デ高作 19－4－4

一 般 社 団 法 人 電 波 産 業 会
デジタル放送システム開発部会
音 声 符 号 化 方 式 作 業 班

高度地上デジタルテレビジョン放送方式
適用技術検討報告
音声符号化方式

2023 年 2 月 27 日

3.5　情報源符号化方式

3.5.1　映像符号化方式

3.5.2　音声符号化方式

3.5.2.1　音声符号化方式選定の基本的考え方

音声符号化方式について、高度地上デジタルテレビジョン放送方式に対する要求条件を基に、現行の地上デジタルテレビジョン放送で採用されている MPEG-2 AAC と比較して、より高効率であるとともに、多様な音声サービスを実現できるオブジェクトベース音響に対応した音声符号化方式を選定した。

3.5.2.1.1　チャネルベース音響

放送局の番組制作では、各音声素材（ナレーションのようなダイアログ、背景音、効果音等）を収音して最終的に一つ（または複数）のチャネル構成（例えば、2ch ステレオや 5.1ch）にまとめて放送される。受信側ではチャネル構成（2ch ステレオの場合は 1ch 目が左チャネル、2ch 目が右チャネル）に対応するスピーカから再生することにより、制作意図のまま番組音声を楽しめる。このように制作時のチャネル構成と受信機側の再生チャネル構成が同一であることを前提に受信側のスピーカからそのまま再生される音声信号を制作/伝送する音響方式をチャネルベース音響と呼ぶ。

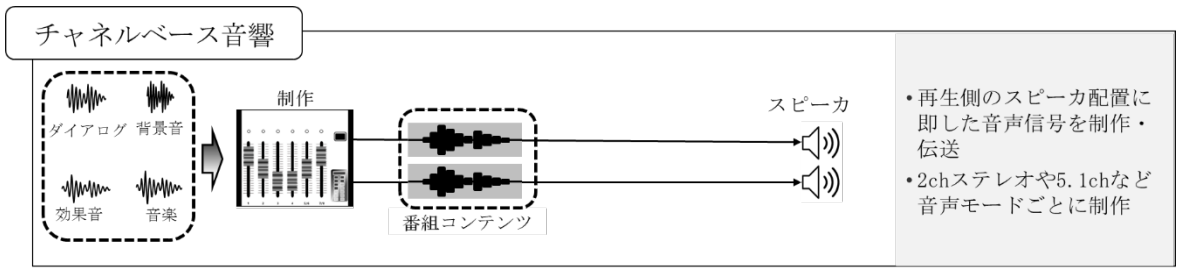


図 3.5.2.1.1-1 チャネルベース音響

3.5.2.1.2　オブジェクトベース音響

放送局の番組制作では、各音声素材（ナレーションのようなダイアログや背景音、音楽や効果音等）を収音して、個別に音声素材と音声素材を再生する位置や音量などを記述した音響メタデータとともに放送される。受信側では、音響メタデータを基に受信機のスピーカ構成に合わせて番組音声を再構成してスピーカから再生することにより、視聴環境や視聴者の好みに合わせて番組音声をカスタマイズして楽しめる。このように視聴環境に応じて信号処理することを前提に音声素材となる音声信号と音響メタデータを制作/伝送する音響方式をオブジェクトベース音響と呼ぶ。

オブジェクトベース音響では、音声素材を個別に受信できるため、視聴環境のスピーカ配置に最適化したり、ナレーションなどダイアログの音量だけを個別に調整することで聞き取りやすく

したり、少ない音声信号数で多言語放送に対応することで様々な言語の話者に正しく情報を伝えたりするなど、視聴環境や視聴者の好みに合わせた音声サービスをきめ細かく提供できる。

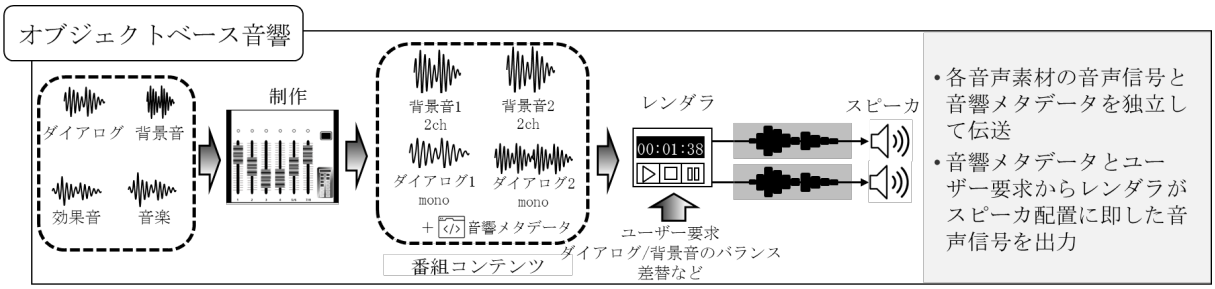


図 3.5.2.1.2-1 オブジェクトベース音響

3.5.2.2　音声入力フォーマット

3.5.2.2.1　標本化周波数

音声信号の標本化周波数は、48kHz とする。

（理由）

・　現状の実運用動向を鑑み、標本化周波数は 48kHz のみとした。
・　48kHz 以外の標本化周波数については、実運用動向から放送局設備へのインパクトが大きいため、また、当面サービスが想定されていないため。

