_T transmission signal, and a 3-channel frontal stereo signal is encoded into M_T and S_T and a third transmission signal T_T. Figure 10 shows how 2- and 3-speaker frontal stereo are encoded to and decoded from three transmission channels, and thereby how conversion between 2 and 3 speakers is achieved via the intermediate (notional) transmission channels.

As will be seen from figure 9, both 2:1 stereo and B-format are encoded into two transmission signals M_T and S_T also used for ordinary 2-speaker stereo, plus a third channel B_T conveying back-stage sound.

As would be expected, 3:1 stereo uses the transmission channels used both for 3:0 and 2:1 stereo; 3:2 stereo uses all 5 transmission channels, including F_T, and 2:2 stereo discards the T_T channel corresponding to the use of 3 frontal speakers. The enhanced B-formats are in a natural correspondence to the m:n stereo 2-stage systems, using the same subsets of transmission channels as shown in figure 9.

3.4 Actual Transmission channels

In the above, we have referred to the transmission signals as "notional". There are many different strategies for choosing actual transmission signals, which depend on such matters as optimising the subjective performance in the presence of transmission signal errors (e.g. those caused by data compression modulation error signals), and on decoder complexity and on operational flexibility.

One option is to use the notional transmission channels as the actual transmission channels. This option is known as hierarchical compatibility matrixing. The big advantage of compatibility matrixing is that it is easily upgradable to adding enhanced reproduction formats, such as those using more than 3 front speakers or more than two rear speakers, or those involving the reproduction of height or even full-sphere (periphonic) directionality. This is because the addition of enhanced reproduction modes simply involves the addition of new transmission channels, without any modification of those used for previous encoding and reproduction modes. This means, in particular, that future enhancements can be added without any loss of services expected by existing users, and without any need to modify the equipment used by existing users. The only precaution needed is that any additional transmission channels should be added in a way that they do not interfere with "existing"equipment.

Compatibility matrixing is also economical in that the addition of enhanced reproduction modes requires the addition of no more than the minimum required number of channels to those used for previous services. Yet another advantage of compatibility matrixing is that when there is severe pressure on available channels (e.g. due to multilingual broadcasts or hearing-impaired services), it provides a graceful route to degrading the directional effect simply by flagging that a channel is no longer available - the fold-down (conversion matrix) is performed automatically simply by muting that transmission channel, and needs no complex signalling that needs to be preserved down a long broadcast or production chain.

Therefore, compatibility matrixing provides both potential for future upgrading of directional reproduction and the operational flexibility to downgrade when there is pressure on audio channels, without having to standardise complex protocols in advance. For this reason, it is the option that is likely to create the fewest problems in production, and may well be ideal for studio and recording applications.

Moreover, providing the "compatibility" of the upconversion and downconversion matrices are well chosen, compatibility matrixing is particularly tolerant of small gain errors in the transmission channels. If the matrices approximately preserve energy (i.e. are fairly close to being orthogonal matrices), any noise errors in the transmission channels are also not going to be unduly exaggerated in reproduction.

The main disadvantage of compatibility matrixing is that it can lead to directional unmasking of noise errors, especially when used with data compression codecs based only on monaural masking. The author describes in ref. [11] methods of minimising the effects of directional unmasking, and notes that previous "compatibility matrixing" based on non-cascadable conversion matrices can cause a marked increase in directional unmasking.

Another strategy that has been proposed is to transmit loudspeaker feeds directly [9,10], for the most complicated available transmission mode, and to "downconvert" to other modes. This approach is operationally much less flexible, since it requires a new set of downmixing coefficients for all possible reproduction modes to be specified every time a new reproduction mode is introduced. In particular, it does not allow upgrading to future improved directional reproduction modes, since every then-existing receiver would need to be equipped with a new reception mode responsive to more channels, plus a yet-unknown set of downmixing coefficients. In principle, one could specify that all present receivers be capable of receiving a stated large number of available audio channels and of also receiving transmitted downmixing coefficients to provide feeds for its own speaker system, but this would still not allow the use of information from any other audio channels that may be made available in the future.

Additionally, this downmixing approach is only good at providing a reasonable degree of directional masking of noise errors when the signals do indeed represent speaker feeds, and may still give poor directional unmasking for systems of encoding representing encoding of direction and azimuth directly, such as Ambisonic systems. This limits the possible enhancements of directional reproduction made with future developments in using the psychoacoustics of directional reproduction.

Downmixing also has the disadvantage that every transmission mode requires use of a different set of matrixing coefficients, depending on the transmitted mode. This complicates switching options in the receiver, requires the use of signalling protocols that have to be maintained down the production and broadcast chain, and makes it much more difficult for broadcasters to use downgraded directional reproduction options when there is pressure on available audio channels, or where attempts to data compress elaborate options extremely heavily produce unacceptable subjective sound quality degradation.

However, whatever transmission channel option is actually used, the work of this paper provides conversion matrices that may be used as the upconversion and downconversion matrices for reception via any transmission system, and for mode conversion anywhere in the production or broadcast chain, without the risk of repeated cascade losses down the chain. Therefore, a standardisation of the conversion matrices R_{ji} is highly desirable, irrespective of the choice of actual transmission channels.

4. THREE CHANNEL SYSTEMS

Rather than attempt to describe the whole 5-channel hierarchy straight away, we shall describe the simplest 3-channel options first. This is both because this gives some insight without the complications of more elaborate cases, and because the three-channel options are very useful in their own right.

In particular, we feel that a detailed examination of the 3-channel surround-sound options reveal useful reproduction modes that have not been adequately regarded in previous studies. When there is pressure on the number of audio channels (say in DAB or DCC applications, or when multilingual or other services limit the number of audio channels available), the three-channel option may prove to be a valuable resource allowing surround-sound when otherwise only stereo would prove possible.

4.1 3-channel frontal stereo

We repeat from refs. [4] and [6] the encoding, decoding and conversion matrices for mono (A_1), 2-channel stereo (A_2) and 3-channel stereo (A_3).

<u>1:0 stereo (mono) encoding</u> E_{11}

$$M_T = C_1$$

<u>2:0 stereo encoding</u> E_{22}

$$\begin{pmatrix} M_T \\ S_T \end{pmatrix} = \begin{pmatrix} 0.7071 & 0.7071 \\ 0.7071 & -0.7071 \end{pmatrix} \begin{pmatrix} L_2 \\ R_2 \end{pmatrix}$$

<u>3:0 stereo encoding</u> E_{33}

$$\begin{pmatrix} M_T \\ S_T \\ T_T \end{pmatrix} = \begin{pmatrix} 0.5000 & 0.7071 & 0.5000 \\ 0.7071 & 0.0000 & -0.7071 \\ 0.5000 & -0.7071 & 0.5000 \end{pmatrix} \begin{pmatrix} L_3 \\ C_3 \\ R_3 \end{pmatrix}$$

The inverse decoding equations are given by:

<u>1:0 stereo (mono) decoding</u> D_{11}

$$C_1 = M_T$$

<u>2:0 stereo decoding</u> D_{22}

$$\begin{pmatrix} L_2 \\ R_2 \end{pmatrix} = \begin{pmatrix} 0.7071 & 0.7071 \\ 0.7071 & -0.7071 \end{pmatrix} \begin{pmatrix} M_T \\ S_T \end{pmatrix}$$

<u>3:0 stereo decoding</u> D_{33}

$$\begin{pmatrix} L_3 \\ C_3 \\ R_3 \end{pmatrix} = \begin{pmatrix} 0.5000 & 0.7071 & 0.5000 \\ 0.7071 & 0.0000 & -0.7071 \\ 0.5000 & -0.7071 & 0.5000 \end{pmatrix} \begin{pmatrix} M_T \\ S_T \\ T_T \end{pmatrix}$$

It is a mere coincidence that, in these three cases, the encoding matrices E_{ii} and the decoding matrices D_{ii} happen to have the same forms. The conversion matrices R_{ji} generated from the above encoding and decoding matrices via fig. 7 are as follows for $i \neq j$:

<u>1:0 to 2:0</u> R_{21}

$$\begin{pmatrix} L_2 \\ R_2 \end{pmatrix} = \begin{pmatrix} 0.7071 \\ 0.7071 \end{pmatrix} \quad C_1$$

<u>1:0 to 3:0</u> R_{31}

$$\begin{pmatrix} L_3 \\ C_3 \\ R_3 \end{pmatrix} = \begin{pmatrix} 0.5000 \\ 0.7071 \\ 0.5000 \end{pmatrix} \quad C_1$$

<u>2:0 to 3:0</u> R_{32}

$$\begin{pmatrix} L_3 \\ C_3 \\ R_3 \end{pmatrix} = \begin{pmatrix} 0.8536 & -0.1464 \\ 0.5000 & 0.5000 \\ -0.1464 & 0.8536 \end{pmatrix} \begin{pmatrix} L_2 \\ R_2 \end{pmatrix}$$

<u>2:0 to 1:0</u> R_{12}

$$C_1 = \begin{pmatrix} 0.7071 & 0.7071 \end{pmatrix} \begin{pmatrix} L_2 \\ R_2 \end{pmatrix}$$

<u>3:0 to 1:0</u> R_{31}

$$C_1 = \begin{pmatrix} 0.5000 & 0.7071 & 0.5000 \end{pmatrix} \begin{pmatrix} L_3 \\ C_3 \\ R_3 \end{pmatrix}$$

<u>3:0 to 2:0</u> R_{23}

$$\begin{pmatrix} L_2 \\ R_2 \end{pmatrix} = \begin{pmatrix} 0.8563 & 0.5000 & -0.1464 \\ -0.1464 & 0.5000 & 0.8536 \end{pmatrix} \begin{pmatrix} L_3 \\ C_3 \\ R_3 \end{pmatrix} .$$

In addition, in refs. [7] and [4-6], we showed that there was a frequency dependent upconversion R_{32} from 2:0 to 3:0 stereo that replaced the zero third transmission channel T_T with a frequency-dependent third channel as illustrated in fig. 11, which gives psychoacoustically optimised subjective upconversion results. This synthesises a third channel T_T from the M_T channel by passing it through an all-pass filter A with gain -1 below 5 kHz and gain +1 above 5 kHz, and adding an overall gain 0.1716, as shown in fig. 11. This yields the frequency dependent upconversion matrix

<u>2:0 to 3:0</u> R_{32} (psychoacoustic)

$$\begin{pmatrix} L_3 \\ C_3 \\ R_3 \end{pmatrix} = \begin{pmatrix} 0.8536 + 0.0607\,A & -0.1464 + 0.0607\,A \\ 0.5000 - 0.0858\,A & 0.5000 - 0.0858\,A \\ -0.1464 + 0.0607\,A & 0.8536 + 0.0607\,A \end{pmatrix} \begin{pmatrix} L_2 \\ R_2 \end{pmatrix} ,$$

where A is an all-pass filter with gain +1 below 5 kHz and gain -1 above 5 kHz (and a smooth phase transition in the 5 kHz region).

Because this psychoacoustic upconversion only alters the third channel T_T of the transmission hierarchy, it does not affect the cascadability of the hierarchy; a similar frequency dependent upconversion of mono is obtained from figure 11 if the S_T signal is omitted.

4.2 3-channel surround hierarchy

A second 3-channel hierarchy based on the 3 notional transmission channels M_T, S_T and B_T arises from considering the four systems A_1 (1:0), A_2 (2:0 stereo), A_4 (2:1 stereo) and A_7 (B-format). Here we have the requirement that the mono presentation should incorporate not just frontal stage sounds, but should also incorporate rear stage sounds, so as to prevent them disappearing in mono, but at a level between 3 and 6 dB down to prevent "ambience" from muddying the mono sound too much. It is desired that the transmission signal B_T convey only rear stage sound without significant cross-talk from the front stage, and it has been found empirically that for large-auditorium reproduction, the front-to-rear stage cross-talk should be at least 20 dB down. This means that the conversion matrix R_{47} from B-format to 2:1 stereo should have a B signal that has a response to as wide a frontal stage as possible that is 20 dB down relative to rear azimuths.

These compatibility requirements yield encoding and conversion matrices as follows. (The matrices for A_1 and A_2 are already given above).

