

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

音声符号化方式の比較検討 最終報告

2022年2月18日

音声符号化方式検討の目的

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

4方式の音声符号化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。

1. MPEG-4 AAC

2. MPEG-H 3D Audio

3. Enhanced AC-3

4. AC-4

今回の報告内容

- ビットレートと品質に関する実験結果
 - 各音声符号化方式の所要ビットレートの算出と方式間の品質差を確認
- OBAにおけるレンダリング技術に関する評価実験結果
 - 符号化方式ごとにレンダリング手法が異なるため、方式間の品質差を確認

ビットレートと品質に関する主観評価実験

実験方法	勧告ITU-R BS.1116-3 隠れ基準付き三刺激二重盲検法* ¹	評点	評価尺度
		5.0	差が分からない
音声フォーマット	22.2マルチチャンネル音響(22.2ch) 7.1.4 5.1chサラウンド(5.1ch) ステレオ モノ	4.0	差が分かるが、気にならない
		3.0	差がわずかに気になる
		2.0	差が気になる
		1.0	差が非常に気になる
音源	音声フォーマットごとに4音源を選出		
符号化方式	4方式（MPEG-4 AAC, MPEG-H 3DA, AC-4* ² , Enhanced AC-3）		
評定者数	音響専門家 19名（ポストスクリーニング* ³ 後）		
実験サイト	NHK技研、東京藝術大学 （ITU-R BS.1116準拠のリスニングルーム）		

*1：非圧縮音を基準として、符号化音と非圧縮音（隠れ基準）をそれぞれ評価させる方法
統計処理に差分評価値（＝符号化音の評価値－隠れ基準の評価値）を用いることで
高品質な（評価値が5.0に近い）符号化音の評価が頭打ちになることを防いでいる

*2：22.2chに対応したプロトタイプ（Level 4相当）と少ないエレメント数で対応するLevel 3により実施

*3：評価結果から信頼度の低い評定者を除く処理
（例えば、原音に低い点をつける/でたらめに回答した分布と統計的な差がない等）

音源/ビットレート

音声フォーマット	使用音源
22.2ch	Applause (自然音)、Tempura (自然音)、Etenraku (音楽)、MyKingdom (ドラマ)
7.1.4	Ave Maria (音楽)、Choir at side (音楽)、Unfold (自然音)、Water burst (自然音)
5.1ch	Atami fireworks festival (自然音)、Applause (自然音)、Jazzquartet (音楽)、Glasses (自然音)
ステレオ	Glockenspiel (単楽器)、Feste Romane - La Befana (音楽)、German Narration - male (音声)、Harpsichord (単楽器)
モノ	Wind chime (単楽器)、Shamisen (単楽器)、Glockenspiel (単楽器)、English Narration - female (音声)

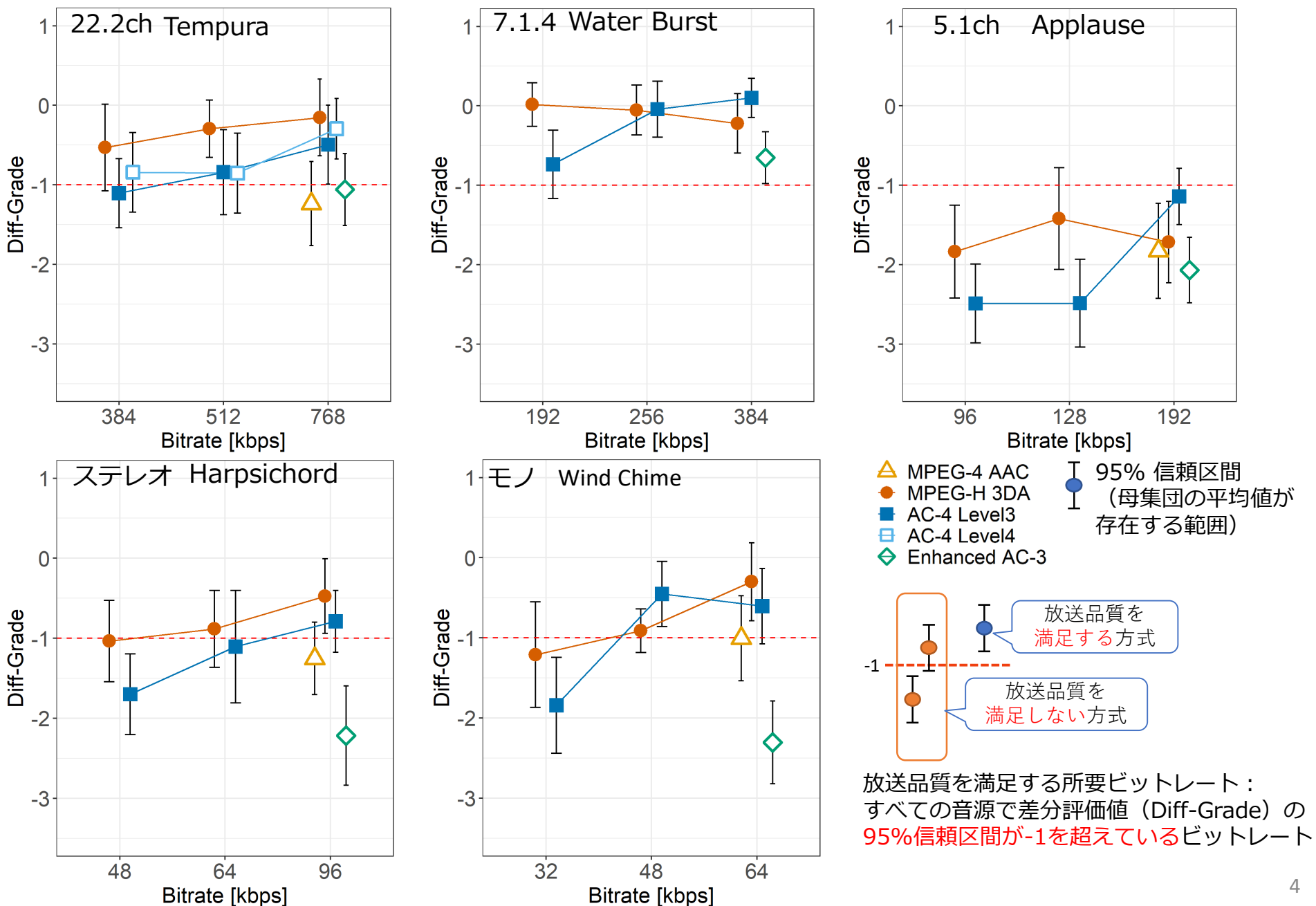
音声フォーマット	22.2ch			7.1.4			5.1ch			ステレオ			モノ		
ビットレート [kbps]	384	512	768	192	256	384	96	128	192	48	64	96	32	48	64
MPEG-4 AAC	X	X	O	X	X	X	X	X	O	X	X	O	X	X	O
MPEG-H 3DA	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Enhanced AC-3	X	X	O	X	X	O	X	X	O	X	X	O	X	X	O
AC-4	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O

音源はARIB 委員（符号化の専門家含）から符号化が難しいと思われる音源を収集し、試聴により決定

ビットレートは先行研究を参考に決定（Oが採用したビットレート）

- ・ AAC、Enhanced AC-3については低いレートでの使用を想定してないため、最上位ビットレートを設定した
- ・ AACは7.1.4に対応していないため設定していない

