参考資料4

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

地上デジタル放送高度化に関わる音声符号化方式の比較検討

2022 年 2 月 18 日

デジタル放送システム開発部会
一般社団法人　電波産業会

まえがき

総務省からの諮問第 2044 号「放送システムに関する技術的条件」（2019 年 6 月 18 日）を受け、情報通信審議会放送システム委員会に地上デジタル放送方式高度化作業班が設置され、技術的条件の検討が始まった。本活動の一環として、2020 年 6 月 22 日に、ARIB に対して映像符号化及び音声符号化方式の高度化に必要な技術的検討の依頼があった。

本依頼を受け、音声符号化方式作業班は、スタジオ開発部会のスタジオ音声作業班と音声品質評価法作業班とともに、次世代音声符号化方式検討 JTG（Joint Task Group）を立ち上げ、音声符号化方式の検討を開始している。2020 年には、総務省周波数逼迫対策技術試験事務において動作検証を実施している音声符号化方式 MPEG-H 3D Audio に加え、放送システム委員会で実施した「次世代地上デジタルテレビジョン方式に関する技術の提案募集」に提案のあった AC- 4 及び Enhanced AC-3 を対象として比較調査を進めた。3 方式についてはいずれもオブジェクトベース音響（以下、OBA）を利用したサービスに対応しており、音声信号の符号化・復号化処理（コアコーダ）と復号した信号のポスト処理（レンダラー）により様々な音声サービスの提供が可能な方式となっている。2021 年には、各音声符号化方式の所要ビットレートを求めるため、コアコーダの主観評価実験とともに、レンダラー後の品質を比較するための主観評価実験を合わせて実施した。

本報告書は、MPEG-H 3DA、AC-4、Enhanced AC-3 の規格調査結果、及び各音声符号化方式のコアコーダとレンダラーの評価実験結果をまとめたものである。

内容

第1章 一般事項

1.1 用語

本報告書で用いる用語の説明

エレメント	エンコーダ内で圧縮処理された 1 音声信号を含むビットストリーム上の信号
オブジェクト	ダイアログや楽器、背景音などの番組音声を構成する 1 個以上の音声信号からなる音声素材。メタデータとともに使用
チャンネル	チャンネルベース音響で使用されるスピーカ配置に則した音声信号
チャンネル コンフィグレーション	スピーカ配置に対応した音声信号の順番、または音声モード
チャンネルベース音響	スピーカ配置の各スピーカと一対一に対応する音声信号で構成される音響方式またはシステム
オブジェクトベース音響	ダイアログや楽器などのオブジェクトに対応する音声信号と再生位置などのメタデータで構成される音響方式またはシステム
シーンベース音響	聴取位置に到来する音波を記録した音声信号と展開次数などのメタデータで構成される音響方式またはシステム
音声モード	チャンネルコンフィグレーションの符号化表現

1.2 略語

本報告書で用いる略語の説明

3D	Three-Dimensional
ISO	International Organization for Standardization
IEC	International Electrotechnical Commission
MPEG	Moving Picture Experts Group
ITU-R	International Telecommunications Union – Radiocommunication Sector
HOA	Higher Order Ambisonics
ETSI	European Telecommunications Standards Institute
EBU	European Broadcasting Union
CENELEC	Comité Européen de Normalisation ELECtrotechnique
ATSC	Advanced Television Systems Committee
DVB	Digital Video Broadcasting
SBTVD	Sistema Brasileiro de Televisão Digital (Forum)
ABNT	Associação Brasileira de Normas Técnicas
HbbTV	Hybrid broadcast broadband TV
TTC	Telecommunication Technology Association
CCTV	China Central TV

DRC	Dynamic Range Compression/Control
AAC	Advanced Audio Coding
MDCT	Modified Discreate Cosine Transform
MAT	Metadata-enhanced Audio Transmission
TNS	Temporal Noise shaping
PNS	Perceptual Noise Substitution
ADTS	Audio Data Transport Stream
LATM	Low Overhead MPEG-4 Audio Transport Multiplex
LOAS	Low Overhead Audio Stream
LFE	Low Frequency Enhancement
USAC	Unified Speech and Audio Coding
OBA	Objects Base Audio
IGF	Intelligent Gap Filling
SBR	Spectral Band Replication
QMF	Quadrature Mirror Filter
MCT	Multichannel Coding Tool
FDP	Frequency Domain Predictor
LTPF	Long Term Post Filter
VBAP	Vector Base Amplitude Panning
MHAS	MPEG-H Audio Stream format
ASI	Audio Scene Information
ISOBMFF	ISO Base Media File Format
MMT	MPEG Media Transport
AHT	Adaptive Hybrid Transform
A-CPL	Advanced Coupling
A-JOC	Advanced Joint Object Coding
A-SPX	Advanced Spectral Extension
ASF	Audio Spectral Frontend
SSF	Speech Spectral Frontend

第2章　　目的と背景

2.1 目的

次世代地上デジタル放送の音声符号化方式として以前より研究開発・技術試験事務等で検討中、あるいは 2020 年 3 月の提案募集に対して応募があった以下の 4 方式に対して機能および性能の比較（オブジェクト符号化の比較を含む）を行う。

MPEG-4 AAC

MPEG-H 3D Audio

Enhanced AC-3

AC-4

2.2　音声符号化方式の標準化経緯

2.2.1　MPEG-H 3D Audio

MPEG-H 3D Audio は、チャンネルベース音響に加え、オブジェクトベース音響、シーンベース音響に対応した音声符号化方式である。標準化は、ISO / IEC（International Organization for Standardization：国際標準化機構 / International Electrotechnical Commission：国際電気標準会議）の JTC1（Joint Technical Committee 1：第 1 合同技術委員会）傘下の SC29/WG11 の呼称である MPEG （Moving Picture Experts Group）において進められた。（2020 年から MPEG Audio Group は単独で SC29/WG6 として活動）

2009 年から、ITU-R（International Telecommunications Union – Radiocommunication Sector）においてマルチチャンネル音響の拡張として先進的音響システム（チャネルベース音響, オブジェクトベース音響, シーンベース音響）の研究が開始され、MPEG でも並行して先進的音響システムに対応した音声符号化方式の標準化に向けた研究が始まった。2013 年に MPEG-H 3DA の requirement が発行され本格的に標準化がスタートし、2015 年に ISO/IEC 23008-3 (MPEG-H Part 3)として MPEG-H 3DA の第 1 版が、2019 年に第 2 版が発行されている。

第 2 版では、処理負荷量に応じて profile（High、Low Complexity）が規定されている。Low Complexity profile（以下 LC profile）は、放送や低ビットレート向けにデコーダ負荷を軽減した符号化ツールセットとして規定されているが、シーンベース音響で用いられる処理負荷の高い HOA（Higher Order Ambisonics）用符号化ツールが包含されているため、より実用的で処理負荷を軽減した profile が望まれていた。そこで、2020 年 11 月に LC profile から HOA と低ビットレート向けの人声に特化した符号化ツールを除外した baseline profile（以下 BL profile）が第 2 版の AMENDMENT 2 として発行された。BL profile の標準化に関連して、ビットストリームの内容を確認するためのリファレンスソフトウェア規格、及びデコーダが規格通り実装されているかを確認するためのコンフォーマンステスト規格の改定作業が

引き続き進められている。また、これまで発行された AMENDMENT や正誤表をまとめ 2022 年に MPEG-H 3DA 第 3 版が発行される予定である。

2.2.2 Enhanced AC-3（E-AC-3）

Enhanced AC-3 は、Dolby Laboratories が開発した音声符号化方式で、チャンネルベース音響とオブジェクトベース音響に対応している。5.1ch サラウンドに対応する AC-3 を高度化・高効率化する音声符号化方式として、ETSI（European Telecommunications Standards Institute：欧州電気通信標準化機構）、EBU（European Broadcasting Union: 欧州放送連合）、CENELEC（Comité Européen de Normalisation ELECtrotechnique：欧州電気標準化委員会）の JTC（Joint Technical Committee）の下で標準化作業が行われ、2005 年に技術仕様が ETSI TS 102 366: “Digital Audio Compression (AC-3, Enhanced AC-3) Standard”として公開されている。また、ATSC（Advanced Television Systems Committee：米国高度化テレビジョンシステム委員会）でも同内容の規格が 2005 年に策定され、ATSC A/52 “Digital Audio Compression (AC-3, E-AC-3)”として技術仕様が公開されている。

また、Enhanced AC-3 には、上記規格との後方互換性を維持しつつ Dolby Atmos として広く知られる高さ方向も含めた 3 次元音響などの追加機能が ETSI TS 103 420: “Backwards-compatible object audio carriage using Enhanced AC-3”として 2016 年に規格化された。

2.2.3 AC-4

AC-4 は、Dolby Laboratories が開発した音声符号化方式で、チャンネルベース音響とオブジェクトベース音響、システムシーンベース音響に対応している。Enhanced AC-3 と同様に、ETSI、EBU、CENELEC の JTC の下で標準化作業が行われた。まず、2014 年 4 月にチャンネルベース音響に関して ETSI から TS 103 190 Part 1 が発行され、その後、オブジェクトベース音響に関する規定を追加する形で TS 103 190 Part 2 が 2015 年 9 月に発行された。現在のバージョンは TS 103 190 Part 1 が 1.3.1 (2018 年 2 月発行)、TS 103 190 Part 2 が 1.2.1（2018 年 2 月発行）である。

2.3 放送規格の採用状況

2.3.1 DVB

DVB（Digital Video Broadcasting:デジタルビデオ放送）は欧州を中心に採用されている国際的なデジタルテレビ放送のための公開標準規格である。DVB では放送や IP ベースでの配信などを目的として映像・音声の符号化方式を利用する際のガイドライン（DVB Document A001）が規定されている。音声符号化方式については MPEG-4 AAC、MPEG-H 3DA、Enhanced AC-3、AC-4 などが採用されており、各方式に対して制約事項が規定されている。MPEG-H 3DA の制約としては、LC profile（レベル 3）となっている。

Ref. DVB BlueBook A001r17 (Draft TS 101 154 V2.7.1)

2.3.2 ATSC

ATSC（Advanced Television Systems Committee：米国高度化テレビジョンシステム委員会）は米国の

デジタルテレビ規格を検討している組織であり、北中米と韓国がこの方式を採用している。ATSC は地上デジタル放送規格 ATSC1.0 と互換性を考慮しない次世代地上デジタル放送規格 ATSC 3.0 の標準化を 2017 年 6 月に完了し、2017 年 2 月に ATSC A/342 part 2 “AC-4 System”、2017 年 10 月に ATSC A/300 “ATSC 3.0 System” 標準規格が承認された。ATSC 3.0 では音声符号化方式に AC-4 または MPEG-H 3DA を採用し、サービス提供地域により符号化方式を選択可能としている。北米（米国、カナダ、メキシコ）では AC-4 を採用し、米国において商用放送が行われ、対応受信機が販売されている。また、2017 年 ATSC 3.0 による放送を開始した韓国の 4K テレビ放送では、音声符号化方式として MPEG-H 3DA を採用し、対応した受信機が販売されている。

Ref.ITU 協会 ITU ジャーナル（Vol.47 No.11）
Ref. *総務省.* “世界情報通信事情 米国”. www.soumu.go.jp.

2.3.3 SBTVD Forum

SBTVD Forum（Sistema Brasileiro de Televisão Digital Forum）はブラジルのデジタル放送規格を開発する機関であり、ブラジル以外に、ペルーやアルゼンチンをはじめとする南米諸国で採用されている。

2006 年に TV2.0 によるデジタル放送を開始し、音声符号化方式として MPEG-4 AAC を採用している。2020 年 5 月に HDR、Immersive Audio や DTV Play（VOD サービス）の拡張を目的に TV 2.5 を開始し、地上波放送規格 ABNT NBR 15602-2:2020 に MPEG-H 3DA、Enhanced AC-3、AC-4 を採用した。TV 2.5 は TV 2.0 との後方互換性を担保するため、MPEG-4 AAC とともに追加された音声符号化方式をサイマルで放送することとしている。サッカー中継などで Enhanced AC-3 を使用した 5.1.2ch（7.1ch）によるサービスを実施している。

TV 3.0 は、TV 2. 0 に代わる新しいオープン TV システムであり 2022 年 9 月の規格発行を目指し提案技術の選定作業を実施している。2020 年 7 月に国際公募（Call for Proposals）が行われ、音声符号化方式には、MPEG-H 3DA、AC-4、AVSA（Audio and Video coding Standard Audio codec：中華人民共和国のオーディオ符号化規格）が提案され技術評価がすすめられていた。その結果、2022 年に放送用の唯一の必須音声符号化技術として MPEG-H 3DA の採用が決まった。

Ref. Fórum SBTVD | TV 3.0 Project (forumsbtvd.org.br)

2.4 各国の採用状況

2.4.1 MPEG-H 3DA

ATSC:
-ATSC3.0 の音声フォーマットとして採用（A/342-3）。韓国 TTC(Telecommunication Technology Association)が国内の地上波 4 K 放送音声方式として、ATSC3.0 の MPEG-H 3DAudio を採用し、2017 年から本放送が開始、現在韓国内で対応 TV 受信機、STB が販売されている。
欧州ＤＶＢ：
-DVB 規格 ETSI TS101 154 v2.3.1、HbbTV 2.0.2 Specification (ETSI TS 102 796)に採用。

-EBU(European Broadcasting Union)が、BBC、France Television、RAI（イタリア）など各国放送局と共同で European Athletics Championships（2018）での初の MPEG-H 3 DA を用いたライブ試験実験を実施、また France television が、2018 年、2019 年の French Tennis Open で MPEG-H 3DA によるライブ試験放送（5. 1 ＋4H、1object）を実施している。

Brazil:

-現行の ISDB-Tb システムの拡張規格である TV2.5 に MPEG-H 3 DA が採用（ABNT NBR 15602-2:2020）された。また現在 SBTVD フォーラムが次世代放送向けに標準化を進めている TV3.0 の放送用唯一の必須オーディオ符号化方式として、2022 年 1 月 MPEG-H 3DAudio の採用が承認された。今後、標準規格文書の策定が行われる予定である。

-Globo が 2019 年開催の世界最大の音楽フェス Rock in Rio において、ISDB-Tb システム上での MPEG-H 3DAudio によるライブ試験放送を実施している。

China:

-CCTV（China Central TV）からの技術提案募集を受けて、AVS 傘下の 3 D Audio Task Group が 2016 年に UHD 放送サービス、ストリーミングサービス用の次世代オーディオフォーマットとして、MPEG-H 3 D Audio を選定し、現在 specification 策定の最終段階に入っている。 2021 年春に CCTV が AVS 8 k Video と MPEG-H 3 DA の組み合わせで IP ベースの放送システムによる実験を行った。また 2022 年の北京オリンピックでも、MPEG-H 3 DA を用いたストリーミング試験放送が計画されている。

3GPP：

-2018 年に 5G の 3D ビデオストリーミングサービス用の唯一のオーディオ方式として採用された。（3GPP Virtual reality profiles for streaming applications： 3GPP TS 26.118 version 15.0.0 Release 15）

360 Reality Audio(360RA):

-MPEG-H 3DAudio ベースに、2019 年に登場した新しい 3 D オーディオフォーマットで、すべてオブジェクト信号で構成される（最大 24 オブジェクト、Baseline profile のレベル 3 に対応）。現在多くのストリーミングサービスで採用され、スマートスピーカ、ワイヤレススピーカー等対応機器も発売されている。その他、放送向けの MPEG-H 3DAudio による各種ライブプロダクションツール（H/W）、ポストプロダクションツール（H/W、S/W）、リアルタイムエンコーダ等が各社よりすでに提供されている。また、数多くの SoC ベンダーから、LC profile、BL profile のデコーダが提供されており、また将来の 22.2ch の市場として有望な日本市場に向けて、各社より level 4 に対応したデコーダも既に開発されている。

MPEG-H 3DAudio のライセンス

-本年(2021 年 6 月)より、MPEG-H 3D Audio の特許プールライセンスが開始されている。(Via Licensing)
Ref. MPEG-H 3D Audio – ViaCorp (via-corp.com)

2.4.2 Enhanced AC-3

Enhanced AC-3 は国外の HD 放送の国際標準として用いられている音声符号化方式である。導入当初は 5.1ch や 7.1ch に代表されるチャンネルベースの音声符号化規格 ETSI TS 102 366: “Digital Audio Compression (AC-3, Enhanced AC-3) Standard”、ATSC A/52 “Digital Audio Compression (AC-3) (E-

AC-3)”として技術が公開され、イギリス、フランス、ドイツ、インドなど 38 か国の地上波放送規格に採用されている。

また、Enhanced AC-3 には、後方互換性を維持しつつ Dolby Atmos として広く知られる立体音響やオブジェクトベース音響に対応するなどの追加機能が ETSI TS 103 420: “Backwards-compatible object audio carriage using Enhanced AC-3”として技術公開され、日本の Hybridcast 規格やブラジルの地上波放送規格 ABNT NBR 15602-2:2020 にも採用されている。高さ方向も含めた高臨場感音声サービスとして、ブラジルではサッカー中継などで本方式による放送を実施している。

対応機器の普及

Dolby Atmos に対応するテレビ受信機、PC/モバイル機器、ホームシアター機器は ETSI TS 103 420 での機能拡張も含めた Enhanced AC-3 のデコードに対応し、国内外で既に広く普及している。国内では、ソニー、パナソニック、FUNAI、REGZA、LG、TCL などから内蔵スピーカのみで立体音響を楽しめる Dolby Atmos 対応テレビ受信機が数多く発売され、非対応のテレビ受信機や STB においてもビットストリームのパススルー出力によって外付けのサウンドバーなどで立体音響を楽しむことができる。テレビ放送受信以外の用途としては、Apple の iOS 機器(iPhone、iPad など）、Android モバイル機器、PC(Windows、Mac)でも多くの機種が対応し、動画配信サービスを立体音響で楽しむために利用されている。

サービスの普及

海外では、放送・ケーブルサービスも含めて既に幅広く利用されており、東京オリンピックも米国や中国などでは Enhanced AC-3 を用いて立体音響で放送・配信された。国内においても Netflix、Disney+、Apple TV+、ひかり TV、U-NEXT、J:COM オンデマンドなどの動画配信サービスや、Apple Music、Amazon Music などの音楽配信サービスなどで採用されている。また、IPTV フォーラムにより Hybridcast 規格として採用され、地上波放送局により実証実験が行われている。

2.4.3 AC-4

放送規格としては、2015 年 3 月に、DVB-T2 や UHD 放送での利用に向けて ETSI TS 101 154 に採用されている。ATSC においては、2017 年 2 月に ATSC A/342 part 2 “AC-4 System”、2017 年 10 月に ATSC A/300 “ATSC 3.0 System” 標準規格が承認され、地域毎に選択する音声符号化方式として、北米（米国、カナダ、メキシコ）は AC-4 を採用した。欧州各国の放送規格においても、NorDig v3.1.1（デンマーク、ノルウェー、フィンランド、アイスランド、スウェーデン、アイルランド)、イタリアの UHD-Book v1.0、ポーランドの DVB-T2 で AC-4 が採用されている。また、ブラジルでも 2020 年 5 月に地上波放送規格 ABNT NBR 15602-2:2020 に AC-4 が採用された。現在、米国において AC-4 での商用放送が行われ、欧州のスペイン、フランス、イタリア、ポーランドでは試験放送が行われている。欧州での最近の主な事例としては、ポーランドで 2021 年の Euro2020(サッカー, 5.1.4ch+3 オブジェクト)、ショパン国際ピアノコンクール、フランスで 2020 年、2021 年などの全仏オープン(テニス)、スペインで 2020 年

に 8K 映像との組み合わせでの試験放送などがある。

次世代放送規格での採用に加えて、テレビ受信機やモバイル機器などでの実用化が既に浸透していることも AC-4 の特長であり、世界 165 カ国で 1 億 1500 万台以上の AC-4 搭載機器が既に出荷されている。日本国内も含めた実用化状況を以下に示す。

実用化状況（4K テレビ受信機）

国外はもとより、国内向けにソニー、パナソニック、FUNAI、REGZA、LG、TCL などから販売されている多数の 4K テレビ受信機が AC-4 デコーダを搭載している。これらの受信機はスピーカバーチャライザ機能を搭載し、テレビ受信機の内蔵スピーカだけでも立体音響を楽しむことができる。国内で出荷される 4K テレビ受信機における AC-4 デコーダ搭載率は高く、例えばソニーからは国内向け全モデルに AC-4 デコーダが搭載されている(2021 年 12 月時点)。

実用化状況（モバイル機器）

国外はもとより、国内でもソニー、シャープ、FCNT、サムスンなどから販売されている多数のスマートフォンが AC-4 デコーダを Android OS の一部として搭載している。米国などでは、3.4.5 章で述べた Immersive Stereo 機能を用いた立体音響による音楽配信サービスが実用化されている。

実用化状況（ホームシアター機器）

AC-4 デコーダを AV アンプやサウンドバーなどのホームシアター機器に実装することなしに、既に広く普及している Dolby Atmos 対応のホームシアター機器で再生することができる。受信機では、AC-4 デコーダ出力を Dolby MAT (IEC 61937-9, Metadata-enhanced Audio Transmission)と呼ばれる形式で HDMI 出力することで、Dolby Atmos に対応する AV アンプやサウンドバーなどが視聴環境に応じたレンダリングと再生をすることができる。22.2ch といった多チャンネル番組についても、Dolby MAT 形式を用いることでダウンミックスせずにホームシアター機器へ伝送することができ、臨場感の高い番組を手軽に楽しむことができる。この機能を採用した BS4K 受信機製品が国内で発表され、22.2ch 番組を家庭で手軽に楽しめるようになると期待されている。

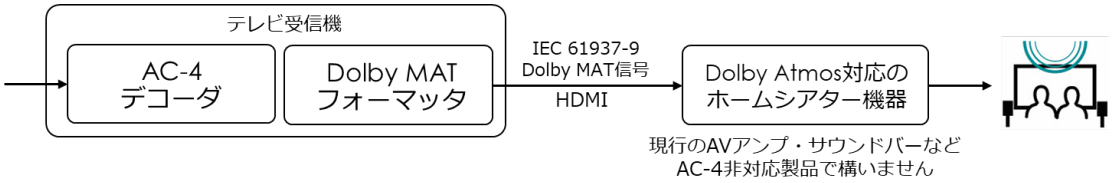


図 1 Dolby MAT を用いたホームシアター機器での再生

実用化状況（SoC）

4K テレビ受信機の多くで AC-4 が搭載されていることからも分かるように、既にテレビ受信機やセットトップボックス向けの主要な SoC ベンダー（MediaTek、Amlogic、Broadcom、Cadence、HiSilicon、LG、

Novatek、Realtek、Samsung、Synaptics など）から AC-4 を搭載した SoC が販売されている。

ライセンスと費用

Dolby Laboratories がライセンスをし、製品開発のための開発キットとサポートを提供している。例えば、テレビ受信機用の開発キットには AAC、HE-AAC、AC-3、Enhanced AC-3 といったデコーダに加え、レンダラーやバーチャライザなどの機能が含まれている。この開発キットを用いることで、世界各国の様々な放送規格に対応したテレビ受信機の開発が容易となるため、国内外のテレビ受信機メーカに幅広く利用されている。最新の開発キットとそのライセンスでは、AC-4 デコーダも追加費用なしに利用可能になり、これが国内外の多くのテレビ受信機とその SoC で AC-4 が既に実装されている背景となっている。製品の開発サポートは、日本も含めた Dolby Laboratories の世界各地のオフィスから提供され、機器間の接続互換性も含めた製品評価・認証も行われている。

2.5 音響方式

2.5.1 チャンネルベース音響

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）のチャンネル構成（例えば、2ch ステレオや 5.1ch）にまとめて放送される。受信側ではチャンネル構成（2ch の場合は左チャンネルと右チャンネル）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ンネル構成と受信機側の再生チャンネル構成が同一であることを前提に受信側のスピーカを直接ドライブする信号を制作/伝送する方式をチャンネルベース音響と呼ぶ。

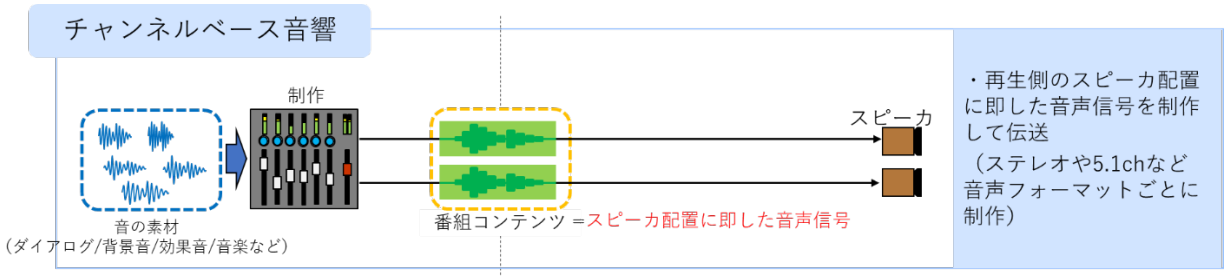


図 2 チャンネルベース音響

2.5.2 オブジェクトベース音響

オブジェクトベース音響ではナレーションや背景音、音楽や効果音を各々個別に位置やレベルなどを記述した音響メタデータとともに伝送し、受信側で音響メタデータを基に番組音声を再構成してスピーカをドライブする信号を出力する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システムでは、ナレーションや背景音、音楽や効果音を個別に受信できるため、家庭のスピーカ配置へ最適化したり、ナレーションなどダイアログの音量を個別に調整することで聞き取りやすくしたり、多言語に対応することで日本語が非ネイティブの方に情報を伝えるなど、試聴環境や

視聴者の好みに合わせた番組音声のサービスをきめ細かく提供できる。

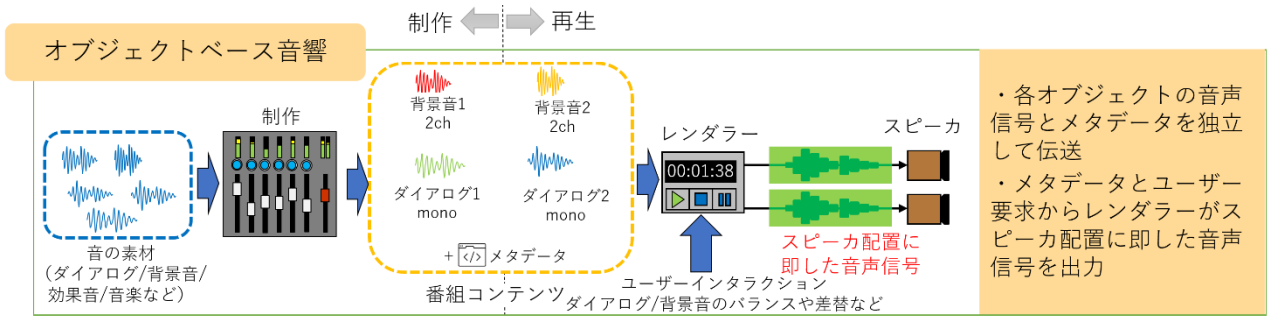


図 3 オブジェクトベース音響

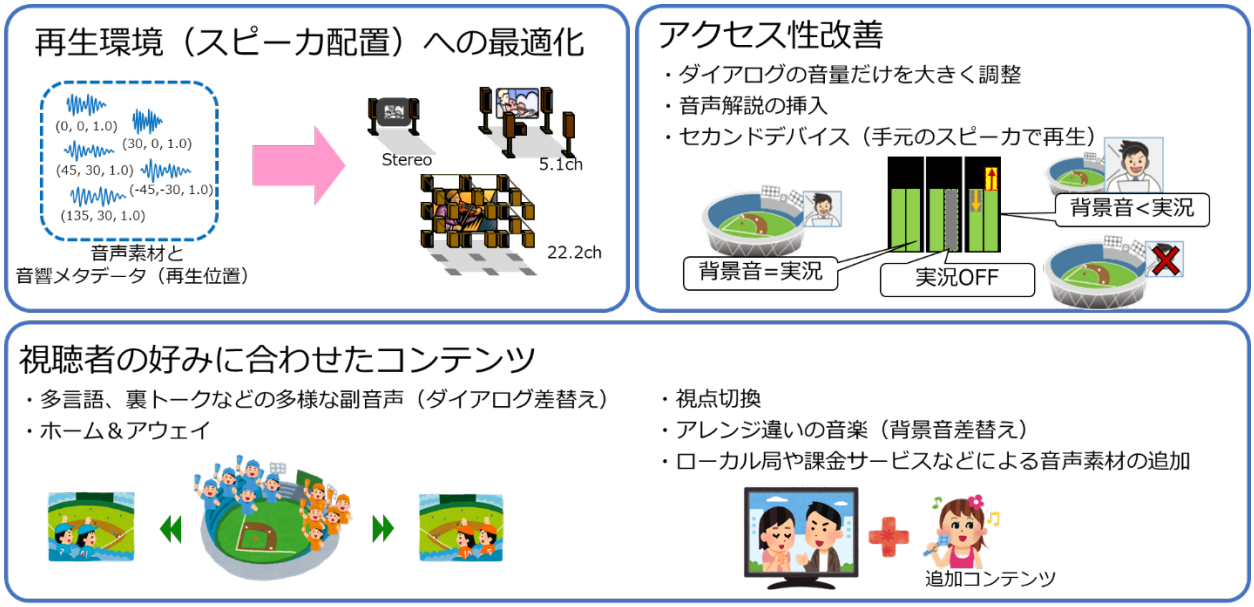


図 4 オブジェクトベース音響で想定される音声サービス

Ref. ISO/IEC JTC1/SC29/WG11/N13411 “Call for proposal for 3DA”

2.5.3 シーンベース音響

チャンネルベースやオブジェクトベースでは素材毎の音声を個別に収録するが、シーンベース音響では到来する音波に対し視聴位置を中心とした球面上の信号として記録するシステムである。記録された信号は、3 次元空間上で球面調和関数展開することが可能で、展開次数のそれぞれの係数を基に 5.1ch や 22.2ch など様々な音声フォーマットに変換して再生する。このように音の場を記録再生する方式をシーンベース音響（Ambisonics）と呼ぶ。展開次数が低いと変換後の再生品質が担保できないため、高次の展開次数で記録再生する HOA（Higher Order Ambisonics）が一般的には用いられる。