2:1 stereo encoding E_{44}

$$\begin{pmatrix} M_T \\ S_T \\ B_T \end{pmatrix} = \begin{pmatrix} 0.7071 & 0.7071 & 0.6396 \\ 0.7071 & -0.7071 & 0.0000 \\ 0.0000 & 0.0000 & 0.9360 \end{pmatrix} \begin{pmatrix} L_{2F} \\ R_{2F} \\ B \end{pmatrix}$$

B-format encoding E_{77}

$$\begin{pmatrix} M_T \\ S_T \\ B_T \end{pmatrix} = \begin{pmatrix} 0.7500 & 0.1768 & 0.0000 \\ 0.0000 & 0.0000 & 0.6638 \\ 0.4500 & -0.3889 & 0.0000 \end{pmatrix} \begin{pmatrix} W \\ X \\ Y \end{pmatrix} .$$

2:1 stereo decoding D_{44}

$$\begin{pmatrix} L_{2F} \\ R_{2F} \\ B \end{pmatrix} = \begin{pmatrix} 0.7071 & 0.7071 & -0.4832 \\ 0.7071 & -0.7071 & -0.4832 \\ 0.0000 & 0.0000 & 1.0683 \end{pmatrix} \begin{pmatrix} M_T \\ S_T \\ B_T \end{pmatrix}$$

B-format decoding D_{77}

$$\begin{pmatrix} W \\ X \\ Y \end{pmatrix} = \begin{pmatrix} 1.0476 & 0.0000 & 0.4762 \\ 1.2122 & 0.0000 & -2.0203 \\ 0.0000 & 1.5064 & 0.0000 \end{pmatrix} \begin{pmatrix} M_T \\ S_T \\ B_T \end{pmatrix} .$$

These encoding and decoding matrices in turn yield the following conversion matrices:

B-format to 2:1 stereo R_{47}

$$\begin{pmatrix} L_{2F} \\ R_{2F} \\ B \end{pmatrix} = \begin{pmatrix} 0.3129 & 0.3129 & 0.4694 \\ 0.3129 & 0.3129 & -0.4694 \\ 0.4808 & -0.4155 & 0.0000 \end{pmatrix} \begin{pmatrix} W \\ X \\ Y \end{pmatrix}$$

2:1 stereo to B-format R_{74}

$$\begin{pmatrix} W \\ X \\ Y \end{pmatrix} = \begin{pmatrix} 0.7408 & 0.7408 & 1.1158 \\ 0.8571 & 0.8571 & -1.1158 \\ 1.0652 & -1.0652 & 0.0000 \end{pmatrix} \begin{pmatrix} L_{2F} \\ R_{2F} \\ B \end{pmatrix}$$

R_{47} and R_{74} are mutually inverse matrix transformations, and one also has that

$$R_{47} = D_{44}E_{77}\ ,\quad R_{74} = D_{77}E_{44}\ , \qquad (4\text{-}2\text{-}1)$$

by the construction of figure 7, since $Z_4 = Z_7 = \{M_T,\ S_T,\ B_T\}$.

The choice of conversion matrices between B-format and 2:1 stereo is a compromise, since a perfect conversion between these two signal formats does not exist: for example a panned frontal stereo stage in a 2:1 signal cannot conform after any matrixing to B-format across a whole frontal sector. However, R_{47} ensures that sounds across a ±50° frontal stage in B-format are reproduced in 2:1 stereo with a crosstalk onto the B back speaker signals of less than -20 dB, which ensures that even under auditorium conditions, the rear speakers are unlikely to be distracting for frontal stage sounds, and even for a ± 60° wide frontal stage, the front-to rear crosstalk is less than -15.4 dB. Rear to front cross-talk is subjectively less critical , and is set at -15.3 dB.

The gain of the Y signal in the above matrices has been selected so that a ± 60° wide Ambisonic B-format sound stage is reproduced in stereo, either from the 2:1 frontal speakers or from 2:0 speakers with a crosstalk front left to right of -28.70 dB. The various conversion matrices involving mono and 2:0 stereo may be computed to be:

2:1 to 2:0 stereo R_{24}

$$\begin{pmatrix} L_2 \\ R_2 \end{pmatrix} = \begin{pmatrix} 1.0000 & 0.0000 & 0.4523 \\ 0.0000 & 1.0000 & 0.4523 \end{pmatrix} \begin{pmatrix} L_{2F} \\ R_{2F} \\ B \end{pmatrix}$$

which incorporates rear stage sounds at a level of - 3.88 dB,

B-format to 2:0 stereo R_{27}

$$\begin{pmatrix} L_2 \\ R_2 \end{pmatrix} = \begin{pmatrix} 0.5303 & 0.1250 & 0.4694 \\ 0.5303 & 0.1250 & -0.4694 \end{pmatrix} \begin{pmatrix} W \\ X \\ Y \end{pmatrix}$$

which incorporates the due rear azimuth $\theta = 180^\circ$ at a level of - 6.02 dB.

2:0 to 2:1 stereo R_{42}

$$\begin{pmatrix} L_{2F} \\ R_{2F} \\ B \end{pmatrix} = \begin{pmatrix} 1.0000 & 0.0000 \\ 0.0000 & 1.0000 \\ 0.0000 & 0.0000 \end{pmatrix} \begin{pmatrix} L_2 \\ R_2 \end{pmatrix}$$

i.e., frontal stereo is reproduced just from the front 2 speakers,

2:0 to B-format R_{72}

$$\begin{pmatrix} W \\ X \\ Y \end{pmatrix} = \begin{pmatrix} 0.7408 & 0.7408 \\ 0.8571 & 0.8571 \\ 1.0652 & -1.0652 \end{pmatrix} \begin{pmatrix} L_2 \\ R_2 \end{pmatrix}$$

The mono matrices R_{14} or R_{17} can be determined from the stereo ones R_{24} or R_{27} simply by taking 0.7071 times the sum of the two matrix rows, giving respectively for R_{14} and R_{17}

$$C_1 = 0.7071\ (L_{2F}+R_{2F}) + 0.6396\ B$$

$$C_1 = 0.7500\ W + 0.1768\ X\ ,$$

and conversely, R_{41} and R_{71} are given respectively by

$$L_{2F} = 0.7071\ C_1\ ,\ R_{2F} = 0.7071\ C_1,\quad B = 0$$

$$W = 1.0476\ C_1,\ X = 1.2122\ C_1,\ Y = 0\quad .$$

4.3 3:0 conversions

The above also allows the conversions to and from 3:0 reproduction to be determined, using figure 7 in association with subsections 4.1 and 4.2. To determine R_{34} or R_{37}, one first reduces 2:1 stereo or B-format to ordinary 2:0 stereo by means of R_{24} or R_{27}, and then upconverts to 3:0 stereo by means of the psychoacoustic R_{32} matrix given earlier. We do not here give the detailed matrices here, since R_{32} preserves all level balances in its input 2:0 signal (so that the relative balance of front and rear encoded sounds in 3:0 reproduction is the same as in 2:0 reproduction), and otherwise, the results are just those of upconverting 2:0 stereo to 3:0 reproduction.

The converse conversions of 3:0 stereo to 2:1 or B-format are also not complicated, again passing through a downconversion to 2:0 stereo before being upconverted. For example R_{43} is simply the process of reproducing the results of the downconversion R_{23} given in section 4.1 via the front two speakers. R_{73} is a little more complicated, and is given by:

3:0 stereo to B-format R_{73}

$$\begin{pmatrix} W \\ X \\ Y \end{pmatrix} = \begin{pmatrix} 0.5238 & 0.7408 & 0.5238 \\ 0.6061 & 0.8574 & 0.6061 \\ 1.0652 & 0.0000 & -1.0652 \end{pmatrix} \begin{pmatrix} L_3 \\ C_3 \\ R_3 \end{pmatrix}\quad .$$

5. B-FORMAT 3-SPEAKER DECODER

One of the surprises in this work was the discovery reported in this section of a particularly good method of reproducing B-format signals W, X and Y via a frontal 3-speaker layout L_3, C_3 and R_3, considerably better than the discarding of one channel by reduction via R_{37} described in section 4.3, which is effectively only a psychoacoustic upconversion (using the 3×2 decoder of fig. 11) of a 2-channel stereo signal.

The existence of an improved 3-speaker decoder for what is essentially a surround-sound 3-channel format, B-format, means that B-format transmissions can serve a dual purpose: they can be used both for surround-sound reproduction and for frontal-stage 3-speaker reproduction giving enhanced central image stability with respect to a visual image. This makes broadcasting in the 3-channel mode derived from B-format particularly attractive.

However, this enhanced 3-speaker decoder does not naturally fit into the cascadable hierarchy, i.e. it is not a route to providing encoding of other 3-speaker stereo signals L_3, C_3 and R_3 into the B-format signals. It therefore must be considered as purely a reproduction option for B-format, entirely separate to standard 3:0 reproduction from the hierarchy. Thus a 3-speaker stereo receiver has the option of either receiving in 3:0 mode according to the hierarchy by responding to the channels M_T, S_T and T_T, or of receiving M_T, S_T and B_T and decoding these in the manner to be described below. This latter option will normally only be invoked if the receiver detects that the transmission signal T_T is not present, or if a B-format transmission is flagged.

The problem of optimal 3-speaker presentation of B-format for frontal stages was earlier raised by Meares[12] in connection with the use of a sound field microphone for HDTV sound production work. Hitherto, this optimum presentation has not been known, and normally the production engineer has used an adjustable left/right symmetrical matrix in an attempt to optimise presentation. However, it turns out that most presentations via 3 speakers from B-format either have an excessive rear pick-up, which is undesirable, or an excessive degree of cross-talk among the three speakers for sounds across a frontal stage.

Here, we develop the methods used in ref. [7] to design the optimal psychoacoustic 3×2 decoder (shown in fig. 11 of this paper) for reproducing 2-speaker stereo via 3 speakers. That reference used theoretical models for localisation of panned sounds, and showed, by a combination of psychoacoustic theory and experiment that this 3×2 stereo decoder is essentially optimal. We shall not repeat the theoretical material of that paper here (see also ref. [13] for the use of the same theory for optimising 3-speaker panpots), but will merely draw on the results of those investigations.

Essentially, the 3×2 decoder shown in fig. 11 is a close approximation to the energy-preserving 3 × 2 decoder shown in fig. 12 (taken from ref. [7]), where the angle parameter ϕ varies with frequency from 35^o below 5 kHz to about 55^o above 5 kHz.

In terms of the notional transmission signals M, S and T, the operation of the decoder of figure 12 can, as shown in ref. [4] section 6 equs. (29) and (30), be represented as a 3-stage matrixing process on the input L_2 and R_2 signals:

(1) <u>An input MS matrix</u> E_{22}

$$\begin{pmatrix} M_2 \\ S_2 \end{pmatrix} = \begin{pmatrix} 0.7071 & 0.7071 \\ 0.7071 & -0.7071 \end{pmatrix} \begin{pmatrix} L_2 \\ R_2 \end{pmatrix}, \qquad (5\text{-}1)$$

(2) A "rotation" by ϕ - 45^o of the M_2 signal among M and T

$$\begin{pmatrix} M \\ T \end{pmatrix} = \begin{pmatrix} \cos(\phi-45^o) \\ \sin(\phi-45^o) \end{pmatrix} M_2 \qquad (5\text{-}2)$$

(3) An output 3-channel transmission decoding matrix D_{33}

$$\begin{pmatrix} L_3 \\ C_3 \\ R_3 \end{pmatrix} = \begin{pmatrix} 0.5000 & 0.7071 & 0.5000 \\ 0.7071 & 0.0000 & -0.7071 \\ 0.5000 & -0.7071 & 0.5000 \end{pmatrix} \begin{pmatrix} M \\ S \\ T \end{pmatrix}, \qquad (5\text{-}3)$$

and fig. 11 implements this version of the decoder, using the fact that for $35^o \leq \phi \leq 55^o$, $\cos(\phi-45^o)$ nearly equals 1, and $\sin(\phi-45^o)$ approximately equals $\tan(\phi-45^o)$ and so equals about $\pm$ 0.1716 for $\phi = 35.26^o$ and $\phi = 54.74^o$ respectively.

Now the frequency dependent decoder of figs. 11 and 12 for 2-channel signals is designed to provide a more stable central image below 5 kHz, and to provide a greater width and stability of edge images above 5 kHz than a frequency-independent decoder with an average value of $\phi = 45^o$. It has been found by listening that the lower frequencies below about 5 kHz are the most important for central imaging, whereas those above 5 kHz are more critical for the sense of stage width.