最も評価が低い（符号化が難しい）音源による結果



ビットレートと品質に関する主観評価実験まとめ

- 所要ビットレートの算出
 - 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

表 放送品質を満足するビットレート

音声符号化方式	音声フォーマット				
	22.2ch	7.1.4* ¹	5.1ch* ²	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

*1：符号化が難しい音源が集まらなかったため参考値

*2：5.1chの場合、実験では放送品質を見出すことができなかった

- 符号化方式間の品質比較（同一ビットレートでの品質比較）
 - MPEG-H 3DA > MPEG-4 AAC（22.2chのTempura）
 - MPEG-4 AAC > Enhanced AC-3（22.2chのApplause、5.1chのAtami fireworks festival、ステレオのGlockenspiel、モノのWind Chime・Glockenspiel）
 - MPEG-H 3DA > AC-4 Level3（22.2ch 384kbpsのApplause）
 - MPEG-H 3DA > AC-4（7.1.4 192bpsのWater burst、5.1ch 128kbpsと192kbpsの花火、ステレオ 48kbpsのFesta Romane）
 - AC-4 > MPEG-H 3DA（モノ 48kbpsのGlockenspiel）
- AACと比較してEnhanced AC-3は品質評価値は同等もしくは低い
- 2ch以上ではAC-4よりMPEG-H 3DAの方が品質評価値が高いものがある
- モノではAC-4が品質評価値が高いものがある
- 実験で設定した最高ビットレートでは、MPEG-H 3DAとAC-4で有意差はない

OBAにおけるレンダリング技術の主観評価実験

- 2つの実験

- 音源位置の再現性能の違い（3D VBAP／Triple Balance Panner）

- 音の位置を変えるためのルールが符号化方式ごとに異なるため、その違いが品質に与える影響について調査
 - オブジェクト音声＋動的/静的メタデータによるレンダリング後の定位品質を評価

- 再生環境（スピーカ配置）への適応方法の違い（Egocentric/Allocentric）

- 22.2ch+4または1 静的オブジェクト/22オブジェクトを
22.2ch/7.1.4/5.1.4に変換した場合の空間印象や定位品質を評価

- 補足

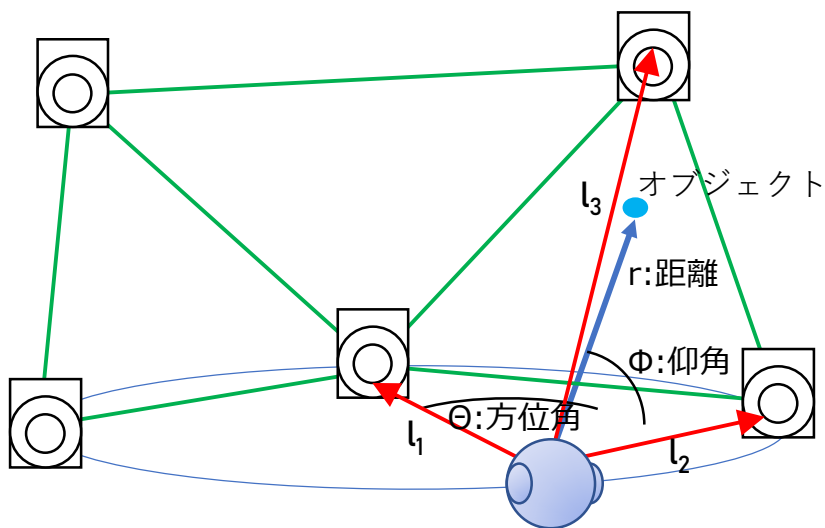
- レンダリング技術の評価については先行研究がないため、評価方法も併せて検討し、評価実験を実施

音源位置の再現性能の違い（音声信号のマッピング）

受信機側のスピーカ配置に即してダイアログなどのオブジェクトをマッピングする技術

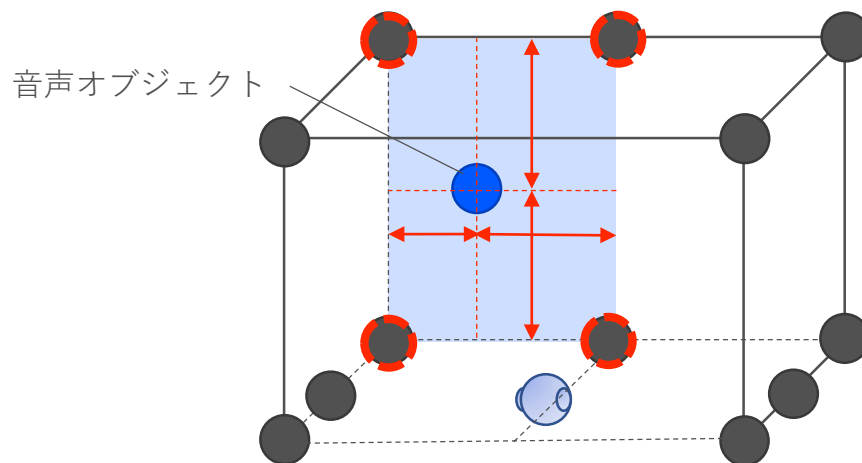
- MPEG-H 3DA : **3D VBAP** (Vector Based. Amplitude Panning)
 - 極座標をベースとし、3スピーカによる領域に分割し領域ごとに各スピーカに分配するゲインをベクトル重み付けにより決定
- AC-4 : **Triple Balance Panner**
 - 直交座標をベースとし、オブジェクトの座標に対しx軸、y軸、z軸の各スピーカに分配するゲインを決定

3D VBAP



音像を**3個のスピーカ**（部分的に4個）に信号を分配し音像を再生

Triple Balance Panner



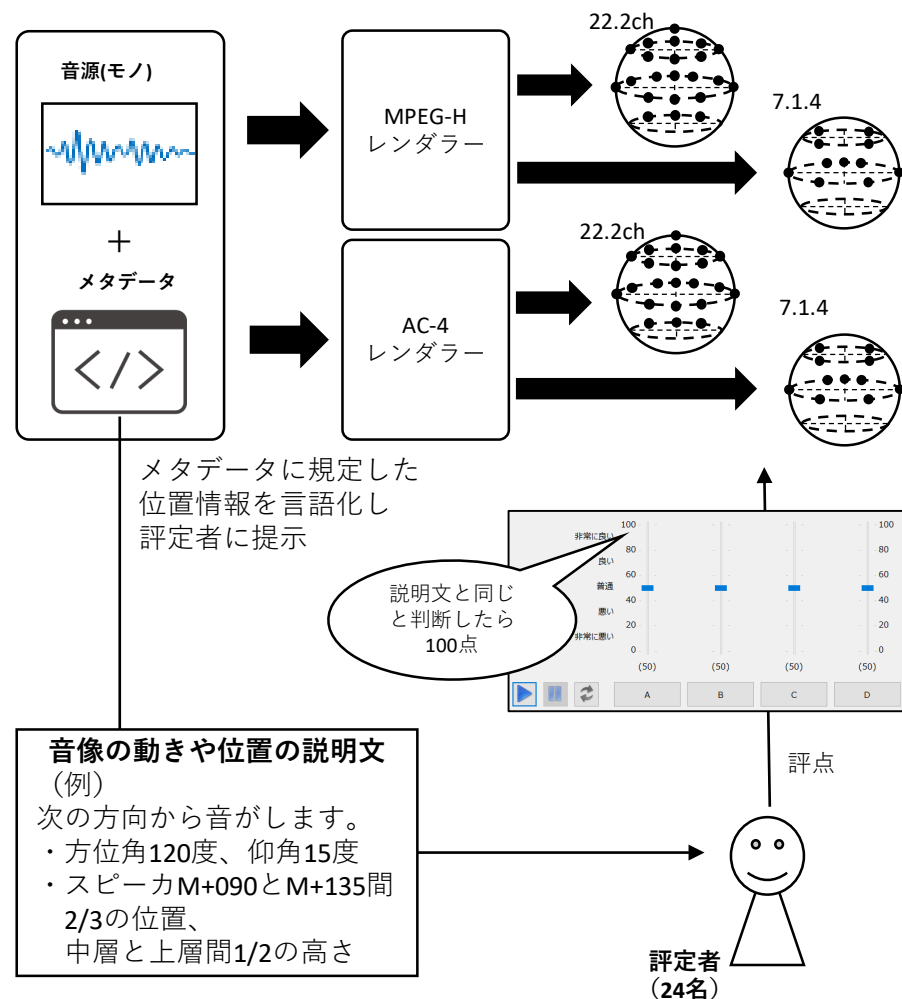
壁面上の音像を**4個のスピーカ※**に信号を分配し音像を再生
(※壁面：最大4個、立方体内部：最大8個)