3.5.2.2.2　入力量子化ビット数

入力量子化ビット数は、16 ビット以上とする。

（理由）

・　現状の実運用動向を鑑み、量子化ビット数は 16 ビット以上とした。

3.5.2.2.3　対応する音声信号

対応する音声信号は、チャネルベース音響とオブジェクトベース音響の音声信号とする。

音声信号のうち、同時に再生される可能性があるすべての音声信号の標本化の時刻は、同一時刻であることとする。

（理由）

・　チャネルベース音響に加え、多様な音声サービスを実現可能なオブジェクトベース音響を採用したため。

3.5.2.2.4　入力音声チャネル数

最大入力音声チャネル数は、56 チャネルとする。

（理由）

・ 22.2 マルチチャネル音響に対応し、且つ、オブジェクトベース音響を用いた音声信号の差し替えによる音声サービスを考慮して、MPEG-H 3D Audio のレベル 4 で規定された最大入力音声チャネル数としたため。

3.5.2.3　音声符号化方式

3.5.2.3.1　MPEG-H 3D Audio

3.5.2.3.1.1　準拠規格

ISO/IEC23008-3 Information technology -High efficiency coding and media delivery in heterogeneous environments - Part 3:3D audio (3rd Edition : 2022)

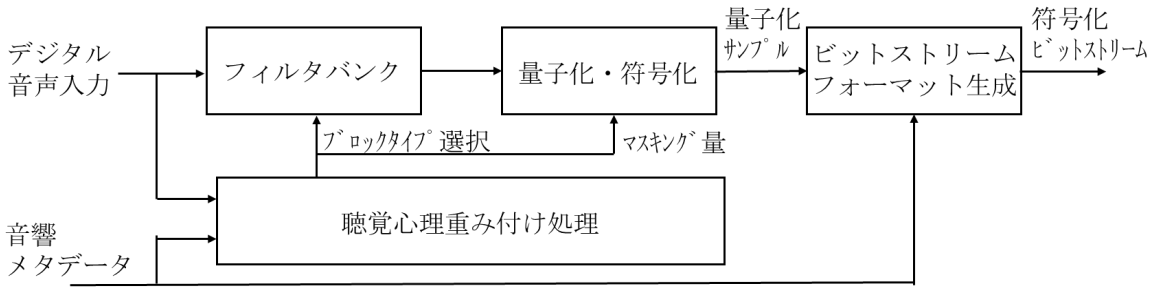
プロファイルは ISO/IEC23008-3 に準拠する Baseline profile とする

（理由）

・ 現行の地上デジタルテレビジョン放送で採用されている MPEG-2 AAC と比較して、より高効率であるとともに、多様な音声サービスを実現できるオブジェクトベース音響に対応しているため。
・ オブジェクトベースによる音声サービスを実施可能であり、且つ、回路規模が最も小さいプロファイルであるため。

3.5.2.3.1.2　音声信号の圧縮手順及び送出手順

圧縮手順は以下の通りとする。



符号化は、時間周波数変換符号化方式及び聴覚心理重み付けビット割当方式を組み合わせたものとする。

1. フィルタバンクは、デジタル音声入力信号を変形離散コサイン変換によって時間から周波数軸へ変換する。この際、フィルタバンクは、入力信号の聴覚心理特性に応じて、変形離散コサイン変換への入力ブロックタイプ及び窓関数を選択する。
2. 聴覚心理重み付け処理は、フィルタバンクへの入力信号に対応して、マスキング量（一の音声信号と他の音声信号を識別できる限界）及びフィルタバンクの入力ブロックタイプを算出する。
3. 量子化及び符号化は、聴覚心理重み付け処理で計算されたマスキング量に基づき、フィ

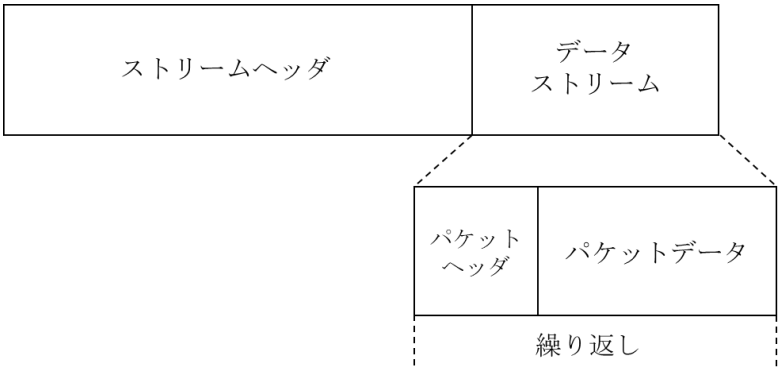
ルタバンクからの出力信号を各ブロックで使用できるトータルビット数を超えない範囲で量子化及び符号化し、量子化サンプルを出力する。

4. 入力音声信号の最大チャネル数は 56 チャネルであり、一つ以上のチャネルベース音響またはオブジェクトベース音響の音声信号によって構成される。音響メタデータは入力音声信号の属性を示す情報であり、各処理で参照、また量子化及び符号化され、ビットストリームに多重化される。
5. 符号化ビットストリームの構成は、3.5.2.3.1.3 節のビットストリーム形式 MHAS (MPEG-H 3D Audio Stream)を使用するものとする。