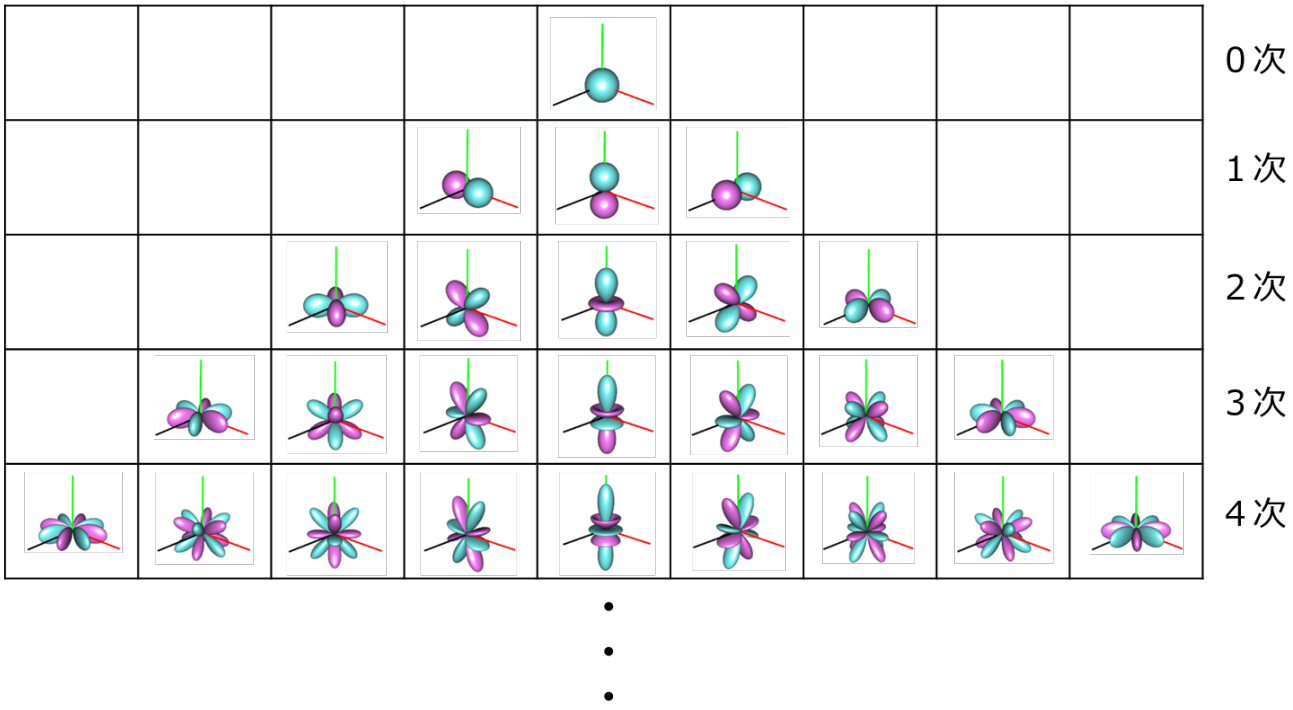
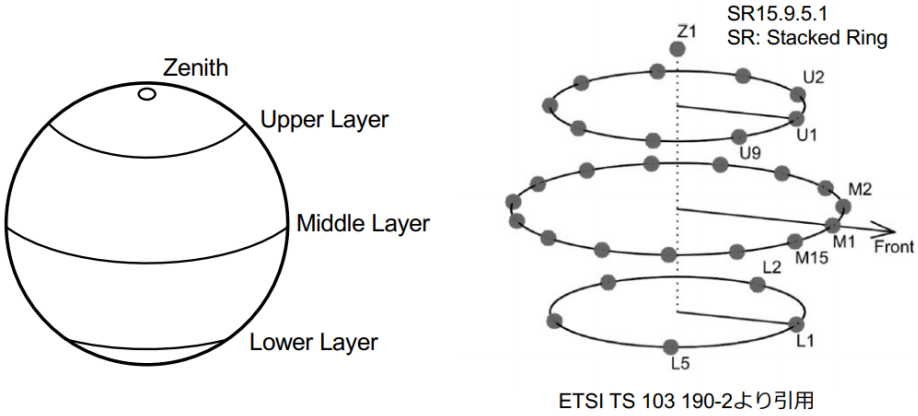


図 5 球面調和関数（4 次まで）

MPEG-H 3DA ではレベルにより異なるが最大 6 次+4 オブジェクト（49+4 エレメント）までの伝送が可能であり、AC-4 では Intermediate Spatial Format と呼ばれる球体の表面上に音源を配置することで音の場を表現しており、最大 30ch のチャンネルベース音響信号として伝送している。



ETSI TS 103 190-2より引用

図 6 AC-4 によるシーンベース音響の信号割り当て

2.6 オブジェクトベース音響に対応した音声符号化方式

オブジェクトベース音響に対応した音声符号化方式の基本的な構成について説明する。符号化処理は概して 3 つのブロックに分かれている。

- コア符号化

  チャンネルベースで制作された背景音信号やダイアログなどのオブジェクト信号、メタデータが入力され、それぞれの信号を圧縮、多重化しビットストリームを生成出力する。コア符号化では圧縮伝送する総信号数をエレメント数と呼んでいる。例えば背景音として 22.2ch、ダイアログ音声として 4 オブジェクトを入力した場合、28 エレメントとなる。最大伝送可能なエレメント数に制約がある場合（MPEG-H 3 DA）や、指定されたビットレート内で制限なくエレメント数を伝送可能な方式（Enhanced AC-3/AC-4）もある。また、符号化の過程でエレメント数を削減する場合もある。

- コア復号

  ビットストリームから背景音信号やダイアログなどのオブジェクト信号、メタデータを復号（デコード）する。一般にビットストリームに記録された信号のうち、同時にデコードできるエレメント数に制限がある。しかし、ビットストリームから復号するオブジェクト信号をユーザーインタラクションにより選択することで同時デコード数より多くのエレメントを利用できる場合がある。

- レンダラー

  コア復号で復号された背景音、ダイアログ音声、メタデータに加え、ユーザーインタラクション情報に基づき再生環境のスピーカ信号を生成する。レンダラーには背景音とダイアログをミックスするミキサー、各信号のダイナミクスを制御する DRC（Dynamic Range Compression/Control）、オブジェクト音声を空間上にマッピングするパンニング、スピーカ配置の信号数に変換するためのチャンネル数変換など、放送用音声卓の基本機能が凝縮したものといえる。

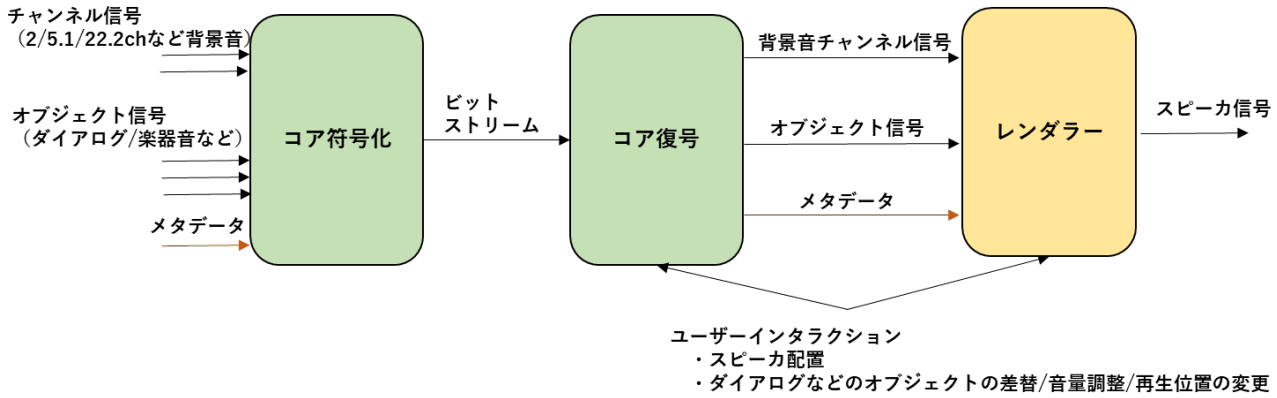


図 7 OBA に対応した音声符号化方式の基本的な構成

第3章　各音声符号化方式の概要

本章では各音声符号化方式の概要について述べる。本報告の目的にある通り、オブジェクトベース音響を含む各方式の比較が目的であるので、シーンベース音響の技術については記述しないこととする。また、複数の profile 規定がある MPEG の音声符号化方式については以下に示す profile の説明にとどめる。

- MPEG-2/4 AAC　　LC profile
- MPEG-H 3DA　　BL profile

3.1 MPEG-2/4 AAC

3.1.1 符号化アルゴリズム

MPEG-4 AAC LC profile の符号化ブロックダイアグラムを図 8 に示す。

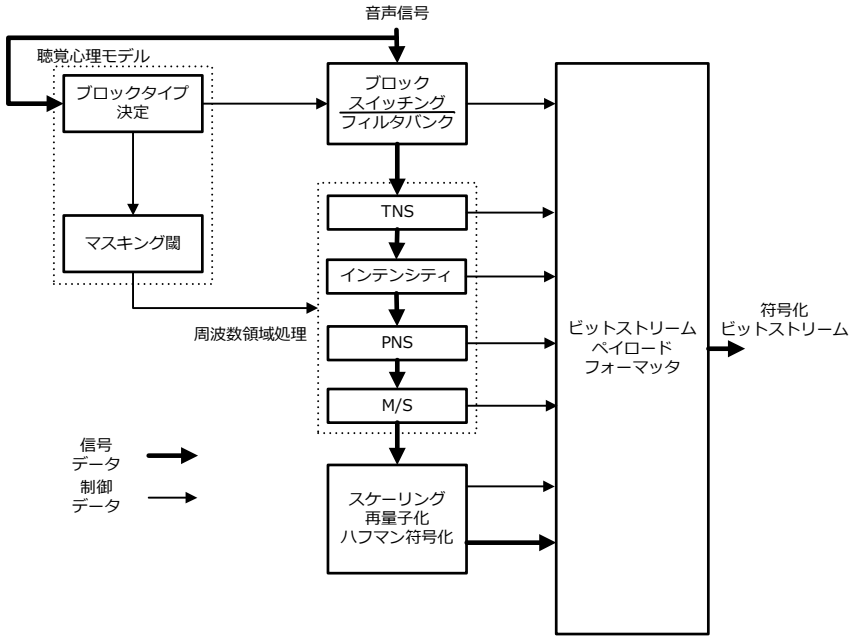


図 8 AAC 符号化ブロックダイアグラム

音声信号は、ブロックスイッチングおよびフィルタバンクにより時間領域信号から周波数領域信号に変換される。変換には MDCT（修正離散コサイン変換）が用いられ、その変換ブロック長は聴覚心理モデルにより決定される。音声信号が急峻に変化する場合には、短い変換ブロック長を適用することでプリエコーを抑圧する。周波数領域処理では効率的に圧縮するため、MDCT 係数を各符号化ツールにより変換処理する。

1. TNS（Temporal Noise Shaping）

   TNS は周波数軸の MDCT 係数を時間軸の信号とみなし、線形予測係数を用いたトランスバーサルフィルタおよび巡回型フィルターにより、波形に含まれる量子化雑音を信号レベルの大きなとこ

ろに集中させることで信号の音質を向上させる。

2. インテンシティ
   高い周波数成分において左右の定位感は聴感上、時間差や位相差よりも音の大きさの影響が大きい性質を利用し、二つのチャンネルの高い周波数量の子化係数を一つの情報にまとめ左右のレベル差を制御データとして伝送することによりビットレートを削減する。
3. PNS（Perceptual Noise Substitution）
   MPEG-4 AAC で用いられる符号化ツールで、ノイズ性の信号に対してスケールファクタバンド（近い周波数の MDCT 係数を纏めたグループ）内の信号をバンド全体に対するノイズとして扱い、そのパワー情報を送る。復号側では、この情報を用いて適正なレベルのノイズを挿入し、音声信号を再構成することによりビットレートを削減する。
4. M/S
   二つのチャンネルをそれぞれ符号化するか、和信号と差信号を代わりに符号化するかをスケールファクタバンドごとに選択する方式。二つの信号間の同位相成分が多い場合に符号化効率を高めることができる。

インテンシティと M/S は二つの音声信号間の冗長成分を削減するステレオ符号化ツールである。これらのツールはモノの信号には適用されないため、AAC ではモノよりも 2ch ステレオで効率的に信号圧縮が可能である。 3 ch 以上のマルチチャンネル音声では、できるだけ 2ch の組み合わせで符号化する方式がとられており、例えば、5.1ch の場合、センターチャンネルはモノ信号として符号化し、レフトチャンネルとライトチャンネル、レフトサラウンドチャンネルとライトサラウンドチャンネルは 2 ch ペアとして符号化する。LFE（Low Frequency Enhancement Channel）は信号帯域を制限したモノ信号として符号化される。

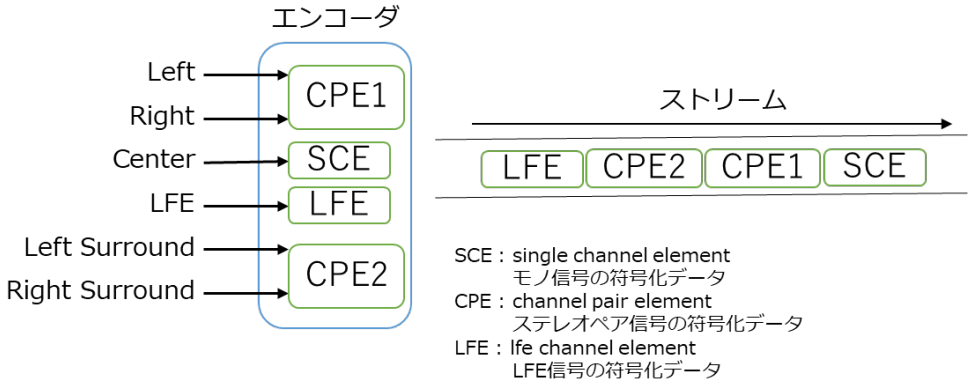


図 9 5.1ch を符号化した場合のストリーム構成

各符号化ツールにより効率化された信号データはスケール変換により再量子化されハフマン符号化による可逆圧縮処理の後に各符号化ツールの制御データとともに多重化してビットストリームとして伝送される。AAC では複数のビットストリーム形式が規定されており、国内のデジタル放送として MPEG-2

AAC では ADTS (Audio Data Transport Stream)形式、MPEG-4 AAC では LATM/LOAS（Low-overhead MPEG-4 Audio Transport Multiplex/Low Overhead Audio Stream）形式が採用されている。

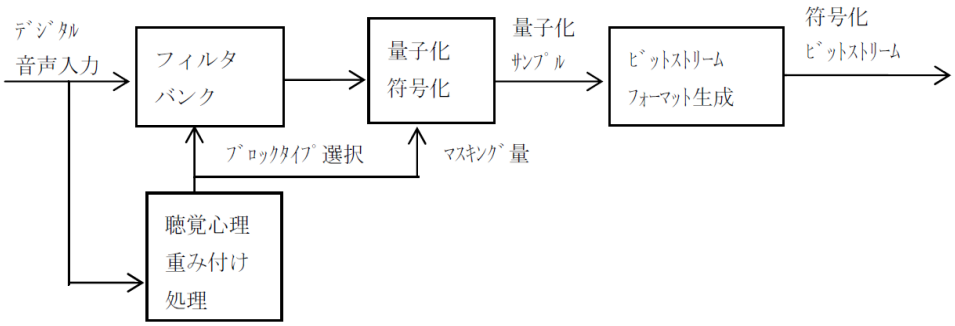
Ref.　APAB　AAC 調整連絡会「MPEG-2 AAC 方式の運用に関する技術解説」2009 年
Ref.　NHK 技研 R＆D/No.155/2016.1 「音声符号化技術の標準化動向」

3.1.2　省令および告示

省令および告示における MPEG-4 AAC の準拠する方式、圧縮手順および送出手順は以下のように記載されている。

> 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し、音声の圧縮手順及び送出手順については、総務大臣が別に告示するところ（4.3 章参照）によるものとする。
>
> （省令第 64 条、第 81 条の 3）

音声の圧縮手順及び送出手順については以下のとおりとする。



注 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。

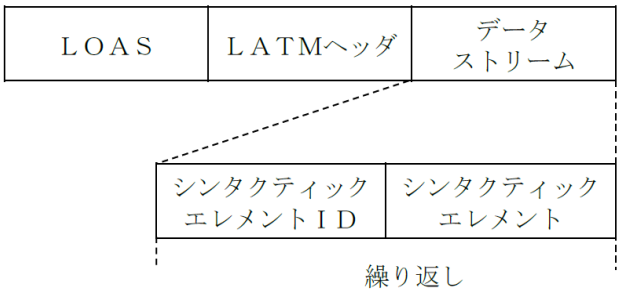
2　聴覚心理重み付け処理は、フィルタバンクへの入力信号に対応して、マスキング量（一の音声信号と他の音声信号を識別できる限界）及びフィルタバンクの入力ブロックタイプを算出する。

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　符号化ビットストリームのチャンネルモードの最大値は、22チャンネル及び低域を強調する2チャンネルとする。

5　ビットストリームの構成は、次の通りLATM/LOAS形式、その他の形式のいずれかとする(*)。

（LATM/LOAS形式のビットストリーム構成）



注 1　LOASは、同期及びISO/IEC 14496-3に規定される音声符号化情報により構成されるものとする。

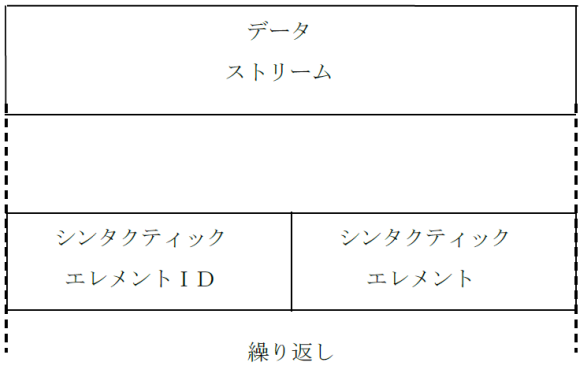
2　LATMヘッダは、ISO/IEC 14496-3に規定される音声符号化情報により構成されるものとする。

3　データストリームは、ISO/IEC 14496-3により符号化される音声データにより構成されるものとする。

4　シンタクティックエレメントIDは、後に続くシンタクティックエレメントの種類又はデータストリームの終了を示すものとする。

5　シンタクティックエレメントは、ISO/IEC 14496-3により符号化される音声データの各構成要素により構成されるものとし、LATMヘッダに記述された回数分繰り返されることとする。

（その他の形式のビットストリーム構成



注 1　データストリームは、ISO/IEC 14496-3により符号化される音声データにより構成されるものとする。

2　シンタクティックエレメントIDは、後に続くシンタクティックエレメントの種類又はデータストリームの終了を示すものとする。

3　シンタクティックエレメントは、ISO/IEC 14496-3により符号化される音声データの各

構成要素により構成される。

（告示別表第 5 号別記 2、別記 3）

3.1.3 ARIB STD-B32

国内のデジタル放送では MPEG-2/4 AAC を運用する際の音声符号化パラメータの制約条件を ARIB STD-B32 で規定している。

表 1 主な運用上の符号化パラメータの制約条件

項目	制約条件
ビットストリーム形式	ADTS（MPEG-2） LATM/LOAS（MPEG-4）
profile	AAC LC
最大符号化チャンネル数	1ADTS あたり最大 5.1 チャンネル（MPEG-2） 1raw_data_block あたり最大 22.2 チャンネル（MPEG-4）
サンプリング周波数	96kHz（注 1）、48kHz、44.1kHz、32kHz、24kHz（注 2）、22.05kHz（注 2）、16kHz（注 2）（MPEG-2） 96kHz（注 1）、48kHz、24kHz（注 1）（MPEG-4） 注 1：V-Low マルチメディア放送のみ 注 2：BS／広帯域 CS デジタル放送においては使用しない。
チャンネルコンフィグレーション	モノ/2ch/3ch/4ch/5ch/5.1ch（MPEG-2） モノ/2ch/3ch/4ch/5ch/5.1ch/ 7.1ch(5/2.1、3/2/2.1、5.1.2ch)/6.1ch/22.2ch（MPEG-4）

また､ARIB 独自の拡張として､ダイアログのレベル制御とダイアログの差し替え制御を追加している。これは、例えば 22.2ch の信号のうち FC（Front Center）や BtFC（Bottom Front Center）をダイアログ専用チャンネルとして番組を制作した場合、それぞれの信号レベルの制御や、音声の切り替えを可能としている。この方式はチャンネルのうち少なくとも一つをダイアログの専用とするため、該当のチャンネルには他の音声信号をミックスすることができないことから制作上の制約となる。MPEG-H 3DA や AC-4 などオブジェクトベースに対応した音声符号化方式ではこのような制約をすることなく同様のサービスが可能となる。

3.2 MPEG-H 3DA

3.2.1 MPEG-H 3DA の概要

MPEG-H 3D Audio は、ISO/IEC23008-3 として、第 1 版が 2015 年に規格化された次世代の先進的

音響システムのための符号化方式である。MPEG では、AAC 規格化以降、HE（High Efficiency）-AAC（ISO/IEC14496-3)、USAC（Unified Speech and Audio Coding：統合音声音響符号化、ISO/IEC23003-3）といったさらなる符号化効率を向上させた符号化方式を規格化している。　MPEG-H 3DAudio は、それらのチャンネルベースの高効率符号化ツール群に加えて、オブジェクトベース音響、シーンベース音響（HOA：Higher Order Ambisonics）に対応し、これらを自由に組み合わせることが可能な方式である。

図 10 には、チャンネル信号とオブジェクト信号を対象とした MPEG-H 3 D Audio のエンコーダおよびデコーダの概略ブロック図を示す。エンコーダでは、チャンネル信号、オブジェクト信号、オブジェクトメタデータはそれぞれ独立に符号化され、ビットストリームが生成される。オブジェクト信号の一部は必要に応じて、プリレンダラー/ミキサー部であらかじめチャンネル信号にミックス処理される。デコーダ側では、復号化されたチャンネル信号はフォーマットコンバーターで再生するスピーカ配置に最適なダウンミックス処理が、復号化されたオブジェクト信号は、復号化されたメタデータ情報をもとにオブジェクトレンダラーでレンダリング処理後、すべての信号がミキサー部でミキシング処理され、DRC（Dynamic Range Control）やラウドネス制御等の処理を経て最終出力を得る。ヘッドフォン再生の場合は、バイノーラルレンダリング処理が施される。

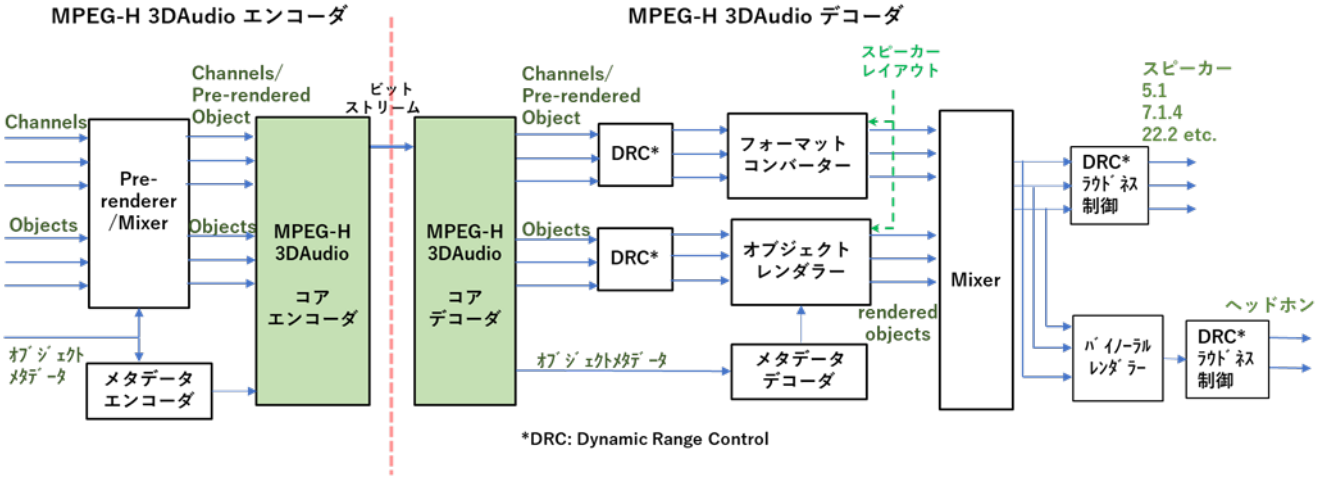


図 10 MPEG-H 3DAudio のエンコーダ/デコーダブロック図

3.2.2 Profile と Level

MPEG 標準では、規格化されたすべてのツールから、処理量と具体的な応用分野に応じて最適なツールのセットを選択し profile として定義している。MPEG-H 3DA では現在下記の 3 つの profile が定義されている。

1. High profile：MPEG-H 3DA で規格化されたすべてのツール群を包含してもので、下記の LC profile、BL profile は、そのサブセットとなる。

2. Low Complexity（LC）profile:実用的な演算量を考慮した符号化ツールセットで、HOA や低ビットレート音声応用に適した音声ツール群を含む。放送、ストリーミング、AR/VR 応用をターゲットとしている。

3. Baseline（BL）profile：LC profile から、HOA と音声ツールを省略した profile で、LC profile のサブセットとなる。主に高音質の放送、ストリーミング応用をターゲットとしている。

また、各 profile には、実装の際の指標となるよう、デコード処理の際のサンプリングレートやチャンネル数等の各種パラメータを制限するためのレベルが 5 段階で定義されている。表 2 に、LC profile、BL profile のレベルの定義を示す。例えば、LC profile のレベル 3 デコーダでは、サンプリング周波数が最大 48kHz まで対応、ビットストリームで伝送されるエレメント数（チャンネル数とオブジェクト数の合計）が最大 32、同時に復号されるエレメント数は最大 16 までとなる。なお BL profile のレベル 3 では、処理量の制限条件をもとで、最大 24 のオブジェクトの同時再生が可能である。また、22.2ch のチャンネル信号を再生するためには、レベル 4 のデコーダが必要である。

表 2　LC profile、BL profile のレベルの定義

レベル	最大ｻﾝﾌﾟﾘﾝｸﾞ周波数	伝送される最大エレメント数 (channel/object)	同時に復号される最大エレメント数 (channel/object)
1	48kHz	10	5
2	48kHz	18	9
3	48kHz	32	16、24**
4 *	48kHz	56	28
5	96kHz	56	28

**BLプロファイルでは、処理量の制限条件の元で
最大24ｵﾌﾞｼﾞｪｸﾄまで可能
*22.2chはレベル 4 で対応

現在、すでに LC profile（レベル 3）を用いて韓国の 4K 地上波放送が開始されており、また音楽ストリーミング分野では BL profile（レベル 3）が実用化されている。

3.2.3 MPEG-H 3DAudio のコア符号化技術

MPEG-H 3DAudio のコア符号化技術では、AAC 以降に開発された様々な符号化技術をベースに、さらなる符号化効率改善のためのツールが追加された。表 3 には、AAC および AAC 以降に開発された符号化ツール（赤字、緑字で示したもの）のうち、MPEG-H 3D Audio の各 profile がどの符号化ツールを採用しているかを示したものである。

図 11 は、コアエンコーダの概略ブロック図、図 12 はその詳細ブロック図を示している。なお、図 11

および図 12 は、Baseline profile に含まれるコア符号化ブロックを示しており、LC profile に含まれる線形予測処理（LPC）ブロックは含まれていない。以下、MPEG-H 3DAudio で採用された代表的な高能率符号化ツールを中心に概説する。