実験方法（音源位置の再現性能）

勧告ITU-R BS.2132 多重刺激（基準音なし、アンカーなし）

- 音源は女声、ヴァイオリン、ノイズの3音源*1
- 評定者に音像の動きや位置に関する説明文（評価の基準）を事前に提示
- 4個の評価音源を聴取
 - 2方式(MPEG-H 3DA/AC-4)×2フォーマット(22.2ch/7.1.4)
- 聴取範囲の指定、および繰り返し聴取が可能
- 各評価音源を5段階尺度で説明文との違い（明瞭感や定位感などの総合印象）を評点

評点	評価語
100～81	非常に良い
80～61	良い
60～41	普通
40～21	悪い
20～ 0	非常に悪い

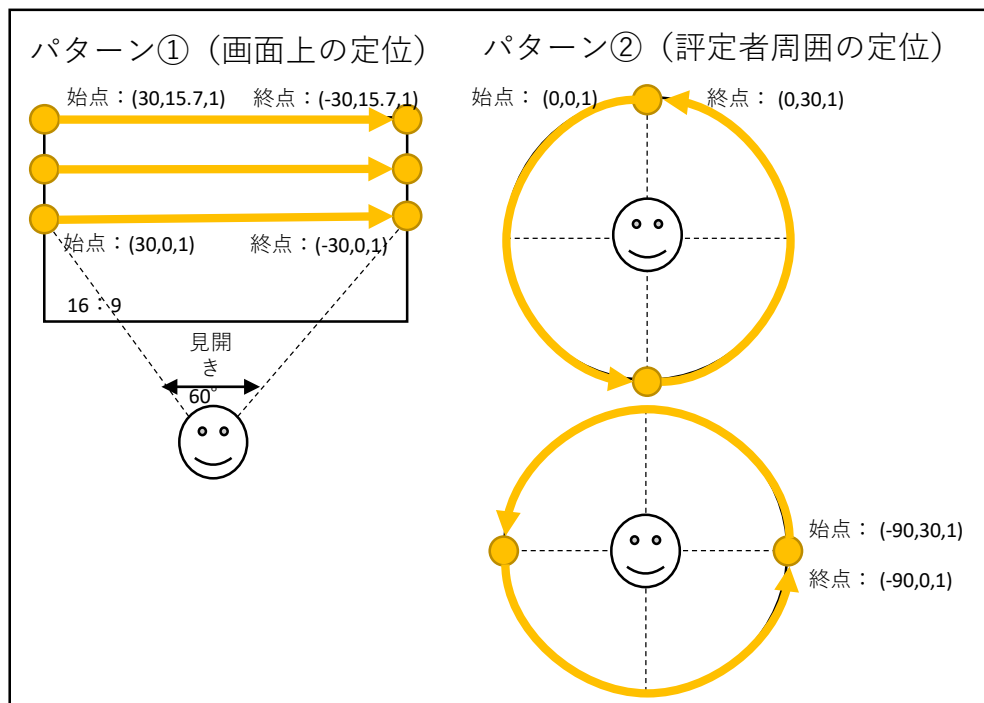


*1：同様の実験では一般的にノイズのような広帯域な音源を使用する。さらに、放送で多く使用される人声に加え楽音を使用した。

評価音源（音源位置の再現性能）

- 動きや位置を規定したメタデータを用いて22.2chと7.1.4のスピーカ配置に変換した際の品質を評価
- 上下と試聴位置周囲の動きと前方から側方/後方、上方まで3次元的な位置を網羅

音声オブジェクトが動くパターン×2



音声オブジェクトが静止しているパターン×7

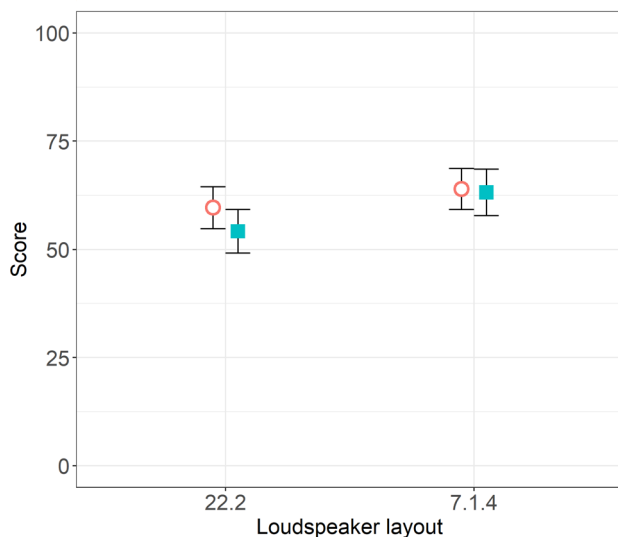
	極座標 (MPEG-H 3DA)	直交座標 (AC-4)	
パターン③	(0.0, 30.0, 1)	(0.00, 1.00, 1.00)	前方
パターン④	(15.0, 7.5, 1)	(-0.50, 1.00, 0.25)	
パターン⑤	(90.0, 22.5, 1)	(-1.00, 0.00, 0.75)	
パターン⑥	(120.0, 15.0, 1)	(-1.00, -0.67, 0.50)	側方 後方
パターン⑦	(165.0, 15.0, 1)	(-0.33, -1.00, 0.50)	
パターン⑧	(45.0, 60.0, 1)	(-0.50, 0.50, 1.00)	天井
パターン⑨	(180.0, 75.0, 1)	(0.00, -0.25, 1.00)	

※極座標：(方位角, 仰角, 距離)

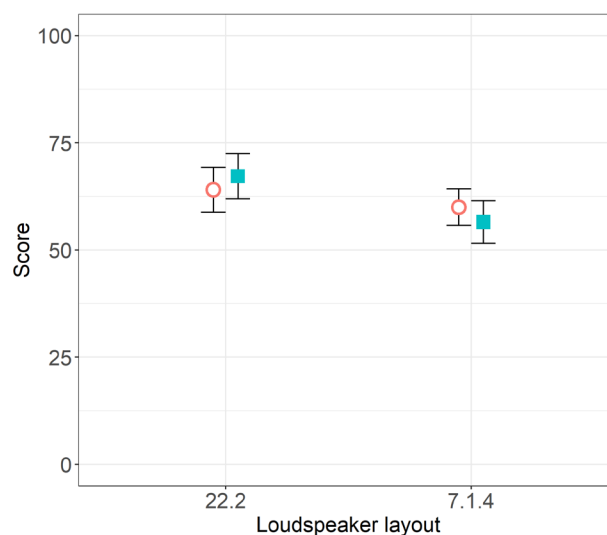
直交座標：(x, y, z)