3.5.2.3.1.3　ビットストリーム形式

送出手順は以下の通りとする。

(MHAS 形式のビットストリーム構成)



1. ビットストリーム構成は、ISO/IEC 23008-3 記載の MPEG-H 3D Audio Stream(MHAS)に準拠するものとする。
2. ストリームヘッダは、ビットストリームおよび後に続くデータストリームの属性、制御等に関する情報を含み、ISO/IEC23008-3 に記載の１つ以上パケットで構成されるパケット列である。
3. データストリームは、データストリームに関するヘッダ情報とパケットデータを含み、ISO/IEC 23008-3 で規定される符号化データのパケット、またはその他のパケットであり、データストリーム毎に繰り返す。
4. パケットヘッダは、パケットの種類を示す識別番号、パケットの長さ等の情報から構成される。
5. パケットデータは、パケットヘッダの情報に基づく入力信号の符号化されたデータ、あるいは符号化データの復号のための制御情報等から構成される。

3.5.2.3.2　AC-4

3.5.2.3.2.1　準拠規格

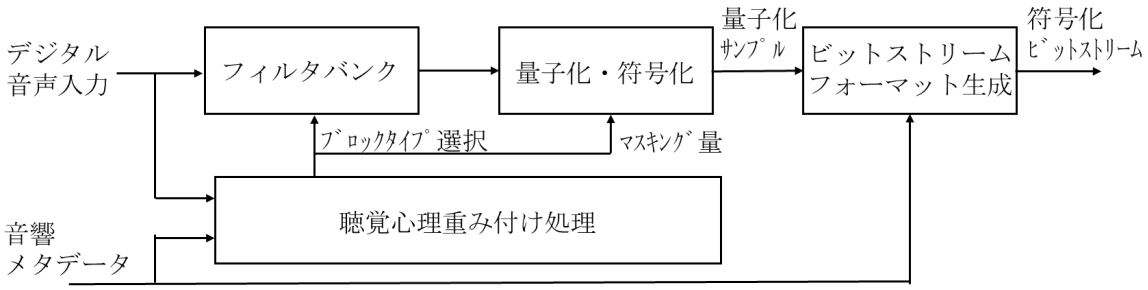
ETSI TS 103 190-2 V1.2.1 (2018-02) Digital Audio Compression (AC-4) Standard; Part 2: Immersive and personalized audio

（理由）

・ 現行の地上デジタルテレビジョン放送で採用されている MPEG-2 AAC と比較して、より高効率であるとともに、多様な音声サービスを実現できるオブジェクトベース音響に対応しているため。

3.5.2.3.2.2　音声信号の圧縮手順及び送出手順

圧縮手順は以下の通りとする。



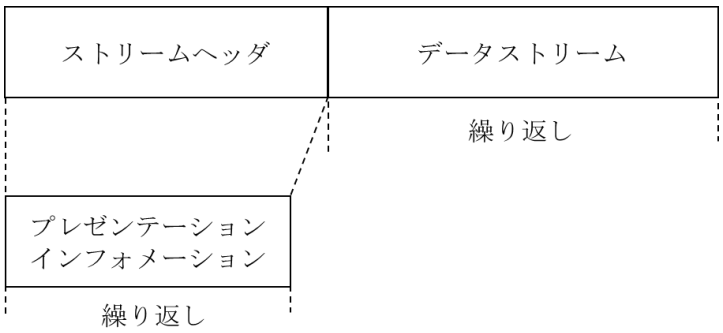
符号化は、時間周波数変換符号化方式及び聴覚心理重み付けビット割当方式を組み合わせたものとする。

1. フィルタバンクは、デジタル音声入力信号を直交ミラーフィルタと変形離散コサイン変換によって時間から周波数軸へ変換する。この際、フィルタバンクは、入力信号の聴覚心理特性に応じて、変形離散コサイン変換への入力ブロックタイプを選択する。
2. 聴覚心理重み付け処理は、フィルタバンクへの入力信号に対応して、マスキング量（一の音声信号と他の音声信号を識別できる限界）及びフィルタバンクの入力ブロックタイプを算出する。
3. 量子化及び符号化は、聴覚心理重み付け処理で計算されたマスキング量に基づき、フィルタバンクからの出力信号を各ブロックで使用できるトータルビット数を超えない範囲で量子化及び符号化し、量子化サンプルを出力する。
4. 入力音声信号の最大チャネル数は 56 チャネルであり、一つ以上のチャネルベース音響またはオブジェクトベース音響の音声信号によって構成される。音響メタデータは入力音声信号の属性を示す情報であり、各処理で参照、また量子化及び符号化され、ビットストリームに多重化される。
5. 符号化ビットストリームの構成は、3.5.2.3.2.3 節のビットストリーム形式 raw AC-4 frame を使用するものとする。