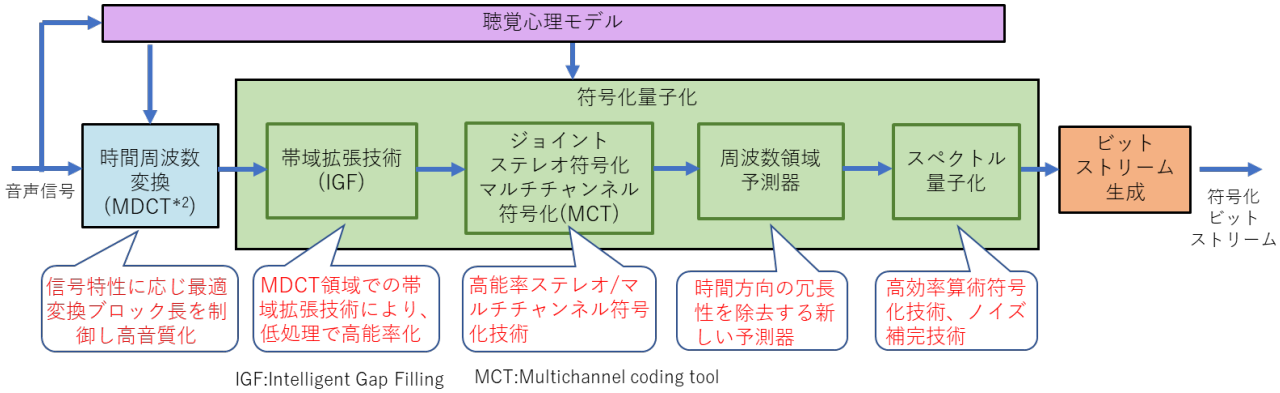


図 11 コアエンコーダの概略ブロック図（BL profile）

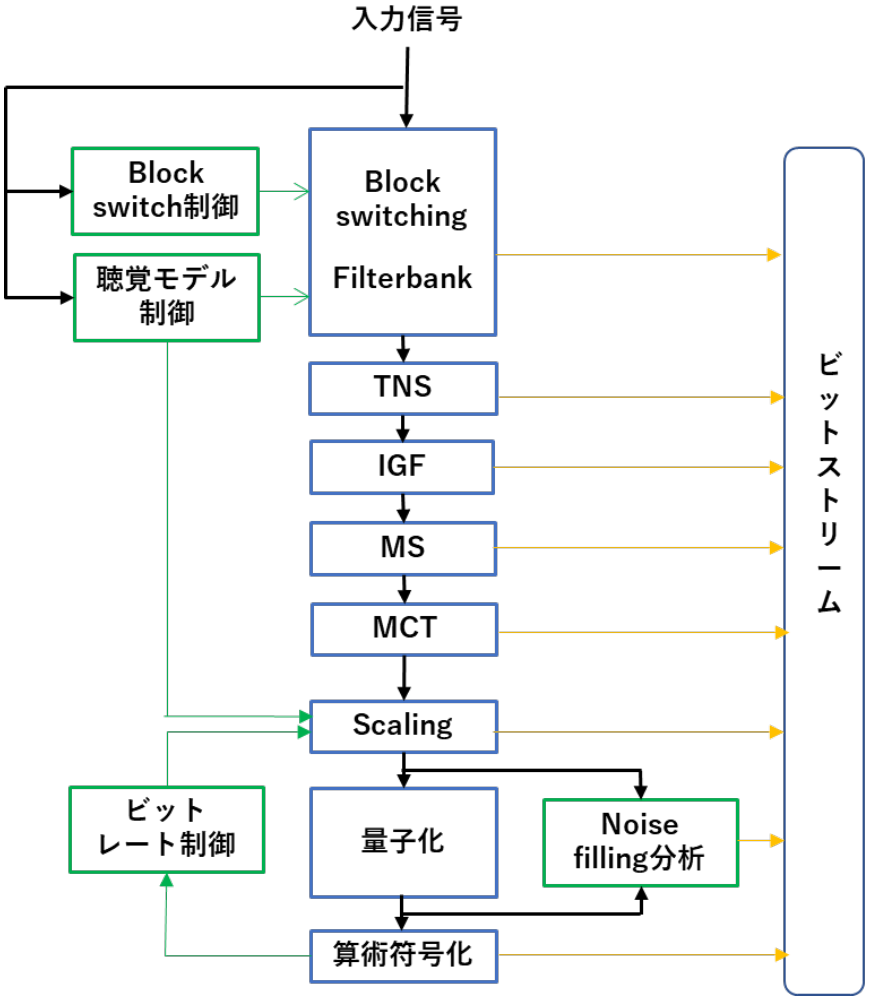


図 12 コアエンコーダの詳細ブロック図（BL profile の符号化ツール）

表 3 各 profile に含まれる符号化ツール

符号化ツール	MPEG-2 MPEG-4 AAC LC	MPEG-H 3DAudio high profile	MPEG-H 3DAudio LC profile	MPEG-H Baseline profile
Block switching	○	○	○	○
Window shape(AAC/MPEG-H)	○	○ ○	○ ○	○ ○
FilterBank(AAC/MPEG-H)	○	○ ○	○ ○	○ ○
TNS (Temporal Noise Shaping)	○	○	○	○
Intensity stereo/Coupling	○			
PNS (Perceptual Noise Substitution)	○			
Noise Filling		○	○	○
MS coding(AAC/MPEG-H)	○	○ ○	○ ○	○ ○
Huffman coding	○			
Arithmetic coding		○	○	○
SBR		○		
Enhanced SBR		○		
IGF (Intelligent Gap Filling)		○	○	○
ACELP/LPD coding tools		○ ○	○ ○	
FDP(Frequency domain predictor)		○	○	○
Multichannel coding tool		○	○	○
LTPF(Long-Term Post-Filter)		○	○	○
Metadata (static/dynamic)		○	○	○
HOA (Higher Order Ambisonics)		○	○	
Format converter/object renderer		○	○	○
DRC and Loudness control		○	○	○

赤字：AAC以降に開発導入されたツール
緑字：MPEG-Hで採用されたツール

1. IGF（Intelligent Gap Filling）

低ビットレート符号化では、オーディオスペクトルの量子化の際、情報量を割り当てることのできない周波数領域(Spectral Gap と呼ぶ)が、特に高周波領域に発生する。IGF は、低周波数領域の符号化スペクトルと補助情報と利用してパラメトリックにこの spectral Gap を埋める技術である。この技術は、いわゆる従来の SBR（Spectral Band Replication）に代表される帯域拡張技術と同様に、低周波スペクトル成分からのコピーにより高周波スペクトルを復元することができる。図 13

に IGF による高域復元の原理（IGF source range から IGF range へのコピー）を模式的に示している。IGF では、最適な低周波側成分を選択するための複数のスペクトルセグメンテーションの手法を採用し、高性能な復元を実現している。また、IGF では、一部の領域（例えばオーディオ信号に強いトーン成分が存在する部分（図中の remaining waveform）については MDCT 変換符号化を施し、復号時に IGF 処理により復元された高周波成分と、復号化された remaining waveform をミックスして出力することにより、より精度の高い復元が可能である。

また、従来の帯域拡大処理は、QMF(Quadrature Mirror Filter)領域上での処理のため、QMF 分析、合成フィルター処理を介する必要があったが、ＩＧＦはすべてＭＤＣＴ領域の処理で完結するため、より効率的な演算量で実現できる。

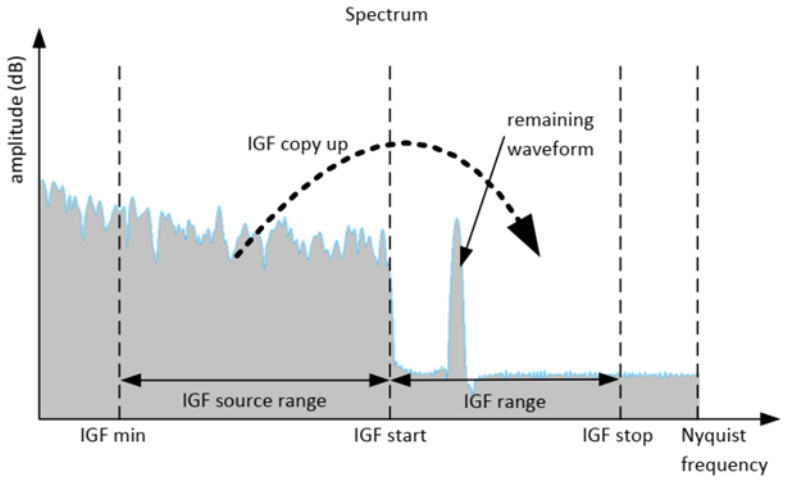


図 13 IGF デコーダの高域復元処理

2. MCT (Multichannel Coding Tool)

MCT（マルチチャンネル符号化ツール）は、マルチチャンネル信号の中で音源の特性に応じて最適なチャンネルペアを選択して、それぞれに最適なステレオ符号化処理を行うことにより、高音質高能率符号化を実現したマルチチャンネル符号化技術である。符号化側では、オーディオフレーム単位ですべてのマルチチャンネル信号間の相関を計算した後、最も相関の高いチャンネルペアを抽出し、最適なステレオ処理を施す。選択されたすべてのチャンネルペアの情報とステレオ符号化処理のための補助情報がビットストリームとして伝送される。復号化側では、伝送された補助情報をもとに、すべてのチャンネルペアに最適なステレオ復号化処理を繰り返し行う。

3. FDP (Frequency Domain Predictor)

FDP は、MDCT 領域で、ハーモニクス性の信号の持つ時間方向の冗長性を前方フレームの MDCT スペクトルを利用して除き、符号化効率を向上させるための符号化ツールである。そのための情報量は、On/OFF フラグ 1 ビットとハーモニクス成分の配置を示すための 8 ビットの情報のみで、符号化のためのオーバーヘッドが少なく、低演算量で実現できる。

4. LTPF (Long Term Post Filter)

   LTPT は、低ビットレート符号化の際に、MDCT 領域での量子化で特にトーナル信号成分に対して生じる時間的変調によるひずみを改善する予測ベースの復号器でのポスト処理ツールである。

3.2.4 MPEG-H 3DAudio のレンダリング技術

MPEG-H 3DAudio では、チャンネルコンテンツ、オブジェクトコンテンツそれぞれに適したレンダラーツールが提供される。チャンネルコンテンツに対しては、たとえば 22.2ch コンテンツを試聴環境の 5.1ch あるいはステレオスピーカで再生する場合のダウンミックス係数等の制御データ情報がビットストリームを介して符号化側から復号化側に伝送され、フォーマットコンバーターでダウンミックス処理される。オブジェクトコンテンツに対しては、オブジェクトレンダラーにより、符号化側から伝送されたメタデータ情報、および試聴環境のスピーカ配置に基づきそれぞれのオブジェクトがターゲットのチャンネル信号にレンダリングされる。これらのすべてのレンダラーの出力は、図 10 に示されるミキサー(Mixer)部で結合され、最終オーディオ出力となる。

1. フォーマットコンバーター(Format Converter)

   MPEG-H 3DA では、予め定義された複数のチャンネル構成のためのダウンミックスマトリクスや制御情報を高効率符号化し、ビットストリームを介して伝送することができる。特に放送用途において、放送事業者、コンテンツ制作者がデコーダ側のダウンミックス処理を制御したい場合に有効である。フォーマットコンバーターは、伝送されたダウンミックス情報から、ターゲットのスピーカ配置に最適なダウンミックスマトリクスを選択しダウンミックス処理を行う。

   ターゲットのスピーカ配置が、伝送されたダウンミックスマトリクス情報のどれとも適合しない場合に対して、MPEG-H 3DA のフォーマットコンバーターは、最適なダウンミックスマトリクスを自動的に生成する機能を有している。入力フォーマットと出力フォーマット間の最適なダウンミックスマトリクスを生成するために、聴覚的な性質を考慮して予め設計されたマッピングルールのリストの中から、各入力チャンネルに対して最適なマッピングルールを選択するアルゴリズムが適用され、ダウンミックスゲインが決定される。その際、実際のスピーカ配置が既定のスピーカ配置からずれている場合の補正処理も考慮されている。

2. オブジェクトレンダラー（Object Renderer）

   MPEG-H 3DA では、オブジェクトの位置は極座標系で記述される。図 14 において例えば、極座標によるオブジェクトの位置 S は、radius(r)、elevation(θ)、angle（ ϕ ）により表される。各オブジェクトは、メタデータによりあらかじめ定義された時間間隔で送られたくるオブジェクトの位置情報、ゲイン情報をもとに、3D-VBAP（3 dimensional -Vector Base Amplitude Panning）アルゴリズムを用いて所望の位置にレンダリングされる。図 15 は、簡単化のために 2 次元の VBAP アルゴリズムを模式的に示したものである。オーディオオブジェクト p は、隣接する 2 つのスピーカ（3 D-VBAP の場合は 3 つのスピーカ）のベクトル l とゲイン g の線形結合により表現することができる。

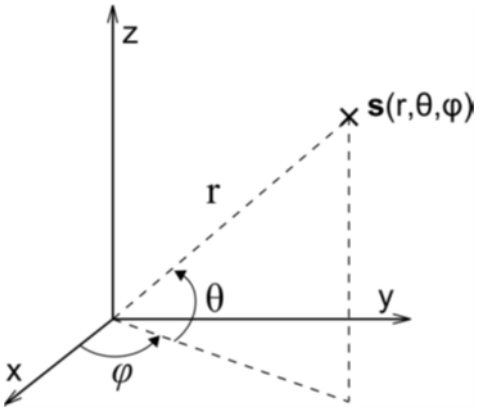


図 14 極座標によるオブジェクトの記述

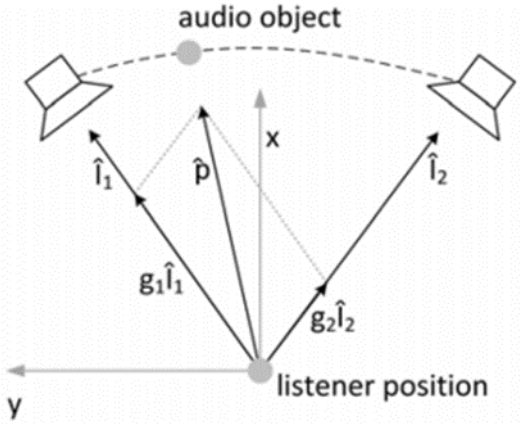


図 15 VBAP アルゴリズム （2D-VBAP の例）

オブジェクトをレンダリングするためには、Delaunay Triangulation Algorithm を用いて、最適な３つのスピーカの頂点による三角メッシュ(Triangle mesh)を決定し、各スピーカに分配する信号ベクトルを生成する。(図 16) ここで MPEG-H 3DA では、三角メッシュが左右のスピーカ、また前後のスピーカに対して対称形となるように設計される。(図 17) これにより、対称位置に配置されるべきオブジェクト同志が非対称な位置にレンダリングされることを避けることができ、任意のスピーカ配置に対して、より高品質なレンダリングを可能としている。

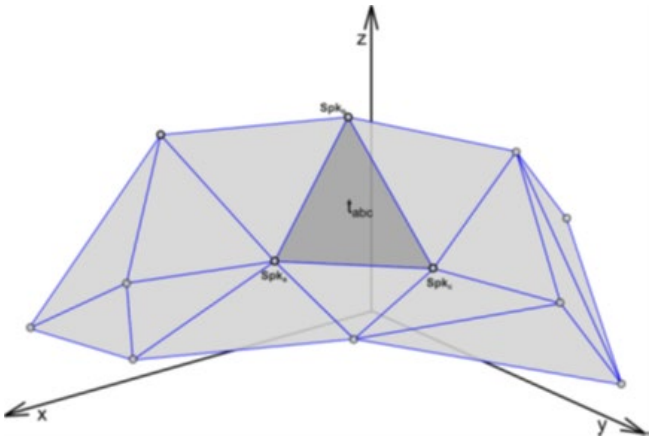


図 16 三角メッシュの例

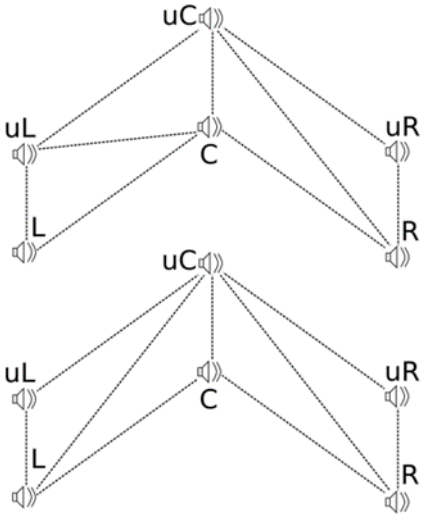


図 17 左右非対称の三角メッシュ（上） 左右対称な三角メッシュ（下）

また、MPEG-H 3DA では、物理的なスピーカによる三角メッシュではカバーできない球体上の領域に対して、バーチャルなイマジナリスピーカーを設定することできる。VBAP アルゴリズムによりイマジナリスピーカーを用いてオブジェクトがレンダリングされた場合は、隣接する物理的に存在するスピーカに等しくエネルギを分配するようにダウンミックスゲインを調整し、ダウンミックス処理される。このイマジナリスピーカーの利用によって、さまざまなスピーカーセットアップに対して、より完全な三角メッシュによる 3 D-VBAP アルゴリズムを実現することが可能となっている。

2.2.5 MPEG-H Audio Stream Format（MHAS）

MPEG-H Audio Stream Format(MHAS)は、MPEG-H エンコーダから出力されるオーディオデータをパケット化して伝送するための自己完結型で柔軟性、拡張性を有する MPEG-H 3DAudio のためのパケット型オーディオストリームフォーマットである。MHAS ストリームは、符号化されたオーディオデー

タ、チャンネル数やサンプリング周波数といった Configuration データ、メタデータ、制御データ等のそれぞれ異なるパケットの列として構成される。MHAS ストリームの例および MHAS のパケットの構造を図 18 に示す。

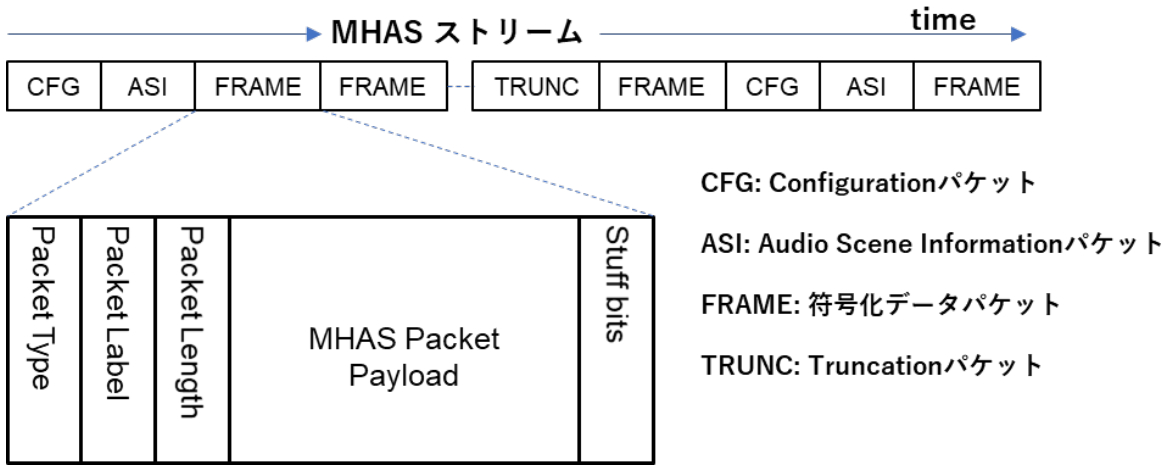


図 18 MHAS ストリームの例と MHAS パケットの構造

ASI（Audio Scene Information パケット）には、追加のメタデータ情報、制御情報などが含まれる。TRUNC パケットには、オーディオサンプルの開始点あるいは終了点からあるサンプル数を破棄するための Truncation 情報が含まれる。この情報により、ビデオのフレーム境界との正確なアラインメントや、オーディオストリームのコンフィギュレーションが途中で切り替わる場合(例えばステレオから 5.1ch への変更など)に、オーディオサンプル単位での正確な変更を可能にしている。パケットには図中に示したもの以外にも、フレーム同期が利用できない場合に使用する SYNC パケット、buffer fullness の情報を含む BINFO パケットなどがあり、必要に応じて挿入される。

MHAS パケットの PacketType は、CFG、ASI 等のパケットのタイプを、PacketLabel は一つの MHAS ストリームの中に含まれる複数の Configuration を区別するためのラベル情報を示しており、例えば、マルチストリームがひとつの MHAS ストリームに含まれる場合にそれらを区別するために使用される。また、MHAS ストリームは、放送、ストリーミング分野で採用されている ISOBMFF(ISO Base Media File Format(ISOBMFF)、MMT（MPEG Media Transport)、MPEG-2Transport Stream などにカプセル化して伝送することも可能である。

3.3 Enhanced AC-3

3.3.1 符号化アルゴリズム

符号化処理の概要を下記ブロック図に示す。Enhanced AC-3 では、入力された PCM 信号に対して、まずチャンネル間などでの信号の類似性を利用する空間符号化とエレメント間冗長度削減処理を用いて情報量の削減を図る。次に、周波数領域への変換を経た後に、周波数帯域間での信号の類似性を利用する帯域拡張処理を用いて情報量を削減する。さらに量子化と量子化値のエントロピ符号化によって情報量を削減し、ビットストリームを形成する。これらの処理は、人間の聴覚特性を模した聴覚心理分析を用いて、

高い主観音質の維持と、情報量削減を最大限両立できるよう制御される。

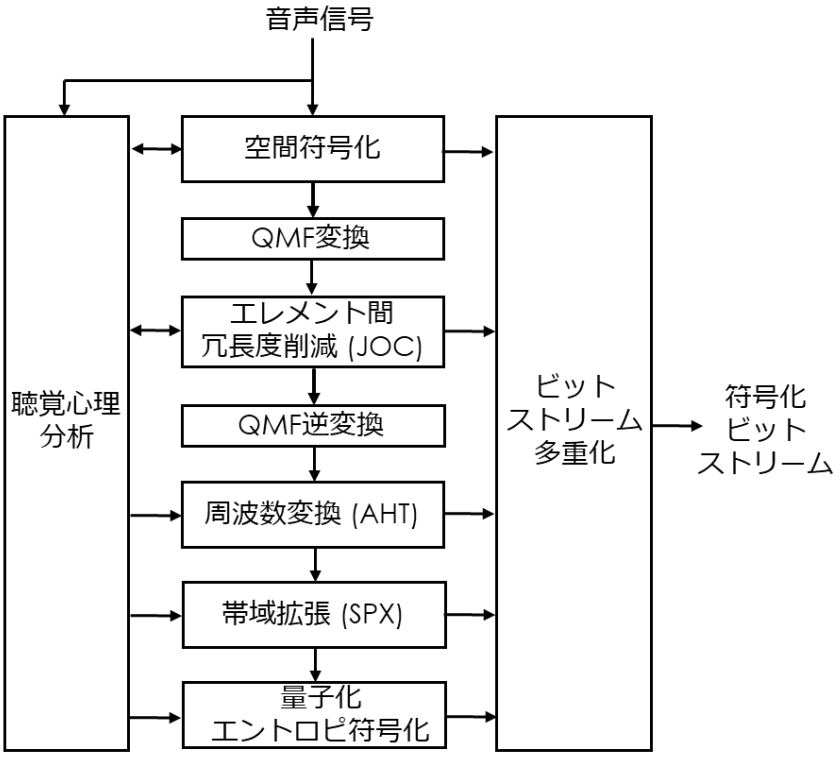


図 19 Enhanced AC-3 符号化ブロックダイアグラム

1. 空間符号化

立体音響で制作された映画や 22.2ch 放送などでは、非常に多くの音声エレメント(チャンネルや動的オブジェクトなどの音源要素)でコンテンツが構成される。空間符号化(Spatial Coding)は、そのような多くのエレメントから構成されるコンテンツを高効率に符号化するために用いられる。例えば、下図の左側は丸印で示された 18 エレメント(18 個の音源要素)で構成される信号である。この図では模式的に、四角形の外周に 7.1ch 相当のチャンネルベースのエレメント、四角形の内側に座標を自由に設定して配置できる効果音などの 11 個のエレメントが存在している。空間符号化では、下図右側に示すように、近接している複数エレメント(破線の円で囲まれたエレメント)を 1 エレメントとして合成することや、一部のエレメントを破線矢印のように複数エレメントに分配して置き換えることで、符号化対象となるエレメント数を低減する。聴覚的に知覚困難な合成や置き換えを選択すること、また、1 エレメント当たりの符号化ビットレートを高めることで、音質向上を図る。最大 16 エレメントの符号化が可能な Enhanced AC-3 においては、22.2ch(24 エレメント)を符号化する際にも、この空間符号化を用いる。エレメント数を低減する手法としてはダウンミックスが広く用いられているが、空間符号化は各エレメントの聴感的重要性に応じて適応的に合成や置き換えを選択する点で優れている。

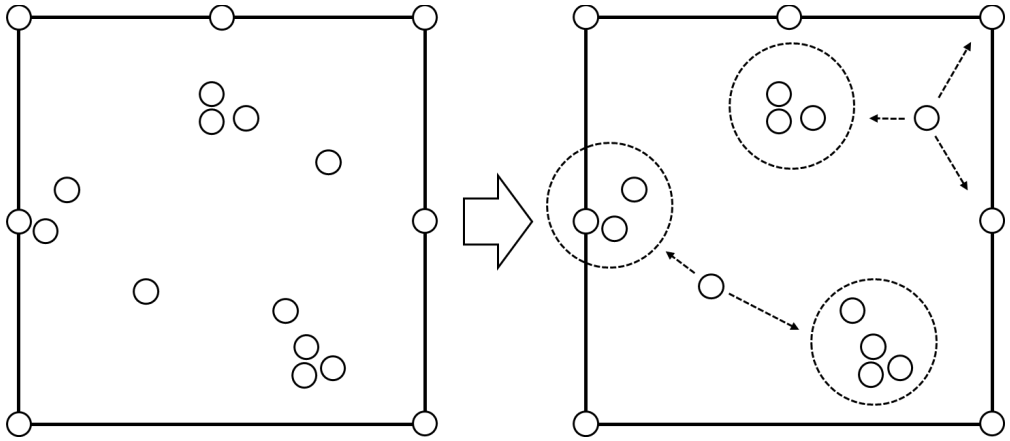


図 20 空間符号化

2. エレメント間冗長度削減
   空間符号化が出力する最大 16 個のエレメントに対し、そのエレメント間の特徴量を抽出して効率的な情報表現をおこなうパラメトリック符号化を用いる。複素 QMF(Quadrature Mirror Filter)信号上で聴覚特性に基づくモデルを用いた分析により、最大 16 個のエレメントを、6 個や 8 個といった、より少ないエレメント数のオーディオ信号と、そこから冗長度削減前の最大 16 個のエレメントへ復元するために必要な補助情報として効率的に符号化する JOC(Joint Object Coding)と呼ぶ符号化ツールを利用する。
3. 周波数変換
   オーディオ信号の定常性に応じて変換ブロック長を適応的に選択して周波数領域へと変換する。変換には MDCT(Modified Discrete Cosine Transform)と DCT(Discrete Cosine Transform)を組み合わせた AHT(Adaptive Hybrid Transform)を用いる。これにより、高い周波数解像度を持つ周波数変換を低演算量に実現する。
4. 帯域拡張
   SPX(Spectral Extension)と呼ぶ符号化ツールを利用し、デコーダにおいて高周波数域の信号を中・低周波数域の信号から高効率に生成できるような補助情報を生成する。
5. 量子化・エントロピ符号化
   周波数領域信号を量子化およびエントロピ符号化する。量子化には信号特性に応じてベクトル量子化を用いることもできるようにし、符号化効率の向上を図る。

3.3.2 Profile と Level

Enhanced AC-3 では、利用できる符号化ツールの集合を表す Profile、および、処理の複雑度を表す Level が単一化されている。ETSI TS 103 420 での機能拡張に対応したデコーダでは、最大 16 エレメントの同時デコードに対応する。22.2ch については、空間符号化を用いて 16 エレメント相当のオーディオ信号に変換した上でエンコードをおこなう。

3.3.3 レンダリングアルゴリズム

Enhanced AC-3 においては、三次元の直交座標系を用いて音声オブジェクトの位置を記述し、トリプルバランスパンナーと呼ぶアルゴリズムを用いてレンダリングをおこなう。このアルゴリズムは AC-4 と同一であり、3.4.3 節でその概要を説明する。

3.3.4 ビットストリーム形式

Enhanced AC-3 は、当初最大 7.1ch に対応する符号化方式として 2005 年に規格化(ETSI TS 102 366)され、2016 年に高さ方向も含めた 3 次元音響機能などを追加する ETSI TS 103 420 が規格化された。この機能追加を後方互換性を保ちながら実現するために、ETSI TS 103 420 で追加される JOC やオブジェクトオーディオなどのメタデータは、ETSI TS 102 366 に用意されている拡張情報領域に格納されている。この拡張情報領域はユーザデータなどを格納する目的で用意され、ETSI TS 103 420 に対応する以前のデコーダでは利用せずに読み飛ばす領域となる。この仕組みにより、ETSI TS 103 420 に対応する以前のデコーダでは、ビットストリーム中の最大 7.1ch のオーディオデータのみをデコードする。また、ETSI TS 103 420 に対応したデコーダでは、これに加えて拡張情報領域に含まれている JOC やオブジェクトオーディオなどのメタデータもデコードし、これを最大 7.1ch のチャンネルベースのオーディオデータと組み合わせて、最大 16 エレメントのオーディオデータをデコードする。

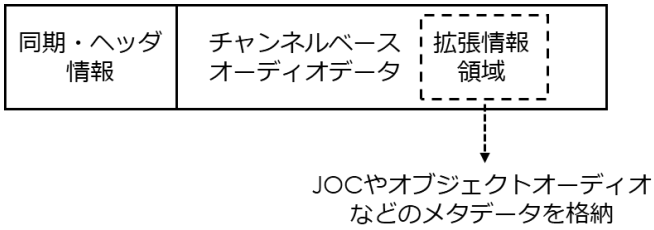


図 21 Enhanced AC-3 のビットストリーム

3.4 AC-4

3.4.1 符号化アルゴリズム

符号化処理の概要を下記ブロック図に示す。AC-4 では、入力された PCM 信号に対して、まずチャンネル間などでの信号の類似性を利用した空間符号化と QMF 領域でのエレメント間冗長度削減処理を適用し、また、周波数帯域間での信号の類似性を利用した帯域拡張処理を用いて情報量の削減を図る。更に、MDCT(Modified Discrete Cosine Transform)による周波数領域への変換を経た後に、量子化と量子化値のエントロピ符号化によって情報量を削減し、ビットストリームを形成する。これらの処理は、人間の聴覚特性を模した聴覚心理分析を用いて、高い主観音質の維持と、情報量削減を最大限両立できるよう制御される。

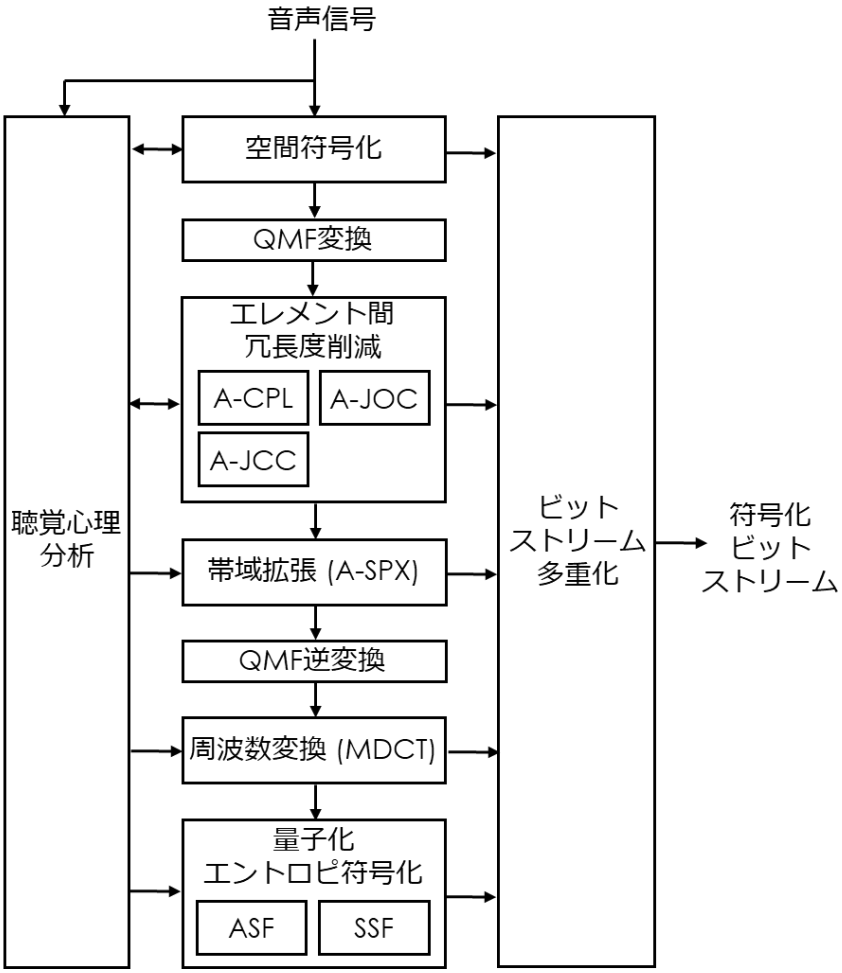


図 22 AC-4 符号化ブロックダイアグラム

以下に、各処理の概要を説明する。

1. 空間符号化

   Enhanced AC-3 と同じく、空間符号化(Spatial Coding)を用いることで、多くのエレメントから構成されるコンテンツの符号化を高効率化する。22.2ch を符号化する場合、Level 3 を超える規定を用いれば空間符号化の利用は必須ではない。その一方、空間符号化を用いることで、北米や欧州各国の放送規格で採用されている AC-4 Level 3 デコーダでも 22.2ch コンテンツを再生できるという大きな利点が生まれる。