Remarkably, studies of frequency-independent 3-speaker matrix decoders for B-format signals show that, if one desires that the total reproduced energy for rear azimuth sounds does not exceed that for frontal azimuths, the stability of central images has to be traded off against that of edge of stage (here defined as azimuth $\pm 60^o$) images, and that the MST matrix defined by

$$\begin{pmatrix} M \\ S \\ T \end{pmatrix} = \begin{pmatrix} 0.4142 & 0.4142 & 0.0000 \\ 0.0000 & 0.0000 & 0.5858 \\ 0.4142 & -0.4142 & 0.0000 \end{pmatrix} \begin{pmatrix} W \\ X \\ Y \end{pmatrix} \qquad (5\text{-}4)$$

when followed by a transmission decoding matrix of equ. (5-3) decodes azimuth $\theta = 0^o$ sounds with exactly the same relative speaker feed gains as does the optimal 3 × 2 psychoacoustic decoder for central images for $\phi = 35.26^o$ (i.e. below 5 kHz), and decodes azimuth $\pm 60^o$ sounds with exactly the same relative gains as does the optimal 3 × 2 psychoacoustic decoder for hard left or right stereo images for $\phi = 54.74^\circ$ (i.e. above 5 kHz).

Thus the frequency-independent 3-speaker decoder from B-format obtained by following an MST matrix as in equ. (5-4) by a transmission decoding matrix D_{33} as in equ. (5-3) achieves, across a ± 60° frontal azimuthal sound stage, the lower-frequency performance for central images and the higher-frequency performance for edge-of-stage images of the psychoacoustic 3×2 decoder. Thus it is already a better performer (across this frontal stage) than the 3×2 decoder.

However, the computed localisation parameters (see refs. [7] and [13]) for central images are still not as good as, say, those for an optimum 3-speaker panning law [13], and those for edge-of-stage images are not as good as for an optimum panpot law even at high frequencies.

5.2 Frequency dependent decoder

Therefore, one seeks a frequency-dependent version of the 3-speaker B-format decoder designed to improve central image localisation further below 5 kHz and to improve edge-of-stage (i.e. azimuths ±60°) localisation above 5 kHz. This is most easily done by a generalisation of fig. 11 shown in fig. 13.

As before, the B-format signal W, X, Y is passed into an MST matrix satisfying equ. (5-4) to derive signals M', S' and T'. However, it is then passed into a frequency-dependent rotation matrix

$$\begin{pmatrix} M \\ T \end{pmatrix} = \begin{pmatrix} \cos(\phi-45^\circ) & -\sin(\phi-45^\circ) \\ \sin(\phi-45^\circ) & \cos(\phi-45^\circ) \end{pmatrix} \begin{pmatrix} M' \\ T' \end{pmatrix} \qquad (5\text{-}5)$$

acting on the M and T signal paths, generalising the 2×1 matrix of equ. (5-2) for the 3×2 case, and then into the output matrix D_{33} of equ. (5-3).

If the T' signal is omitted, this is simply the psychoacoustic 3×2 decoder of figs. 11 and 12, if ϕ varies from 35.26° below 5 kHz to 54.74° above 5 kHz. With the inclusion of the T' signal, the effective value of ϕ for central images is decreased by another 9.74° and the effective value of ϕ for edge of stage images is increased by another 9.74°. This has the effect of making central images almost optimally localised below 5 kHz according to the optimum 3-speaker panpot law of ref. [13], and of improving the edge-of-stage localisation further above 5 kHz.

Thus the decoder of fig. 13, with ϕ varying from approximately 35.26° below 5 kHz to approximately 54.74° above 5 kHz, is very close to a subjectively optimal frequency-dependent 3-speaker stereo decoder for B-format for frontal stage images within azimuths ± 60°. Moreover, because the energy of signals is preserved by all three blocks in figure 13 (which are all proportional to orthogonal matrices), there is no frequency-dependence in the total reproduced energy, only in the way it is distributed among the 3 loudspeakers.

Figure 14 shows a practical approximation to the frequency-dependent rotation matrix of equ. (5-5) for $\phi = 35.26^o$ at low frequencies below 5 kHz and $\phi = 54.74^o$ above 5 kHz, similar to the approximation in fig. 11. This approximation is based on $\cos(\phi-45^o)$ being put equal to 1 and on the all-pass networks, as before, having gains -1 below 5 kHz and + 1 above 5 kHz, and both gains being equal to about 0.1716.

5.3 General 3-speaker B-format decoder

While the above frequency-dependent 3-speaker decoder for B-format is substantially optimal for the azimuthal stage $-60^o \leqslant \theta \leq 60^o$, it is subjectively a poor performer for encoded azimuths near the rear $\theta = 180^o$, since the reproduction is predominantly of a "T" channel component which is very "phasey" and poorly localised. While this impaired rear-stage quality is prevented from getting out of hand by the rear sounds being reproduced with no higher energy gain than the front, it is desirable to reduce the level of these rear stage sounds further if unpleasant side effects are to be minimised.

There are two main types of modification to the decoder of figs. 13 and 14 which reduce the gain of the decoder to rear sounds while retaining the advantages across the frontal azimuthal stage. These are shown in fig. 15.

The first, and easiest to understand, modification is to fit the T' signal with a user-adjustable attenuator between the MST matrix and the rotation matrix. When the attenuation is faded down to zero gain, the decoder acts as a psychoacoustic 3 × 2 stereo decoder acting on the stereo signal whose sum is proportional to W+X and whose difference is proportional to 1.4142 Y, and as T' is faded up to gain 1, the decoder has performance intermediate between this and the full B-format 3-speaker decoding. Because the faded T' component has very low gain to frontal sounds, the effect of this fader is proportionately to reduce the amount of rear stage sound reproduced, with a proportionate reduction in unpleasantness, and some reduction in localisation quality across the frontal stage., but little alteration of stage width.

The second modification is a transformation of the input B-format signal that we term "forward dominance", also user-adjustable. Forward dominance requires a little bit of explanation, although we do not attempt the full theory here. Signals encoded to some (not necessarily known) azimuth θ in B-format are characterised by the fact that the channel gains satisfy the equation

$$2W^2 = X^2 + Y^2, \qquad (5\text{-}6)$$

and any linear transformation of W, X and Y that preserves this relationship produces a new B-format signal, although its gains and distribution of azimuths may have been altered. The most obvious transformation of B-format that preserves the relationship (5-6) is the action of a 2 × 2 rotation matrix on X and Y, which has the effect of rotating the whole azimuthal sound stage. However, another less obvious transformation (which is actually a transformation in the group of Lorentz Transformations - see ref. [14] for the associated mathematical theory, which is usually encountered in the theory of Special Relativity) is the following transformation, known as a Lorentz boost to mathematical physicists, although we shall term it a forward dominance transformation of B-format:

$$W' = \tfrac{1}{2}(\lambda + \lambda^{-1})\ W + 8^{-\frac{1}{2}}(\lambda - \lambda^{-1})\ X$$

$$X' = \tfrac{1}{2}(\lambda + \lambda^{-1})\ X + 2^{-\frac{1}{2}}(\lambda - \lambda^{-1})\ W$$

$$Y' = Y \qquad (5\text{-}7)$$

for a positive parameter λ. It may be checked by simple algebra that if equ. (5-6) holds, then we also have that

$$2W'^2 = X'^2 + Y'^2\ , \qquad (5\text{-}8)$$

so that the transformed signals W', X' and Y' are also B-format signals, with modified encoded azimuth θ' and signal gains. It may be shown that the modified azimuth θ' is given in terms of the original encoded azimuth θ by

$$\cos\theta' = \frac{\mu + \cos\theta}{1 + \mu\cos\theta} \qquad (5\text{-}9)$$

where

$$\mu = (\lambda^2 - 1)/(\lambda^2 + 1)\ . \qquad (5\text{-}10)$$

The effect on gain of the forward dominance transformation (5-7) is to increase the amplitude gain of due front azimuth sounds by a factor λ and to multiply the amplitude gain of due rear sounds by λ^{-1}, resulting in a relative increase of due front gain over due rear gain of λ^2. This gain factor (measured in dB) will be termed the dominance gain, so that $\lambda = 2^{\frac{1}{2}}$ corresponds to a dominance gain of about 6.02 dB, i.e. a relative reduction of rear signals by 6 dB.

Figure 16 shows the change in azimuthal distribution of sounds, calculated via equs. (5-9) and (5-10) caused by a dominance gain of 6dB; it will be seen that the front azimuthal stage is narrowed in width by a factor of about λ^{-1} and that the rear azimuthal stage is widened in width by about a factor λ.

Thus applying a forward dominance transformation at the input of a 3-speaker decoder from B-format, as in figure 15, has the effect (for dominance gain $\lambda^2 > 1$) of decreasing the relative level of rear azimuthal sounds in the speaker outputs by a factor λ^2 and of narrowing the reproduced frontal stage by a factor λ^{-1}, but of otherwise giving a similar localisation quality, since the transformed input W', X', Y' is still a B-format signal.

Thus, by adjustment of both the forward dominance gain λ^2 and the T' attenuation, the relative level of rear-azimuth sounds in the output of the 3-speaker decoder of fig. 15 can be reduced, and the stage width narrowed (or even a little widened for $\lambda < 1$), while still retaining much of the psychoacoustic benefit of this decoder for frontal stage sounds. These adjustments are particularly valuable for providing 3-speaker stereo from the B-format output of a sound field microphone, for production work. For domestic reproduction, fixed values of these parameters may be used which, say reproduce rear stage sounds about 6 dB down relative to frontal azimuth sounds, for example by using a 4 dB attenuation of the T' signal and a 2 dB forward dominance gain.

It will be noted that the total energy from the decoder of fig. 15 from any azimuthal encoded direction in B-format remains frequency-independent in gain, despite the frequency-dependent rotation matrix, since the rotation matrix itself has a total energy gain (unity) which is frequency independent.

5.4 n-speaker stereo from B-format

Finally, it is worth noting that the same decoder architecture shown in figure 15 can also be used to provide n-speaker frontal stereo reproduction from B-format for $n \geqslant 3$, by replacing the output 3×3 transmission decoding matrix D_{33} by an $n \times 3$ transmission decoding matrix D_{n3} as described in ref. [4]; as shown there and in ref. [7], this essentially has the effect of "upconverting" the decoder output from 3 to n-speaker reproduction while hardly changing the reproduced balance or stereo localisation quality

6. 5-CHANNEL HIERARCHY MATRICES

This part of the paper gives basic details about the rest of the 5-channel hierarchy not already specified in section 4 above. Before giving the details of the encoding and decoding matrices E_{ii} and D_{ii}, we would like to make two points clear.

Firstly, the matrices given are provisional in nature, and a slightly amended future choice may give a slightly better "compatibility" performance for the conversion matrices R_{ji}. Nevertheless they form a starting point for refinement that is probably in the right ball park.

Secondly, there are so many conversion matrices R_{ji} to consider here that we have not attempted to list them exhaustively. It is hoped to be able to do this, along with a "psychoacoustic analysis" based on a study of localisation parameters [7,13,15] for each of a large number of reproduction methods once a more comprehensive suite of analysis software has been written by the author. However, initial design studies over a period of several months have analysed enough reproduction options in detail to ensure that more comprehensive analysis will only lead to refinement of the hierarchy, rather than to any major change.

6.1 3:1 stereo and BE-format

The two simplest cases to add to the hierarchy are 3:1 stereo and BE-format. In particular, 3:1 format is simply a natural combination of the earlier equations for 3:0 and 2:1 stereo.