実験結果（音源位置の再現性能）－動き/静止(前方・側方)

パターン①（画面上の動き）



パターン②（評定者周囲の動き）



Renderer

3DA

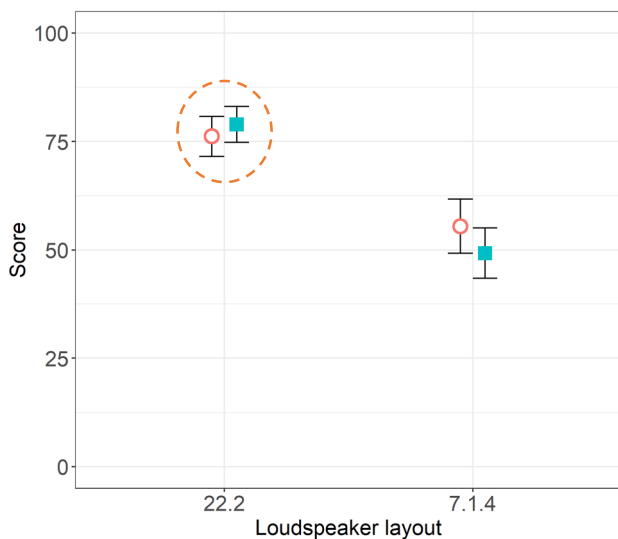
AC4

● : 95%信頼区間

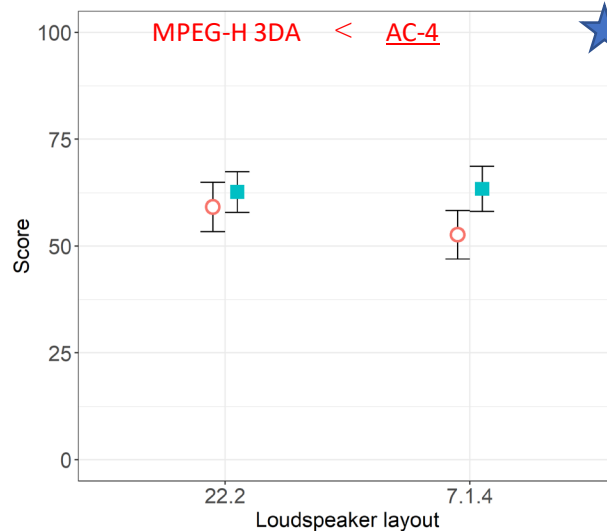
★ : 分散分析により符号化方式間の
主効果が認められ、
且つTukeyの検定で有意差が認め
られた実験条件

○ : 実スピーカ単体による再生

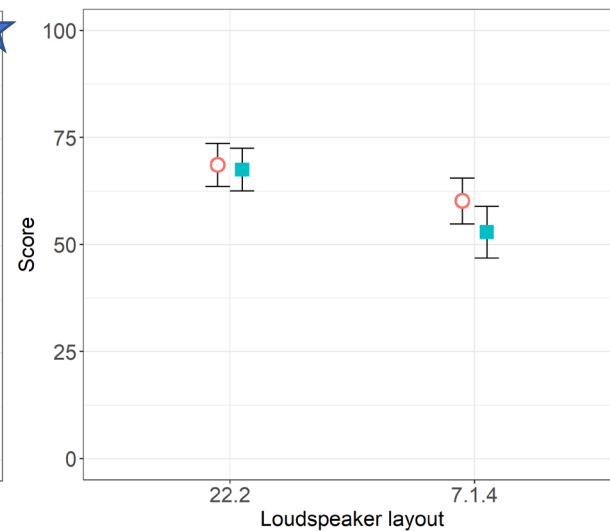
パターン③（方位角0、仰角30）



パターン④（方位角15、仰角7.5）

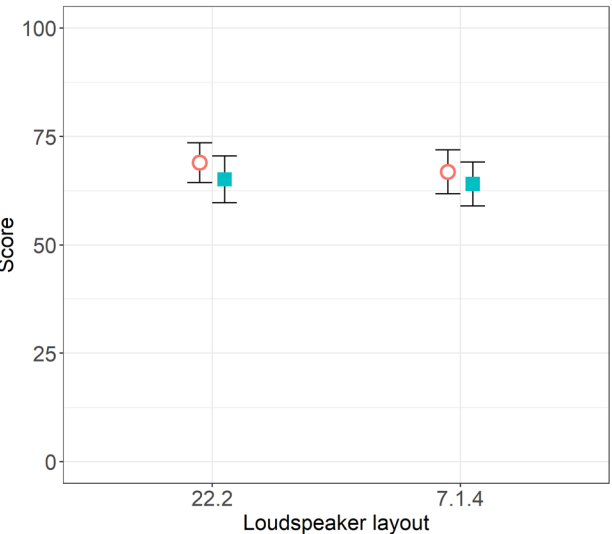


パターン⑤（方位角90、仰角22.5）

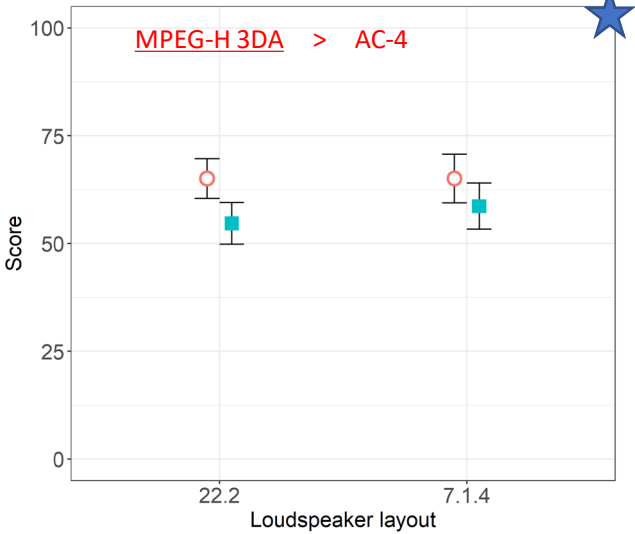


実験結果（音源位置の再現性能） – 静止(側・後方・天井)