2. エレメント間冗長度削減

   空間符号化が出力する複数のエレメントに対し、そのエレメント間の特徴量を抽出して効率的な情報表現をおこなうパラメトリック符号化を用いる。複素 QMF(Quadrature Mirror Filter)領域で分析をし、ステレオ・マルチチャンネル信号のチャンネル間相関を利用する A-CPL(Advanced Coupling)、複数オブジェクトを聴覚特性に基づくモデルを用いて効率的に符号化する A-

JOC(Advanced Joint Object Coding)、A-JOC のチャンネルベース版に相当する A-JCC(Advanced Joint Channel Coding)といった符号化ツールを利用する。

3. 帯域拡張

A-SPX(Advanced Spectral Extension)と呼ぶ符号化ツールを利用し、デコーダにおいて高周波数域の信号を中・低周波数域の信号から生成できるような補助情報を生成する。信号の特性やビットレートに応じて、下図(b)のように特定の周波数帯域や、下図(c)のように特定の時間領域に対してこの処理を選択的に適用することで、より複雑な特性の信号を高効率に符号化できるようになった。

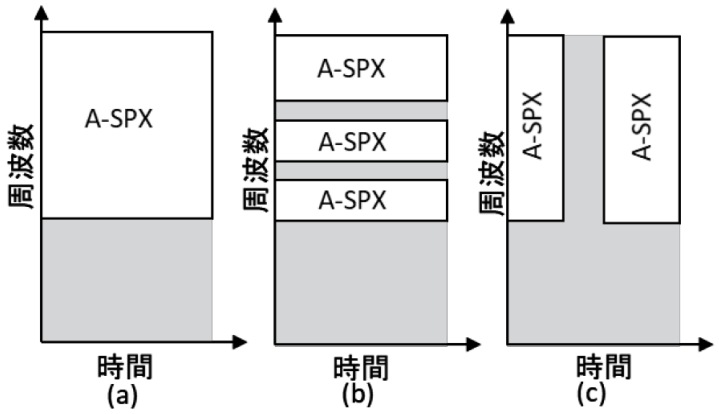


図 23 帯域拡張（A-SPX）

4. 周波数変換

オーディオ信号の定常性に応じて変換ブロック長を適応的に選択して周波数領域の信号へと変換する。変換には AAC 等と同様に MDCT(Modified Discrete Cosine Transform)を用いる一方、AAC LC (Low Complexity) Profile での 1024 点と 128 点の 2 種類の変換ブロック長に対して、AC-4 では 2048 点、1024 点、512 点、256 点、128 点を下図のように多様に組み合わせて利用することで、複雑に特性が変化する非定常的信号を効率的に符号化できるようになった。

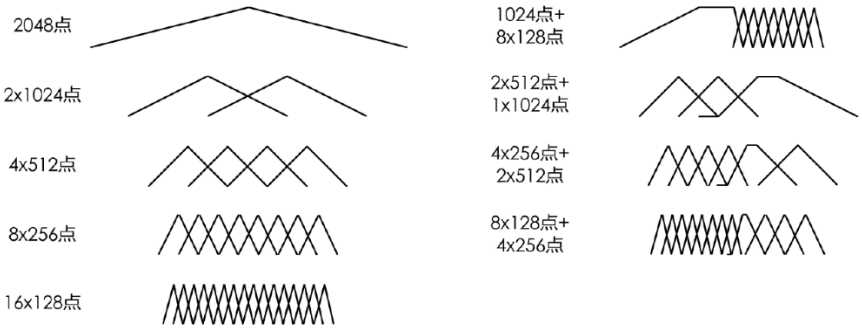


図 24 適応的変換ブロック長

5. 量子化・エントロピ符号化

MDCT 係数を量子化およびエントロピ符号化する処理を、AC-4 ではスペクトラルフロントエン

ドと呼ぶ。一般的なオーディオコンテンツに適した ASF(Audio Spectral Frontend)に加えて、人声の高圧縮に適した SSF(Speech Spectral Frontend)の 2 種類を備える。両者共に MDCT 領域での処理とすることで、ASF と SSF をシームレスに切り替えて利用できる。

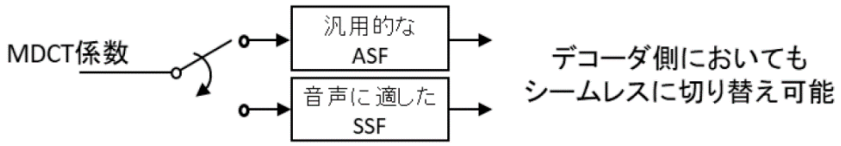


図 25 ASF と SSF の切り替え

3.4.2 Profile と Level

利用できる符号化ツールの集合を表す Profile については、AC-4 では一つの Profile に統一しており、Profile 間での互換性問題は生じない。処理の複雑度を表す Level については、Level 3 が広く用いられ、北米や欧州各国の放送規格においても Level 3 の利用を規定している。この Level 3 に対応する AC-4 デコーダでは、最大 17 のエレメントと 1 つの LFE チャンネルの同時デコードに対応する。

22.2ch については、AC-4 規格(ETSI TS 103 190 Part 2)においてもサポートをしているが、1 チャンネルを 1 エレメントとして独立に符号化する場合には、Level 3 を超える Level 4 の規定と利用が必要となる。一方、AC-4 では符号化アルゴリズムの説明に記載した空間符号化を用いることで、Level 3 においても 22.2ch コンテンツを符号化することが可能である。北米や欧州各国と共通の AC-4 Level 3 デコーダを用いることで、22.2ch サービスに対応する受信機実装コストの低減と普及の促進を図ることができるという利点がある。

3.4.3 レンダリングアルゴリズム

AC-4、および、Enhanced AC-3 においては、三次元の直交座標系を用いて音声オブジェクトの位置を記述し、トリプルバランスパンナーと呼ぶアルゴリズムを用いてレンダリングをおこなう。同様なレンダラーは、家庭用だけでなく、映画での立体音響技術として知られる Dolby Atmos での映画制作や上映でも利用されており、高品質・高臨場感コンテンツの制作・再生に多くの実績がある。

三次元の直交座標系

再生環境における横方向（例えば、左前方スピーカから右前方スピーカへの方向）を x 軸、奥行き方向（例えば、前方左スピーカから後方左スピーカへの方向）を y 軸、高さ方向を z 軸とする三次元の直交座標系を用いて、音声オブジェクトの位置(x, y, z)を記述する。座標の例としては、左前方(L)スピーカが(-1.0, 1.0, 0.0)、右前方(R)スピーカが(1.0, 1.0, 0.0)、左後方(Lrs)スピーカが(-1.0, -1.0, 0.0)、左前方の天井スピーカ(Lfh)スピーカが(-1.0, 1.0, 1.0)と記述される。

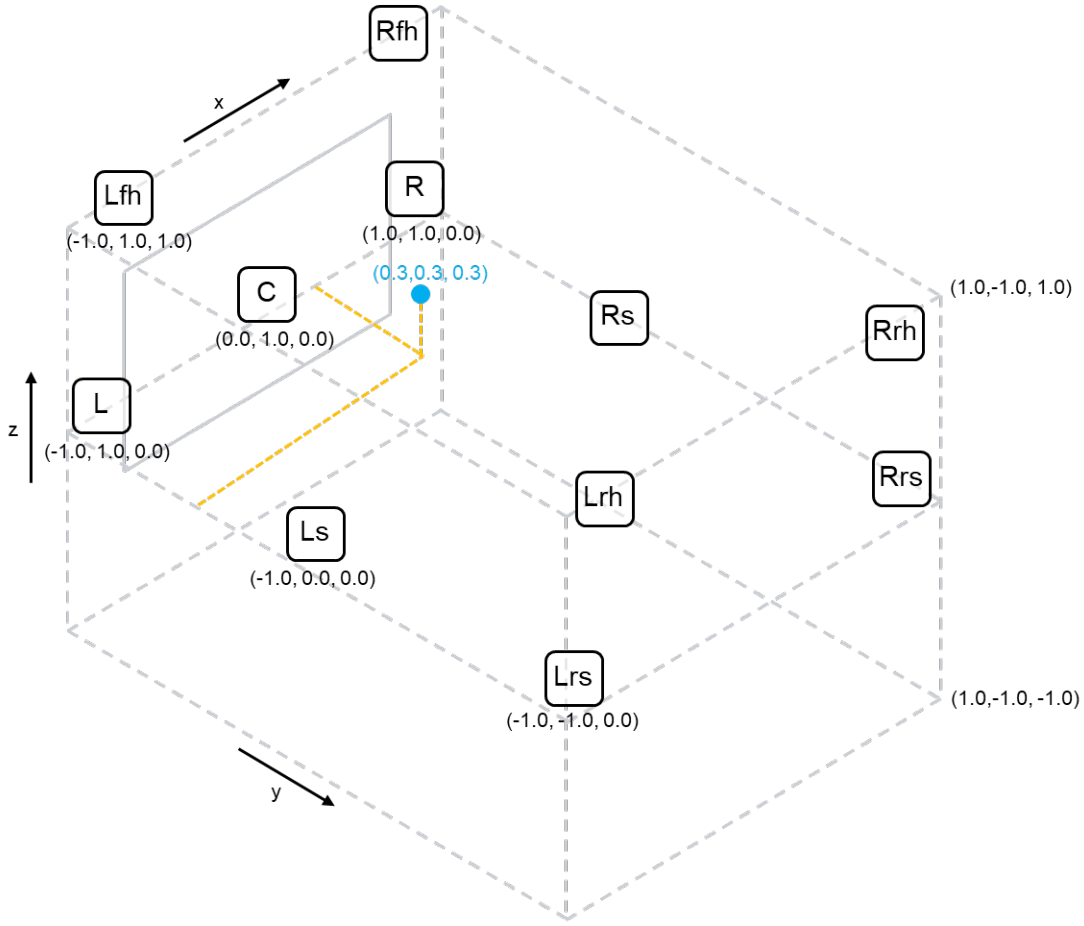


図 26 三次元の直交座標系

トリプルバランスパンナー

従来のステレオや 5.1ch サラウンド制作では、デュアルバランスパンナーが広く用いられてきた。例えば、センター(C)スピーカ(0.0, 1.0, 0.0)と右(R)スピーカ(1.0, 1.0, 0.0)の間の(x, 0.0, 0.0)に位置する点音源 p がある場合を仮定する。この点音源 p を左右スピーカによって再生する場合、デュアルバランスパンナーでは、

左(L)スピーカ出力 $= p * \cos(x * 0.5\pi)$

右(R)スピーカ出力 $= p * \sin(x * 0.5\pi)$

として、点音源 p のエネルギを維持しながら、左右スピーカからの距離に応じて信号 p を分配する。

トリプルバランスパンナーは、このデュアルバランスパンナーを 3 次元に拡張したものである。以下では、上図のように(0.3, 0.3, 0.3)に位置する点音源を例に説明する。

トリプルバランスパンナーでは、まず、上図のような再生環境空間を z 軸方向(高さ方向)に平面(plane)に分割する。例えば、7.1.4ch 環境であれば L/C/R/Ls/Rs/Lrs/Rrs で構成される第 1 平面と、Lfh/Rfh/Lrh/Rrh で構成される第 2 平面に分割する。そして、点音源 p の上下両方向に近接するこれらの 2 平面に対してデュアルバランスパンナーを z 軸方向に用いて、両平面に対する出力を z ゲイン

として計算する。つまり、点音源 p (0.3, 0.3, 0.3)においては、
第 1 平面に対する z ゲイン $= \cos(0.3 * 0.5\pi) = 0.8910$
第 2 平面に対する z ゲイン $= \sin(0.3 * 0.5\pi) = 0.4540$
となる。
次に、各平面において、y 軸方向でスピーカを行(row)として分割する。7.1.4ch 環境では、第一平面において L/C/R が第 1 行、Ls/Rs が第 2 行、Lrs/Rrs が第 3 行となり、第 2 平面において、Lfh/Rfh が第 1 行、Lrh/Rrh が第 2 行となる。これら両平面において、点音源 p に隣接する行に対する出力を y ゲインとして計算する。第 1 平面では、
第 1 行に対する y ゲイン $= \sin(0.6 * 0.5\pi) = 0.8090$
第 2 行に対する y ゲイン $= \cos(0.6 * 0.5\pi) = 0.5878$
となり、第 2 平面についても同様な計算をおこなう。
更に、各行における x 軸方向に近接するスピーカに対しても同様な計算をおこない、x ゲインを算出する。第 1 平面の第 1 行では、
L スピーカに対する x ゲイン $= \sin(0.6 * 0.5\pi) = 0.8090$
C スピーカに対する x ゲイン $= \cos(0.6 * 0.5\pi) = 0.5878$
となり、これを各平面の各行において計算する。
最後に各スピーカに対する x ゲイン、y ゲイン、z ゲインを乗じて最終ゲインを得る。(0.3,0.3, 0.3)に位置する点音源 p の例では、8 スピーカの最終ゲインが非ゼロとなり、これらのスピーカを用いて点音源 p を定位させる。また、より一般的な、点音源 p が xy 平面上、xz 平面上、yz 平面上の何れかに位置する場合は、最大 4 スピーカの最終ゲインが非ゼロとなる。

3.4.4 ビットストリーム形式

AC-4 のエレメントビットストリームは sync frame (synchronization frame)と呼ばれるデータが連続する形で構成されている。各 sync frame は AAC の ADTS 形式と同様に、フレームの先頭を検知するための同期語(sync word)で始まり、誤り検出のための CRC(Cyclic Redundancy Code)や圧縮されたオーディオ信号やメタデータを含む raw AC-4 frame が続く。この raw AC-4 frame の特長として、1 つ以上の substream と TOC(Table of Contents)と呼ばれる目次的な情報を含むことが挙げられ、以下にビットストリーム構成の概要を図示する。

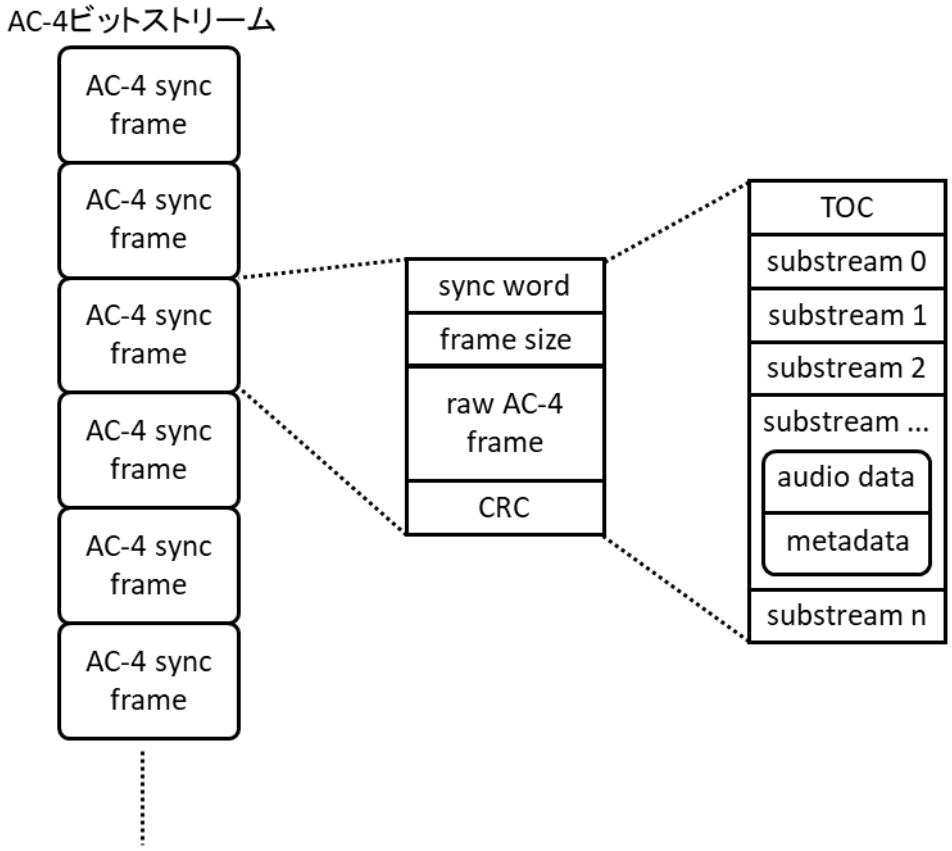


図 27 AC-4 のビットストリーム

raw AC-4 frame の各 substream には、モノラルやマルチチャンネルのオーディオ信号やオーディオオブジェクト信号が符号化されている。TOC はその目次的な情報として、各 substream の内容やビットストリーム中での位置を記載している。また、TOC には presentation info と呼ばれる情報を複数記載することができ、各 presentation info は substream をどのように組み合わせて視聴者に提示するかを示している。

presentation と substream を用いた複数言語番組の例を下図に示す。ここでは、3 言語(日本語、英語、韓国語)のナレーションと、ナレーションの無い 5.1ch の背景音が一つのビットストリームに 4 つの substream として符号化されている。

- substream 0: 5.1ch の背景音
- substream 1: モノラルの日本語ナレーション
- substream 2: モノラルの英語ナレーション
- substream 3: モノラルの韓国語ナレーション

更にこのビットストリームには 4 つの presentation info が含まれ、例えば視聴者が presentation1 (図中の presentation info 1)を選択した場合には、5.1ch の背景音(substream 0)とモノラルの日本語ナレーシ

ョン(substream 1)を組み合わせるようデコーダを動作させる。視聴者が英語や韓国語を選択した場合も同様に、共通の 5.1ch の背景音(substream 0)と英語(substream 2)または韓国語(substream 3)のどちらかを組わせてデコーダを動作させる。また、ナレーションが無く、背景音(substream 0)だけの presentation を含めることもできる。このビットストリーム構成を採ることで、複数言語の番組などを効率的に伝送することできるようになる。

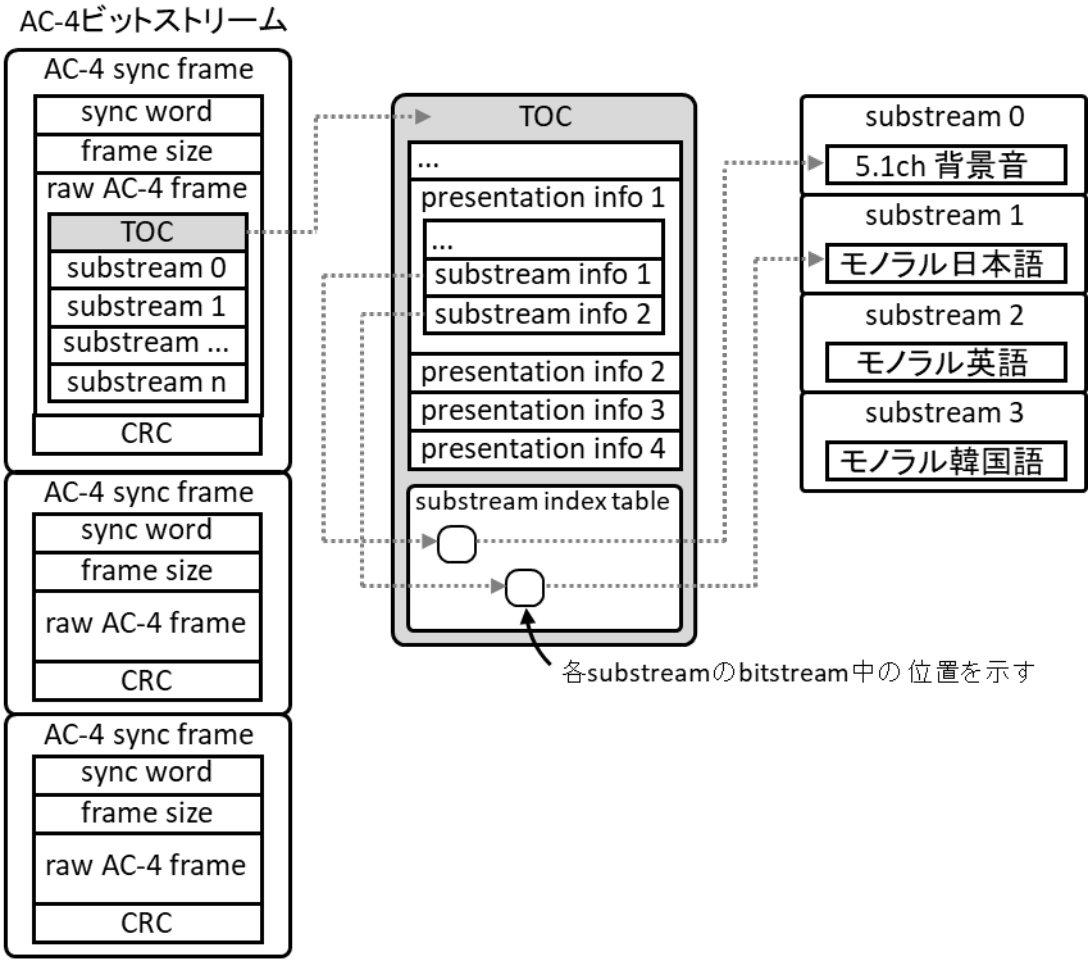


図 28 複数言語番組のストリーム例

3.4.5 その他の機能

スマートフォンやタブレット端末などのステレオオーディオ出力を主用途とする機器向けに Immersive Stereo という符号化ツールが用意されている。これは、5.1ch や高さ方向も含めた立体音響を効率的に扱うためのツールであり、音楽配信サービスなどで利用されている。エンコーダ側では下図に示すように、AC-4 エンコーダの前処理として、IMS レンダラー(Immersive Stereo Renderer)が入力オーディオ信号を分析し、ステレオ信号と IMS メタデータを生成する。この IMS メタデータは、ステレオ信号からヘッドフォン、および、ステレオスピーカ用にバーチャライズ処理を施したステレオ信号へと変換するための情報を含み、通常のステレオ信号をエンコードした AC-4 ビットストリームに埋め込まれる。デ

コーダ側では、通常のステレオ信号に加えて、IMS メタデータを利用することでヘッドフォン、および、ステレオスピーカ用にバーチャライズ処理を施したステレオ信号も得ることができる。この Immersive Stereo は、モバイル機器向けサービスでのビットレートとデコーダ演算量の低減に有用である。ビットレートとしては、MUSHRA 評価における”Good”に対しては 64kbps、”Excellent”に対しては 112kbps を推奨としている。デコーダ演算量としては多チャンネル信号をデコード・バーチャライズする場合と比較して 1/3～1/4 程度に低減でき、バッテリー動作をするモバイル機器での利用に適している。これらの特長から、国外の音楽配信サービスなどで既に利用が始まっている。

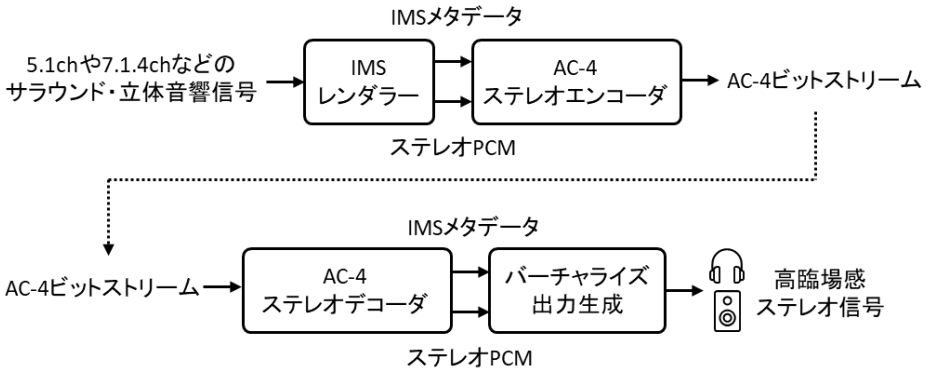


図 29 Immersive Stereo

第4章　音声符号化方式の比較

提案のあった各音声符号化方式について比較をおこなう。

4.1　オブジェクトベース音響対応状況

ARIB でダイアログ制御の拡張について規格化されているが MPEG-2/4 AAC は基本的にチャンネルベースの音声符号化方式である。Enhanced AC-3 や AC-4 については当初チャンネルベースで開発された音声符号化方式であるが、伝送チャンネル数やメタデータ、レンダラーなどの機能を拡張することでオブジェクトベースに対応している。MPEG-H 3DA はオブジェクトベース音響やシーンベース音響に対応することを目的に開発した音声符号化方式である。OBA を想定した場合、MPEG-H 3DA、Enhanced AC-3、AC-4 を選択することとなる。

4.2　各符号化方式のコア符号化の違いについて

第 2 章で記述したように音声符号化方式は多くの圧縮ツールの集合体である。ここでは、提案手法のコア符号化における圧縮ツールの違いについて簡単に述べる。

今回提案されている符号化方式はいずれも周波数領域で圧縮処理をおこなう変換符号化として分類できる。これは、時間信号である音声信号を短時間フレームに区切り、フレームごとに周波数変換を行うことで聴覚心理モデルを適用し、マスキング閾に基づくビット削減を可能とするためである。一般には MDCT 係数を再量子化して伝送する。また、開発時期を考えると、最新の符号化ツールやそれらを用いた全体的なチューニングの点において、MPEG-H 3DA と AC-4 は、MPEG-2/4 AAC、Enhanced AC-3 に比べ符号化品質が高いことが期待できる。

MPEG-2/4 AAC や MPEG-H　3 DA では音声信号を MDCT 変換し、全ての圧縮ツールを MDCT 上で処理しているが、Enhanced AC-3/AC-4 では、一度 QMF によるフィルタバンクで帯域信号に変換しエレメント間の冗長性を削除したのち（AC-4 ではさらに帯域拡張処理を行い）、再度時間信号に変換してからMDCT 変換を行う。

Enhanced AC-3/AC-4 では QMF 変換の前段に最大の同時デコード数を上回る信号が入力された際に、時間領域で空間符号化と呼ばれるエレメント数変換を行う機能がある。これを用いることで ATSC をはじめとする海外の次世代放送規格と同じ Level 3 デコーダで 22.2ch などのマルチチャンネル音響に対応できる利点がある一方、入力信号のエレメント数を維持する場合は MPEG-H 3DA と同様に Level 3 を超える規定を用いる必要がある。MPEG-H　3 DA では基本的に既定以外の音声モードが入力された場合は個別に符号化方法を指示する方法をとっている。しかし、放送の運用では高品質な運用が求められるため最大デコード数内でサービスを行うことが前提となるように、音声モードを ARIB の規定で制約することとなる。

高い周波数成分を高効率に圧縮する帯域拡張処理は MPEG-H　3 DA と Enhanced AC-3/AC-4 で採用

されているが、MPEG-H 3DA と Enhanced AC-3 は MDCT 上で、AC-4 では QMF 上で帯域拡張処理を行っている。高域になるほど聴覚心理モデルによる臨界帯域内の MDCT 係数の数が多くなるため、帯域拡張処理は圧縮効率に大きな影響があると考えられる。MPEG-2/4 AAC では採用されていないため、ほかの符号化方式に比べ品質は低くなると考えられる。

チャンネル間もしくはエレメント間の冗長性削除のための圧縮ツールがそれぞれ利用されている。MPEG-2/4 AAC では固定された 2ch 間の冗長性を削除するジョイントステレオや M/S を採用しているが、MPEG-H 3DA ではマルチチャンネル中で最も相関の高いチャンネルペアを探索し最適なステレオ符号化ツールを適用する MCT が採用されている。Enhanced AC-3/AC-4 では QMF 上でエレメント間冗長性削除ツールを採用し伝送エレメント数を入力信号数よりも削減するとともに、補助情報によりデコーダ側で入力信号数に復元する。これらから、22.2ch のような信号数の多い音声モードの場合にマルチチャンネルに対応した冗長性の削減が期待できる MPEG-H 3DA や Enhanced AC-3/AC-4 の符号化効率が高いと期待できる。

人声に特化した符号化ツールを採用しているのが AC-4 である。MPEG-H 3DA でも LC profile では採用されているが、今回放送品質を主なターゲットと想定して選択された BL profile には採用されていない。この符号化ツールの主観的な効果については今回確認していない。

以上の違いにより、放送高度化の要求条件を達成できる符号化方式は MPEG-H 3DA と AC-4 であると考えられる。ただし、チューニングによって品質は大きく変化するため、あくまでも可能性とすべきである。

4.3 エレメント数・同時再生数

各音声符号化方式のエレメント数、同時再生数を比較する。AAC はデコード可能な同時再生数とエレメント数が同じであるが、Enhanced AC-3、AC-4、MPEG-H 3DA についてはデコード可能な同時再生数を超えるエレメント数をエンコード可能である。

表 4 同時再生数、伝送チャンネル数

	MPEG-2/4 AAC	MPEG-H 3DA	Enhanced AC-3	AC-4
同時再生数 （同時デコード数）	MPEG-2: 8 (ex.7.1ch) MPEG-4 : 24 (ex.22.2ch)	レベル 3 : 16 (ex. 7.1.4+4obj) 24 (モノオブジェクト) レベル 4 : 28 (ex. 22.2ch+4obj)	16 (ex. 7.1.4+4obj)	レベル 3 ：18 (ex.7.1.4+6obj)
エレメント数	MPEG-2: 8 MPEG-4 : 24	レベル 3 : 32 レベル 4 : 56	16＋3 ステレオ	規定なしでビットレートの上限まで