<u>3:1 stereo encoding</u> E_{55}

$$\begin{pmatrix} M_T \\ S_T \\ B_T \\ T_T \end{pmatrix} = \begin{pmatrix} 0.5000 & 0.7071 & 0.5000 & 0.6396 \\ 0.7071 & 0.0000 & -0.7071 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.9360 \\ 0.5000 & -0.7071 & 0.5000 & 0.0000 \end{pmatrix} \begin{pmatrix} L_{3F} \\ C_{3F} \\ R_{3F} \\ B \end{pmatrix}$$

<u>3:1 stereo decoding</u> D_{55}

$$\begin{pmatrix} L_{3F} \\ C_{3F} \\ R_{3F} \\ B \end{pmatrix} = \begin{pmatrix} 0.5000 & 0.7071 & 0.5000 & -0.3198 \\ 0.7071 & 0.0000 & -0.7071 & -0.4832 \\ 0.5000 & -0.7071 & 0.5000 & -0.3198 \\ 0.0000 & 0.0000 & 0.0000 & 1.0683 \end{pmatrix} \begin{pmatrix} M_T \\ S_T \\ T_T \\ B_T \end{pmatrix}$$

<u>BE-format encoding</u> E_{88}

$$\begin{pmatrix} M_T \\ S_T \\ B_T \\ T_T \end{pmatrix} = \begin{pmatrix} 0.7500 & 0.1768 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.6638 & 0.0000 \\ 0.4500 & -0.3889 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & -1.0000 \end{pmatrix} \begin{pmatrix} W \\ X \\ Y \\ E \end{pmatrix}$$

<u>BE-format decoding</u> D_{88}

$$\begin{pmatrix} W \\ X \\ Y \\ E \end{pmatrix} = \begin{pmatrix} 1.0476 & 0.0000 & 0.4762 & 0.0000 \\ 1.2122 & 0.0000 & -2.0203 & 0.0000 \\ 0.0000 & 1.5064 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & -1.0000 \end{pmatrix} \begin{pmatrix} M_T \\ S_T \\ B_T \\ T_T \end{pmatrix}$$

It will be seen that the BE-format equations are essentially the B-format equations, with the E signal being put equal to -T. This is because the gain law across the frontal azimuthal stage of -E is very similar to the gain law of the T_T signal for 3-speaker stereo when sounds are panned across the 3-speaker stereo stage (see, for example, ref. [13] and compare with fig. 5a of this paper).

6.2 2:2 stereo and BF-format

The main innovation in the 5-channel hierarchy considered here as compared with previous proposals is the nature of the fifth notional transmission channel F_T. For m:n stereo systems, transmission channels of the form "mono", "overall stereo difference", "rear mono" and "rear difference" have been considered. While the first three of these correspond to our M_T, S_T and B_T, our signal F_T roughly represents a frontal stage stereo difference signal <u>minus</u> a rear stage stereo difference signal, with a slight extra gain for the rear difference.

The reason for this kind of choice is that it leads to downconversion matrices when it is omitted that retain a component of difference signals across both the front and rear stages, rather than causing the elimination of all difference signal across the rear stage in prior downmixing proposals such as in [9,10,12]. Not only does this help to maintain a true all-round surround sound effect using as few as three transmission channels, but it turns out to allow the largest possible number of supported directional encoding systems in the hierarchy. The choices involved in the F_T signal are most easily discussed in connection with 2:2 stereo ("quadraphonics"), although this system is not the most useful in HDTV applications.

Essentially, the transmission signals M_T and B_T between them convey the sum signals $L_{2F}'+R_{2F}'$ and $L_{2B}+R_{2B}$ of the front and rear 2:2 stereo stages, and the S_T signal conveys (by analogy with its role in B-format encoding) a linear combination

$$a_1\,S_F' + a_2\,S_B = S_T \tag{6-1}$$

of the front-stage difference $S_F' = L_{2F}'-R_{2F}'$ and rear stage differene $S_B = L_{2B}-R_{2B}$, with positive coefficients a_1 and a_2. The F_T transmission signal is required to convey a second linear combination

$$b_1\,S_F' - b_2\,S_B = F_T \tag{6-2}$$

with positive coefficients b_1, b_2 of the front and rear stage difference signals.

In such a system, we look at the downconversion caused by omitting the F_T signal. From equs (6-1) and (6-2), we have:

$$S_F' = (b_2\,S_T + a_2\,F_T)/(a_1b_2 + a_2b_1) \tag{6-3}$$

$$S_B = (b_1\,S_T - a_1\,F_T)/(a_1b_2 + a_2b_1)\ , \tag{6-4}$$

so that if the F_T signal is replaced by zero as a part of a downconversion process, the modified difference signals across the front and rear stages available for reproduction become:

$$S_F'' = b_2S_T/(a_1b_2+a_2b_1) = (a_1b_2S_F' + a_2b_2S_B)/(a_1b_2+a_2b_1) \tag{6-5}$$

and

$$S_B'' = b_1S_T/(a_1b_2+a_2b_1) = (a_1b_1S_F' + a_2b_1S_B)/(a_1b_2+a_2b_1). \tag{6-6}$$

Thus, from equs. (6-5) and (6-6), we see that in order that downconversion should result in a greater difference component across the frontal stage than across the rear stage, one should have $b_2 > b_1$. While this is a provisional choice, studies based on results via a number of different reproduction modes suggest that we should put:

$$b_2 = 1.25\ b_1. \tag{6-7}$$

This has led us to the provisional encoding equations:

2:2 stereo encoding $E_{11\ 11}$

$$\begin{pmatrix} M_T \\ S_T \\ B_T \\ F_T \end{pmatrix} = \begin{pmatrix} 0.7071 & 0.7071 & 0.4523 & 0.4523 \\ 0.7071 & -0.7071 & 0.7071 & -0.7071 \\ 0.0000 & 0.0000 & 0.6619 & 0.6619 \\ 0.7071 & -0.7071 & -0.8839 & 0.8839 \end{pmatrix} \begin{pmatrix} L_{2F}' \\ R_{2F}' \\ L_{2B} \\ R_{2B} \end{pmatrix}$$

This specific choice leads to quite a high level of crosstalk of the rear-stage difference signal onto the reproduced 2:0 stereo difference, but also ensures that rear-stage sounds are "folded down" into 2:0

stereo with an increased width as compared to front stage sounds, which is desirable for a sound stage used primarily for "ambience" effects. Some reduction in the gain (say by a factor 0.8) of the way the rear difference signal is incorporated into S_T may be preferable, but the above is a starting point for $E_{11\ 11}$.

The F signal of BF-format has similarities with the signal F_T proposed above, also being a "front difference minus rear difference" signal, and leads to the encoding matrix:

<u>BF-format encoding</u> E_{99}

$$\begin{pmatrix} M_T \\ S_T \\ B_T \\ F_T \end{pmatrix} = \begin{pmatrix} 0.7500 & 0.1768 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.6638 & 0.0000 \\ 0.4500 & -0.3889 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.6638 \end{pmatrix} \begin{pmatrix} W \\ X \\ Y \\ F \end{pmatrix}.$$

This is essentially B-format encoding, plus a gain adjustment on F to match the Y gain in S_T.

6.3 3:2 stereo and BEF-format

These two cases combine results for 3:1 and 2:2 stereo cases, and for BE- and BF-formats, so that we give the encoding matrices without further explanation.

<u>3:2 stereo encoding</u> E_{66}

$$\begin{pmatrix} M_T \\ S_T \\ B_T \\ T_T \\ F_T \end{pmatrix} = \begin{pmatrix} 0.5000 & 0.7071 & 0.5000 & 0.4523 & 0.4523 \\ 0.7071 & 0.0000 & -0.7071 & 0.7071 & -0.7071 \\ 0.0000 & 0.0000 & 0.0000 & 0.6619 & 0.6619 \\ 0.5000 & -0.7071 & 0.5000 & 0.0000 & 0.0000 \\ 0.7071 & 0.0000 & -0.7071 & -0.8839 & 0.8839 \end{pmatrix} \begin{pmatrix} L_{3F}' \\ C_{3F}' \\ R_{3F}' \\ L_{2B} \\ R_{2B} \end{pmatrix}$$

<u>BEF-format encoding</u> $E_{10\ 10}$

$$\begin{pmatrix} M_T \\ S_T \\ B_T \\ T_T \\ F_T \end{pmatrix} = \begin{pmatrix} 0.7500 & 0.1768 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.6638 & 0.0000 & 0.0000 \\ 0.4500 & -0.3889 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & -1.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.6638 \end{pmatrix} \begin{pmatrix} W \\ X \\ Y \\ E \\ F \end{pmatrix}$$

7. EXTENSIONS OF THE HIERARCHY

Apart from essentially uninteresting modes such as 1:1 or 1:2 stereo, the above 5-channel hierarchy is capable of further future development by the addition of yet more notional transmission channels. The best understood such extension is to 4:0 and 5:0 stereo, shown in figs. 17 to 19, and detailed in refs. [4] to [7].

This involves adding additional notional transmission channels T_{4T} and T_{5T} for handling the respective additional information from the use of four and five frontal-stage speakers. In refs. [4] and [7], it was shown that the associated encoding matrices are:

<u>4:0 stereo encoding</u> $E_{12\,12}$

$$\begin{pmatrix} M_T \\ S_T \\ T_T \\ T_{4T} \end{pmatrix} = \begin{pmatrix} 0.3998 & 0.5832 & 0.5832 & 0.3998 \\ 0.6206 & 0.3389 & -0,3389 & -0.6206 \\ 0.5832 & -0.3998 & -0.3998 & 0.5832 \\ 0.3389 & -0.6206 & 0.6206 & -0.3389 \end{pmatrix} \begin{pmatrix} L_4 \\ L_5 \\ R_5 \\ R_4 \end{pmatrix}$$

<u>5:0 stereo encoding</u> $E_{13\,13}$

$$\begin{pmatrix} M_T \\ S_T \\ T_T \\ T_{4T} \\ T_{5T} \end{pmatrix} = \begin{pmatrix} 0.3394 & 0.4786 & 0.5579 & 0.4786 & 0.3394 \\ 0.5692 & 0.4381 & 0.0000 & -0.4381 & -0.5692 \\ 0.5570 & -0.0373 & -0.6138 & -0.0373 & 0.5570 \\ 0.4381 & -0.5551 & 0.0000 & 0.5551 & -0.4381 \\ -0.2730 & 0.5191 & -0.5585 & 0.5191 & -0.2730 \end{pmatrix} \begin{pmatrix} L_6 \\ L_7 \\ C_5 \\ R_7 \\ R_6 \end{pmatrix}$$

and the decoding matrices $D_{12\,12}$ and $D_{13\,13}$ of these are simply given by the transposes of the above matrices.

As was shown in refs. [4] and [7], use of the encoding matrices for n:0 stereo given here ensures that the upconversion (shown in fig. 19) from one number of stereo speakers to a greater number will substantially preserve the psychoacoustics of the stereo illusion via the larger number of loudspeakers.

The above extra notional transmission channels T_{4T} and T_{5T} can also be used for m:1 and m:2 stereo systems in an obvious way.

Height portrayal for sounds can also be added to the five-channel hierarchy by adding a transmission channel Z_T to convey the full-sphere Z signal shown in figure 4. As noted in [3], full-sphere periphonic systems working from periphonic B-format signals W,X,Y,Z has already been demonstrated (the first public demonstration using ambisonic decoding technology was at the 1980 AES convention in London), and the lack of height portrayal of sounds is currently one of the most notable remaining defects in surround-sound technology.

We expect a future enhanced hierarchy to involve at least eight notional transmission channel signals including M_T, S_T, T_T, B_T, F_T, T_{4T}, Z_T and possibly T_{5T}, and that such an hierarchy may support between fifteen and twenty directional encoding modes. The design of such extremely complex hierarchies of sound reproduction systems is only feasible based on the kind of methodology used in this paper.

8. GENERAL DISCUSSION

The motivation of the work in this paper may be summarised as two key considerations: operational flexibility and the desire for optimum subjective results.

As regards operational flexibility, there is currently a stark choice being faced in standardisation work in sound for HDTV. Either the industry can be locked forever into a single mode of sound directional encoding and reproduction, based on methods originally developed for the cinema around 1940 and optimised for large auditorium work, or a more flexible range of mutually compatible options can be adopted capable of optimising the results under domestic conditions.

Virtually no fundamental research has been done on the directional sound systems proposed for HDTV, certainly nothing comparable to the intensive work on psychovisual aspects and optimisation encountered on the visual side of HDTV. Although the work reported by Meares [9,12] and Theile [10] is fine work of its kind, it is essentially ad-hoc in nature, and the experimental work does little more than test the results of a few ad-hoc proposal with virtually none of the design parameters resulting from proper scientific investigations. For example, this work does not use optimised directional panning technology, since such technology had not been studied or developed at the time, and production trials [9,12] were based on ad-hoc methods of producing signals on equipment designed only for 2:0 stereo use.

The purpose of this paper is not to propose final standards (although we believe that the equations quoted may not be all that different from the best possible), but to indicate that it is possible to design elaborate hierarchies of systems for HDTV and other applications that can take on board very large numbers of directional encoding and reproduction modes, including m:n front/rear stereo systems and fully 360° surround sound systems based on azimuthal directional coding, and capable of being decoded psychoacoustically.

Such complex design work has to be based on quite elaborate theoretical methods, and the methods of this paper, particularly those of sub-section 3.2 based on figs. 6 to 8, allow the constuction of cascadable hierarchies of conversion matrices, allowing different directional encoding systems to be converted for reproduction via any other.