3.5.2.3.2.3　ビットストリーム形式

送出手順は以下の通りとする。

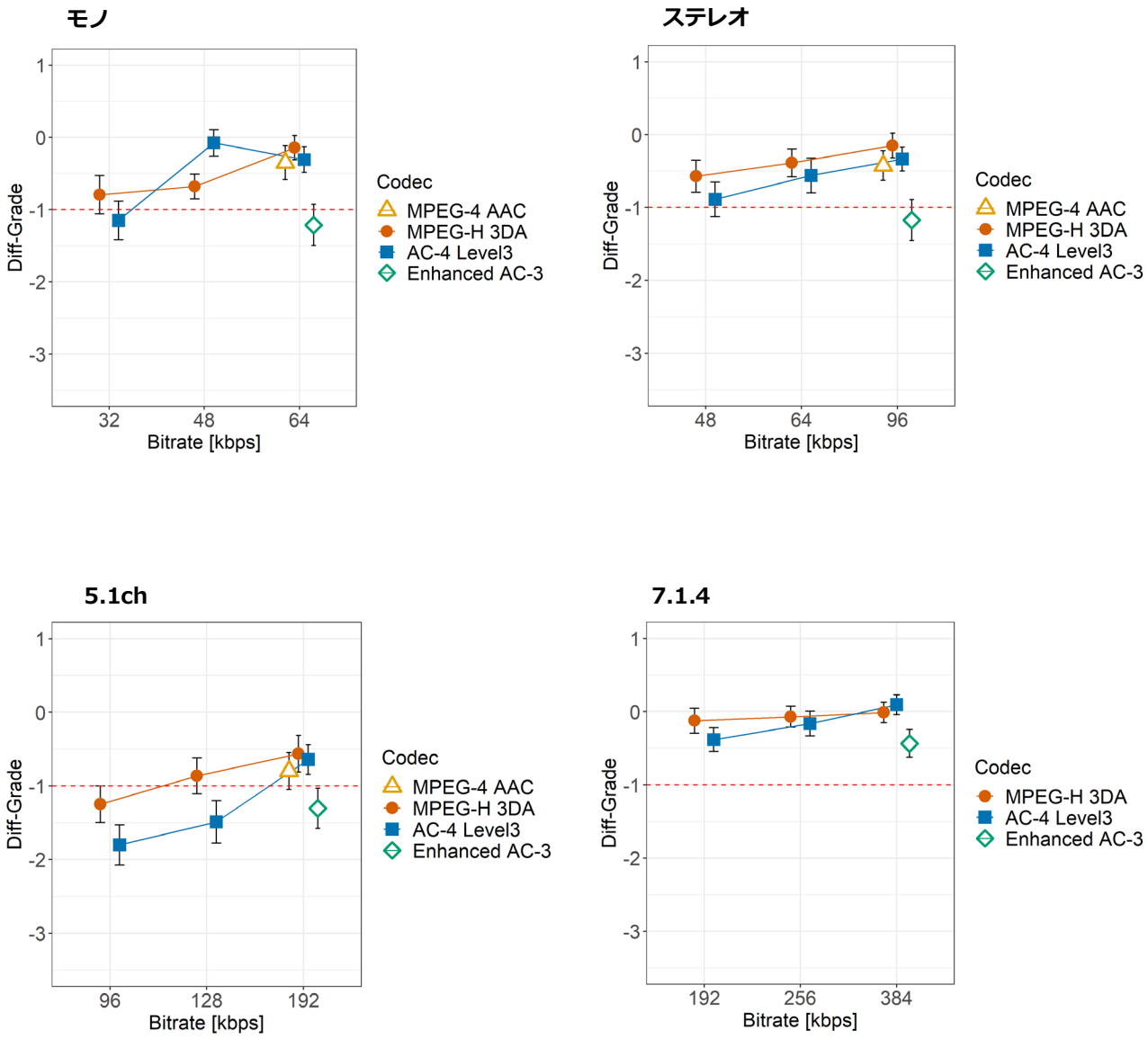
（raw AC-4 frame 形式のビットストリーム構成）



1. ビットストリーム構成は、 ETSI TS 103 190-2 に規定される raw AC-4 frame に準拠するものとする。
2. ストリームヘッダは、ETSI TS 103 190-2 に規定される形式で伝送する音声素材の組み合わせを示すプレゼンテーションインフォメーションを含むものとする。
3. データストリームは、ETSI TS 103 190-2 に規定される形式で音声素材が符号化されているものであり、伝送する音声素材の数だけ繰り返して含むものとする。
4. 音声素材の組み合わせを示すプレゼンテーションインフォメーションは、ETSI TS 103 190-2 に規定される形式で音声素材の組み合わせの数だけ繰り返して含むものとする。

参考１　音声符号化方式の品質比較のための主観評価実験結果

各音声符号化方式の音質を比較するために、同じ音源を同一ビットレートで符号化・復号した評価音を用いて、勧告 ITU-R BS.1116 に基づく主観評価実験を行い、放送品質を満たすビットレート（所要ビットレート）を確認した。主観評価実験では音声フォーマット毎に 4 音源を用いた。4 音源の評価結果の平均を図 1 に示す。また、全ての音源で 95%の信頼区間が差分評価値で-1.0 を上回るビットレートを所要ビットレートとした場合の各音声フォーマットの所要ビットレートを表 1 に示す（‘－’は「該当なし」）。