4.4　対応するチャンネルコンフィグレーション

各音声符号化方式が既定で対応可能なチャンネルコンフィグレーションについて示す。チャンネルコンフィグレーションの表記法はスピーカ数と配置を識別するため 2 種類用意した。

既定以外のチャンネルコンフィグレーションについては、同時再生数の上限の範囲内に限り MPEG-H 3 DA、Enhanced AC-3、AC- 4 ともにオブジェクト音声と位置情報を利用することで対応可能である。

AC-4 は 7.1ch 以上の LFE チャンネルの有無を分けて識別しているが MPEG-H 3DA では 7.1ch 以上の LFE がないチャンネルコンフィグレーションの既定はない。そのため、LFE がないチャンネルコンフィグレーションを伝送する際 LFE のデータを 0 として符号化する。

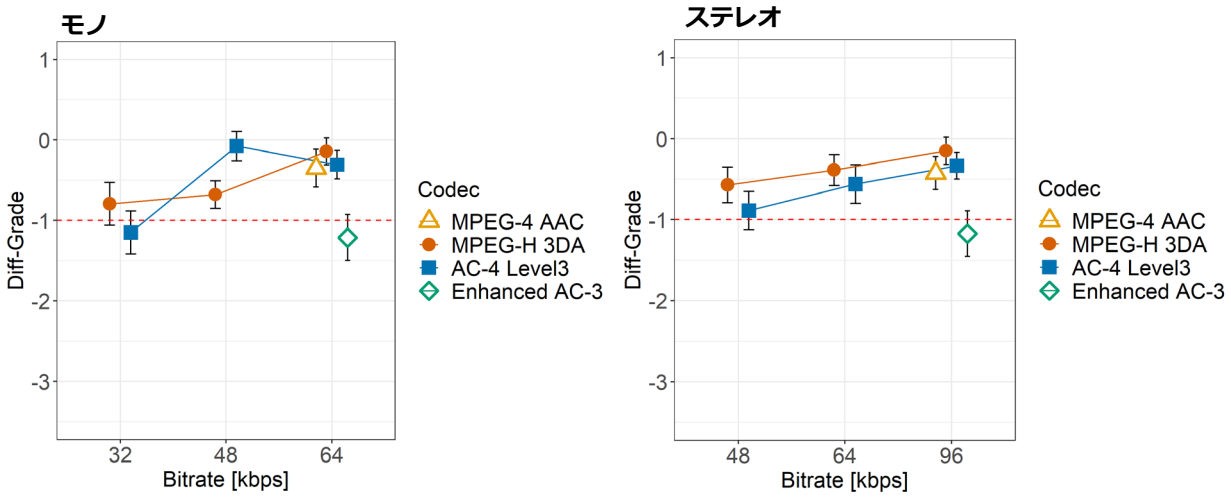
表 5 既定で対応するチャンネルコンフィグレーション

スピーカ.LFE（中層.LFE.上層）	中層（前/後）.LFE＋上層＋下層	MPEG-2/4 AAC	MPEG-H 3DA	Enhanced AC-3	AC-4
1.0	1.0	○	○	○	○
2.0	2.0	○	○	○	○
3.0	3.0	○	○	○	○
4.0	3/1.0	○	○	○	
5.0	3/2.0	○	○	○	○
5.1	3/2.1	○	○	○	○
7.1	5/2.1	○	○	○	○
7.0	5/2.0				○
1+1*			○	○	
3.1	2/1.1		○		
4.0	2/2.0		○		
6.1	3/3.1	○**	○		
7.1	3/4.1	○**	○		○
7.0	3/4.0				○

22.2	5/5.2+9H+3B	○**	○		○
7.1（5.1.2）	3/2.1+2H	○**	○		○
7.0	3/2.0+2H				○
10.2	3/4.2+3H		○		
9.1（5.1.4）	3/2.1+4H		○		
11.1（5.1.6）	3/2.1+6H		○		
13.1（7.1.6）	3/4.1+6H		○		
11.1（7.1.4）	3/4.1+4H		○		○
11.0	3/4.0+4H				○
13.1（9.1.4）	5/4.1+4H		○		○
13.0	5/4.0+4H				○
*dual mono **MPEG-4 AAC のみ					

4.5　品質とビットレート

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を確認した（実験の詳細は別紙 1 を参照）。図 31 に実験結果、表 6 に各音声フォーマットの所要ビットレート示す。



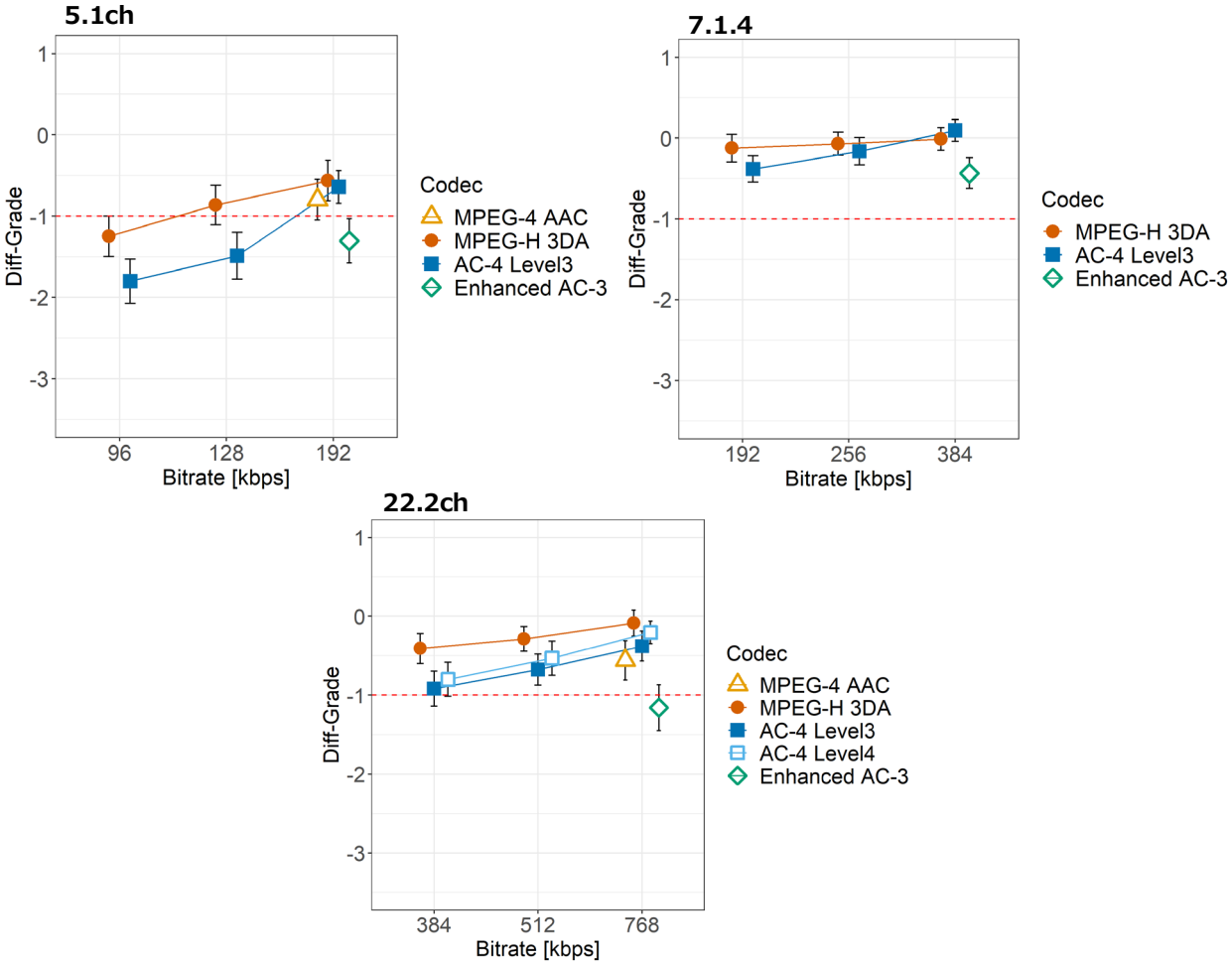


図 301 コアコーダの評価結果（4 音源の平均）

表 6 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 チャンネル以上のマルチチャン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 7 に示す。MPEG-4 AAC の所要ビットレートは先行研究による。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、チャンネル数の多い 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。

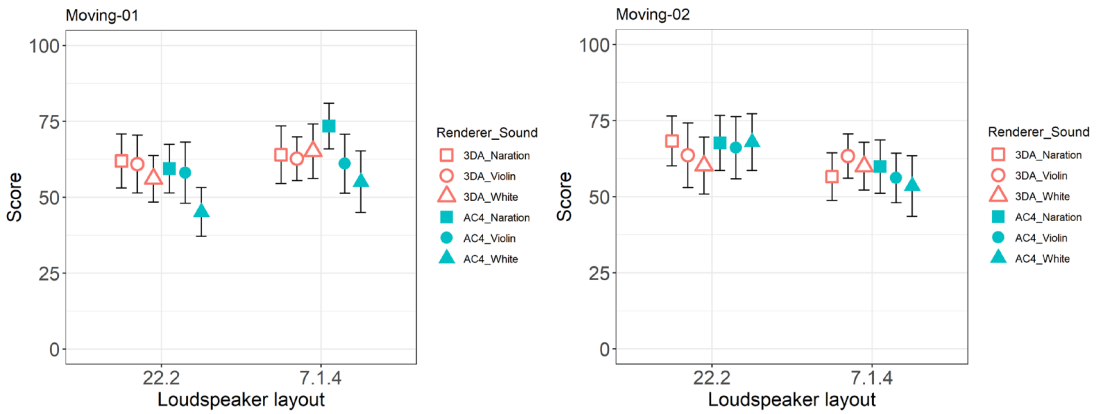
表 7 想定される放送サービスに必要なビットレート

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%))

4.6 品質とレンダリング機能

4.6.1 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。



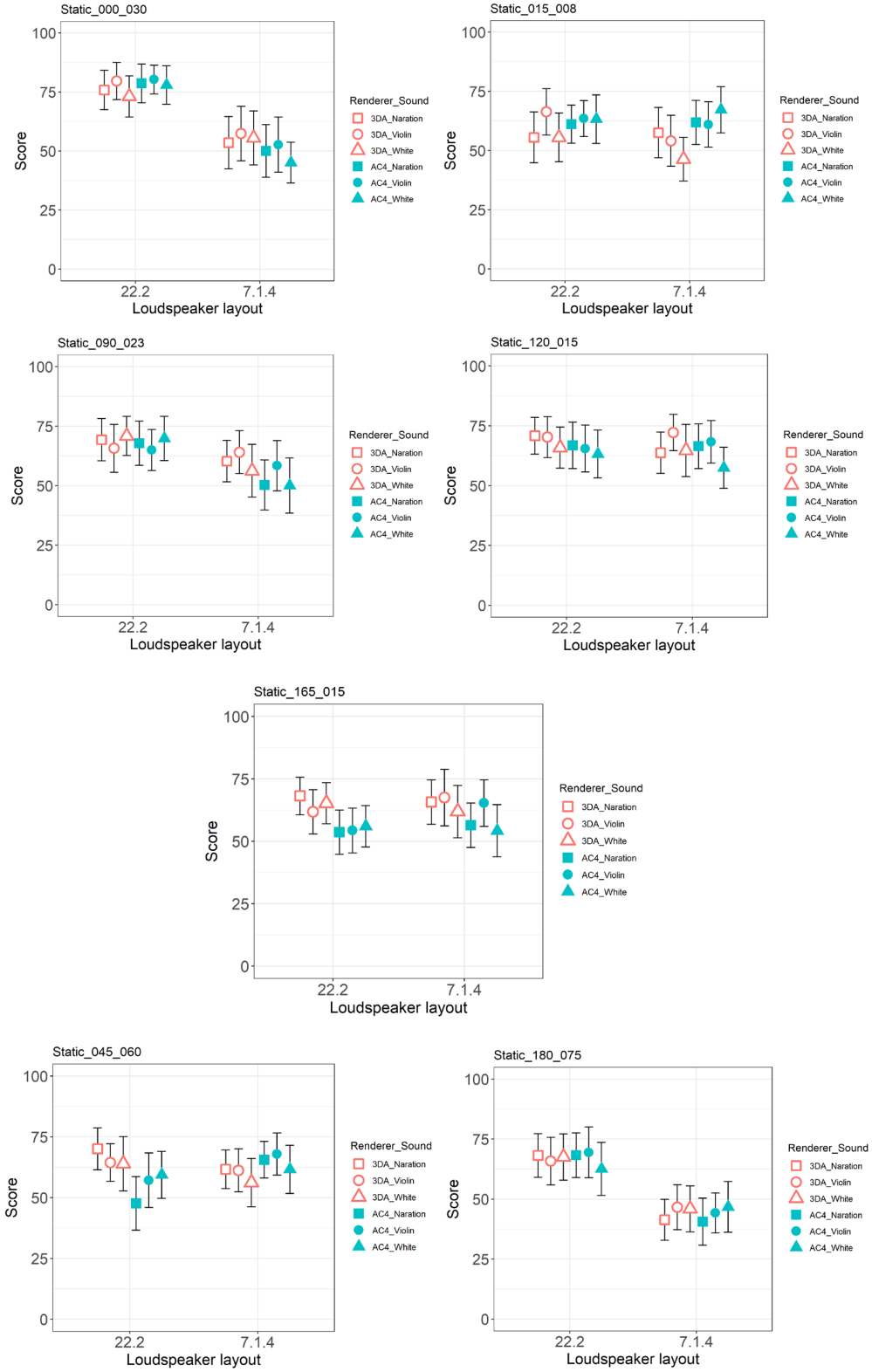


図 312 パンニング則の実験結果（上段：動き①，②、下段：静止 7 種類）

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。

4.6.2 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 の主チャンネル 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。

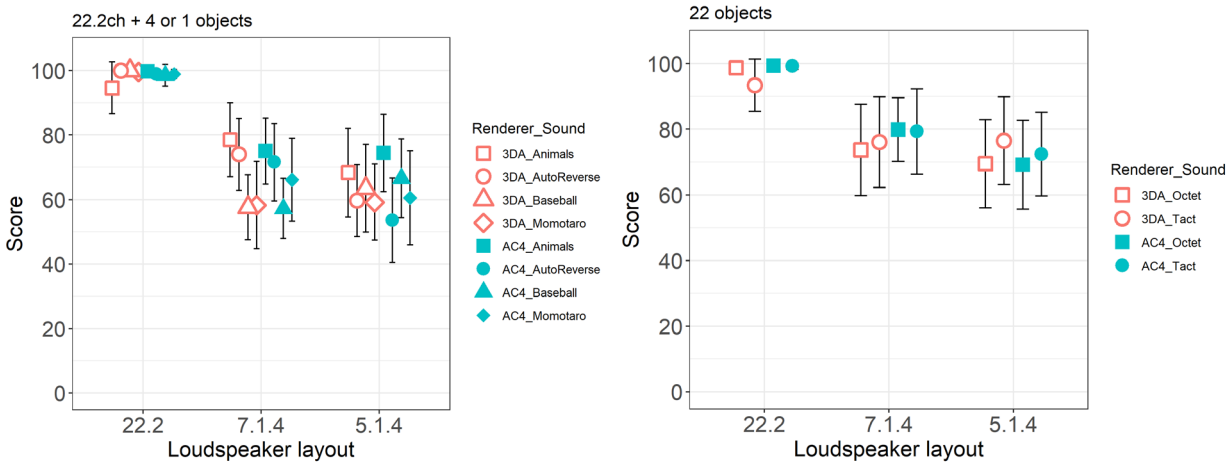


図 32 再生環境適応の実験結果

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。

第5章　まとめ

総務省からの諮問第 2044 号「放送システムに関する技術的条件」を受け、音声符号化方式の高度化の選定に必要な比較検討を実施した。総務省周波数逼迫対策技術試験事務において動作検証を実施している音声符号化方式 MPEG-H 3D Audio に加え、放送システム委員会で実施した「次世代地上デジタルテレビジョン方式に関する技術の提案募集」に提案のあった AC- 4 及び Enhanced AC-3 を対象として、現在国内の音声符号化方式として採用されている MPEG-2/4 AAC をレファレンスとして比較調査を進めた。MPEG-2/4 AAC 以外の 3 方式についてはいずれもオブジェクトベース音響（以下、OBA）を利用したサービスに対応しており、音声信号の符号化・復号化処理（コアコーダ）と復号した信号のポスト処理（レンダラー）により様々な音声サービスの提供が可能な方式である。また、今回の比較検討では、各音声符号化方式の放送品質を満たす所要ビットレートを求めるための主観評価実験とともに、OBA 特有の機能であるレンダラーの品質を比較するための主観評価実験を合わせて実施した。各音声符号化の方式間の違いや、各国での採用実績、各主観評価実験の詳細については前章までにまとめているが、地上放送高度化の音声符号化方式として MPEG-H 3 DA または AC-4 が候補となる。

MPEG-H 3 DA、AC-4 ともに ATSC をはじめとする放送方式として採用されており、北米や韓国で放送サービスが開始されているとともに受信機がすでに実用化され販売されている状況である。SoC（System on Chip）ベンダーからデコーダが供給されており、どちらの方式でも受信機の実装・普及に問題はないと思われる。

機能面については MPEG-H 3 DA、AC-4 ともにコア符号化とレンダラーで構成されており、オブジェクト符号化による基本的な音声サービス（オブジェクト差替/明瞭化/音量調整）に対応している。AC-4 は 22.2ch に報告時では正式対応していないが、実施した評価実験にプロトタイプを提供している。また、レンダリング機能では、位置情報の記述法に極座標（MPEG-H 3 DA）と直交座標（AC-4）と違いがあるほか、スピーカ配置を変換する場合にも、聴取者からみた角度が保持されるようにレンダリング（MPEG-H 3DA）する方式に対し、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（AC-4）と違いがある。これは、端的に説明すると絶対的な音声の位置を保持して再生するか、できるだけスピーカ単体で再生するかの違いになる。これらの方式の相互変換に相当な工夫が必要となることが実験を通じて明確になったことから、一方の符号化方式を選択することが望ましいと考える。

コアコーダ品質とビットレートの主観評価結果では、一部の音源やビットレートで一方の符号化方式が品質に有意な差があるものが見られた。また、レンダリング機能の主観評価結果では、符号化方式の差異があるとは言えないが、レンダリングの考え方の違いによる印象の違いがみられた。

別紙１　音声符号化方式の品質比較のための主観評価実験

1. はじめに

近年、欧米やアジア諸国においてオブジェクトベース音響（Object-based Audio; OBA）に対応した音声符号化方式である MPEG-H 3DA や AC-4 が次世代放送方式に採用されている。これまでに各音声符号化方式の所要ビットレートが報告されているが[1-5]、同じ音源、同じ実験条件で MPEG-H 3DA と AC-4 の品質を確認した例はない。そこで、本章では、地上放送高度化で検討されている音声符号化方式 4 方式（MPEG-4 AAC, MPEG-H 3DA, AC-4, Enhanced AC-3）の符号化音の品質を同じ音源、同じ実験条件で主観評価実験により確認する。

オブジェクトベース音響では、ダイアログ制御などの視聴者による音声オブジェクトごとのレベルバランスの調節が可能となる。通常は複数の音声オブジェクトが同時に再生されるとしても、ユーザ制御による調節値や音声オブジェクトの組合せによっては、音声オブジェクト単体で聴取される可能性がある。そこで、本章では、各音声オブジェクトを構成する音声フォーマットごとに所要ビットレートを確認し、その加算和でオブジェクトベース音響による音声放送サービスの所要ビットレートを求める。ここで、多言語放送のダイアログなどの i 個のモノオブジェクトと、環境音や音楽などの k_x 個の N_x チャンネルの背景音オブジェクト（チャンネル数は制作環境に依存）から構成される場合、オブジェクトベース音響の所要ビットレートは、モノオブジェクトと背景音オブジェクトの所要ビットレートにオブジェクト数を乗じたものの加算和として、

OBA の所要ビットレート ＝ （モノオブジェクトの所要ビットレート）$\times i +$
（N_1チャンネルの背景音の所要ビットレート）$\times k_1 +$
（N_2チャンネルの背景音の所要ビットレート）$\times k_2 +$
・・・　＋（音響メタデータの所要ビットレート）

と算出される。

2. 実験方法

勧告 ITU-R BS.1116-3[6]に規定される隠れ基準付き三刺激二重盲検法により、音声フォーマットごとに音声符号化・復号音の音質を主観評価する。

2.1. 評価音源

実験に用いる音声フォーマットは、22.2 マルチチャンネル音響（System H）、7.1.4（System J）、5.1ch サラウンド（System B）、ステレオ（System A）、およびモノ（Mono）の 5 種類とし、各音声フォーマットのスピーカ配置は勧告 ITU-R BS.2051-2[7]に規定されるスピーカ位置とした。

実験に用いる音源は、音源長 15－20 [s]程度、サンプリング周波数 48 [kHz]、量子化ビット数 24 [bits] の PCM 音源とし、音声フォーマットごとに 4 音源とした（Table 1）。評価音源は、音声符号化方式 4 方式（MPEG-4 AAC、MPEG-H 3DA、AC-4、Enhanced AC-3）にて圧縮符号化・復号したものを評価音源とした。ビットレートについては、先行研究[1,2,4,5,8]を参考に、放送品質を担保する（勧告 ITU-R BS.1116 に準拠した実験で、評点 4.0 以上）値を高ビットレートとし、中、低ビ

ットレートの 3 段階を用いた（Table 2)。但し、先行研究より MPEG-4 AAC と Enhanced AC-3 は AC-4 や MPEG-H 3DA の高ビットレートよりも明らかに所要ビットレートが高いため高ビットレートのみ、MPEG-4 AAC は System J に対応していないため除外、AC-4 は符号化時のエレメント数によって Level 4（エレメント数 24)、Level 3（エレメント数 16）の 2 方式で System H を符号化できるため 2 条件で評価音源を作成した。その結果、System H の評価音源は 44 個、System J は 28 個、それ以外は 32 個となった（Table 2)。また、モノの評価音源は、静的なメタデータを有するオブジェクトベース音響の音声信号として符号化した。

Table 1　実験に使用した音源

音声フォーマット	使用音源
22.2 マルチチャンネル音響 （System H）	Applause（自然音)、Tenpura（自然音)、Etenraku（音楽)、MyKIngdom（ドラマ）
7.1.4 （System J）	Ave Maria（音楽)、Choir at side（音楽)、Unfold（自然音)、Water burst（自然音）
5.1ch サラウンド （System B）	Atami fireworks festival（自然音)、Applause（自然音)、Jazzquartet（音楽)、Glasses（自然音）
ステレオ （System A）	Glockenspiel（単楽器)、Feste Romane - La Befana（音楽)、German Narration – male（音声)、Harpsichord（単楽器）
モノ （Mono）	Wind chime（単楽器)、Shamisen（単楽器)、Glockenspiel（単楽器)、English Narration - female（音声）

Table 2　音声符号化方式とビットレート

音声フォーマット	System H			System J			System B			System A			Mono		
ビットレート [kbps]	384	512	768	192	256	384	96	128	192	48	64	96	32	48	64
MPEG-4 AAC	X	X	O	X	X	X	X	X	O	X	X	O	X	X	O
MPEG-H 3DA	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
AC-4	O※	O※	O※	O	O	O	O	O	O	O	O	O	O	O	O
Enhanced AC-3	X	X	O	X	X	O	X	X	O	X	X	O	X	X	O
音源数	44			28			32			32			32		

＜凡例＞O：使用条件、X：未使用条件

※ AC-4 は、System H のみ Level3/Level4 の 2 種類の方式で符号化した

2.2.　実験設備

日本放送協会放送技術研究所・音響評価室、東京藝術大学千住キャンパス・スタジオ B の 2 箇所の実験設備を使用した。評価者は正面を向いて評価するものとし、聴取位置での平均音圧レベルは 75dB（A 特性）程度とした。各評価室の諸元を Table 3 に示す。

Table 3　本実験で使用した評価室

評価室	日本放送協会 放送技術研究所・音響評価室（勧告 ITU-R BS.1116-3 準拠）	東京藝術大学　千住キャンパス・スタジオ B（勧告 ITU-R BS.1116-3 にほぼ準拠）
容積	6.4m(W)×8.0m(D)×4.5m(H)	6.8m(W)×6.6m(D)×4.5m(H)
スピーカ	Fostex 社製 3-way スピーカ 24 個(勧告 ITU-R BS.1116-3 準拠)	KS-digital 社製 同軸 2-way スピーカ 22 個（勧告 ITU-R BS.1116-3 準拠）
スピーカ配置	半径 2.8m の円柱上に配置	半径 2.7m の球上に配置
	上層：高さ 2.9m（仰角 31 度）	上層：高さ 2.8m（仰角 30 度）
	28 度	28 度
	中層：高さ 1.4m（仰角 0 度）	中層：高さ 1.5m（仰角 0 度）
	下層：高さ 0m（俯角 29 度）	下層：高さ 0.25m（俯角 30 度）
	27 度	28 度
低域効果チャンネル用スピーカ	2 個を前方左右に開き角 75 度、中央から 2.75m の位置の床上に設置	2 個を前方左右に開き角 60 度、中央から 2.8m の位置の床上に設置

2.3. 実験手順

評価方法は、勧告 ITU-R BS.1116-3[6]に準拠した隠れ基準付き三刺激二重盲検法とし、音声フォーマットごとに音声符号化・復号音の音質を評価する。この方法では、評定者は、基準音 REF（非圧縮音）と 2 個の異なる評価音 A、B をそれぞれ聴取し、基準音に対して評価音 A、B それぞれの音質の違いの度合いを 5 段階劣化尺度（Table 4）に沿って回答する。

Table 4　5 段階劣化尺度

尺度	評点
基準音との違いを検知できない	5.0
基準音との違いを検出できるが気にならない	4.0
基準音との違いがやや気になる	3.0
基準音との違いが気になる	2.0
基準音との違いが非常に気になる	1.0

評価音 A、B には、一方に基準音 REF と同じ非圧縮音（隠れ基準音）、もう一方に圧縮符号化・復号音がランダムに提示される。評定者は評価音 A、B のいずれかに 5.0 点（違いを検知できない）、もう一方に 4.9 点以下を 0.1 点単位で採点する。評定者は、基準音 REF と評価音 A、B を好きなポイントで自由に切り替え、何度でも聴取することが可能である。

実験では音声フォーマットや符号化音の提示順による影響を除くため、5 種類の音声フォーマット、音声符号化方式およびビットレート、評価音 A、B の提示順をランダムにした（Figure 1）。

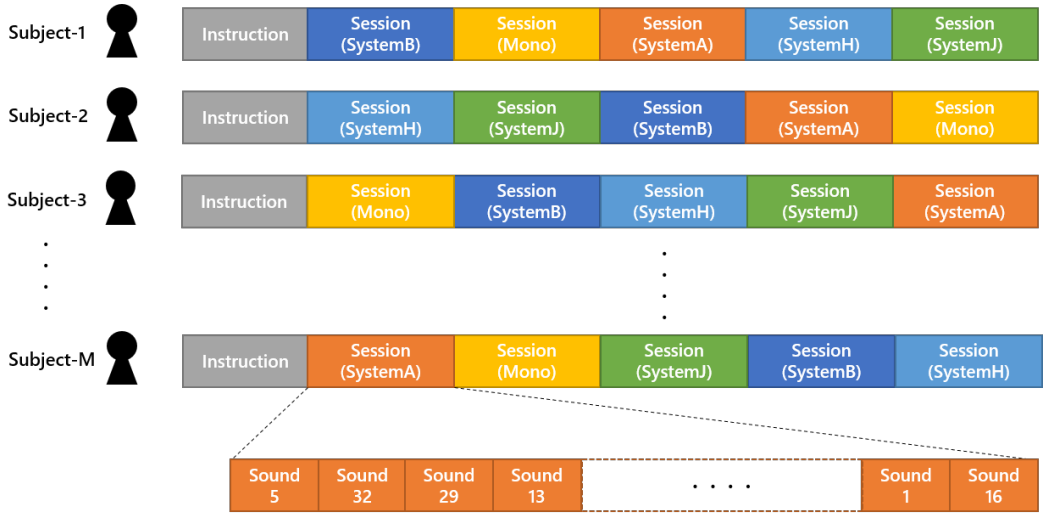


Figure 1　音声フォーマット・評価音源のランダム提示のイメージ

評価者は 20 歳代から 40 歳代までの正常な聴力を有する男女 34 名とした。評定者は、実験前に主観評価用ソフトウェアの操作方法に関する教示を受け、Table 1 に示す音源とは別の音源を圧縮符号化した音を用いて評定作業のトレーニングを受けた。

3.　実験結果

3.1.　各音源の差分評価値

各評価音源の実験結果を Figure 2 - Figure 22 に示す。図の各点は差分評価値の平均を、エラーバーは 95%信頼区間を示す。差分評価値は、隠れ基準音の評点から符号化音の評点の差を取った値である。この値は、正しく隠れ基準音より符号化音が低く評点された場合は負の値をとり、誤って隠れ基準音より符号化音が高く評点された場合は正の値をとる。また、隠れ基準音に対して 4.0 点以下を 10%以上採点した 12 名の評定者、符号化音に対して 4.0 点以下に 10%以上採点できなかった 3 名の評定者については除外し、計 19 名の評定者の結果を集計の対象とした。