While mathematical, this ability to construct cascadable hierarchies of systems is of the greatest operational importance in production and broadcast work, since it allows free conversion between different sound reproduction methods with the minimum of fuss. In particular, repeated conversions down a long production or broadcast chain do not cause continuing degradation of the reproduced directional effect. Therefore, at the very minimum, it is desirable for the industry to standardise on a set of conversion matrices between directional encoding systems, and to ensure that these conversion matrices are not just "compatible" for material prepared with suboptimal panning technology that happened to be initially available, but that they are compatible on a broad range of material and that they are fully cascadable down a production chain.

The conversion matrices described in Meares [9] do not satisfy these requirements, and are not fully cascadable, as well as not incorporating many directional encoding systems.

A particular problem is that of ensuring compatibility between the cinema industry and domestic applications. Large cinema auditoria are more hostile environments to the reproduction of subtle directional effects than the typical home, since interspeaker time delays are larger, the audience area is larger, and the acoustics are often more dominant. For this reason, cinema sound reproduction can accept somewhat cruder standards of directional reproduction than are possible in the domestic environment. It is unwise to limit HDTV sound to such relatively crude results in the home, although one must ensure that:
(i) HDTV sound can be converted for reproduction in the cinema with directional results that are adequate for cinema applications. In particular, it is important to prevent centre-front-stage sounds from appearing from surround loudspeakers at significant levels.
(ii) Cinema sound must be convertible to HDTV sound formats to give a reasonable recreation of the sound intended in the cinema in the home, for purposes of broadcast or home video releases of films.

Although in the best of all possible worlds, separate sound mixes would be done for home and cinema release, realistically, a single mix must often serve for both, and interconversion matrices satisfying the different needs of the two markets must be provided.

The difference of the home environment also arises from the different nature of the programme material produced for the cinema and home entertainment. Cinema releases generally have a relatively high production budget, and are of a theatrical nature in which the sound is contrived for dramatic effect. However, many broadcast and video-only productions, especially as HDTV technology becomes cheaper, will either be less contrived, for example capturing the "live" sound of an actual event as naturally as possible in documentary, news and many musical applications, or will aim at less dramatic and more functional aims - for example that of separating out the sounds in a chat or quiz show so as to maximise clarity and intelligibility.

In such "non-theatrical" applications, it will often be desired to use effects that would be merely distracting in a "theatrical" cinema presentation - for example to have a large number of musical or backgound sound effect lines distributed around the whole of the 360^{o} sound stage. It is this kind of required operational flexibility that makes it necessary to consider other directional encoding modes for non-cinema use. A future paper [8] considers in detail the technology required for creating an optimum illusion of a full 360° sound stage around a listener while having the improved frontal stage image stability across a listening area necessary for the matching of sound and visual images for HDTV.

Besides achieving an appropriate directional illusion, another function of enhanced directional sound systems is that of improving perceived sound quality and intelligibility, and of reducing listening fatigue. Possibly the main reason why stereo has replaced mono in domestic use is not its directional effect (which is poor in typical badly arranged domestic situations), but the enhanced intelligibility it gives of different sounds in a complicated mix, and its lower listening fatigue. Any system of directional encoding for HDTV should be evaluated not only for directional accuracy, but also for such improved "non-directional" perceptual qualities. It is in this area that ambisonic and other "psychoacoustic" technologies of decoding have particular advantages, for by rendering several different localisation cues mutually consistent, they have a side effect of reducing listening fatigue and of giving an improved and more "relaxed" perceptual sound quality.

8. WHAT SHOULD BE STANDARDISED?

In designing transmission systems, one has a choice of several things that might be standardised. One can standardise the directional encoding system itself (e.g. as 2:0 or 3:2 stereo or B-format), one can standardise the transmission channel signals and their method of encoding and decoding, or one can standardise the conversion matrices.

Since we have shown that it is possible to design conversion matrices between various different directional coding systems, it is actually not strictly neccessary to standardise the encoding system itself, since that system can still be used to "carry" the signals for other encoding systems via a conversion matrix. In particular, nominally 3:2 stereo signals can be used to convey signals for any of the directional encoding systems shown in fig. 9 of this paper, via the conversion matrices R_{ji} derived in this paper. From this point of view, standards establishing 3:2 signal formats do no more than provide a framework for conveying signals for other systems.

Similarly, standards for 3:2 transmission provide no more than a standardised method of conveying signals, and do not restrict the choice of directional encoding method used to source or to reproduce the sound.

There is, however, a preferred transmission method based on using the "notional transmission signals" M_T, S_T, T_T, B_T, F_T of this paper as actual transmission signals, because they are operationally simple, allowing the minimum of signalling flags, simple system conversion simply by muting of channels, simple conversion when limited audio channel capacity forces downconversion, and simple enhancement to future production modes. It is therefore recommended that standards be sought for such "notional transmission signals", and methods of recording and transmission, for example by cable and satellite links, be devised. One advantage of such standards is that if an audio channel is lost, the remaining channels can easily be reallocated to provide a next-best downconverted reproduction.

In some cases, such a use of "notional transmission" signals is not feasible, and either direct "3:2 speaker feed" signals are required, or a coded complex of signals representing them is transmitted as a single package in a standardised format.

However, a second set of standards applicable to this case should address the conversion matrices. The conversion matrices R_{ji} between any any two systems A_i and A_j should be standardised, in order to ensure that repeated or cascaded conversions do not produce degraded or unpredictable results down a production or broadcast chain, using the cascadable hierarchy method of section 3 of this paper.

If such standards are adopted industry wide, then once a mixing engineer has produced a mix in any given directional encoding system A_1 which he or she is happy sounds alright (i.e. is compatible) via other reproduction encoding modes via the standardised conversion matrices, then any user of that signal later in the chain will not need to recheck the compatibility with different modes of this mix, even after it has been subjected to many conversion stages.

If such conversion matrices are not standardised, then every mix will have to be subjected to a full range of subjective compatibility checks every time it is converted to a different mode. In view of the number of possible modes, this will be extremely time consuming.

The standardisation of conversion matrices saves the trouble, time and expense of such repeated compatibility checking of mixes. It does not prevent mixing engineers from using non-standard conversions if desired for artistic or production reasons (e.g. to alter the balance within a mix), but there will then be an onus placed on the engineer to check the compatibility of the resulting modified mix through all the standard conversion matrices.

9. CONCLUSIONS

This paper has described a 5-channel hierarchy supporting 11 different methods of directional coding, including systems of front & rear stereo and various azimuthal 360° directional coding systems including B-format ambisonics, ensuring not only mutual compatibility between modes, but also ensuring that repeated conversions in a long broadcast or production chain do not cause increasing cascade losses or degradation.

One method of transmitting signal in this hierarchy are via "notional" transmission signals M_T, S_T, T_T, B_T, F_T which may be added to to enhance the reproduction mode, or subtracted from to downconvert to a "simpler" mode. Such transmission signals offer considerable operational flexibility in a broadcast and production environment.

However, the work of this paper is not confined to the case when these transmission signals are used, since the resulting "conversion matrices" between different directional encoding systems are useful in their own right, not merely as a method of converting between different directional encoding modes, but as a means of avoiding severe production problems in a long production or broadcast chain.

While the mathematical methods used in deriving such "cascadable" conversion matrices are a little abstract, the results yield a considerable simplification in operational practice, and we strongly recommend the search for a standard for such conversion matrices between proposed directional coding systems. The equations given in this paper form a starting point that is probably fairly close to such a standardised system.

Due to lack of space and time in preparation of this paper, we have not given all the 110 conversion matrices arising out of our hierarchy in explicit form (although they can all be computed from the data in this paper using the methods of section 3.) It is hoped to make this detailed information available, along with detailed theoretical psychoacoustic calculations of performance, once a suite of software currently under development makes this feasible, and those wishing to get this information when available should contact the author.

This paper does, however, give an initial proposal for such an hierarchy of conversion matrices, specified by its "transmission encoding matrices" E_{ii}. It is expected that there be minor changes in the matrix coefficients as more is learned about the tradeoff between different compatibilities, but it is believed that these changes should not be large, and that the overall structure of the transmission hierarchy, shown in figure 9, should remain unchanged.

Besides the details of such a complicated system of conversion matrices between systems, this paper has presented a number of new proposals for encoding and decoding. In particular, it has been shown that there is a particularly attractive 3-channel surround-sound proposal based on B-format, suitable for use when the available number of audio channels is restricted. In particular, this proposal is not only capable of full 360° portrayal of directional sound, and of enhanced front-stage image stability [8] as compared to prior ambisonic methods, but we have described a new optimised method of reproducing such B-format signals via a frontal stereo stage with 3 or more loudspeakers, so as to achieve subjectively optimised reproduction results. This 3-speaker "psychoacoustic" B-format decoder can in particular be used for 3-speaker results from a soundfield microphone.

Thus we have identified the existence of a 3-channel mode for use with TV not only capable of full surround sound, but also particularly suited to 3-speaker stereo presentation. Hitherto, no system using less than 4 channels has been known meeting those needs, and such a system may be particularly suited for example to the provision of enhanced directional sound via existing "stereo" media such as TV stereo or CD, where it may be difficult to find "room" for more than one extra audio channel.

Overall, the problem of working in a world where there are many methods of handling directional audio requires a much more comprehensive system design than has hitherto been attempted. This paper presents the first such attempt to include both front/rear "stereo" and azimuthal coding systems in a unified hierarchical framework, although it is an extension of earlier work [4-7] on front-stage stereo systems and earlier work on Ambisonics [3], and builds up from hierarchical ideas originally developed in connection with Cooper and Shiga's UMX system [16] and the author's old work on periphony [1].

While the details presented in this paper may evolve further as more becomes known, it is hoped that the benefits of a proper hierarchical system design become clear to the professional audio community. Not only will such a design simplify operational problems in day-to-day use, but it will also allow new enhanced directional sound reproduction technologies to be added to the hierarchy as and when they become available or desirable. Three examples: 4-speaker front-stage stereo, 5-speaker front-stage stereo and full-sphere (periphonic) directional sound systems were presented in this paper to indicate such possible future enhancements that can be added. We expect 4-speaker frontal stages and the benefits of portrayal of sound elevation eventually to become of practical significance, although it is impossible to predict on what time scale.

The hierarchical system design of this paper is aimed not only at the immediate needs of HDTV, but has also addressed the problem of ensuring that directional sound via all other media with audio, notably cinema and audio-only media, are fully compatible, allowing exchange of audio between media. While HDTV is the means of pioneering new developments in directional sound, the ramifications of any decisions on standards taken there on other media have to be recognised and to form a part of the engineering decisions taken. It is hoped that the material in this paper contributes to this wider process.

Acknowledgements

I would like to thank many people who have offered me support and encouragement in this work, particularly Dr. Geoffrey Barton, whose interaction contributed greatly to the development of the above results, and David Meares, whose interest and discussions over more than 20 years in this area has proved immensely useful. Thanks are also due to Peter Craven and Stanley Lipshitz for their constant interest and support.

Patent note

Some of the detailed methods described in this paper are the subject of various patent applications.

REFERENCES

[1] M.A. Gerzon, "Periphony: With-Height Sound Reproduction", J. Audio Eng. Soc., vol. 21 no. 1 pp. 2-10 (1973 Jan./Feb.)

[2] M.A. Gerzon, "Criteria for Evaluating Surround Sound Systems", J. Audio Eng. Soc., vol. 25 no. 6, pp. 400-408 (1977 June)

[3] M.A. Gerzon, " Ambisonics in Multichannel Broadcasting and Video", J. Audio Eng. Soc., vol. 33 no. 11, pp. 859-871 (1985 Nov.)

[4] M.A. Gerzon, "Hierarchical Transmission System for Multispeaker Stereo", Preprint 3199 of the 91st Audio Engineering Society Convention, New York (1991 Oct.)

[5] M.A. Gerzon, "Production Approaches to Multispeaker Stereo", Studio Sound (1992 April and May) to be published

[6] M.A. Gerzon, " Problems of Upward and Downward Compatibility in Multichannel Stereo Systems", to be published

[7] M.A. Gerzon, "Optimum Reproduction Matrices for Multispeaker Stereo", Preprint 3180 of the 91st Audio Engineering Society Convention, New York (1991 Oct.)