パターン⑥（方位角120、仰角15）



パターン⑦（方位角165、仰角15）



Renderer

○ 3DA

■ AC4

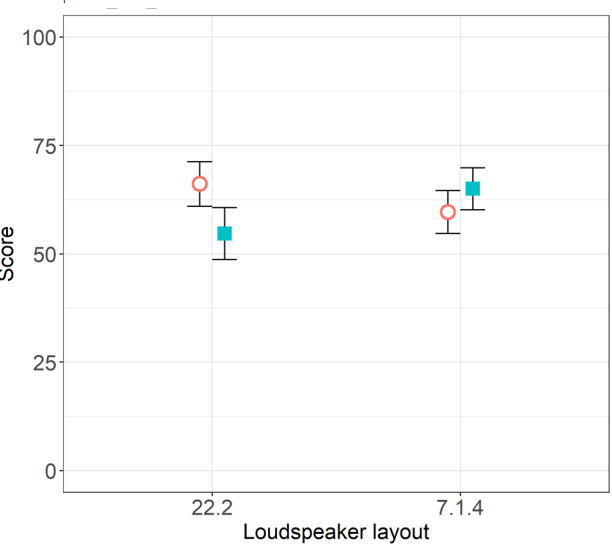
● : 95%信頼区間

★ : 分散分析により符号化方式間の主効果が認められ、且つTukeyの検定で有意差が認められた実験条件

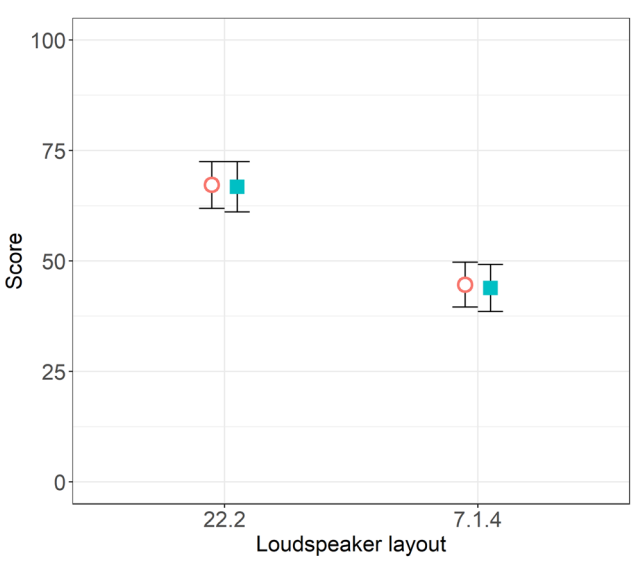
方向、スピーカ配置によって品質が高いパンニング則が異なる

➡常にどちらかの符号化方式が優れているとは言えない

パターン⑧（方位角45、仰角60）



パターン⑨（方位角180、仰角75）



再生環境への適応 (Egocentric/Allocentric)

スピーカ配置に対応してレンダリングする場合の基本的な考え方

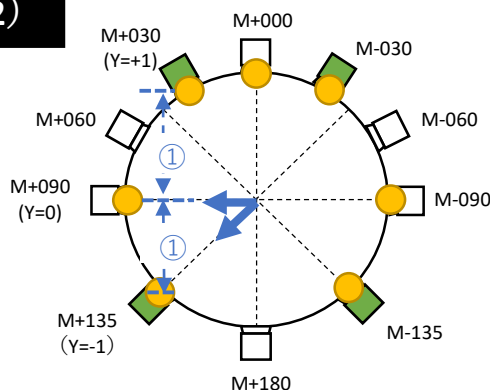
- MPEG-H 3DA : Egocentric

極座標をベースとし、聴取者からみた角度が保持されるようにレンダリング

- AC-4 : Allocentric

直交座標をベースとし、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

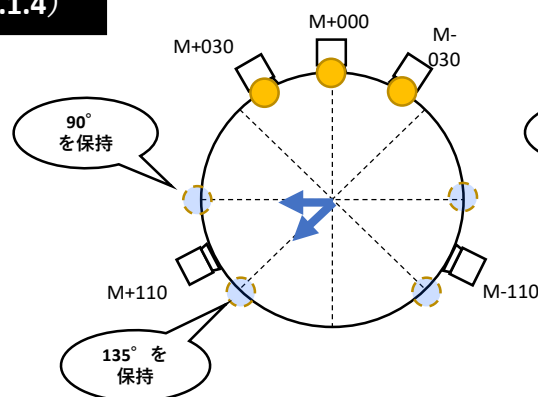
制作時 (22.2)



- : AC-4で部屋の四隅に設置するスピーカ
- : 単一スピーカで再生される音声オブジェクト
- : 複数スピーカで再生される音声オブジェクト (ファンタム)

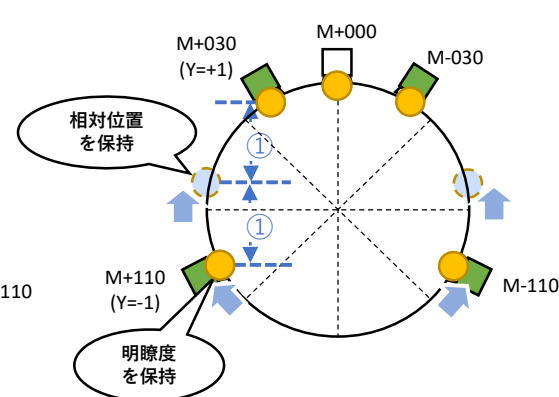
再生時 (5.1.4)