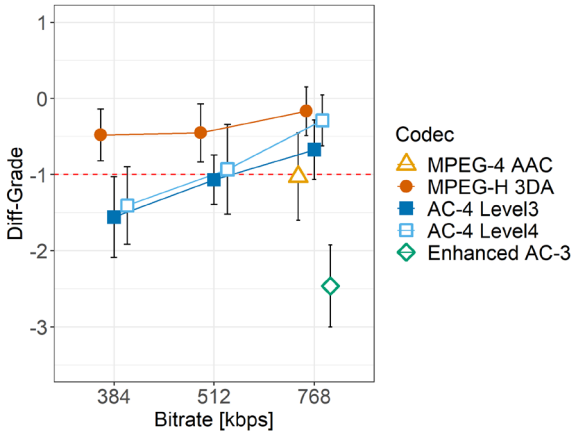


Figure 2　評価結果
（22.2 マルチチャンネル音響、Applause）

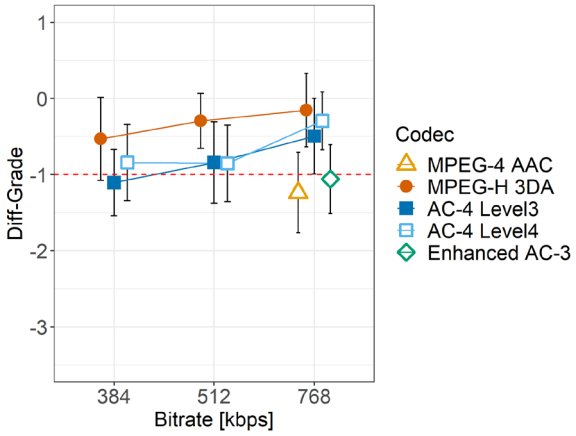


Figure 3　評価結果
（22.2 マルチチャンネル音響、Tempura）

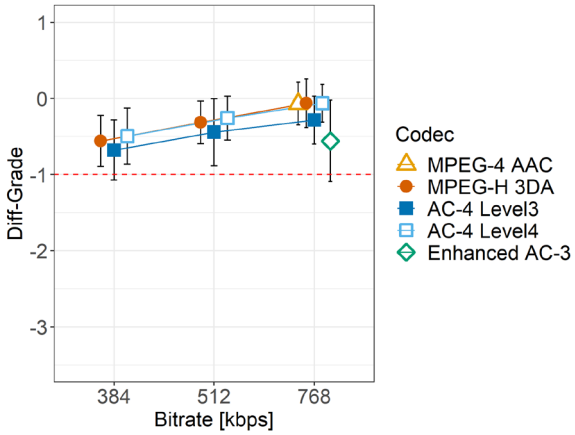


Figure 4　評価結果
（22.2 マルチチャンネル音響、Etenraku）

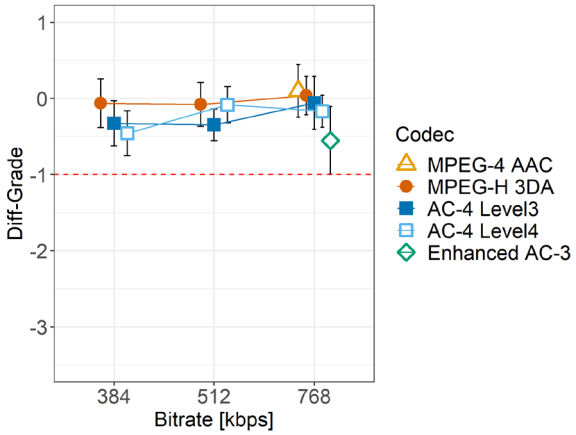


Figure 5　評価結果
（22.2 マルチチャンネル音響、MyKIngdom）

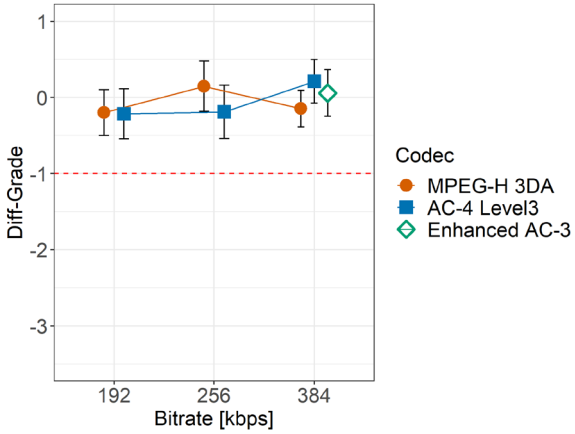


Figure 6　評価結果（7.1.4、Ave Maria）

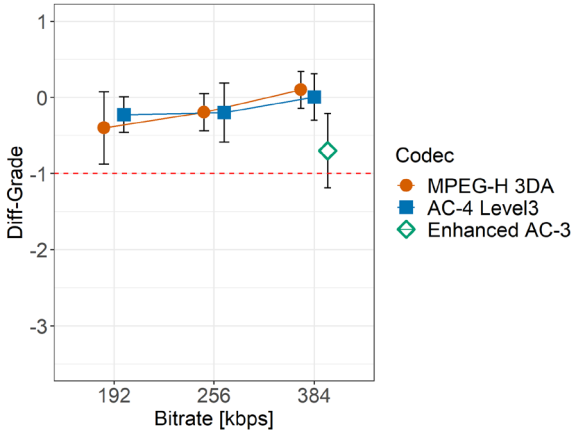


Figure 7　評価結果（7.1.4、Choir at side）

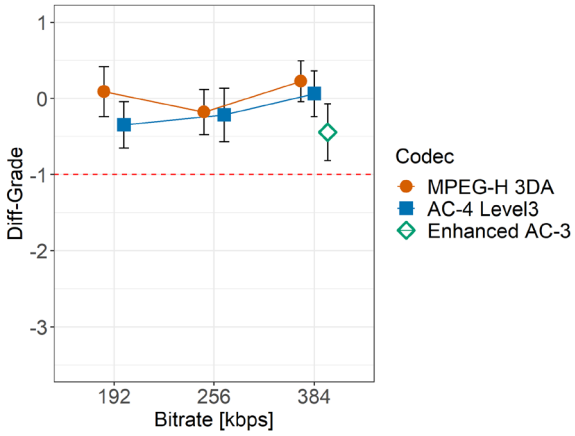


Figure 8　評価結果（7.1.4、Unfold）

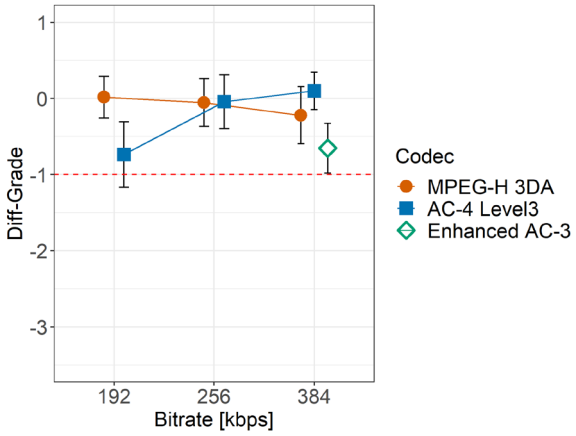


Figure 9　評価結果（7.1.4、Water Burst）

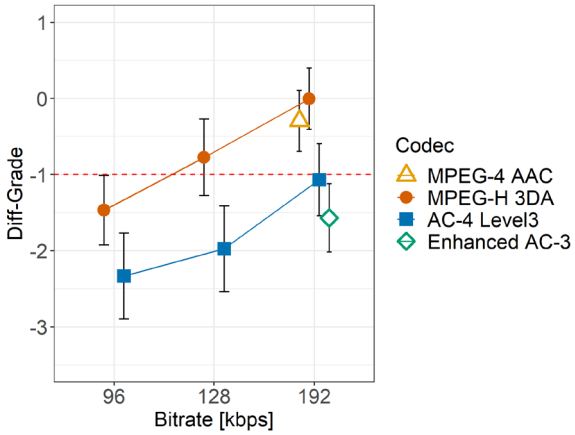


Figure 10　評価結果（5.1ch サラウンド、Atami fireworks festival）

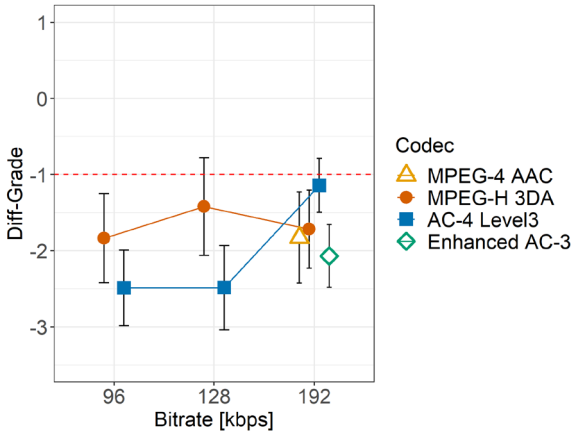


Figure 11　評価結果（5.1ch サラウンド、Applause）

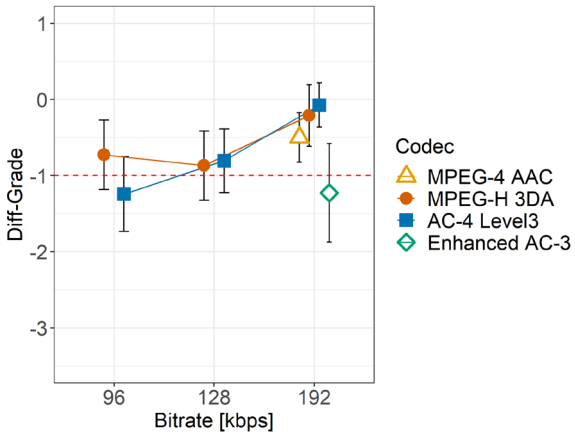


Figure 12　評価結果（5.1ch サラウンド、Jazzquartet）

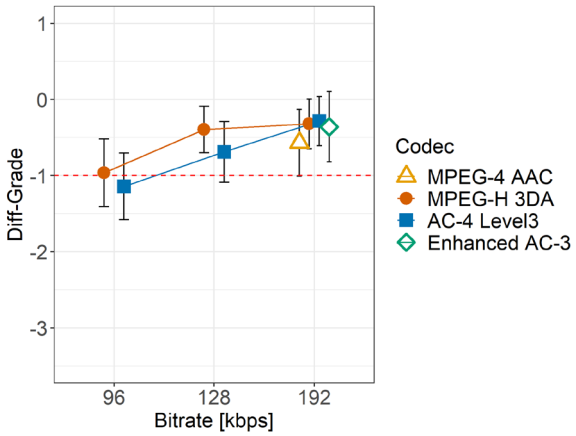


Figure 13　評価結果（5.1ch サラウンド、Glasses）

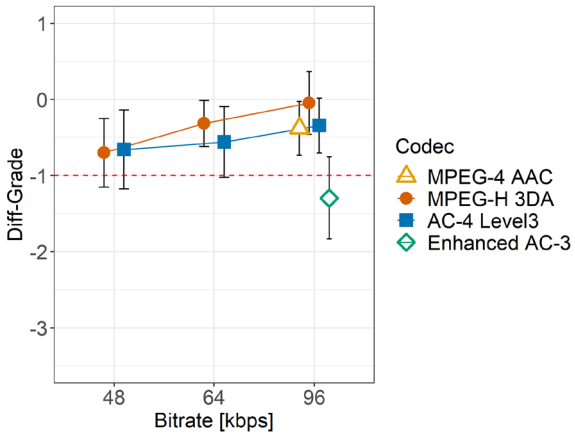


Figure 14　評価結果（ステレオ、Glockenspiel）

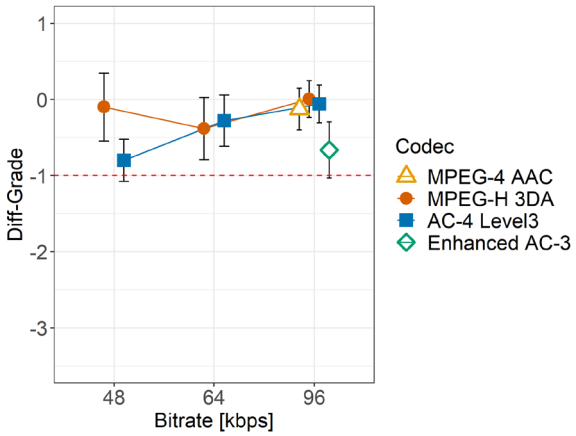


Figure 15　評価結果（ステレオ、Feste Romane - La Befana）

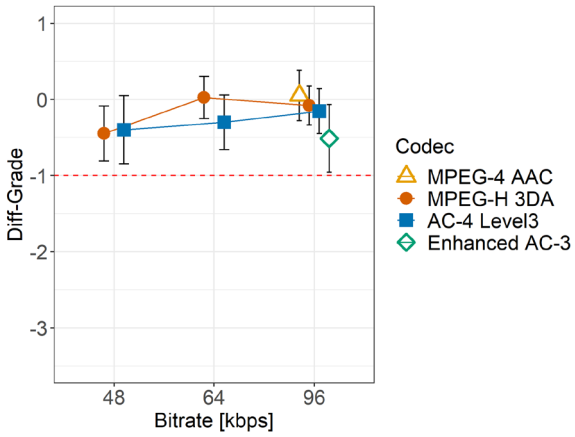


Figure 16　評価結果（ステレオ、German Narration – male）

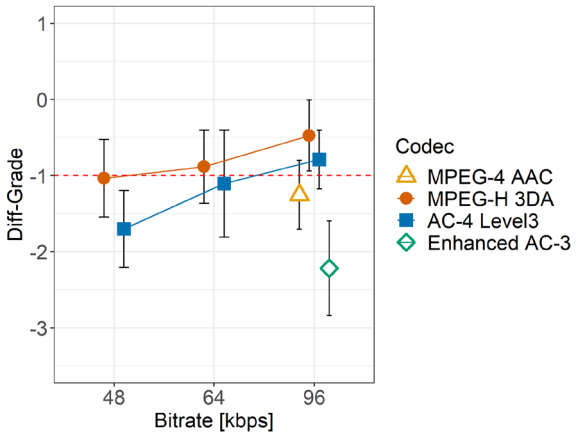


Figure 17　評価結果（ステレオ、Harpsichord）

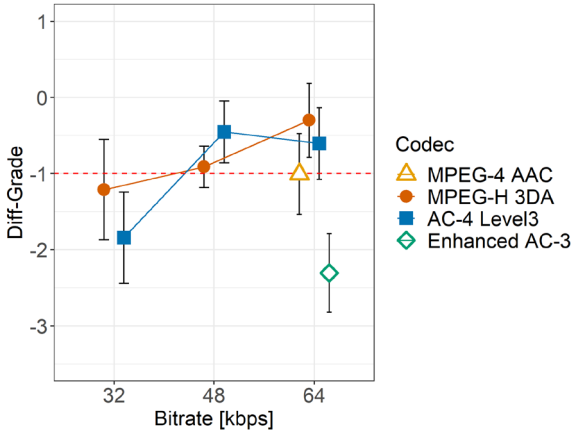


Figure 18　評価結果（モノ、Wind chime）

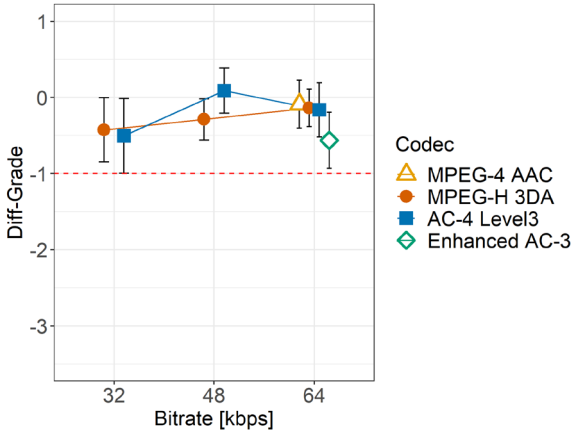


Figure 19　評価結果（モノ、Shamisen）

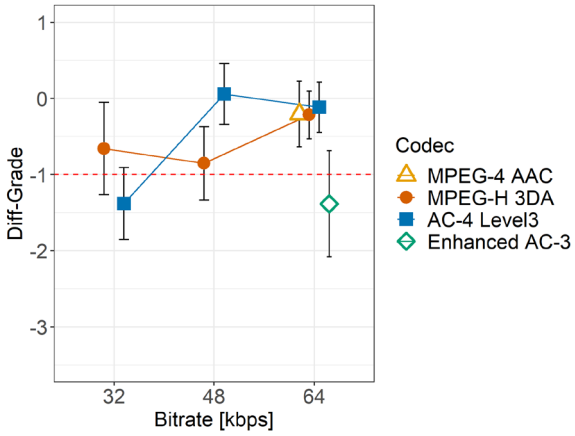


Figure 20　評価結果（モノ、Glockenspiel）

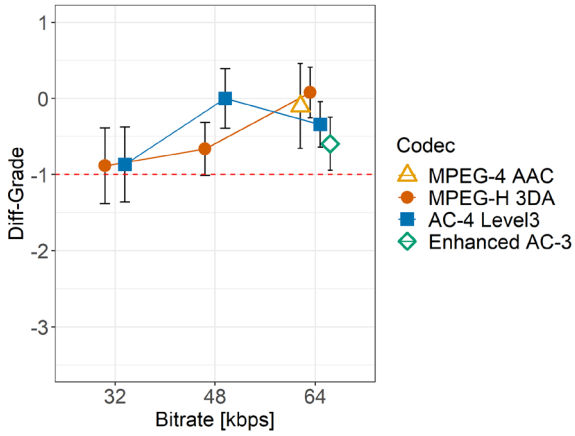


Figure 21　評価結果（モノ、English Narration - female）

放送品質を満たすビットレートは、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であるとすると Table 5 となる。

Table 5　MPEG-H 3DA と AC-4 の各音声フォーマットの所要ビットレート

音声符号化方式	音声フォーマット				
	System H	System J	System B	System A	Mono
MPEG-H 3DA	512 kbps	192 kbps	該当なし	96 kbps	64 kbps
AC-4	768 kbps	256 kbps	該当なし	該当なし	48 kbps

4.　考察

4.1.　各評価音源の音質の差

各音声符号化方式におけるビットレート間の平均値の違いを統計的に解析するために、帰無仮説を「各群（各符号化条件）の差分評価値の平均間に差がない」として、Tukey 検定を行う。本検定により算出された P 値を Tables 6 - 25 に示す。検定における有意水準は 5%とし、帰無仮説が棄却された組は＊印を併記する。また MPEG-H 3DA と AC-4 の同一ビットレートの組の P 値を示す箇所は灰色でハイライトする。これより、MPEG-H 3DA と AC-4 の同一ビットレートについては、22.2 マルチチャンネル音響の音源“Applause”(384 kbps)や 7.1.4 の音源“Water Burst”(192 kbps)、5.1ch サラウンドの音源“Atami fireworks festival”(128 kbps, 192 kbps)、ステレオの音源“Feste Romane - La Befana”(48 kbps)において平均値に統計的に有意な差があるが、その他についてはいずれの音声フォーマットにおいても同一ビットレートの場合は両者の差分評価値の平均に統計的に有意な差はない。統計的有意差がある例はいずれも放送品質を満たすビットレートではなく、放送品質を満たすビットレートにおいては、MPEG-H 3DA と AC-4 の音質に差があるとは言えない。

一方、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（22.2 マルチチャンネル音響の音源“Tenpura”）、有意に評点が低かったのは Enhanced AC-3（22.2 マルチチャンネル音響の音源“Applause”、5.1ch サラウンドの音源“Atami fireworks festival”、ステレオの音源“Glockenspiel”、モノの音源“Wind chime”と“Glockenspiel”）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。

Table 6　Tukey 検定による P 値（22.2 マルチチャンネル音響、Applause）

		MPEG-H 3DA			AC-4 Level4			AC-4 Level3			Enhanced AC-3
		384 kbps	512 kbps	768 kbps	384 kbps	512 kbps	768 kbps	384 kbps	512 kbps	768 kbps	768 kbps
MPEG-4 AAC	768 kbps	0.780	0.726	0.159	0.976	1.000	0.358	0.809	1.000	0.986	0.000*
MPEG-H 3DA	384 kbps		1.000	0.995	0.090	0.922	1.000	0.020*	0.692	1.000	0.000*
	512 kbps			0.997	0.072	0.891	1.000	0.015*	0.633	1.000	0.000*
	768 kbps				0.003*	0.306	1.000	0.000*	0.113	0.854	0.000*
AC-4 Level4	384 kbps					0.898	0.014*	1.000	0.990	0.369	0.026*
	512 kbps						0.572	0.608	1.000	0.999	0.000*
	768 kbps							0.002*	0.277	0.974	0.000*
AC-4 Level3	384 kbps								0.877	0.129	0.108
	512 kbps									0.968	0.000*
	768 kbps										0.000*

※有意水準 5%としたときに統計的に有意である組の P 値に＊印を併記

※MPEG-H 3DA と AC-4 の同一ビットレートの組の P 値を示す箇所は灰色でハイライト

Table 7　Tukey 検定による P 値（22.2 マルチチャンネル音響、Tenpura）

		MPEG-H 3DA			AC-4 Level4			AC-4 Level3			Enhanced AC-3
		384 kbps	512 kbps	768 kbps	384 kbps	512 kbps	768 kbps	384 kbps	512 kbps	768 kbps	768 kbps
MPEG-4 AAC	768 kbps	0.511	0.121	0.035*	0.979	0.982	0.121	1.000	0.979	0.432	1.000
MPEG-H 3DA	384 kbps		1.000	0.984	0.997	0.996	1.000	0.787	0.997	1.000	0.864
	512 kbps			1.000	0.832	0.815	1.000	0.298	0.832	1.000	0.388
	768 kbps				0.546	0.523	1.000	0.111	0.546	0.993	0.159
AC-4 Level4	384 kbps					1.000	0.832	0.999	1.000	0.992	1.000
	512 kbps						0.815	0.999	1.000	0.990	1.000
	768 kbps							0.298	0.832	1.000	0.388

AC-4 Level3	384 kbps								0.999	0.717	1.000
	512 kbps									0.992	1.000
	768 kbps										0.806

※有意水準 5%としたときに統計的に有意である組の P 値に＊印を併記

※MPEG-H 3DA と AC-4 の同一ビットレートの組の P 値を示す箇所は灰色でハイライト

Table 8　Tukey 検定による P 値（22.2 マルチチャンネル音響、Etenraku）

		MPEG-H 3DA			AC-4 Level4			AC-4 Level3			Enhanced AC-3
		384 kbps	512 kbps	768 kbps	384 kbps	512 kbps	768 kbps	384 kbps	512 kbps	768 kbps	768 kbps
MPEG-4 AAC	768 kbps	0.614	0.994	1.000	0.787	0.999	1.000	0.279	0.895	0.998	0.614
MPEG-H 3DA	384 kbps		0.995	0.599	1.000	0.975	0.599	1.000	1.000	0.987	1.000
	512 kbps			0.993	1.000	1.000	0.993	0.911	1.000	1.000	0.995
	768 kbps				0.774	0.999	1.000	0.267	0.886	0.998	0.599
AC-4 Level4	384 kbps					0.996	0.774	1.000	1.000	0.998	1.000
	512 kbps						0.999	0.800	1.000	1.000	0.975
	768 kbps							0.267	0.886	0.998	0.599
AC-4 Level3	384 kbps								0.996	0.857	1.000
	512 kbps									1.000	1.000
	768 kbps										0.987

※有意水準 5%としたときに統計的に有意である組の P 値に＊印を併記

※MPEG-H 3DA と AC-4 の同一ビットレートの組の P 値を示す箇所は灰色でハイライト

Table 9　Tukey 検定による P 値（22.2 マルチチャンネル音響、MyKIngdom）

		MPEG-H 3DA			AC-4 Level4			AC-4 Level3			Enhanced AC-3
		384 kbps	512 kbps	768 kbps	384 kbps	512 kbps	768 kbps	384 kbps	512 kbps	768 kbps	768 kbps
MPEG-4 AAC	768 kbps	0.999	0.998	1.000	0.188	0.998	0.964	0.580	0.506	0.999	0.057
MPEG-H 3DA	384 kbps		1.000	1.000	0.688	1.000	1.000	0.969	0.947	1.000	0.366
	512 kbps			1.000	0.739	1.000	1.000	0.980	0.964	1.000	0.417
	768 kbps				0.350	1.000	0.995	0.786	0.722	1.000	0.130
AC-4 Level4	384 kbps					0.755	0.941	1.000	1.000	0.671	1.000
	512 kbps						1.000	0.983	0.969	1.000	0.434
	768 kbps							0.999	0.998	1.000	0.722

AC-4 Level3	384 kbps								1.000	0.964	0.990
	512 kbps									0.941	0.995
	768 kbps										0.350

※有意水準 5%としたときに統計的に有意である組の P 値に＊印を併記

※MPEG-H 3DA と AC-4 の同一ビットレートの組の P 値を示す箇所は灰色でハイライト

Table 10　Tukey 検定による P 値（7.1.4、Ave Maria）

		MPEG-H 3DA		AC-4			Enhanced AC-3
		256 kbps	384 kbps	192 kbps	256 kbps	384 kbps	384 kbps
MPEG-H 3DA	192 kbps	0.634	1.000	1.000	1.000	0.433	0.876
	256 kbps		0.790	0.584	0.667	1.000	0.999
	384 kbps			1.000	1.000	0.600	0.955
AC-4	192 kbps				1.000	0.386	0.842
	256 kbps					0.466	0.896
	384 kbps						0.990

※有意水準 5%としたときに統計的に有意である組の P 値に＊印を併記

※MPEG-H 3DA と AC-4 の同一ビットレートの組の P 値を示す箇所は灰色でハイライト

Table 11　Tukey 検定による P 値（7.1.4、Choir at side）

		MPEG-H 3DA		AC-4			Enhanced AC-3
		256 kbps	384 kbps	192 kbps	256 kbps	384 kbps	384 kbps
MPEG-H 3DA	192 kbps	0.978	0.363	0.991	0.980	0.619	0.870
	256 kbps		0.879	1.000	1.000	0.980	0.350
	384 kbps			0.818	0.870	1.000	0.018*
AC-4	192 kbps				1.000	0.960	0.430
	256 kbps					0.978	0.363
	384 kbps						0.056

※有意水準 5%としたときに統計的に有意である組の P 値に＊印を併記

※MPEG-H 3DA と AC-4 の同一ビットレートの組の P 値を示す箇所は灰色でハイライト

Table 12　Tukey 検定による P 値（7.1.4、Unfold）

	MPEG-H 3DA		AC-4			Enhanced AC-3
	256 kbps	384 kbps	192 kbps	256 kbps	384 kbps	384 kbps

MPEG-H 3DA	192 kbps	0.873	0.995	0.398	0.789	1.000	0.177
	256 kbps		0.492	0.986	1.000	0.918	0.883
	384 kbps			0.114	0.383	0.988	0.036*
AC-4	192 kbps				0.996	0.476	0.999
	256 kbps					0.851	0.940
	384 kbps						0.227

※有意水準 5%としたときに統計的に有意である組の P 値に＊印を併記

※MPEG-H 3DA と AC-4 の同一ビットレートの組の P 値を示す箇所は灰色でハイライト

Table 13　Tukey 検定による P 値（7.1.4、Water burst）

		MPEG-H 3DA		AC-4			Enhanced AC-3
		256 kbps	384 kbps	192 kbps	256 kbps	384 kbps	384 kbps
MPEG-H 3DA	192 kbps	1.000	0.942	0.019*	1.000	1.000	0.055
	256 kbps		0.989	0.046*	1.000	0.994	0.119
	384 kbps			0.261	0.985	0.790	0.478
AC-4	192 kbps				0.040*	0.006*	1.000
	256 kbps					0.996	0.106
	384 kbps						0.019*

※有意水準 5%としたときに統計的に有意である組の P 値に＊印を併記

※MPEG-H 3DA と AC-4 の同一ビットレートの組の P 値を示す箇所は灰色でハイライト

Table 14　Tukey 検定による P 値（5.1ch サラウンド、Atami fireworks festival）

Atami fireworks festival		MPEG-H 3DA			AC-4			Enhanced AC-3
		96 kbps	128 kbps	192 kbps	96 kbps	128 kbps	192 kbps	192 kbps
MPEG-4 AAC	192 kbps	0.009*	0.815	0.986	0.000*	0.000*	0.252	0.003*
MPEG-H 3DA	96 kbps		0.388	0.000*	0.140	0.771	0.919	1.000
	128 kbps			0.260	0.000*	0.007*	0.984	0.221
	192 kbps				0.000*	0.000*	0.027*	0.000*
AC-4	96 kbps					0.954	0.003*	0.268
	128 kbps						0.102	0.914
	192 kbps							0.780

※有意水準 5%としたときに統計的に有意である組の P 値に＊印を併記

※MPEG-H 3DA と AC-4 の同一ビットレートの組の P 値を示す箇所は灰色でハイライト

Table 15　Tukey 検定による P 値（5.1ch サラウンド、Applause）

		MPEG-H 3DA			AC-4			Enhanced AC-3
		96 kbps	128 kbps	192 kbps	96 kbps	128 kbps	192 kbps	192 kbps
MPEG-4 AAC	192 kbps	1.000	0.946	1.000	0.575	0.585	0.534	0.997
MPEG-H 3DA	96 kbps		0.939	1.000	0.595	0.605	0.514	0.998
	128 kbps			0.991	0.060	0.062	0.994	0.605
	192 kbps				0.370	0.379	0.740	0.975
AC-4	96 kbps					1.000	0.005*	0.935
	128 kbps						0.005*	0.939
	192 kbps							0.161

※有意水準 5%としたときに統計的に有意である組の P 値に＊印を併記

※MPEG-H 3DA と AC-4 の同一ビットレートの組の P 値を示す箇所は灰色でハイライト

Table 16　Tukey 検定による P 値（5.1ch サラウンド、Jazzquartet）

		MPEG-H 3DA			AC-4			Enhanced AC-3
		96 kbps	128 kbps	192 kbps	96 kbps	128 kbps	192 kbps	192 kbps
MPEG-4 AAC	192 kbps	0.995	0.924	0.979	0.220	0.972	0.849	0.244
MPEG-H 3DA	96 kbps		1.000	0.679	0.679	1.000	0.378	0.713
	128 kbps			0.367	0.918	1.000	0.151	0.934
	192 kbps				0.018*	0.503	1.000	0.021*
AC-4	96 kbps					0.832	0.004*	1.000
	128 kbps						0.236	0.857
	192 kbps							0.005*

※有意水準 5%としたときに統計的に有意である組の P 値に＊印を併記

※MPEG-H 3DA と AC-4 の同一ビットレートの組の P 値を示す箇所は灰色でハイライト

Table 17　Tukey 検定による P 値（5.1ch サラウンド、Glasses）

		MPEG-H 3DA			AC-4			Enhanced AC-3
		96 kbps	128 kbps	192 kbps	96 kbps	128 kbps	192 kbps	192 kbps
MPEG-4 AAC	192 kbps	0.817	0.998	0.983	0.388	1.000	0.963	0.993