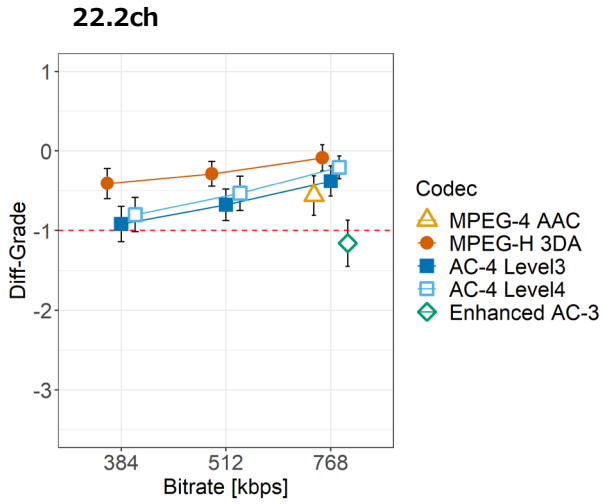


図 1 コアコーダの評価結果（4 音源の平均）

表 1 MPEG-H 3DA と AC-4 の放送品質を満たすビットレート

音声符号化方式	音声フォーマット				
	22.2ch	7.1.4	5.1ch	2ch（ステレオ）	1ch（モノ）
MPEG-H 3DA	512 kbps	192 kbps	—	96 kbps	64 kbps
AC-4	768 kbps	256 kbps	—	—	48 kbps

放送品質を満たすビットレートが確認できた MPEG-H 3DA と AC-4 では、2 チャネル以上のマルチチャネル音響方式の場合 MPEG-H 3DA の方が、モノの場合 AC-4 の方が放送品質を満たすビットレートが低いことが分かった。但し、同じビットレートで評点の差を検定した結果、放送品質での利用が想定されるビットレートにおいては両音声符号化方式間に統計的な有意差はみられなかった。今回の実験では、先行研究に基づき MPEG-H 3DA と AC-4 の所要ビットレートと言われているビットレートを中心に実験を行ったため、MPEG-4 AAC と Enhanced AC-3 では放送品質を満たすビットレートを確認することはできなかった。同じビットレートで比較して、MPEG-4 AAC よりも有意に評点が高かったのは、MPEG-H 3DA（22.2ch の天ぷら）、有意に評点が低かったのは Enhanced AC-3（22.2ch の拍手、5.1ch の花火、2ch のグロッケン、1ch のウィンドチャイム・グロッケン）であった。MPEG-H 3DA か AC-4 を用いることで、従来よりもビットレートを低く設定しても同等もしくはより高い品質が担保される。

次に、多言語放送や裏トークなどの副音声を想定し、2 ヵ国語放送と 4 ヵ国語放送を 22.2ch と 2ch を実施した場合に放送品質を満たすビットレートを表 2 に示す。MPEG-4 AAC の所要ビットレートは、複数ストリームを想定し、先行研究によるチャネルベース音響の所要ビットレートを言語数で乗じた値となっている。MPEG-H 3DA と AC-4 は 22.2ch か 2ch の背景音と言語数分のモノのダイアログから構成されるオブジェクトベース音響の所要ビットレートで各音声素材の所要ビットレートはチャネルベース音響と同じ所要ビットレートを用いている。現行の 4K8K 衛星放送で使

用されている MPEG-4 AAC と比較して、オブジェクトベース音響に対応した音声符号化方式を用いることにより、チャネル数の多い 22.2ch では現行放送の 1.5～3 割、2ch でも 6～8 割程度のビットレートで 2～4 ヵ国語放送のサービスが可能となる。

表 2 想定される放送サービスに必要なビットレート

音声符号化方式	放送サービスの例			
	22.2ch2 ヵ国語	22.2ch4 ヵ国語	ステレオ 2 ヵ国語	ステレオ 4 ヵ国語
MPEG-4 AAC	2,800 kbps	5,600 kbps	288 kbps	576 kbps
MPEG-H 3DA	640 kbps (23%)	768 kbps (14%)	224 kbps (78%)	352 kbps (61%)
AC-4	864 kbps (31%)	960 kbps (17%)	(> 192 kbps (67%))	(> 288 kbps (50%))