MPEG-H 3DA



視聴位置からの音の絶対位置は保持されるが明瞭度が下がることもある

AC-4

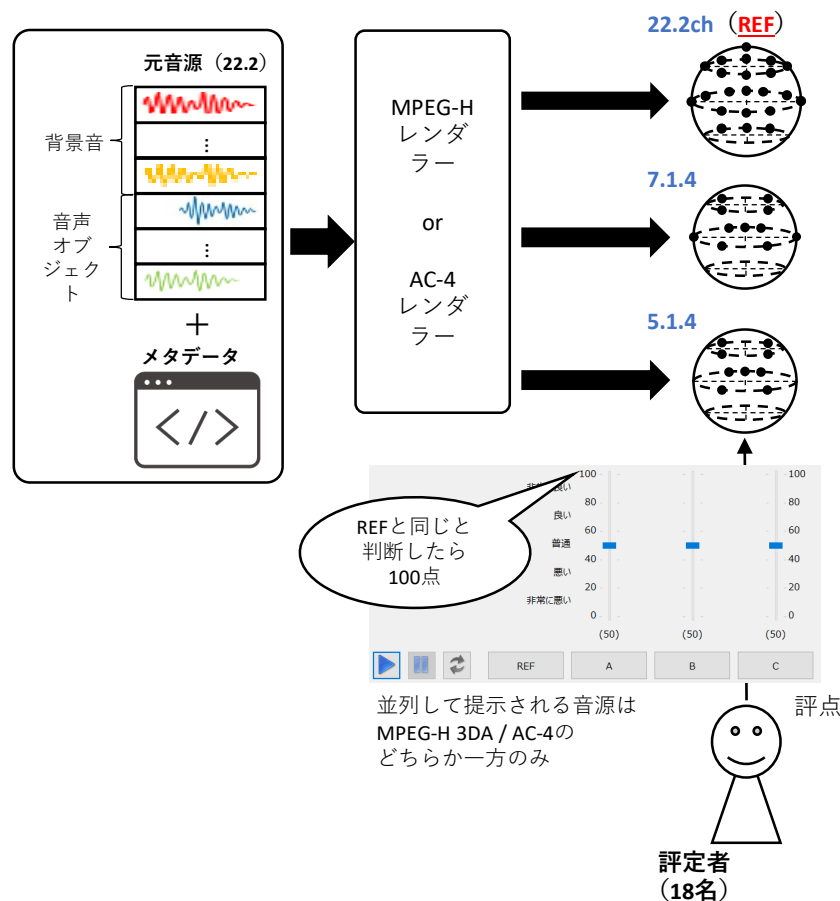


基準となる4隅のスピーカの信号については絶対的な位置が変わっても明瞭度を保持するが、再生音像の絶対位置がずれる場合がある

実験方法（再生環境への適応）

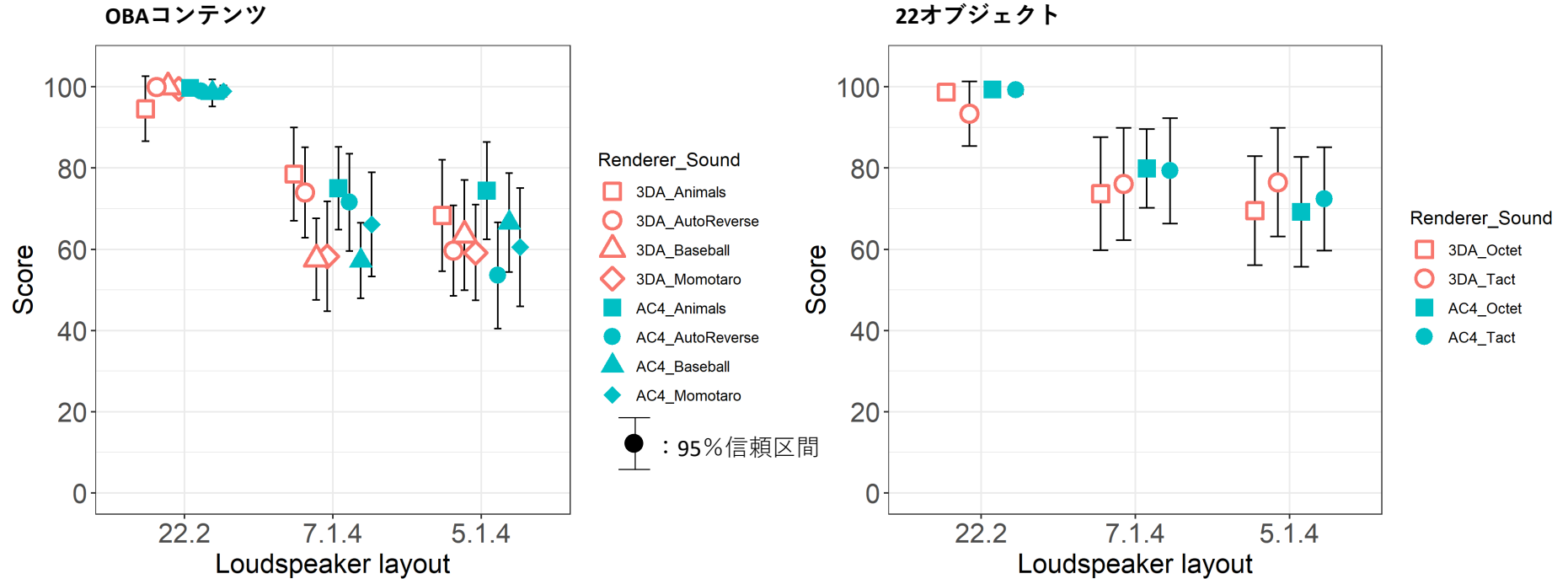
勧告ITU-R BS.2132 多重刺激（基準音あり、アンカーなし）

- 音源位置が明確な複数音源で構成されている番組選択
 - 22.2ch+オブジェクト音声で構成（OBA）
 - 22.2chのスピーカ信号を22オブジェクトとして構成
- 3個の評価音源をそれぞれREFと比較
 - ✓ REF [基準音] に22.2chのレンダリング音が提示される
 - ✓ 評価音の1つにREFと同じ22.2chのレンダリング音 [隠れ基準音]、他の2つに7.1.4/5.1.4のレンダリング音がランダムに提示される
- 聴取範囲の指定、および繰り返し聴取が可能
- 各評価音源を5段階尺度で基準音との印象差を評点



評点	評価語
100～81	非常に良い
80～61	良い
60～41	普通
40～21	悪い
20～ 0	非常に悪い

実験結果（再生環境への適応）



● 分散分析の結果、スピーカ配置、音源の主効果*¹が見られ、**符号化方式間に主効果が見られなかった。**

➡ **どちらかの符号化方式が優れてるとは言えない**

*¹：主効果がある場合、評価結果に影響する。主効果がない場合、評価結果に影響を与えない。

音源内容

Animal：複数の動物の鳴き声とナレーション（オブジェクト）とジャングルの環境音など（22.2ch）

AutoReverse：動き回るドラマの登場人物のセリフ（オブジェクト）と環境音など（22.2ch）

BaseBall：実況やヤジなど（オブジェクト）会場の歓声（22.2ch）

Momotaro：桃太郎と犬、サル、キジのセリフ（オブジェクト）と音楽などの背景音（22.2ch）

Octet：異なる方向から聞こえる弦楽器による八重奏（22.2chミックスをオブジェクトで実験）

Tact：オーケストラに加えて様々な効果音が動き回る（22.2chミックスをオブジェクトで実験）

レンダリング技術の評価実験まとめ

- 音源位置の再現性能
 - 一部に分散分析により音声符号化方式間の主効果があり、Tukey検定で有意差が認められた
 - パターン④：AC-4
 - パターン⑦：MPEG-H 3DA
- ⇒一方の音声符号化方式が優れた評価とはなっていない
- 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
- ⇒座標系を制作者の意図を保持したまま相互変換することは困難であり、運用にあたって**どちらか一方の方式を採用することが望ましい**
- 再生環境への適応
 - 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

音像定位やコンテンツから受ける印象は、スピーカ配置、音源による影響が大きく
定位方向・スピーカ配置によって、品質の高い符号化方式は異なる

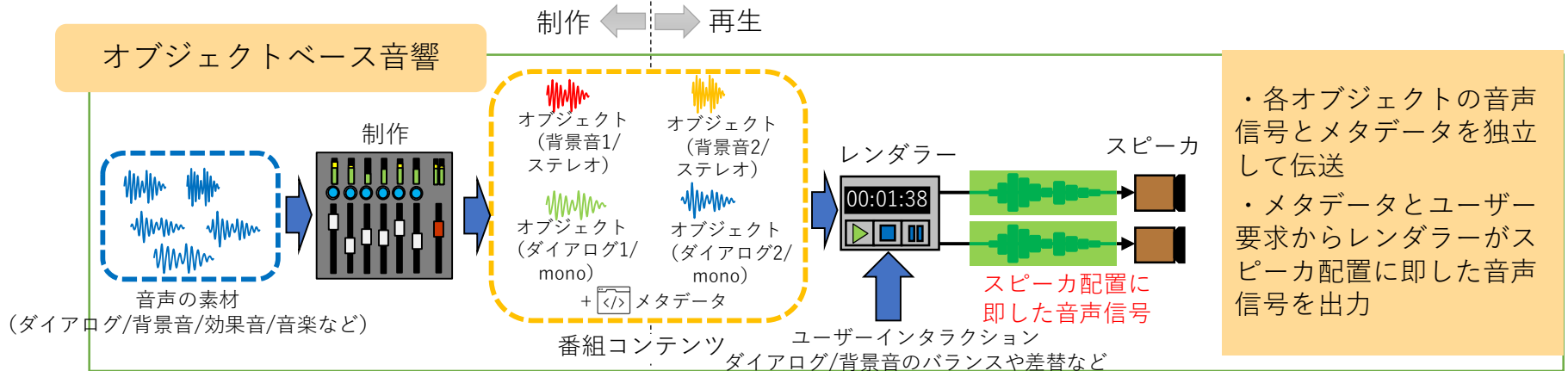
➡ 常にMPEG-H 3DAとAC-4のどちらかの方式が優れているとは言えない
(但し、ダイアログの明瞭度が保持される方が好まれる傾向がある)