MPEG-H 3DA	96 kbps		0.400	0.246	0.998	0.970	0.185	0.318
	128 kbps			1.000	0.103	0.955	1.000	1.000
	192 kbps				0.050*	0.865	1.000	1.000
AC-4	96 kbps					0.689	0.034*	0.072
	128 kbps						0.796	0.918
	192 kbps							1.000

※有意水準 5%としたときに統計的に有意である組の P 値に＊印を併記

※MPEG-H 3DA と AC-4 の同一ビットレートの組の P 値を示す箇所は灰色でハイライト

Table 18　Tukey 検定による P 値（ステレオ、Glockenspiel）

Glockenspiel		MPEG-H 3DA			AC-4			Enhanced AC-3
		48 kbps	64 kbps	96 kbps	48 kbps	64 kbps	96 kbps	96 kbps
MPEG-4 AAC	96 kbps	0.955	1.000	0.947	0.979	0.999	1.000	0.041*
MPEG-H 3DA	48 kbps		0.890	0.333	1.000	1.000	0.922	0.456
	64 kbps			0.983	0.938	0.991	1.000	0.021*
	96 kbps				0.421	0.652	0.972	0.001*
AC-4	48 kbps					1.000	0.959	0.365
	64 kbps						0.996	0.191
	96 kbps							0.028*

※有意水準 5%としたときに統計的に有意である組の P 値に＊印を併記

※MPEG-H 3DA と AC-4 の同一ビットレートの組の P 値を示す箇所は灰色でハイライト

Table 19　Tukey 検定による P 値（ステレオ、Feste Romane - La Befana）

		MPEG-H 3DA			AC-4			Enhanced AC-3
		48 kbps	64 kbps	96 kbps	48 kbps	64 kbps	96 kbps	96 kbps
MPEG-4 AAC	96 kbps	1.000	0.943	0.999	0.059	0.997	1.000	0.248
MPEG-H 3DA	48 kbps		0.908	1.000	0.043*	0.993	1.000	0.195
	64 kbps			0.659	0.580	1.000	0.827	0.916
	96 kbps				0.010*	0.908	1.000	0.063
AC-4	48 kbps					0.284	0.024*	0.999
	64 kbps						0.975	0.675
	96 kbps							0.128

※有意水準 5%としたときに統計的に有意である組の P 値に＊印を併記
※MPEG-H 3DA と AC-4 の同一ビットレートの組の P 値を示す箇所は灰色でハイライト

Table 20　Tukey 検定による P 値（ステレオ、German Narration - male）

German Narration - male		MPEG-H 3DA			AC-4			Enhanced AC-3
		48 kbps	64 kbps	96 kbps	48 kbps	64 kbps	96 kbps	96 kbps
MPEG-4 AAC	96 kbps	0.416	1.000	0.999	0.548	0.814	0.989	0.263
MPEG-H 3DA	48 kbps		0.488	0.777	1.000	0.999	0.918	1.000
	64 kbps			1.000	0.624	0.867	0.995	0.322
	96 kbps				0.877	0.982	1.000	0.609
AC-4	48 kbps					1.000	0.967	1.000
	64 kbps						0.999	0.987
	96 kbps							0.802

※有意水準 5%としたときに統計的に有意である組の P 値に＊印を併記
※MPEG-H 3DA と AC-4 の同一ビットレートの組の P 値を示す箇所は灰色でハイライト

Table 21　Tukey 検定による P 値（ステレオ、Harpsichord）

		MPEG-H 3DA			AC-4			Enhanced AC-3
		48 kbps	64 kbps	96 kbps	48 kbps	64 kbps	96 kbps	96 kbps
MPEG-4 AAC	96 kbps	0.999	0.967	0.353	0.909	1.000	0.893	0.122
MPEG-H 3DA	48 kbps		1.000	0.752	0.567	1.000	0.997	0.023*
	64 kbps			0.941	0.294	0.998	1.000	0.006*
	96 kbps				0.015*	0.627	0.986	0.000*
AC-4	48 kbps					0.696	0.171	0.826
	64 kbps						0.986	0.041*
	96 kbps							0.002*

※有意水準 5%としたときに統計的に有意である組の P 値に＊印を併記
※MPEG-H 3DA と AC-4 の同一ビットレートの組の P 値を示す箇所は灰色でハイライト

Table 22　Tukey 検定による P 値（モノ、Wind chime）

	MPEG-H 3DA	AC-4	Enhanced AC-3

		32 kbps	48 kbps	64 kbps	32 kbps	48 kbps	64 kbps	64 kbps
MPEG-4 AAC	64 kbps	0.999	1.000	0.435	0.220	0.733	0.937	0.005*
MPEG-H 3DA	32 kbps		0.987	0.137	0.581	0.340	0.633	0.033*
	48 kbps			0.623	0.119	0.879	0.986	0.002*
	64 kbps				0.000*	1.000	0.986	0.000*
AC-4	32 kbps					0.002*	0.009*	0.872
	48 kbps						1.000	0.000*
	64 kbps							0.000*

※有意水準 5%としたときに統計的に有意である組の P 値に＊印を併記

※MPEG-H 3DA と AC-4 の同一ビットレートの組の P 値を示す箇所は灰色でハイライト

Table 23　Tukey 検定による P 値（モノ、Shamisen）

		MPEG-H 3DA			AC-4			Enhanced AC-3
		32 kbps	48 kbps	64 kbps	32 kbps	48 kbps	64 kbps	64 kbps
MPEG-4 AAC	64 kbps	0.848	0.990	1.000	0.655	0.995	1.000	0.490
MPEG-H 3DA	32 kbps		0.999	0.925	1.000	0.376	0.954	0.999
	48 kbps			0.998	0.985	0.753	0.999	0.944
	64 kbps				0.779	0.980	1.000	0.625
AC-4	32 kbps					0.203	0.837	1.000
	48 kbps						0.963	0.118
	64 kbps							0.698

※有意水準 5%としたときに統計的に有意である組の P 値に＊印を併記

※MPEG-H 3DA と AC-4 の同一ビットレートの組の P 値を示す箇所は灰色でハイライト

Table 24　Tukey 検定による P 値（モノ、Glockenspiel）

		MPEG-H 3DA			AC-4			Enhanced AC-3
		32 kbps	48 kbps	64 kbps	32 kbps	48 kbps	64 kbps	64 kbps
MPEG-4 AAC	64 kbps	0.859	0.491	1.000	0.010*	0.992	1.000	0.009*
MPEG-H 3DA	32 kbps		0.999	0.873	0.347	0.356	0.708	0.337
	48 kbps			0.513	0.738	0.103	0.319	0.728
	64 kbps				0.011*	0.990	1.000	0.010*
AC-4	32 kbps					0.000*	0.004*	1.000

	48 kbps						0.999	0.000*
	64 kbps							0.004*

※有意水準 5%としたときに統計的に有意である組の P 値に＊印を併記

※MPEG-H 3DA と AC-4 の同一ビットレートの組の P 値を示す箇所は灰色でハイライト

Table 25　Tukey 検定による P 値（モノ、English Narration - female）

English Narration - female		MPEG-H 3DA			AC-4			Enhanced AC-3
		32 kbps	48 kbps	64 kbps	32 kbps	48 kbps	64 kbps	64 kbps
MPEG-4 AAC	64 kbps	0.105	0.482	0.998	0.120	1.000	0.989	0.647
MPEG-H 3DA	32 kbps		0.994	0.018	1.000	0.041	0.532	0.969
	48 kbps			0.150	0.996	0.268	0.946	1.000
	64 kbps				0.021	1.000	0.807	0.250
AC-4	32 kbps					0.048	0.571	0.977
	48 kbps						0.925	0.409
	64 kbps							0.986

※有意水準 5%としたときに統計的に有意である組の P 値に＊印を併記

※MPEG-H 3DA と AC-4 の同一ビットレートの組の P 値を示す箇所は灰色でハイライト

4.2.　先行研究との関係

勧告 ITU-R BS.1548-4[8]に基づけば、Table 3 に示す 5 段階劣化尺度で評点が 4.0 以上（差分評価値で-1.0 以上）であれば、放送品質を満たすと言える。本実験では、MPEG-H 3DA の 7.1.4（System J）の所要ビットレートは 192 kbps であるが、5.1ch サラウンド（System B）は 192 kbps でも該当なしという結果になっており、チャンネル数が少ない 5.1ch サラウンドの所要ビットレートの方が高くなっている。また、準備した 4 音源のうち 3 音源でどのビットレートでも放送品質を満たすという結果に達している。一方、先行研究[1,2]においては、384 kbps となっており、本実験ではクリティカルな音源が準備できなかったものと考えられる。7.1.4（System J）については先行研究の結果を参照することが妥当と考えられる。本実験では、マルチチャンネル音響において MPEG-H 3DA の方が AC-4 よりも所要ビットレートが低くなった。先行研究[1,2]においては、7.1.4 の MPEG-H 3DA の所要ビットレートが 384 kbps であるのに対し、AC-4 は 288 kbps となっている。本実験では同一ビットレートでは MPEG-H 3DA と AC-4 の品質差は統計的に有意ではないため、MPEG-H 3DA の所要ビットレートの方が高い場合は、MPEG-H 3DA の所要ビットレートと同じかそれ以上とする方が妥当と思われる。

また、MPEG-4 AAC の 22.2 マルチチャンネル音響の音源 Applause や Tenpura の 768 kbps の差分評価値が-1.0 付近にあり、MPEG-H 3DA や AC-4 に対してあまり見劣りしないようにみえる。先

行研究[4]によると、Applause の 704 kbps の差分評価値は-1.5 から-2.0 の間にあり、本実験の 95%信頼区間の下限付近となる。本実験の評点は先行研究よりも若干高い点数が付けられていると考えられる。しかし、MPEG-4 AAC は 1 Mbps でも-1.0 点を超えており、オンライン処理を考慮して 1.2 Mbps となっていることから、現行放送の 1.4 Mbps という設定は安全を見越した数値と言える。MPEG-H 3DA の 22.2 マルチチャンネル音響の所要ビットレートは先行研究を加味して 768 kbps とするか、MPEG-4 AAC の所要ビットレートを 1 Mbps として比較することが妥当であろう。

4.3. 想定される放送サービスと所要ビットレート

先行研究[1-5]で示された所要ビットレートが今回の実験結果を上回る場合は先行研究の結果を用いるとすると、次世代音声符号化方式の所要ビットレートは Table 26 （上段）のように推定される。また、2 ヵ国語や 4 ヵ国語の多言語放送を想定したビットレートを下段に示す。

System H を放送する場合、現行放送の MPEG-4 AAC の 1.4 Mbps と比較して次世代音声符号化方式は 768 kbps と約半分程度のビットレートであるが、4 ヵ国語放送の場合、MPEG-4 AAC の 5.6 Mbps に対し、次世代音声符号化方式は 1,024 kbps と 1/5 以下になる。オブジェクトベース音響の機能を用いることでダイアログ制御などのユーザ制御が可能となるだけではなく、非常に高効率に圧縮できると言える。

Table 26　次世代音声符号化方式の各音声フォーマットの所要ビットレート

	音声フォーマット				
	System H	System J	System B	System A	Mono
単独	768 kbps	384 kbps	208 kbps※1	96 kbps	64 kbps
2 カ国語	896 kbps	512 kbps	336 kbps	224 kbps	-
4 ヵ国語	1,024 kbps	640 kbps	464 kbps	352 kbps	-

5. まとめ

本実験結果から、MPEG-H 3DA では 22.2 マルチチャンネル音響で 512 kbps、7.1.4 で 192 kbps、5.1ch サラウンドで該当なし、ステレオで 96 kbps、モノで 64 kbps となり、AC-4 では 22.2 マルチチャンネル音響で 768 kbps、7.1.4 で 256 kbps、5.1ch サラウンドで該当なし、ステレオで該当なし、モノで 48 kbps と推定された。マルチチャンネル音響では MPEG-H 3DA の方が、モノでは AC-4 の方が高効率と言える。しかし、放送品質を満たすビットレートにおいては同一ビットレートにおいて統計的に有意な差があるとは言えなかった。

現行の 4K/8K 衛星放送の音声符号化方式である MPEG-4 AAC と比較し、22.2ch マルチチャンネル音響において MPEG-H 3DA や AC-4 は 1/2 程度、高効率に圧縮しても同程度の音質であることが確認された。また、2 ヶ国語放送や裏トークなどのダイアログ差替えを想定し、4 個のモノオブジェクトと組み合わせる場合、22.2ch マルチチャンネル音響の所要ビットレートは MPEG-H 3DA や AC-4 で 1,024 kbps であり、現行 4K8K 衛星放送の音声符号化方式である MPEG-4 AAC で 4 ヵ国語

を放送する場合と比較し、1/5 程度、高効率に圧縮することが可能であると考えられる。オブジェクトベース音響を用いることで高効率に多言語放送などのサービスを実施することができる。

本実験では、復号音の音質差だけを評価したが、実際に視聴者が聴取する音はレンダリング後の音である。各音声符号化方式のレンダリング手法の違いによって視聴者が聴取する音声品質には差がある可能性があり、今後、各音声符号化方式のレンダラーの品質（音像の定位精度や明瞭度など）についても確認する必要がある。

6. 参考文献

[1] MPEG2017/N16584, “MPEG-H 3D Audio Verification Test Report,” 2017.

[2] MPEG2017/N19407, “MPEG-H 3D Audio Baseline Profile Verification Test Report,” 2020.

[3] ATSC, “ATSC 3.0 Audio Testing Report,” 2015.

[4] Sugimoto etc., “Bit rate of 22.2 multichannel sound signal meeting broadcast quality,” AES Convention Paper 9096, 2014

[5] Sugimoto etc., “Required bit rate of 22.2 multichannel audio signal compressed by MPEG-H 3D Audio to meet broadcast quality,” Acoust. Sci. & Tech. 39, 3, 2018.

[6] ITU, Recommendation ITU-R BS.1116-3, 2015-2.

[7] ITU, Recommendation ITU-R BS.2051-2, 2018

[8] ITU, Recommendation ITU-R BS.1548-X, 20XX

[9] EBU, “ETSI TS 103 190-1 V1.3.1,” 2018-2.

[10] ISO, “ISO/IEC 23008-3:2019,” 2019.

別紙２　レンダラー方式の品質比較のための主観評価実験

1. はじめに

オブジェクトベース音響では、視聴者は、再生環境に応じた音声フォーマットにレンダリングされたコンテンツを聴取することが可能となる。これにより、従来のチャンネルベース音響よりも低コストで個人適用したサービスが提供できることが期待される一方で、制作時とは異なるスピーカ配置にレンダリングされることで、信号加算やマスキング解除によって劣化が顕著になる可能性が懸念されている。また、レンダリング手法の違いにより、モノ音声オブジェクトの定位やスピーカ配置が異なる場合の空間印象に違いが生じる可能性があるが、地上放送高度化において検討されているMPEG-H 3DA、AC-4 のレンダリング品質を、同じ音源、同じ実験条件で確認した事例はない。これまでに ITU-R 等では、レンダリング技術性能を計る主観評価法の検討が継続してなされてきているが、主観評価手法の確立には至っていない。

本章では、今回挑戦的な取り組みとして独自に検討された、レンダリング方式２方式（MPEG-H 3DA、AC-4）によるレンダリング音の品質（空間印象や定位精度など）を確認するための主観評価手法、およびその主観評価実験結果について解説する。なお、今回の実験結果は、あくまで提案されたレンダリング技術性能の一側面を示しているものであり、これらでレンダリング技術の全体性能を必ずしも推し量れるわけではないことに留意すべきである。

2. 実験方法

本実験はパンニング則評価実験、および Egocentric/Allocentric 評価実験の２つを実施した。前者は、MPEG-H レンダラーと AC-4 レンダラーそれぞれのパンニング則、3D-VBAP とトリプルバランスパンナーによる音像定位精度を評価するものである。後者は、制作時と異なるスピーカ配置にレンダリングされた場合における、MPEG-H レンダラーの Egocentric 思想と AC-4 レンダラーの Allocentric 思想が総合印象（音像定位や包まれ感などの空間印象を含む）に与える影響を評価するものである。２つの実験では、共通の実験設備を使用し、評定者は 20 歳代から 50 歳代までの正常な聴力を有する男女 24 名とした。評定者は各実験に関する主観評価用ソフトウェアの操作方法に関する教示を受けてから評定作業を実施した。

2.1. 実験設備

日本放送協会放送技術研究所・音響評価室の実験設備を使用した。評価者は正面を向いて評価するものとし、聴取位置での平均音圧レベルは 75dB（A 特性）程度とした。評価室の諸元を Table 1 に示す。

Table 1　評価室諸元

評価室	日本放送協会 放送技術研究所・音響評価室（勧告 ITU-R BS.1116-3 準拠）
容積	6.4m(W)×8.0m(D)×4.5m(H)

スピーカ	Fostex 社製 3-way スピーカ 24 個（勧告 ITU-R BS.1116-3 準拠）
スピーカ配置	半径 2.5m の円柱上に配置
	上層：高さ 2.9m（仰角 31 度）
	中層：高さ 1.4m（仰角 0 度）
	下層：高さ 0m（俯角 29 度）
低域効果チャンネル用スピーカ	2 個を前方左右に開き角 75 度、中央から 2.75m の位置の床上に設置

2.2. パンニング則評価実験

勧告 ITU-R BS. 2132-0 [1] に規定される多重比較法（参照音源なし、アンカーなし）により、指定する音像の動きや位置のパターンに関する説明文を基準として、音響フォーマットごとのレンダリング音の定位精度を主観評価する。

2.2.1. 評価音源・ADM

実験では、音像が動くパターン 2 種類と、静止するパターン 7 種類の計 9 種類の ADM を用いて評価音源を作成した。ただし、ADM で指定する座標値は、MPEG-H レンダラー、AC-4 レンダラーの設計思想や仕様に合わせて座標系、時刻を Table 2 の通り指定した。

Table 2　ADM における座標・時刻の指定方法

レンダラー	座標系	時刻
MPEG-H レンダラー	極座標系	サンプル表記
AC-4 レンダラー	直交座標系	時刻表記

(1) 画面上の移動（パターン①）

聴取者の正面にある幅 60 度、高さ 36 度の画面上を、音が画面左端から画面右端まで高さを保ったまま等速度（10 秒間）で 3 回移動する。1 回目は画面 1/2 の高さ、2 回目は画面 3/4 の高さ、3 回目は画面上端の高さとした。ADM で指定した座標値の代表点を Table 3 に示す。代表点間は画面上を等速度で移動するように座標値を 1024 サンプル（約 0.02133 秒）ごとに指定した。

(2) 聴取者の周囲の移動（パターン②）

音が聴取者の周囲を 12 秒間で反時計回りに高さを変えながら 2 回移動する。1 周目は始点をスピーカ M+000、終点をスピーカ U+000 とし、2 周目は始点をスピーカ U-090、終点をスピーカ M-090 とした。ADM で指定した代表点の座標値を Table 3 に示す。代表点間は座標値を 1024 サンプル（約 0.02133 秒）ごとに線形に補間した。

(3) 静止のパターン（パターン③ － パターン⑨）

Table 3 に示す位置で、3 秒程度音がする。

Table 3　各パターンの座標値

パターン	極座標系	直交座標系
パターン①※	(30.0, 0.0, 1)→(30.0, 0.0, 1) →(-30.0, 0.0, 1)	(-1.0, 1.0, 0.0)→(0.0, 1.0, 0.0) →(1.0, 1.0, 0.0)
	(30.0, 8.0, 1)→(30.0, 9.2, 1) →(-30.0, 8.0, 1)	(-0.9, 1.0, 0.3)→(0.0, 1.0, 0.3) →(0.9, 1.0, 0.3)
	(30.0, 15.7, 1)→(30.0, 18.0, 1) →(-30.0, 15.7, 1)	(-0.8, 1.0, 0.6)→(0.0, 1.0, 0.6) →(0.8, 1.0, 0.6)
パターン②※	(0.0, 0.0, 1)→(30.0, 2.5, 1) →(60.0, 5.0, 1)→(90.0, 7.5, 1) →(180.0,15.0,1)→(-90.0,22.5,1) →(-60.0, 25.0, 1)→(-30.0, 27.5, 1) →(0.0, 30.0, 1)	(0.0, 1.0, 0.0)→(-1.0, 1.0, 0.125) →(-1.0,0.0,0.250)→(-1.0,-1.0,0.375) →(0.0,-1.0, 0.500)→(1.0,-1.0, 0.625) →(1.0, 0.0, 0.750)→(1.0, 1.0, 0.875) →(0.0, 1.0, 1.0)
	(-90.0, 30.0, 1)→(-60.0, 27.5, 1) →(-30.0, 25.0, 1)→(0.0, 22.5, 1) →(30.0, 20.0, 1)→(60.0, 17.5, 1) →(90.0, 15.0, 1)→(180.0,7.5,1) →(-90.0, 0.0, 1)	(1.0, 0.0, 1.0)→(1.0, 1.0, 0.875) →(0.0, 1.0, 0.750)→(-1.0, 1.0, 0.625) →(-1.0,0.0, 0.500)→(-1.0,-1.0, 0.375) →(0.0,-1.0, 0.250)→(1.0,-1.0, 0.125) →(1.0, 0.0, 0.0)
パターン③	(0.0, 30.0, 1)	(0.00, 1.00, 1.00)
パターン④	(15.0, 7.5, 1)	(-0.50, 1.00, 0.25)
パターン⑤	(90.0, 22.5, 1)	(-1.00, 0.00, 0.75)
パターン⑥	(120.0, 15.0, 1)	(-1.00, -0.67, 0.50)
パターン⑦	(165.0, 15.0, 1)	(-0.33, -1.00, 0.50)
パターン⑧	(45.0, 60.0, 1)	(-0.50, 0.50, 1.00)
パターン⑨	(180.0, 75.0, 1)	(0.00, -0.25, 1.00)

※ 極座標系では（方位角、仰角、距離）の組合せにより、直交座標系では（X 座標、Y 座標、Z 座標）の組合せにより音像の位置を指定する。

実験に用いる音声フォーマットは、22.2 マルチチャンネル音響（System H）、7.1.4（System J）の 2 種類とし、各音声フォーマットのスピーカ配置は勧告 ITU-R BS.2051-2 [2] に規定されるスピーカ位置とした。

評価音源は、モノ音源と、9 種類の ADM（2.2.1 項）を入力として、2 方式（MPEG-H 3DA、AC-4）によりレンダリングした音源を使用した。AC-4 については ITU-R BS.2127-0 としてオープンソース化されているレンダラーソフトウェアを使用した。実験に用いるモノ音源は、音源長 6－48 [s]程度、サンプリング周波数 48 [kHz]、量子化ビット数 24 [bits] の PCM 音源とし、3 種類の音

源を用いた（Table 4 ）。

Table 4　パンニング則評価実験に使用したモノ音源

項目	使用音源	音源長
パターン①	シルクロード（人声）	約 48 秒
パターン②	ヴァイオリン（単楽器）	約 32 秒
パターン③～⑨	ホワイトノイズ（人工音）	約 6 秒

2.2.2. 実験手順

評価方法は勧告 ITU-R BS.2132-0[1] に準拠した多重比較法（参照音源なし、アンカーなし）とし、音響フォーマットごとのレンダリング音の定位精度を主観評価する。評定者は、事前に音像の動きや位置についての説明文が提示される。この説明文を基準として、4 個の異なる評価音 A、B、C、D をそれぞれ聴取し、各音源が説明文の通りであるかを 5 段階連続品質尺度（Table 5 ）に沿って回答する。ただし、音像の動きや位置示す説明文は、音像の位置を方位角・仰角による表記と、複数スピーカとの位置関係による表記とを併記した（Figure 1 ）。このとき、各スピーカは Figure 2 の通り呼称し、実スピーカにラベリングを行った（Figure 2 ）。また、パターン①における「画面」は、想定する寸法の画面を模した紐を張り、その場所を提示した（Figure 3 ）。

Table 5　5 段階連続品質尺度

尺度	評点
非常に良い	100 – 81 点
良い	80 – 61 点
普通	60 – 41 点
悪い	40 – 21 点
非常に悪い	20 – 0 点

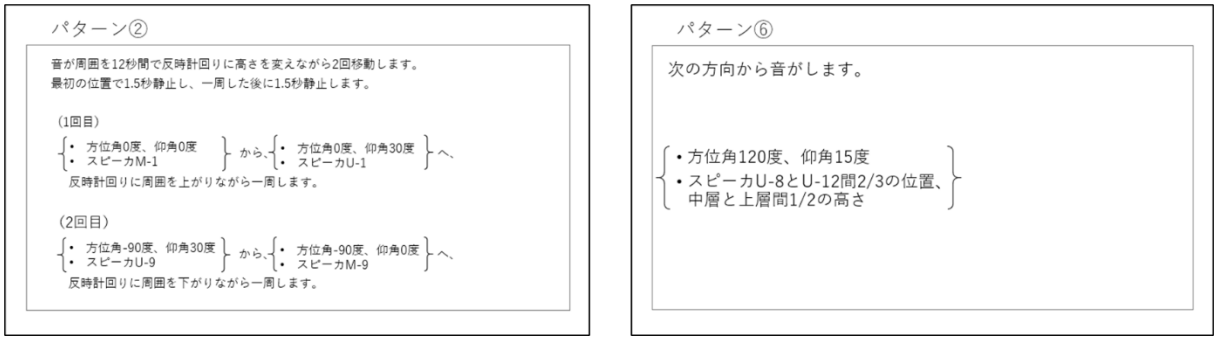


Figure 1　音像の動き・位置についての説明文例（左図：パターン②、右図：パターン⑥）

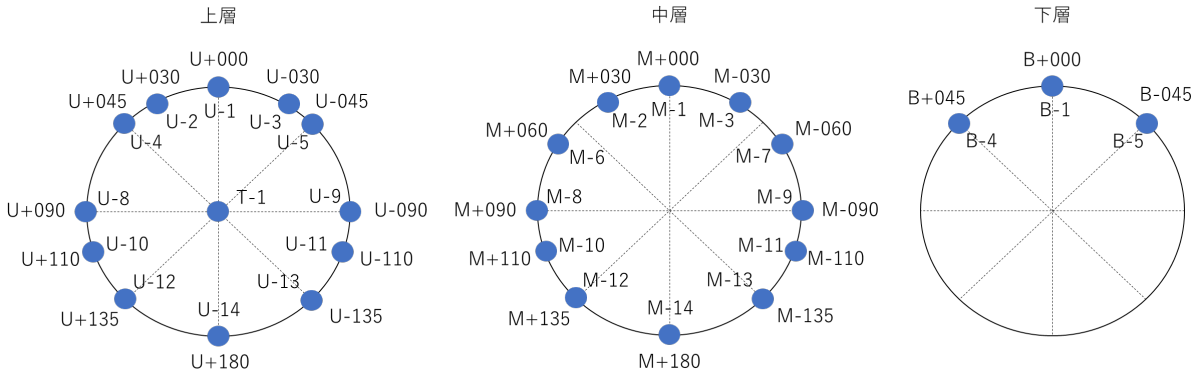


Figure 2　スピーカラベリング

※　上図の円の外側に勧告 ITU-R BS.2051-2 [2] に規定されるスピーカラベルを記載し、内側に本実験で使用するスピーカラベルを記載した。

Figure 3　実験設備（模擬画面、スピーカラベル）

評価音 A、B、C、D には、MPEG-H レンダラーによるレンダリング音（22.2 マルチチャンネル音響、7.1.4）、および AC-4 レンダラーにレンダリング音（22.2 マルチチャンネル音響、7.1.4）がランダムに提示される。評定者は、評価音 A、B、C、D を好きなポイントで自由に切り替え、何度でも聴取することが可能である。評定者は、評価音 A、B、C、D に 0 点以上 100 点以下を 1 点単位で採点する。このとき、評定者が説明文の通りである判断した評価音には 100 点を採点するよう教示した。ただし、複数の評価音に 100 点を採点することを可能とし、いずれの評価音も説明文の通りでないと判断した場合は全ての評価音に 100 点未満を採点することを可能とした。

2.3.　Egocentric/Allocentric 評価実験

勧告 ITU-R BS.1534-3 [3] に規定される多重比較法（参照音源あり、アンカーなし）により、スピーカ位置が異なる音響フォーマット間で、レンダリング思想（Egocentric/Allocentric）による音質・音像定位精度・空間印象の違いを主観評価する。

2.3.1. 評価音源・ADM

実験に用いる音声フォーマットは、22.2 マルチチャンネル音響（System H）、7.1.4（System J）、5.1.4（System D）の 3 種類とし、各音声フォーマットのスピーカ配置は勧告 ITU-R BS.2051-2 [2] に規定されるスピーカ位置とした。

本実験では、オブジェクトベース音響コンテンツを想定する背景音（22.2 マルチチャンネル音響）と音声オブジェクトから構成される音源と、符号化時のエレメント数が限られた場合を想定する 22 個の音声オブジェクトから構成される音源の 2 種類を評価した。実験に使用した音源は、音源長 15 – 20 [s]程度、サンプリング周波数 48 [kHz]、量子化ビット数 24 [bits] の PCM 音源とし、オブジェクトベース音響コンテンツを 4 音源、22 音声オブジェクトコンテンツを 2 音源とした（Table 5 ）。22 音声オブジェクトコンテンツは、符号化時のエレメント数が限られる場合で、各チャンネルトラックを音声オブジェクトとして送出するケースを想定する。評価音源は、レンダリング方式 2 方式（MPEG-H 3DA、AC-4）により前述の 3 音声フォーマットにレンダリングしたものとした。このとき、オブジェクトベース音響コンテンツの ADM で指定した音声オブジェクトの座標値は Table 6 の通りとし、22 音声オブジェクトの ADM では 22.2 マルチチャンネル音響のスピーカ（LFE を除く）の座標値を指定し、直交座標系については勧告 ITU-R BS.2127-0 [4] に準拠した。

Table 5　Egocentric/Allocentric 評価実験で使用した音源

項目	使用音源	構成
オブジェクトベース音響コンテンツ	桃太郎（ドラマ）	・ 背景音 ・ 音声オブジェクト（ダイアログ）×4
	生き物の世界（ドラマ）	・ 背景音 ・ 音声オブジェクト（ダイアログ）×1 ・ 音声オブジェクト（動物の鳴き声）×3
	野球中継	・ 背景音 ・ 音声オブジェクト（ダイアログ）×4
	オートリバースの恋（ドラマ）	・ 背景音 ・ 音声オブジェクト（ダイアログ、動的）×1
22 音声オブジェクトコンテンツ	八重奏（楽曲）	・ 音声オブジェクト×22
	光るタクト（ドラマ）	・ 音声オブジェクト×22