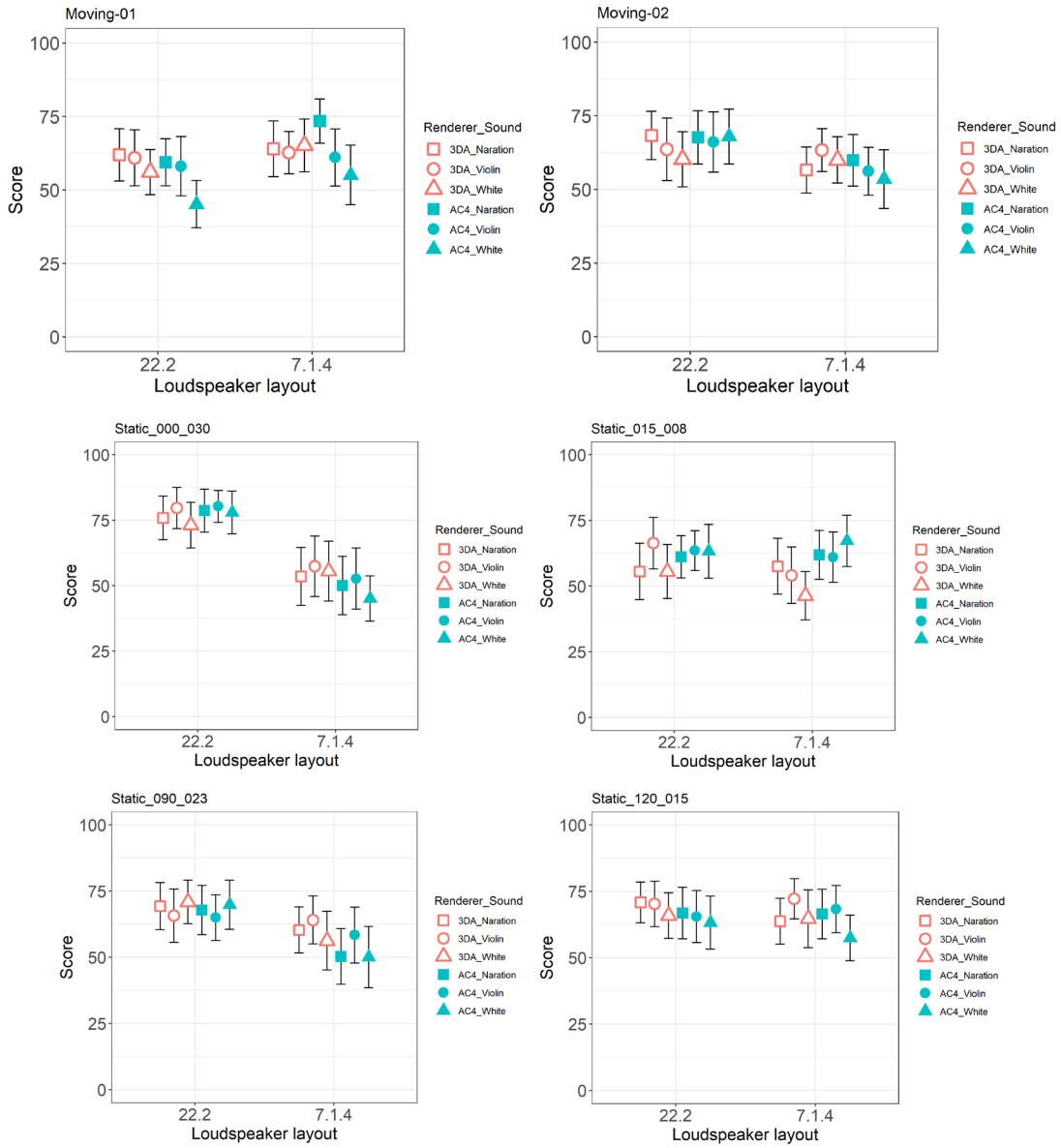
参考文献：

地上デジタル放送方式高度化に関わる適用技術検討作業　最終報告　音声符号化方式の比較検討（情報通信審議会 情報通信技術分科会 放送システム委員会 地上デジタル放送方式高度化作業班（第 13 回） 参考資料 4)

参考 2 レンダラ方式の品質比較のための主観評価実験結果

1. パンニング則

オブジェクトベース音響では、音声信号と再生位置を記述したメタデータをセットで伝送し、再生環境のスピーカ配置に合わせて再生信号を作成（レンダリング）する。MPEG-H 3DA では、位置情報を極座標で記述し、3D-VBAP アルゴリズム（3 スピーカで再生）を用い、Enhanced AC-3 及び AC-4 では、位置情報を直交座標で記述し、トリプルバランスパンナーアルゴリズム（最大 8 スピーカを指定）を用いる。パンニング則による印象差を確認するために、9 種類の位置情報、3 種類の音源、計 27 種類のオブジェクトをサラウンドサークルの中心位置で聴取する主観評価実験を行った。9 種類の位置情報は、画面上を異なる高さで左右に移動する動き、聴取者の周囲を周りながら上昇・下降する動きの 2 動作、7 方向（前方 2 方向、側方 2 方向、後方 1 方向、上方 2 方向）の静止位置である。スピーカ配置は、22.2ch と 7.1.4 の 2 種類とした。実験は多重刺激で、0-100 で回答させた。