各符号化方式の比較表

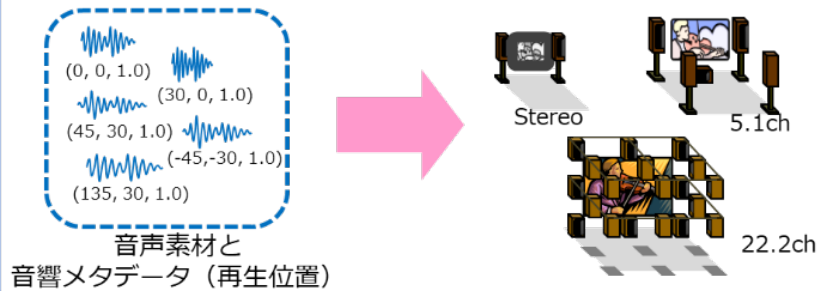
項目	地上デジタルテレビジョン方式の高度化の要求条件	MPEG-2/4 AAC	MPEG-H 3D Audio	Enhanced AC-3	AC-4
放送規格		ARIB DVB SBTVD	ATSC DVB SBTVD	ATSC DVB SBTVD	ATSC DVB SBTVD
主な採用国 ※ 放送サービスを実施 無印は規格採用国		日本※ ブラジル※	ブラジル 韓国※	ブラジル※ 欧州※	ブラジル 米国※、カナダ、メキシコ 欧州
同時再生可能な音声信号数 ※括弧内はサービス例	音声入力フォーマット及び符号化方式 音質	MPEG-2：8 (例 7.1ch) MPEG-4：24 (例 22.2ch)	レベル3： 16 (例 7.1.4+4obj)、 24 (モノオブジェクトのみ) レベル4： 28 (例 22.2ch+4obj)	16 (例 7.1.4+4obj)	レベル3： 18 (例 7.1.4+6obj) レベル4：参考プロトタイプ 28(例 22.2ch+4obj)
再生機能	高機能化/多様化サービス	MPEG-4：チャンネル差替/ 音量調整 (機能制限・制作制約あり)	オブジェクト差替/明瞭化/ 音量調整	オブジェクト差替/明瞭化/ 音量調整 (機能制限あり)	オブジェクト差替/明瞭化/ 音量調整
放送品質を満足するビットレート *1：実験で満足するレートが見いだせなかった *2：実験音源が符号化しやすい音源であったため参考		22.2ch：*1 7.1.4：(未対応) 5.1ch：*1 ステレオ：*1 モノ：64kbps	22.2ch：512kbps 7.1.4：192kbps*2 5.1ch：*1 ステレオ：96kbps モノ：64kbps	AACと比較して 同程度～やや劣化	22.2ch：768kbps 7.1.4：256kbps*2 5.1ch：*1 ステレオ：*1 モノ：48kbps
レンダリング技術		—	3D-VBAP/Egocentric	Triple Balance Panner/ Allocentric	Triple Balance Panner/ Allocentric
ネット連携サービス/ モバイル規格ほか	高機能化/多様化	3GPP/DASH-IF/ Hybridcast	3GPP/DASH-IF/HbbTV	DASH-IF/HbbTV/ Hybridcast	DASH-IF/HbbTV
HDMI伝送規格	共通性/ インターフェース	IEC 61937-6 IEC 61937-11	IEC 61937-13	IEC 61937-3	IEC 61937-14 (IEC61937-9への変換も可能)
実用化状況		国内の衛星放送用TV、 地上放送用TV	韓国発売のTV、STB、 サウンドバー、AVR	国内のTV、STB、スマート フォン、サウンドバー、 AVR	国内のTV、スマートフォン (MAT変換によりデコーダを 搭載しないサウンドバー、 AVRで再生可能)
ライセンス		RANDライセンス	RANDライセンス	RANDライセンス	RANDライセンス

- 參考資料

オブジェクトベース音響（OBA）とサービスイメージ

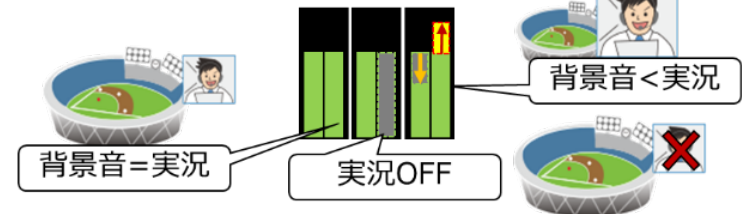


再生環境（スピーカ配置）への最適化



アクセシビリティ改善

- ・ダイアログの音量だけを大きく調整
- ・音声解説の挿入
- ・セカンドデバイス（手元のスピーカで再生）



視聴者の好みに合わせたコンテンツ

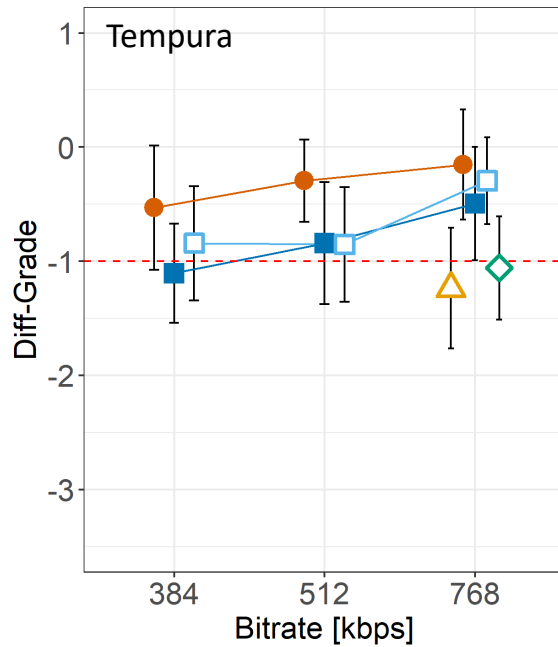
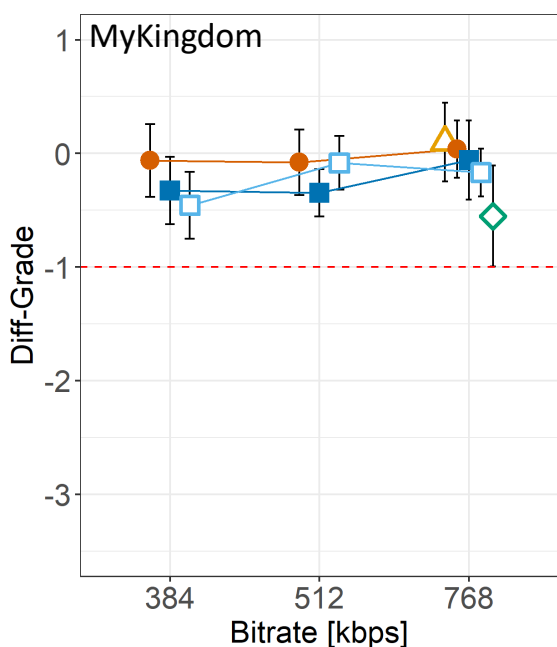
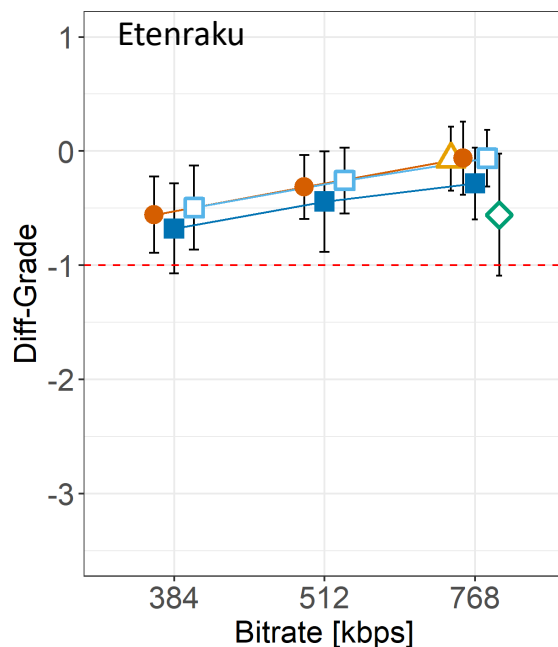
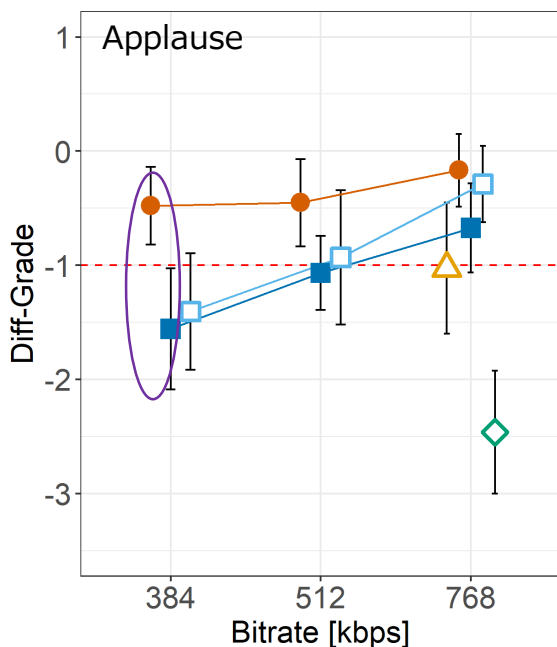
- ・多言語、裏トークなどの多様な副音声（ダイアログ差替え）
- ・ホーム&アウェイ