Table 6　オブジェクトベース音響コンテンツにおける音声オブジェクトの座標値

音源	音声 オブジェクト	極座標系	直交座標系
桃太郎	桃太郎	(-30.0, 0.0, 1)	(1.0, 1.0, 0.0)
	犬	(45.0, -30.0, 1)	(-1.0, 1.0, -1.0)
	猿	(-90.0, 30.0, 1)	(1.0, 0.0, 1.0)
	雉	(135.0, 30.0, 1)	(-1.0, -1.0, 1.0)
生き物の世界	ナレーション	(0.0, 30.0, 1)	(0.0, 1.0, 1.0)
	ゴリラ	(90.0, 0.0, 1)	(-1.0, 0.0, 0.0)
	ピューマ	(45.0, 30.0, 1)	(-1.0, 1.0, 1.0)
	カバ	(-60.0, 0.0, 1)	(1.0, 0.414, 0.0)
野球中継	実況	(0.0, 30.0, 1)	(0.0, 1.0, 1.0)
	PA	(0.0, 90.0, 1)	(0.0, 0.0, 1.0)
	売子	(120.0, 0.0, 1)	(-1.0, 0.67, 0.0)
	ヤジ	(-120.0, 15.0, 1)	(1.0, -0.67, 0.50)
オートリバースの恋	少女	(45.0, 0.0, 1) →(90.0, 0.0, 1) →(45.0, 0.0, 1)	(-1.0, 0.67, 0.0) →(-1.0, 0.0, 0.0) →(-1.0, 0.67, 0.0)

2.3.2. 実験手順

評価方法は、勧告 ITU-R BS.1534-3 [3] に準拠した多重比較法（参照音源あり、アンカーなし）とし、スピーカ位置が異なる音響フォーマット間の音質・音像定位精度・空間印象の違いを主観評価する。評定者は基準音 REF（22.2 マルチチャンネル音響のレンダリング音）と 3 個の異なる評価音 A、B、C をそれぞれ聴取し、基準音に対して評価音 A、B、C それぞれの総合印象の違いを 5 段階連続品質尺度（Table 5 、パンニング則評価実験と同様の尺度）に沿って回答する。ただし、基準音および評価音は、MPEG-H レンダラーあるいは AC-4 レンダラーのいずれか一方によるレンダリング音のみが提示され、2 種類の方式によるレンダリング音が混在しないものとした。評価音 A、B、C のうち、1 つに基準音 REF と同じ 22.2 マルチチャンネル音響のレンダリング音、他の 2 つに 7.1.4 のレンダリング音、5.1.4 のレンダリング音がランダムに提示される。評定者は評価音のいずれかに 100 点、他の 2 つに 100 点以下 1 点単位で採点する。評定者は、基準音 REF と評価音 A、B、C を好きなポイントで自由に切り替え、何度でも聴取することが可能である。また、音源ごとの印象差の評価を行うために、評定者には、少なくとも最も低く評点した音源については、REF との印象差をどこに感じたかについて用紙に記入させた。

3. 実験結果

3.1. パンニング則評価実験

各音像パターンの実験結果を Figure 4 – Figure12 に示す。図の各点は、評定点の平均値をエラー

バーは 95%信頼区間を示す。

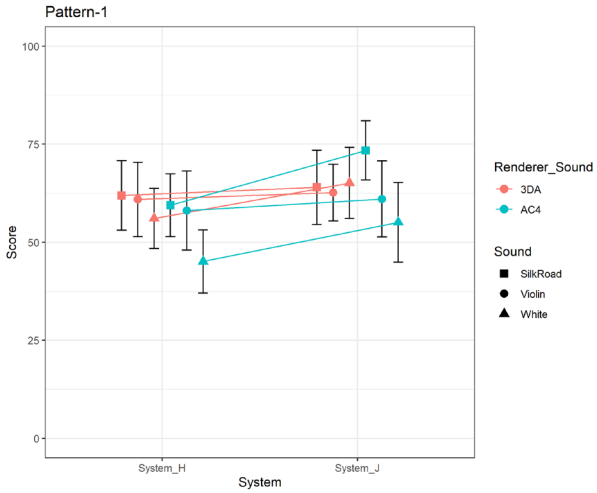


Figure 4　評価結果
（パターン①：画面上の移動）

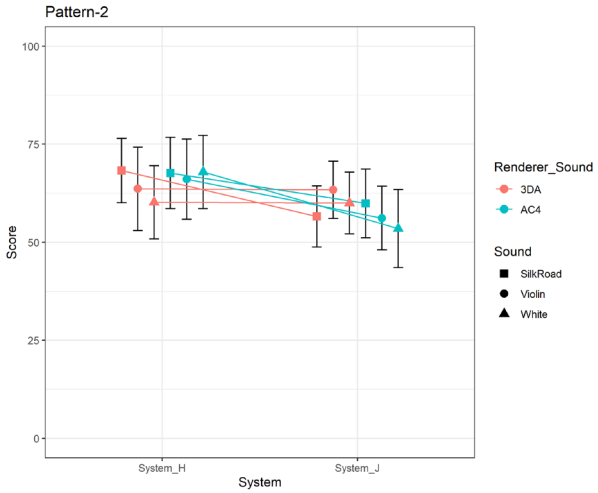


Figure 5　評価結果
（パターン②：聴取者の周囲の移動）

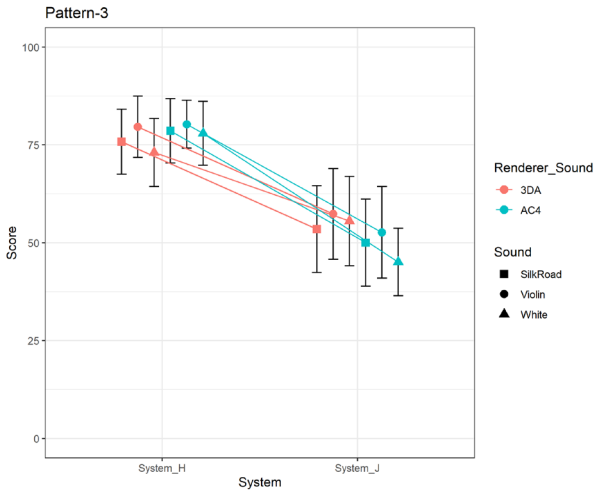


Figure 6　評価結果
（パターン③：静止(0.0, 30.0, 1)）

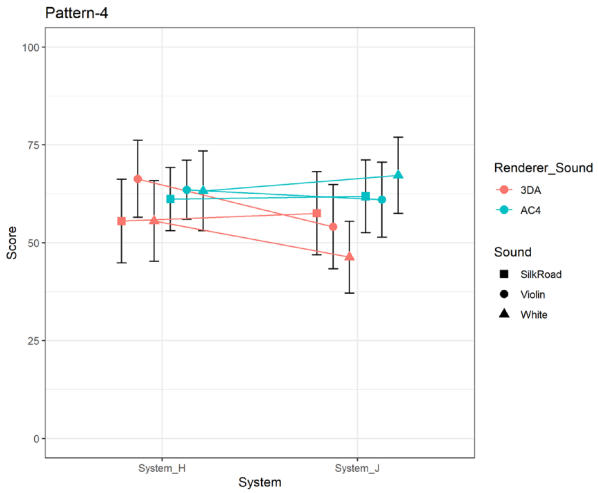


Figure 7
（パターン④：静止(15.0, 7.5, 1)）

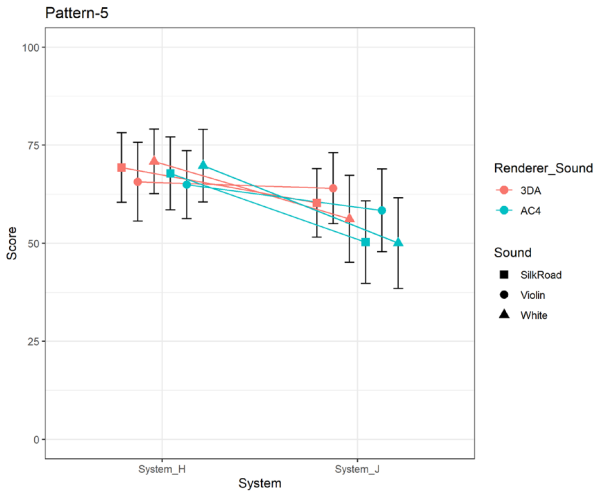


Figure8　評価結果
（パターン⑤：静止(90.0, 22.5, 1)）

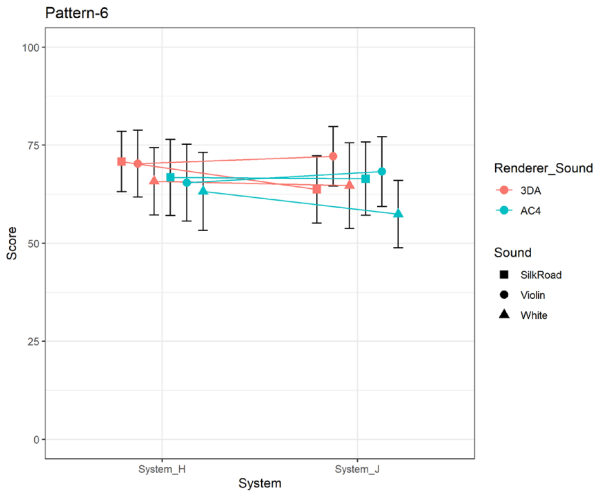


Figure9　評価結果
（パターン⑥：静止(120.0, 15.0, 1)）

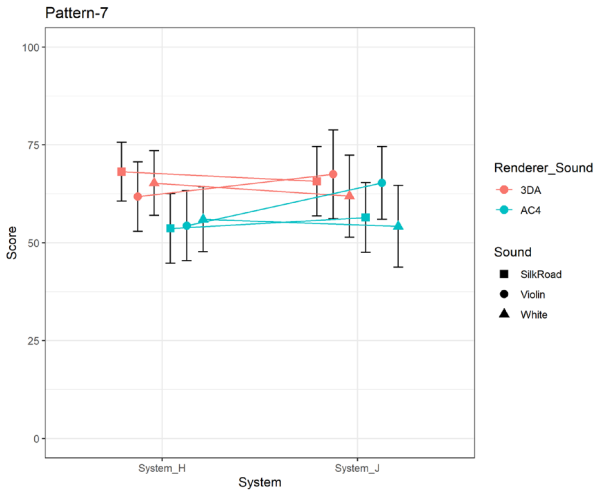


Figure10　評価結果
（パターン⑦：静止(165.0, 15.0, 1)）

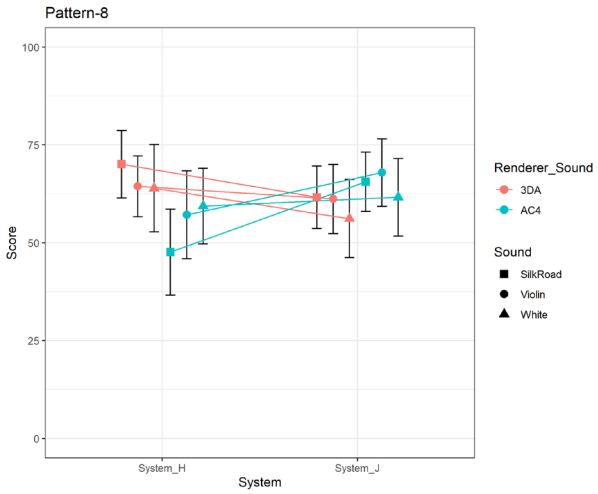


Figure11　評価結果
（パターン⑧：静止(45.0, 60.0, 1)）

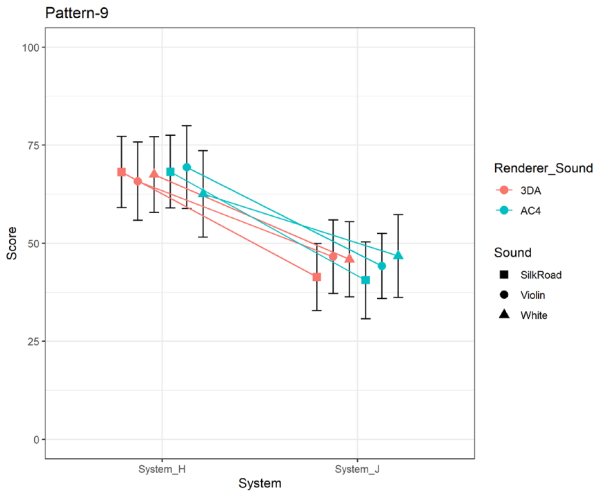


Figure12　評価結果
（パターン⑨：静止(180.0, 75.0, 1)）

3.2.　Egocentric/Allocentric 評価実験

オブジェクトベース音響コンテンツの実験結果を Figure13 に、22 音声オブジェクトの実験結果を Figure14 に示す。図の各点は、評定点の平均値をエラーバーは 95%信頼区間を示す。ただし、隠れ基準音に対して全 12 試行の 15%以上で 90 点以下を採点した 6 名の評定者は除外し、計 18 名の評定者の結果を集計の対象とした。

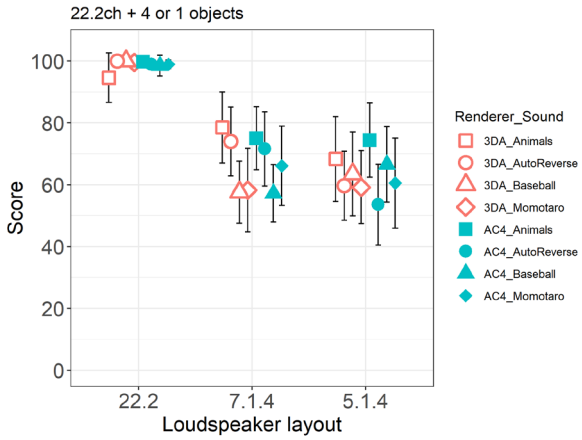


Figure13　評価結果
（オブジェクトベース音響コンテンツ）

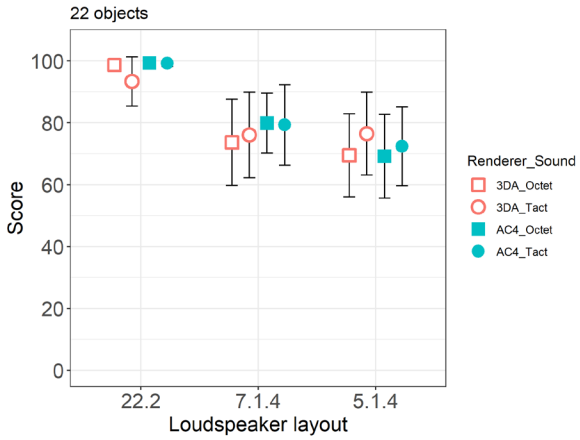


Figure14　評価結果
（22 音声オブジェクト）

4.　考察

各レンダリング方式における評定値の平均値の違いを統計的に解析するために、3 因子（レンダリング方式、音源、音声フォーマット）に対する三元配置分散分析を行った。このとき帰無仮説は「3 因

子の水準間の平均値に差がない」とし、検定における有意水準は 5%とした。

4.1. パンニング則評価実験

3 因子（レンダリング方式、音源、音声フォーマット）に対する三元配置分散分析を行った。各パターンに対する分散分析の結果を Table 7 – Table 1 5 に示す。帰無仮説が棄却された要因については、p 値に＊印を併記する。

Table 7　分散分析結果（パターン①）

変動要因	自由度	自由度（残差）	F 値	P 値	p<.05
レンダリング方式	1	23	1.287	0.268	
音声フォーマット	1	23	7.558	0.011	*
音源	2	46	7.74	0.001	*
レンダリング方式×音声フォーマット	1	23	1	0.328	
レンダリング方式×音源	2	46	3.383	0.043	*
音声フォーマット×音源	2	46	1.576	0.218	
レンダリング方式×音声フォーマット×音源	2	46	0.923	0.405	

Table 8　分散分析結果（パターン②）

変動要因	自由度	自由度（残差）	F 値	P 値	p<.05
レンダリング方式	1	23	0.004	0.948	
音声フォーマット	1	23	5.151	0.033	*
音源	2	46	1.021	0.368	
レンダリング方式×音声フォーマット	1	23	2.499	0.128	
レンダリング方式×音源	2	46	0.317	0.73	
音声フォーマット×音源	2	46	0.273	0.763	
レンダリング方式×音声フォーマット×音源	2	46	1.668	0.2	

Table 9　分散分析結果（パターン③）

変動要因	自由度	自由度（残差）	F 値	P 値	p<.05
レンダリング方式	1	23	2.35	0.139	

音声フォーマット	1	23	20.756	0.000141	*
音源	2	46	3.805	0.03	*
レンダリング方式×音声フォーマット	1	23	7.859	0.01	*
レンダリング方式×音源	2	46	0.303	0.74	
音声フォーマット×音源	2	46	0.005	0.995	
レンダリング方式×音声フォーマット×音源	2	46	1.513	0.231	

Table１０　分散分析結果（パターン④）

変動要因	自由度	自由度（残差）	F 値	P 値	p<.05
レンダリング方式	1	23	8.223	0.009	*
音声フォーマット	1	23	1.024	0.322	
音源	2	46	1.281	0.288	
レンダリング方式×音声フォーマット	1	23	3.621	0.07	
レンダリング方式×音源	2	46	4.001	0.025	*
音声フォーマット×音源	2	46	1.089	0.345	
レンダリング方式×音声フォーマット×音源	2	46	1.342	0.271	

Table１１　分散分析結果（パターン⑤）

変動要因	自由度	自由度（残差）	F 値	P 値	p<.05
レンダリング方式	1	23	7.153	0.014	*
音声フォーマット	1	23	7.351	0.012	*
音源	2	46	0.387	0.681	
レンダリング方式×音声フォーマット	1	23	2.391	0.136	
レンダリング方式×音源	2	46	0.21	0.811	
音声フォーマット×音源	2	46	1.855	0.168	
レンダリング方式×音声フォーマット×音源	2	46	0.089	0.915	

Table１２　分散分析結果（パターン⑥）

変動要因	自由度	自由度（残差）	F 値	P 値	p<.05

レンダリング方式	1	23	1.658	0.211	
音声フォーマット	1	23	0.871	0.36	
音源	2	46	6.918	0.002	*
レンダリング方式×音声フォーマット	1	23	0.101	0.754	
レンダリング方式×音源	2	46	0.325	0.724	
音声フォーマット×音源	2	46	1.443	0.247	
レンダリング方式×音声フォーマット×音源	2	46	1.198	0.311	

Table１３　分散分析結果（パターン⑦）

変動要因	自由度	自由度（残差）	F 値	P 値	p<.05
レンダリング方式	1	23	9.996	0.004	*
音声フォーマット	1	23	0.865	0.362	
音源	2	46	0.871	0.426	
レンダリング方式×音声フォーマット	1	23	0.954	0.339	
レンダリング方式×音源	2	46	1.524	0.229	
音声フォーマット×音源	2	46	2.062	0.139	
レンダリング方式×音声フォーマット×音源	2	46	0.205	0.816	

Table１４　分散分析結果（パターン⑧）

変動要因	自由度	自由度（残差）	F 値	P 値	p<.05
レンダリング方式	1	23	2.512	0.127	
音声フォーマット	1	23	0.396	0.535	
音源	2	46	0.666	0.518	
レンダリング方式×音声フォーマット	1	23	13.025	0.001	*
レンダリング方式×音源	2	46	3.181	0.051	
音声フォーマット×音源	2	46	0.607	0.549	
レンダリング方式×音声フォーマット×音源	2	46	2.437	0.099	

Table１５　分散分析結果（パターン⑨）

変動要因	自由度	自由度	F 値	P 値	p<.05

		（残差）			
レンダリング方式	1	23	0.226	0.639	
音声フォーマット	1	23	19.057	0.000226	*
音源	2	46	0.309	0.735	
レンダリング方式×音声フォーマット	1	23	0.015	0.904	
レンダリング方式×音源	2	46	0.279	0.758	
音声フォーマット×音源	2	46	0.951	0.394	
レンダリング方式×音声フォーマット×音源	2	46	1.222	0.304	

これらの結果を見ると、パターン④、⑤、⑦においてレンダリング方式に主効果があり、パターン①においてレンダリング方式と音源間で、パターン③、⑧においてレンダリング方式と音声フォーマット間で有意な交互作用があった。その他パターンではレンダリング方式による単一要因、交互作用による統計的に有意な平均値の差は認められなかった。

詳細な分析のため、レンダリング方式に関連する要因について平均値の差に有意な項目を含むパターンについては事後検定として、Tukey 検定を行った。このとき帰無仮説は「各群の評定値の平均値間に差がない」とし、検定における有意水準は 5%とした。その結果を Table１６ –Table２１に示す。本検定結果、および実験結果のグラフより、パターン④においては AC-4 が、パターン⑦では MPEG-H 3DA の方が評定点の平均値が統計的に有意に高いことが言える。ただし、全体の傾向としては、レンダリング方式の傾向よりも音源や音声フォーマットによる差の方が大きく、パンニング則による空間印象や定位精度については、MPEG-H 3DA レンダラーと AC-4 レンダラーは総合的には同程度であると考えられる。

Table１６　Tukey 検定の結果（パターン①）

要因	群 1	群 2	p 値	p<.05
レンダリング方式	MPEG-H 3DA	AC4	0.212	
音声フォーマット	System H	System J	0.00775	*
音源	シルクロード	ヴァイオリン	0.383	
	シルクロード	ホワイトノイズ	0.00638	*
	ヴァイオリン	ホワイトノイズ	0.186	

Table１７　Tukey 検定の結果（パターン③）

要因	群 1	群 2	p 値	p<.05
レンダリング方式	MPEG-H 3DA	AC4	0.519	
音声フォーマット	System H	System J	5.06E-13	*

音源	シルクロード	ヴァイオリン	0.629	
	シルクロード	ホワイトノイズ	0.878	
	ヴァイオリン	ホワイトノイズ	0.339	

Table１８　Tukey 検定の結果（パターン④）

要因	群 1	群 2	p 値	p<.05
レンダリング方式	MPEG-H 3DA	AC4	0.00925	*
音声フォーマット	System H	System J	0.29	
音源	シルクロード	ヴァイオリン	0.785	
	シルクロード	ホワイトノイズ	0.957	
	ヴァイオリン	ホワイトノイズ	0.612	

Table１９　Tukey 検定の結果（パターン⑤）

要因	群 1	群 2	p 値	p<.05
レンダリング方式	MPEG-H 3DA	AC4	0.122	
音声フォーマット	System H	System J	0.0000275	*
音源	シルクロード	ヴァイオリン	0.913	
	シルクロード	ホワイトノイズ	0.998	
	ヴァイオリン	ホワイトノイズ	0.889	

Table２０　Tukey 検定の結果（パターン⑦）

要因	群 1	群 2	p 値	p<.05
レンダリング方式	MPEG-H 3DA	AC4	0.00121	*
音声フォーマット	System H	System J	0.442	
音源	シルクロード	ヴァイオリン	0.919	
	シルクロード	ホワイトノイズ	0.858	
	ヴァイオリン	ホワイトノイズ	0.63	

Table２１　Tukey 検定の結果（パターン⑧）

要因	群 1	群 2	p 値	p<.05
レンダリング方式	MPEG-H 3DA	AC4	0.259	
音声フォーマット	System H	System J	0.471	
音源	シルクロード	ヴァイオリン	0.9	
	シルクロード	ホワイトノイズ	0.954	
	ヴァイオリン	ホワイトノイズ	0.745	

なお、本評価実験結果、Figure4～Figure12 のグラフの値を見る際には、本実験が隠れ参照音源および隠れアンカー信号のない多重音声比較法を用い、Table 5 の 5 段階連続品質尺度により評価されている点に留意すべきである。例えば、パターン③：静止(0.0, 30.0, 1)のレンダリング条件で System H（２２．２ｃｈ）にレンダリングする場合、オブジェクトの再生位置は、上層センタースピーカのみとなり、2 つのレンダリング方式とも上層センタースピーカからのみレンダリングされた音声が再生されるはずである。その場合、100 点に近い得点が得られることが期待できるが、Figure 6 の示す平均値の値はそれよりも低い値になっている。これは、隠れ参照音源、隠れアンカー音源を使用しない評価法を用いていることが一原因と考えられ、今回評価で得られたその他の Figure 群を含む各得点が、5 段階連続品質尺度の印象語(非常に良い、よい、普通、悪い、非常に悪い)と一致しているかどうかは慎重にみる必要がある。

4.2. Egocentric/Allocentric 評価実験

3 因子（レンダリング方式、音源、音声フォーマット）に対する三元配置分散分析を行った。オブジェクトベース音響コンテンツに対する分散分析の結果を Table２２に、22 音声オブジェクトに対する分散分析の結果を Table２３に示す。帰無仮説が棄却された要因については、p 値に＊印を併記する。これらの結果を見ると、音声フォーマットに主効果があり、さらに音声フォーマットと音源の要因間で有意な交互作用があったが、レンダリング方式に有意な平均値の差は認められなかった。従って、いずれのレンダリング方式（Egocentric、Allocentric のいずれの思想）であっても、制作環境と異なる音声フォーマットにレンダリングされた場合の印象差は同程度であると考えられる。

Table２２　分散分析結果（オブジェクトベース音響コンテンツ）

変動要因	自由度	自由度（残差）	F 値	P 値	p<.05
レンダリング方式	1	15	0.391	0.541	
音声フォーマット	2	30	60.156	3.18E-11	*
音源	3	45	6.574	0.000883	*
レンダリング方式×音声フォーマット	2	30	0.023	0.977	
レンダリング方式×音源	3	45	1.362	0.266	
音声フォーマット×音源	6	90	4.925	0.000213	*
レンダリング方式×音声フォーマット×音源	6	90	0.851	0.534	

Table２３　分散分析結果（22 音声オブジェクト）

変動要因	自由度	自由度	F 値	P 値	p<.05

		（残差）			
レンダリング方式	1	15	3.057	0.101	
音声フォーマット	2	30	17.61	0.00000873	*
音源	1	15	1.179	0.295	
レンダリング方式×音声フォーマット	2	30	1.369	0.27	
レンダリング方式×音源	1	15	0.059	0.812	
音声フォーマット×音源	2	30	1.472	0.246	
レンダリング方式×音声フォーマット×音源	2	30	0.412	0.666	

また、22.2 マルチチャンネル音響との印象差として気になった点についての代表的なコメントを Table２４、Table２５に示す。MPEG-H 3DA では生き物の世界のナレーションや、八重奏などで音の明瞭度に関するコメントがあった。これは Egocentric 思想により角度を保持するためのファンタムに起因するものと考えられるが、定位の差異に関するコメントも多く見受けられた。一方、AC-4 ではコメントや背景音に含まれる音のレベルが変わって聞こえるコメント等が見受けられた。これは Allocentric 思想により背景音と音声オブジェクトの空間上の位置関係が変化したことで空間マスキングが解除されたことが要因であることが考えられる。ただ、全体としては MPEG-H 3DA、AC-4 に共通のコメントが多く見受けられ、双方の思想の違いによる印象差よりも再生時のスピーカ配置の違いに起因する定位の違いが印象差として感じることが多かったと考えられる。

Table２４　印象差に関する代表的なコメント（オブジェクトベース音響コンテンツ）

コンテンツ	MPEG-H 3DA	AC4
桃太郎	✓ 犬の台詞の定位が異なる ✓ 猿の台詞の定位が異なる	✓ 犬の台詞の定位が異なる ✓ 猿の台詞の定位が異なる ✓ 雉の声が左横に定位がずれる ✓ （背景音に含まれる）鳥の鳴き声が大きく聞こえた
生き物の世界	✓ ナレーションの定位が異なる ✓ ナレーションの音色が異なる、こもって聞こえる ✓ ピューマの鳴き声の定位が異なる	✓ ナレーションの定位が異なる ✓ ナレーションの音色が異なる ✓ 後半の鳴き声（カバ）の定位が異なる
野球中継	✓ 実況の定位が異なる ✓ 実況の音色が異なる、低く聞こえる、レベルが高い	✓ 実況の定位が上側に寄って聞こえる ✓ 実況の音色が異なる、低く聞こえる

	✓ 守備のアナウンスの定位が異なる ✓ ヤジが右側方に近づいた	✓ 守備のアナウンスの定位が異なる ✓ ヤジが右側方に近づいた
オートリバースの恋	✓ 開始の女性の台詞がセンター寄りに 15 度ずれている（M-2 に寄っている） ✓ 左側方の女性の台詞の定位が異なる ✓ 音色が異なる	✓ 足音の動きが異なる ✓ 開始の女性の台詞の定位が異なる ✓ 左側方の女性の台詞の定位が異なる ✓ 背景音のボーカルの声が良く聞こえる

Table２５　印象差に関する代表的なコメント（22 音声オブジェクトコンテンツ）

コンテンツ	MPEG-H 3DA	AC4
八重奏	✓ 明瞭度が下がる、音がこもる ✓ ヴァイオリンの広がり、音色が異なる	✓ 空間の広がり、リバーブ感が異なる ✓ チェロの音量が大きい ✓ ヴァイオリンの広がりが異なる
光るタクト	✓ 物音が全体的に上から聴こえる ✓ 残響音が異なる	✓ 物音の定位が左横にずれる ✓ 定位ずれを感じる、定位が真ん中に寄る ✓ 残響音が異なる ✓ トロンボーンの音色が異なる

5. まとめ

本実験の結果から、音源や定位のパターンによっては定位精度、空間印象に関する評定点に統計的に有意な差が出る場合もあったが、レンダリング方式の間に統計的に有意な差があるとは言えなかった。これはパンニング則評価実験で提示した単一音源についても、Egocentric/Allocentric 評価実験で提示したコンテンツ音源でも同様である。また、コメント上で双方の思想の違いによる影響とみられる印象差についてのコメントが見受けられたが、全体としては双方に共通するスピーカ配置の違いによる影響も大きいと考えられる。

6. 参考文献

[1] ITU, Recommendation ITU-R BS.2123-0, 2019
[2] ITU, Recommendation ITU-R BS.1534-3, 2015-10
[3] ITU, Recommendation ITU-R BS.2051-2, 2018-7

[4] ITU, Recommendation ITU-R BS.2127-0, 2019-6