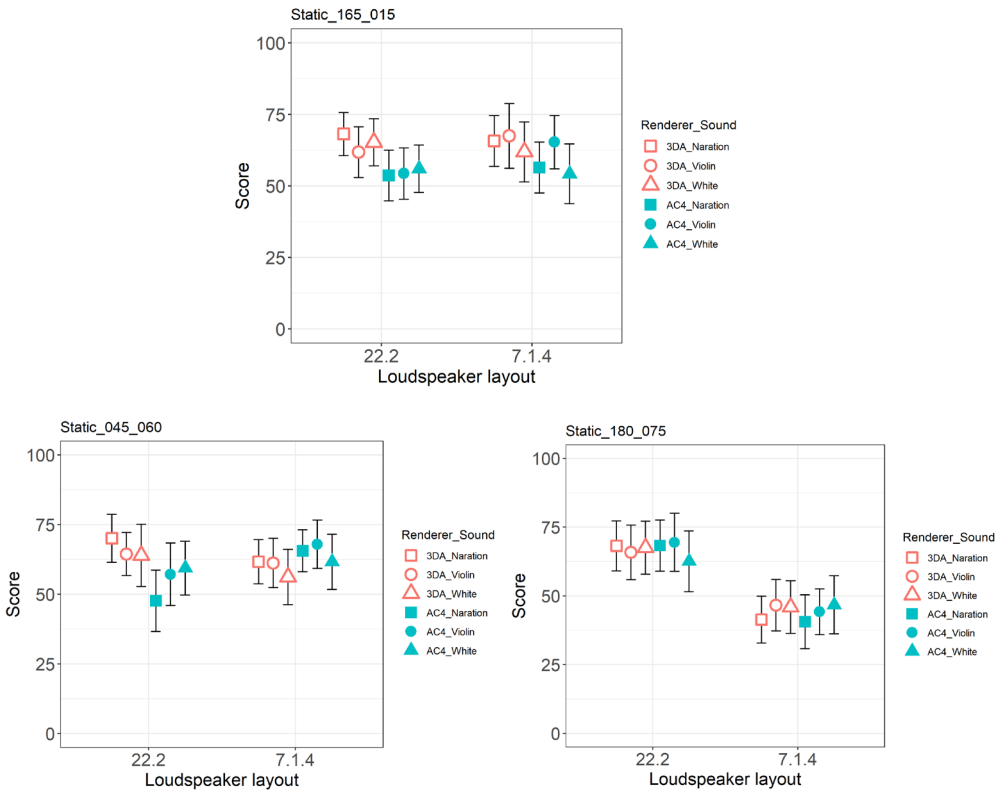


図 1-1 パンニング則の実験結果（上段：動き①，②、下段：静止 7 種類）

音源やスピーカ配置による品質差と比較して、パンニング則による品質差は小さく音声符号化方式に差があるとは言えない。一方、画面上に音像を移動させるという実験を実施するときに、座標系の違いや設計思想の違いにより、同一条件で実施することが困難であった。極座標を採用する MPEG-H 3DA では、聴取位置から見た音像位置をスピーカ配置によらず、角度（例：方位角 15 度、仰角 7.5 度）で指定する。一方、直交座標を採用する Enhanced AC-3 及び AC-4 では、スピーカとの相対位置で音像位置を指定するため、スピーカ配置に依存して想定される音像の位置が変化する。今回は、中層は方位角 30 度と 135 度、上層を方位角 45 度と 135 度を基準とした。このため、上層の方位角 30 度にスピーカを配置する 5.1.4 は実験条件から除外した。座標系を制作者の意図を保持したまま変換することは困難であり、どちらか一方の方式を採用することが望ましい。

2. 再生環境（スピーカ配置）への適応

制作時のスピーカ配置と再生時のスピーカ配置が異なるとき、MPEG-H 3DA では聴取者からみた角度が保持されるようにレンダリングする（Egocentric）のに対し、Enhanced AC-3 及び AC-4 では基準となる四隅のスピーカとの相対位置が保持されるようにレンダリングする（Allocentric）。この設計思想の違いが聴感に与える影響を調べるために主観評価実験を実施した。コンテンツは 22.2ch の背景音に 4 個の静的なオブジェクトが配置されたコンテンツ 3 種類、動的なオブジェクトが 1 個配置されたコンテンツ 1 種類、22.2ch の主チャネル 22 個をオブジェクトとするコンテンツ 2 種類とした。評価はそれぞれのパンニング則で 22.2ch のスピーカ配置にオブジェクトをレンダリングさせた音源（背景音はダウンミックス係数を指定）を参照刺激とし、参照刺激からの

印象差を 0-100 点で評価させた。評価音は、22.2ch（隠れ基準）、7.1.4、5.1.4 の 3 種類である。

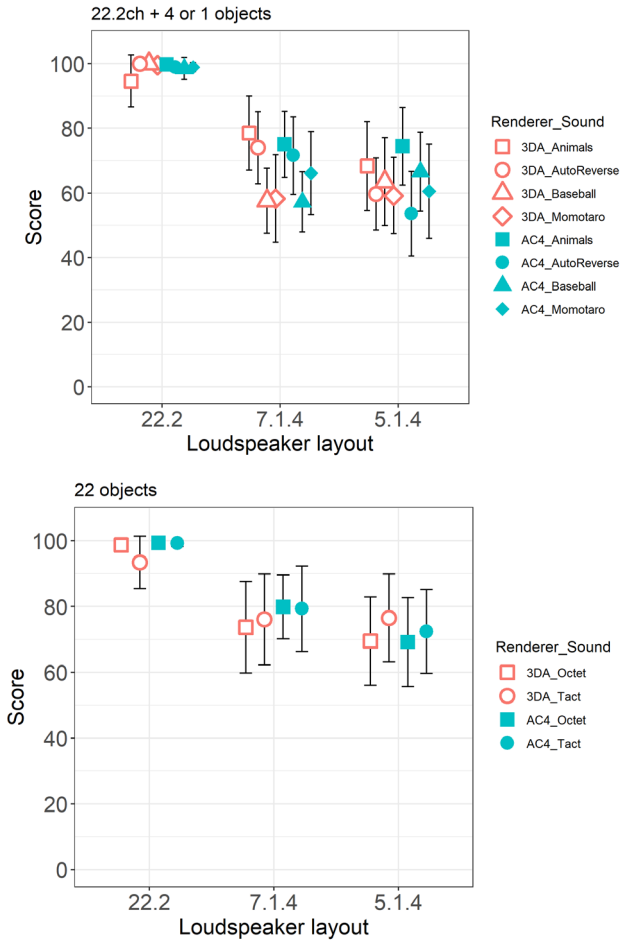


図 2-1 再生環境適応の実験結果

音源やスピーカ配置による品質差と比較して、レンダリング手法による差は小さく、音声符号化方式による差があるとは言えない。音像位置を重視するのか、方向が多少変化しても明瞭度を重視するのかは、番組制作意図にも関連し、評価が分かれる点であろう。