[8] M.A. Gerzon, " Ambisonic Decoders for HDTV", to be presented at the 92nd Audio Engineering Society Convention, Vienna (1992 March)

[9] D.J. Meares, "High Quality Sound for High-Definition Television", Proc. 10th International AES Conference "Images of Audio", London, (1991 Sept.) pp. 163- 177

[10] G. Theile, " HDTV Sound Systems: How Many Channels?", Proc. 10th AES International Conf. "Images of Audio", London (1991 Sept.) pp. 147-162

[11] M.A. Gerzon, "Directional Masking Coders for Multichannel Subband Audio Data Compression Systems", Preprinted at the 92nd Audio Engineering Society Convention, Vienna (1992 March)

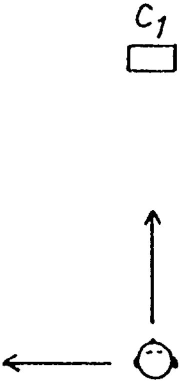
[12] D.J. Meares "High Definition Sound for High Definition Television", Proc. 9th AES International Conference "Television Sound Today and Tomorrow", Detroit Michigan (1991 Feb.) pp. 187-215

[13] M.A. Gerzon, "Panpot Laws for Multichannel Stereo", Preprinted at the 92nd Audio Engineering Society Convention, Vienna (1992 Mar.)

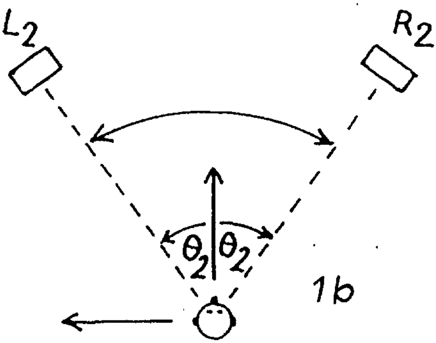
[14] I.M. Gel'fand, R.A. Minlos & Z. Ya Shapiro, "Representations of the Rotation and Lorentz Groups and their Apllications", The Macmillan Company, New York, 1963

[15] M.A. Gerzon, "General Metatheory of Auditory Localisation", Preprinted at the 92nd Audio Engineering Society Convention, Vienna, (1992 March)

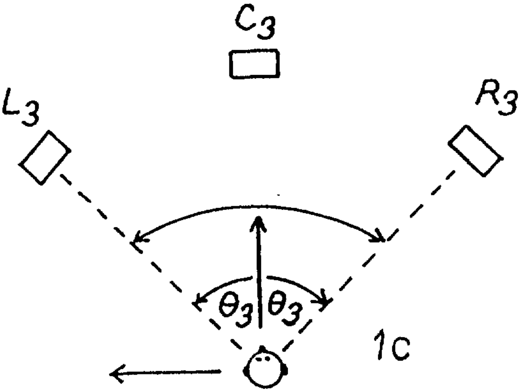
[16] D.H. Cooper and T. Shiga, "Discrete-Matrix Multichannel Stereo", J. Audio Eng. Soc., vol. 20 no. 6, pp. 346-360 (1972 June)


C_1

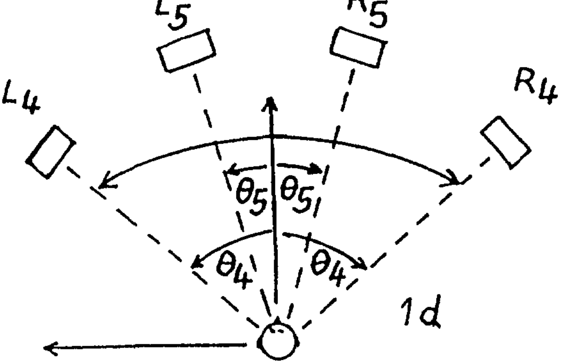
1a


L_2
R_2
θ_2
θ_2

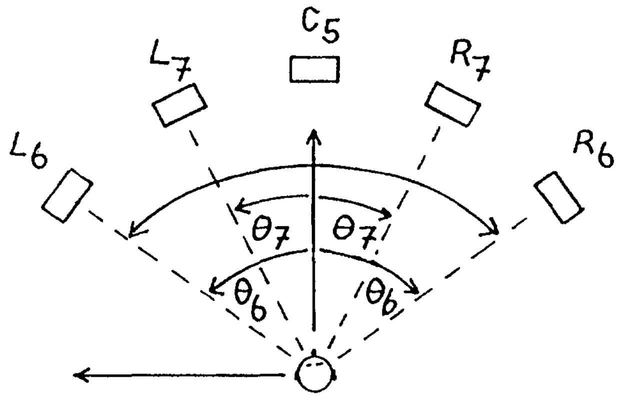
1b


C_3
L_3
R_3
θ_3
θ_3

1c


L_5
R_5
L_4
R_4
θ_5
θ_5
θ_4
θ_4

1d



1e

Figure 1. Loudspeaker layouts for front-stage stereo using in figs. la to le respectively from one to five loudspeakers indicating angles and speaker symbols.

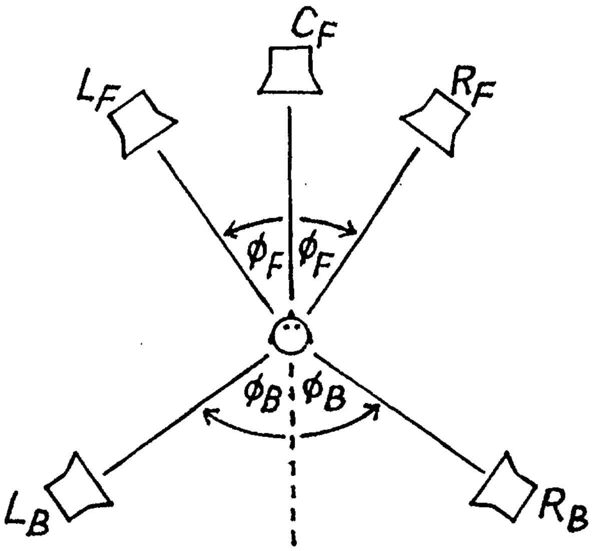


Figure 2. 5-speaker 3:2 stereo layout illustrating front and rear stereo stages.

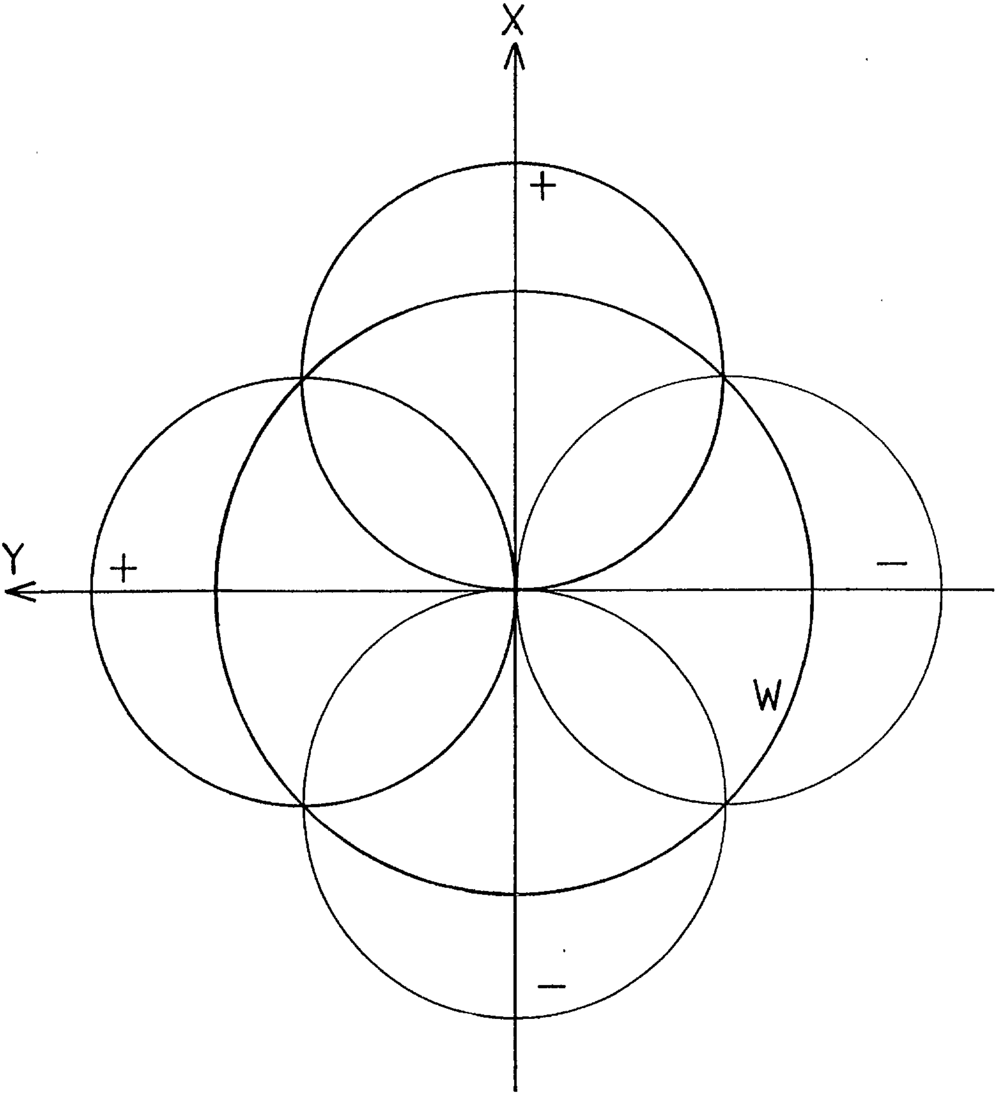


Figure 3. Polar directional patterns for horizontal B-format directionally encoded signals W, X and Y.

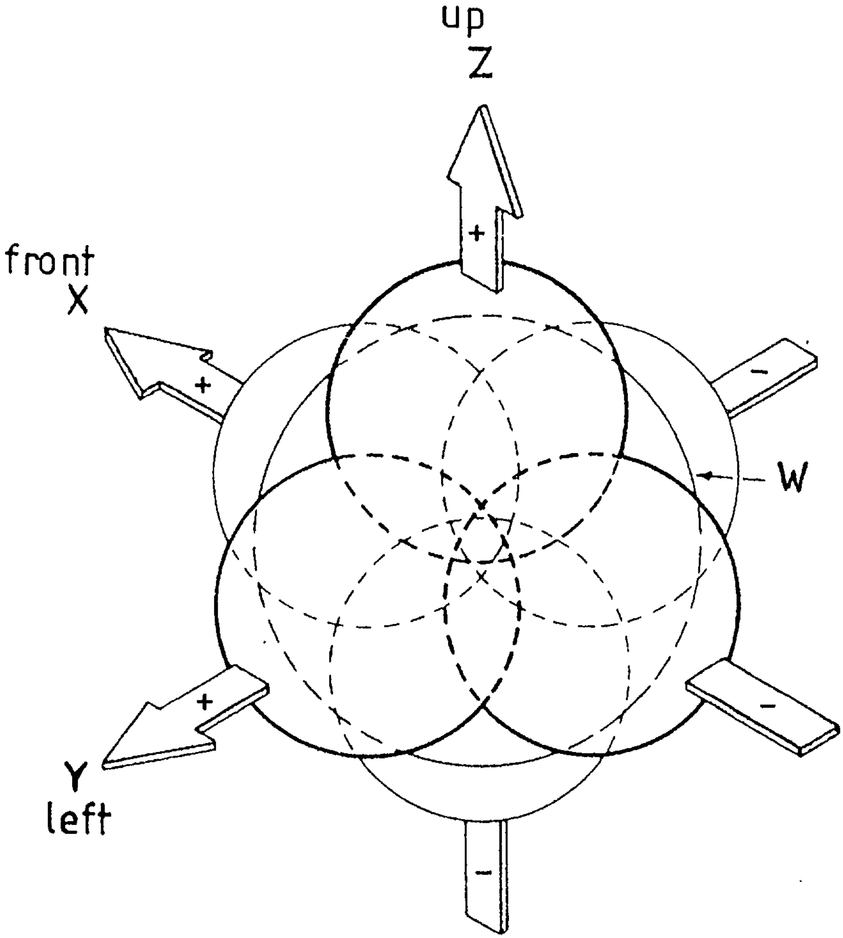


Figure 4. B-format directional polar patterns for W, X, Y and Z for full-sphere directionality.