今回の実験において、音響メタデータは勧告 ITU-R BS.2076 に規定される音響定義モデル(ADM)を使用した。ADM には使用できる記述子や設定できるパラメータに自由度があり、想定される動作が異なったり、音声符号化方式によっては動作しなかったりと運用上の課題が散見された。各社が独自の ADM プロファイルを公表しているが、EBU では制作用プロファイルが規格化され、ITU-R では放送用プロファイルが検討中である。実運用に則したプロファイルやメタデータのテンプレートなど、円滑な設備整備・番組交換を行うには、運用規定によるメタデータの仕様の明確化が求められる。

参考文献：

地上デジタル放送方式高度化に関わる適用技術検討作業　最終報告　音声符号化方式の比較検討

（情報通信審議会 情報通信技術分科会 放送システム委員会 地上デジタル放送方式高度化作業班（第 13 回） 参考資料 4）

参考 3　最大入力音声チャネル数の考え方について

入力音声チャネル数は最大 56ch とした。これは、MPEG で回路規模を規定するためのパラメータであるレベル 4 の最大入力音声チャネル数（＝56ch）に対応したものである。放送開始時期が明確になることで各放送事業者が必要とする最大入力音声チャネル数が具体化するものと思われる。そこで、高度広帯域衛星デジタル放送（4K8K 衛星放送）の音声符号化方式で採用されている、22.2 マルチチャネル音響に対応し、且つ、オブジェクトベース音響によるサービスを想定した場合の最大入力音声チャネル数について検討を行った。

オブジェクトベース音響では以下のサービスが新たに想定されている。

・再生環境への最適化
　様々なスピーカレイアウトに受信機で対応
・アクセス性改善
　受信機でダイアログを聴きやすく調整
・視聴者の好みに合わせた再生
　多言語音声・解説音声・背景音などのダイアログの差し替え、ホーム＆アウェイなど

上記、想定サービスを前提にオブジェクトベース音響によるコンテンツ事例と入力音声チャネル数の関係を表 1 にまとめた。オブジェクトベース音響の場合、多言語放送や緊急放送などの音声信号、2ch ステレオのサイマル放送なども音声オブジェクトとして同一エンコーダに入力することになり、いくつかのコンテンツによっては、音声信号の合計数は 56ch を超える場合も想定されることが分かった。実運用を考慮していないことを鑑み、MPEG で規定されているレベル 4 で制約を与えることが妥当であるとした。

表 1　オブジェクトベース音響のコンテンツ事例と入力音声チャネル数

コンテンツ名	サービス事例	ダイアログ	背景音	Total （Stereo×2+5.1×2）
国際放送21言語	国際放送21言語		42ch(2ch)：Stereo×42カ国語	42ch
	OBAによる国際放送21言語	22ch(2ch)：21カ国語（mono）、緊急放送	2ch(2ch)：背景音	28ch(+Stereo x 2) 70ch(+Stereo x 42)
スポーツ番組（解説）	異なる解説で聞きたい	6ch(2ch)：解説（普通、英語、裏トーク、専門家、音声解説）、緊急放送	24ch(24ch)：背景音（22.2）	46ch
ドキュメンタリ（吹き替え）	出演者が日本語を話さないときに吹き替えで聞きたい、本人の声で聞きたい	7ch(2ch)：ダイアログ（日本語、英語、吹き替え、オリジナル、裏トーク、音声解説）、緊急放送	24ch(24ch)：背景音（22.2）	47ch
音楽	歌手グループの誰か一人の歌声を聞きたい	11ch(3ch)：歌声（全員、歌手A～D：Stereo）、緊急放送	24ch(24ch)：伴奏（22.2）	51ch
音楽	アレンジ違いの楽曲（伴奏）を聞きたい。ナレーションはOFFにしたい。	15ch(14ch)：メインとなる楽器・演奏（7.1.4）、ナレーション（日本語、英語）、緊急放送	36ch(12ch)：アレンジの異なる伴奏3種類（7.1.4）	67ch
スポーツ番組（Home & Away）	スポーツ番組で応援しているチームよりの背景音や解説を聞きたい（＋裏トーク＋二カ国語）	7ch(2ch)：Home（日本語、英語、裏トーク、無）、Away（日本語、英語、裏トーク、無）、緊急放送	48ch(24ch)：Home（22.2背景）、Away（22.2背景）	71ch

参考 4　AC-4 の要求条件の適合性に関する補足

欧米各国で採用されている AC-4 Level 3 で 22.2 マルチチャネル音響（22.2ch）を符号化するためには、聴感的に重要度の低いチャネルを適応的にまとめて、最大 17.1 個の座標付きモノ信号として符号化する。復号側では、これを 22.2ch にレンダリングして出力を得る。ARIB での主観音質評価試験では良好な結果を得た。

更なる高音質化などのために、22.2ch をそのまま符号化する必要がある場合、同時最大デコードチャネル数が AC-4 Level 3 では不足する。放送開始時期が明確になることで各放送事業者が必要とする最大入力音声チャネル数や同時最大デコードチャネル数などが具体化するものと思われ、これに適合する形で、現在の AC-4 規格上で"Reserved"となっている Level 4 として定義する必要がある。これは、符号化・復号アルゴリズムはそのままに、対応する最大チャネル数の上限を明確化するものである。