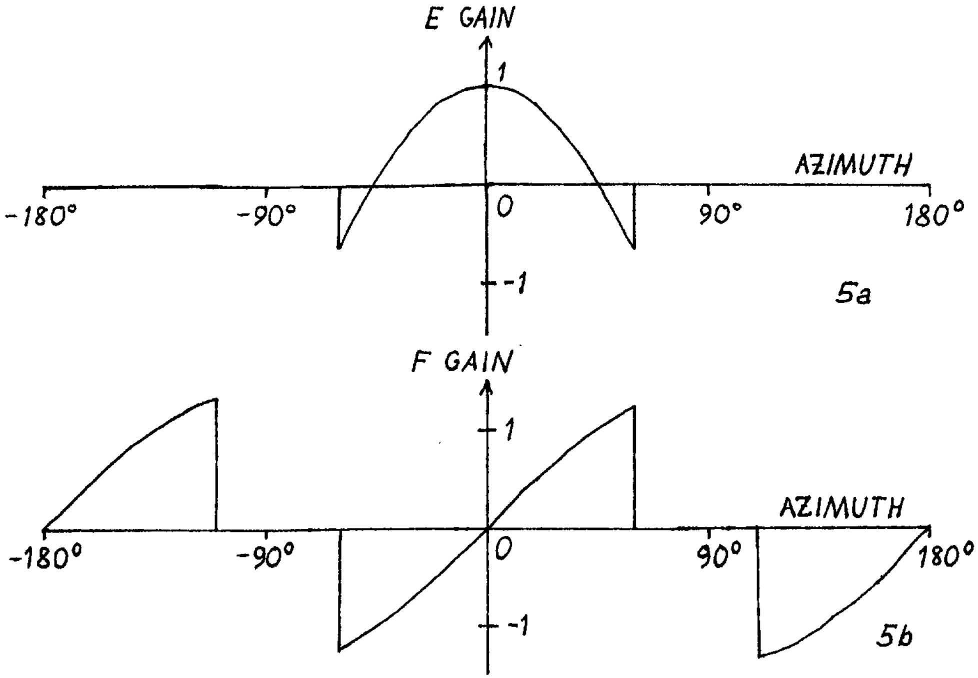


Figure 5. Gain as a function of direction azimuth θ for the signals E (figure 5a) and F (figure 5b) for $\theta_S = 60^{\circ}$, $\theta_B = 70^{\circ}$, $k_g = 3.25$ and $k_d = k_f = k_b = 1$.

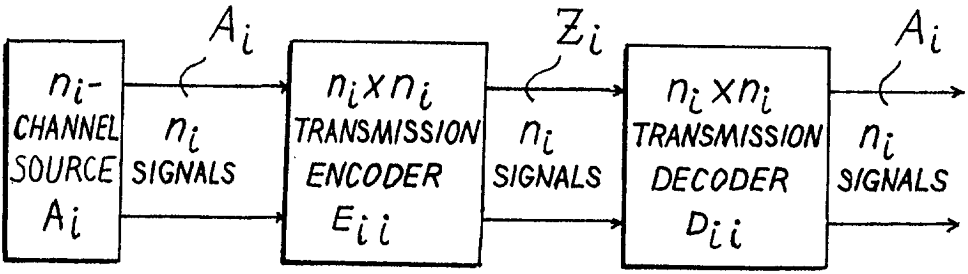


Figure 6. Showing the encoding E_{ii} to and decoding D_{ii} from notional transmission channel signals Z_i for signals of a directional encoding system A_i.

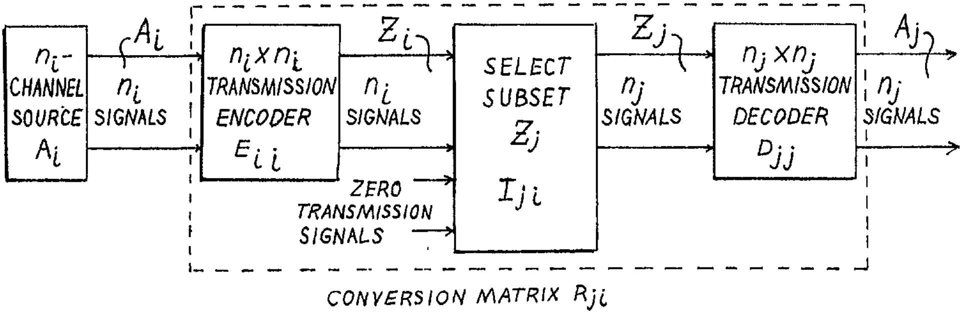

n_i-CHANNEL SOURCE A_i
A_i
n_i SIGNALS
$n_i \times n_i$ TRANSMISSION ENCODER E_{ii}
Z_i
n_i SIGNALS
ZERO TRANSMISSION SIGNALS
SELECT SUBSET Z_j I_{ji}
Z_j
n_j SIGNALS
$n_j \times n_j$ TRANSMISSION DECODER D_{jj}
A_j
n_j SIGNALS
CONVERSION MATRIX R_{ji}

Figure 7. Schematic showing the construction of an $n_j \times n_i$ conversion matrix R_{ji} from a system A_i to a system A_j via intermediate notional transmission channels.

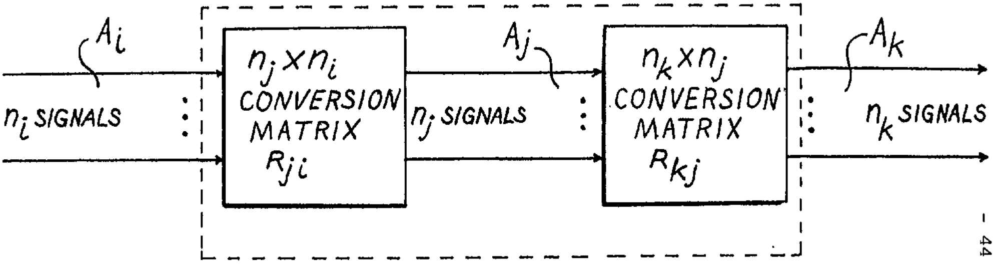

A_i
n_i SIGNALS
$n_j \times n_i$ CONVERSION MATRIX R_{ji}
A_j
n_j SIGNALS
$n_k \times n_j$ CONVERSION MATRIX R_{kj}
A_k
n_k SIGNALS

Figure 8. Showing the cascade of a conversion matrix R_{kj} with a conversion matrix R_{ji} from systems A_i to A_j to A_k.

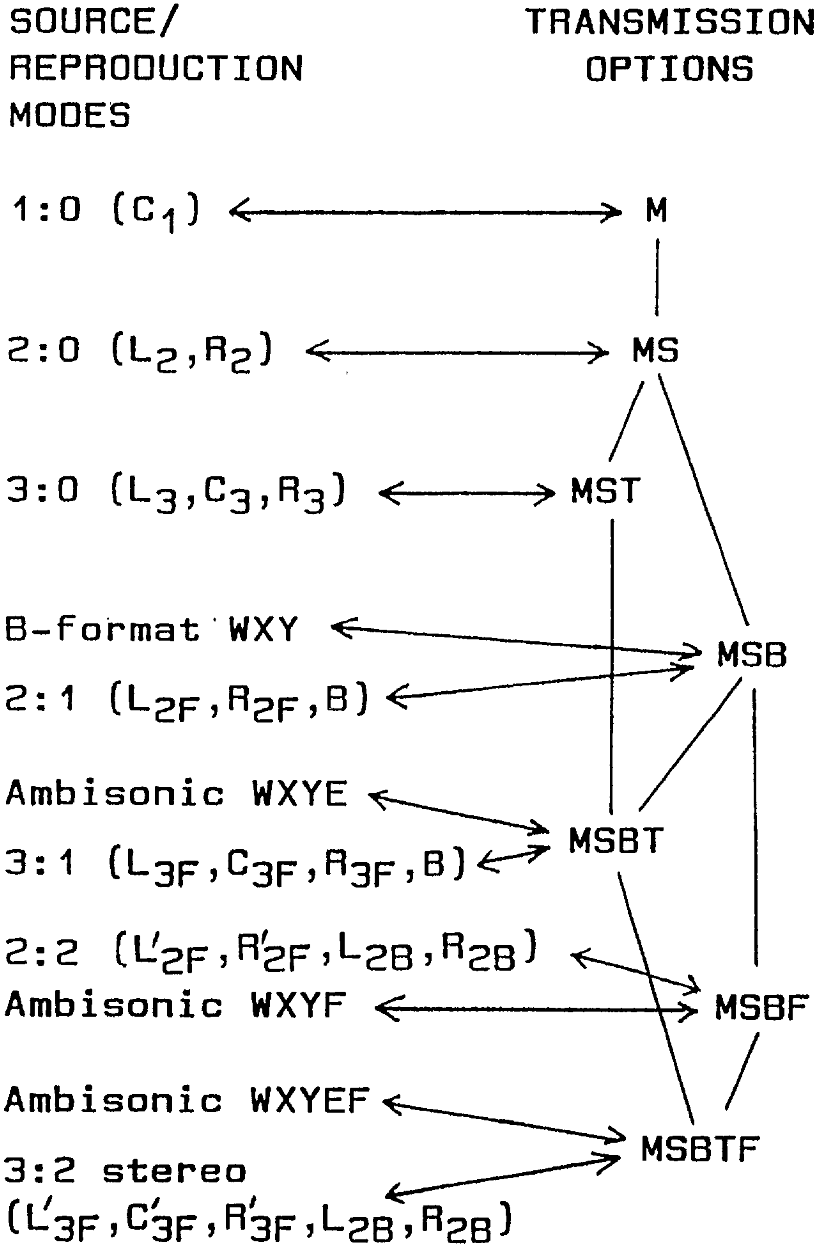

SOURCE/ REPRODUCTION MODES
TRANSMISSION OPTIONS
1:0 (C_1)
M
2:0 (L_2,R_2)
MS
3:0 (L_3,C_3,R_3)
MST
B-format WXY
2:1 (L_{2F},R_{2F},B)
MSB
Ambisonic WXYE
3:1 (L_{3F},C_{3F},R_{3F},B)
MSBT
2:2 (L'_{2F},R'_{2F},L_{2B},R_{2B})
Ambisonic WXYF
MSBF
Ambisonic WXYEF
3:2 stereo
(L'_{3F},C'_{3F},R'_{3F},L_{2B},R_{2B})
MSBTF

Figure 9. Showing the hierarchical structure of the 5-channel cascadable hierarchy of 11 directional encoding systems described in this paper.

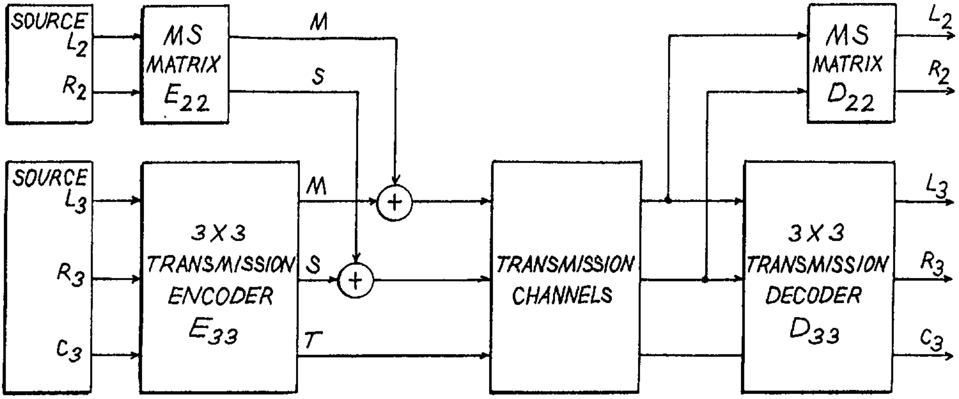


Figure 10. The hierarchy for mono, 2-speaker and 3-speaker frontal stage stereo.

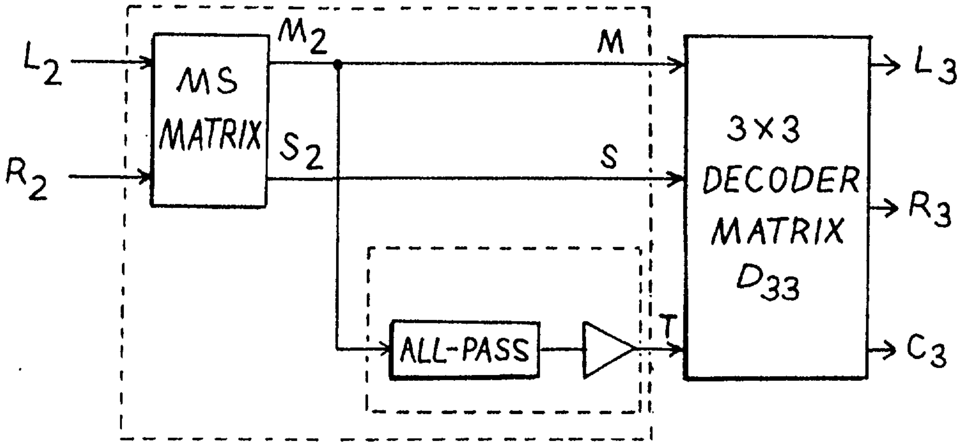


Figure 11. Frequency-dependent psychoacoustically optimised 3-speaker decoder for 2-speaker stereo material, using a synthetic third transmission channel T.

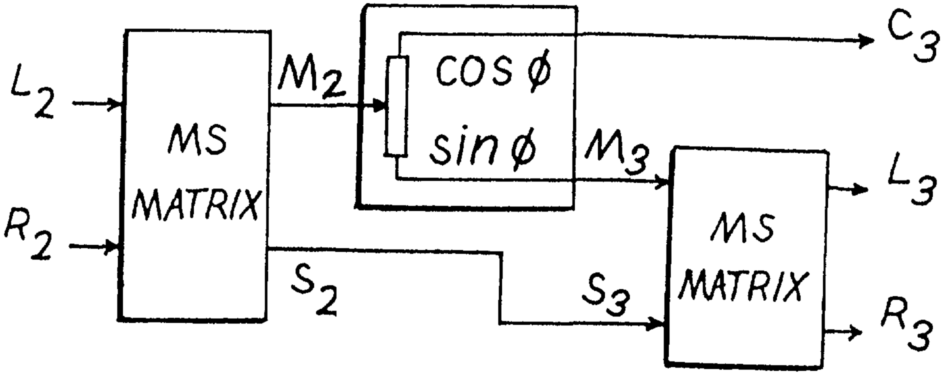


Figure 12. 3 × 2 energy-preserving reproduction matrix decoder for converting 2-speaker stereo into 3-speaker feeds (from ref. [7]).

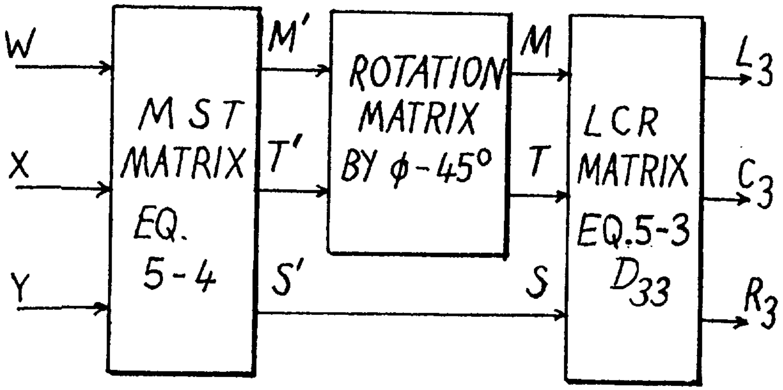


Figure 13. Frequency-dependent 3-speaker decoder for B-format, using frequency-dependent rotation matrix.

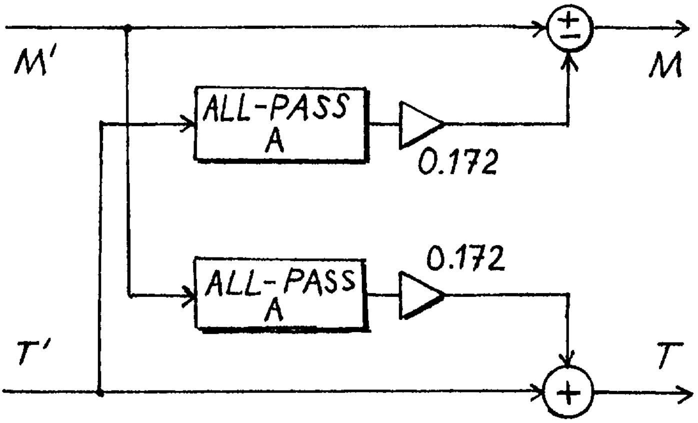


Figure 14. Practical form of the rotation matrix in figure 13, with gains 0.1716 and all-pass networks with gain -1 below 5 kHz and + 1 above 5 kHz.

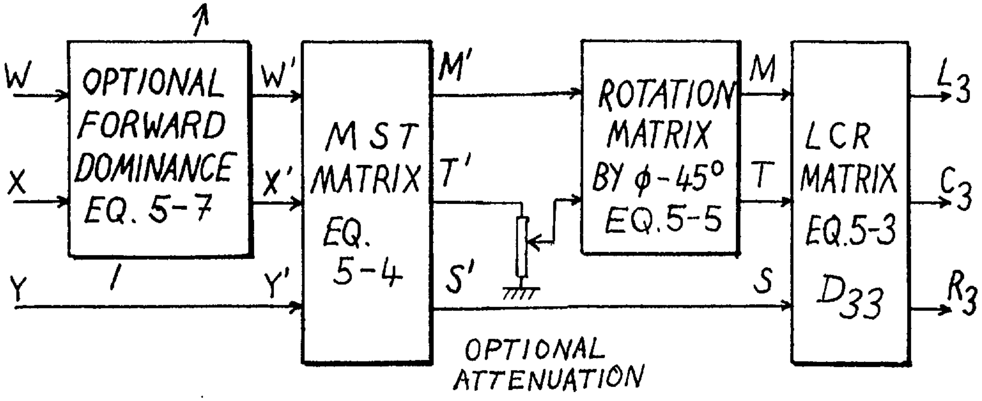

W
X
Y
OPTIONAL FORWARD DOMINANCE EQ. 5-7
W'
X'
Y'
M S T MATRIX EQ. 5-4
M'
T'
S'
OPTIONAL ATTENUATION
ROTATION MATRIX BY ϕ-45° EQ.5-5
M
T
S
LCR MATRIX EQ.5-3 D_{33}
L_3
C_3
R_3

Figure 15. General 3-speaker decoder for B-format with frequency-dependent rotation matrix, incorporating initial user-adjustable forward dominance transformation and T-channel attenuation before the rotation matrix.

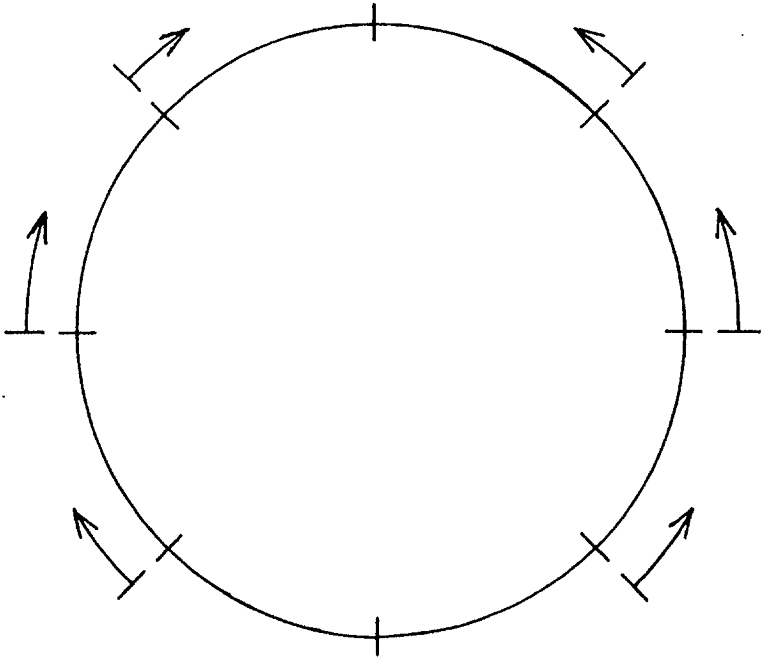


Figure 16. Effect on azimuthal distribution of 6 dB forward dominance, i.e. with $\lambda = 2^{\frac{1}{2}}$.

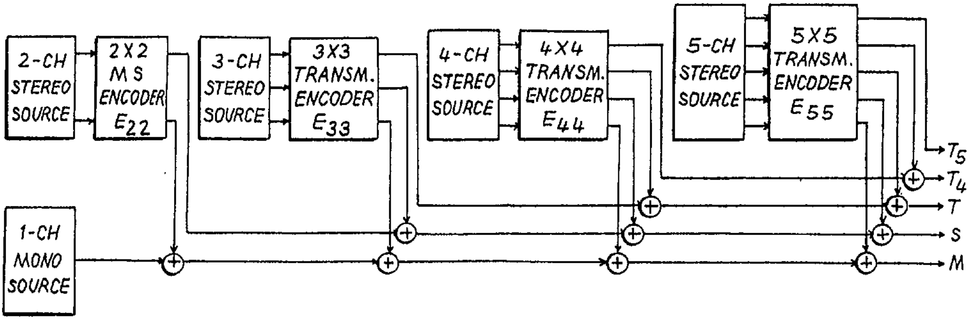


Figure 17. Encoding n-speaker frontal stage stereo into transmission channel signals for n = 1 to 5, taken from ref. [4]. Note that numbering conventions for the E_{ii} matrices differ from earlier in this paper.

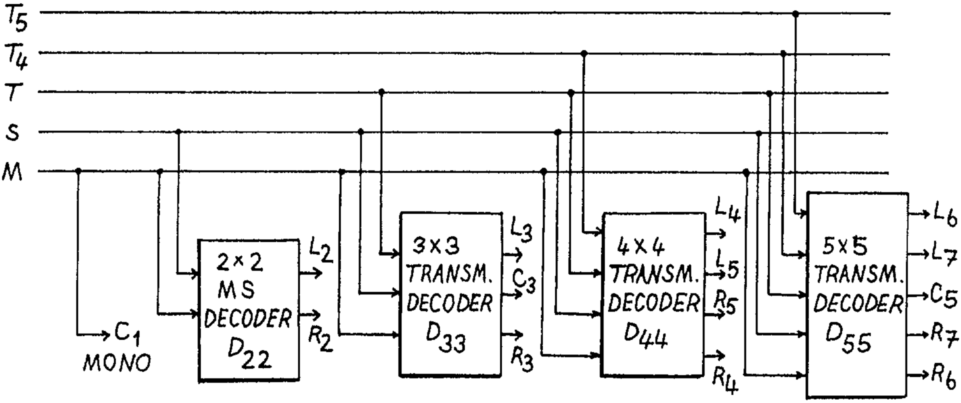


Figure 18. The Decoding inverse to that of fig. 17 for n-speaker frontal stage stereo for n = 1 to 5. Again, the numbering conventions for the D_{ii} matrices differ from earlier in the paper.

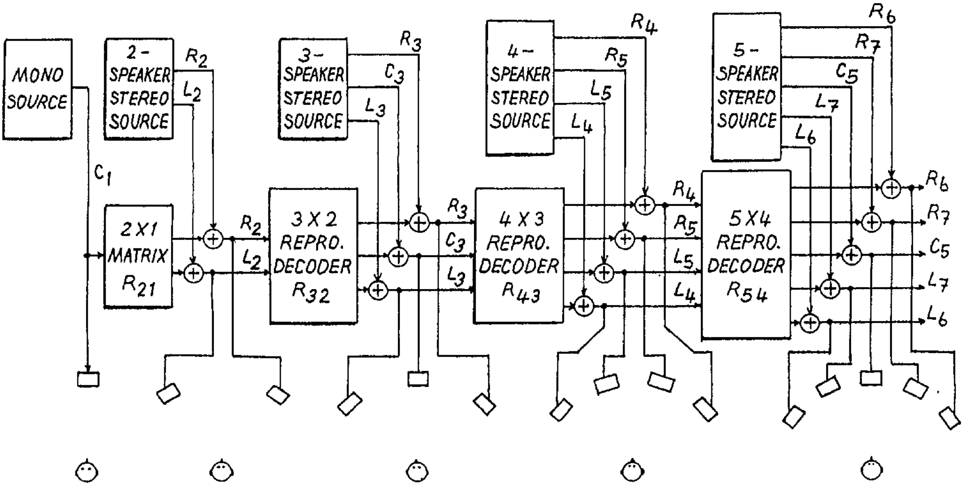


Figure 19. Showing the cascadable hierarchy of upconversions for n-speaker frontal stage stereo, taken from ref. [7]. The numbering conventions for R_{ji} differ from earlier in this paper, but are consistent with figs. 16 & 17.