- ・視点切替
- ・アレンジ違いの音楽（背景音差替え）
- ・ローカル局や課金サービスなどによる音声素材の追加



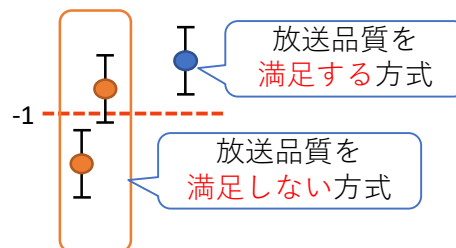
ビットレートと品質の結果 (22.2ch)



Codec

- △ MPEG-4 AAC
- MPEG-H 3DA
- AC-4 Level3
- AC-4 Level4
- ◇ Enhanced AC-3

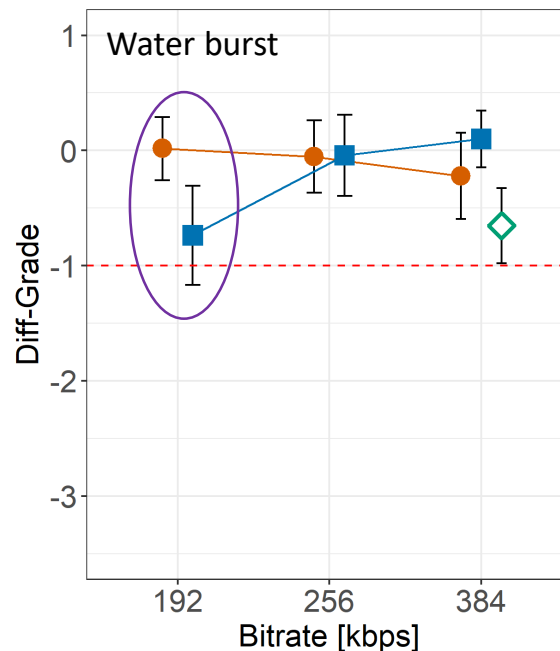
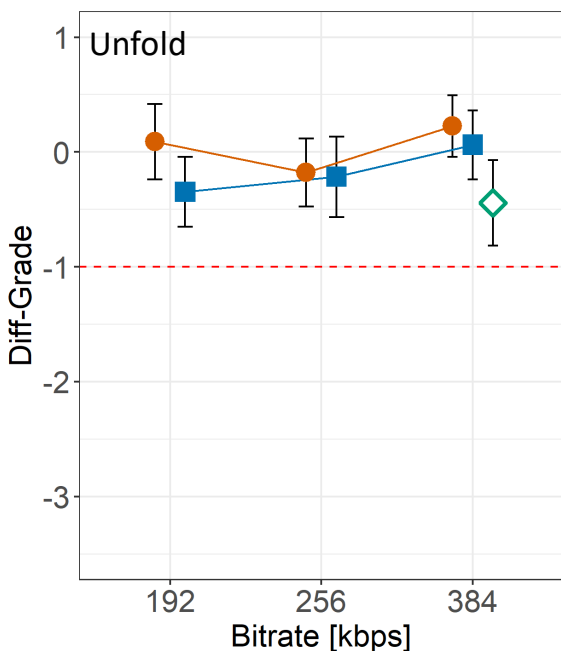
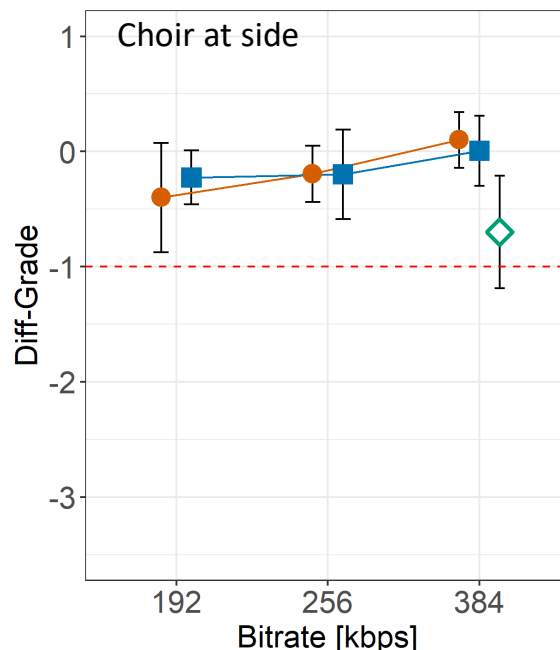
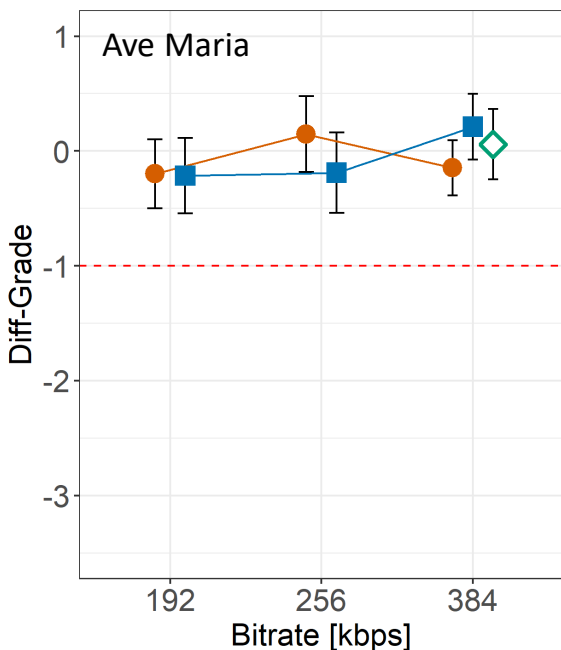
95% 信頼区間
(母集団の平均値が
存在する範囲)



放送品質を満足する所要ビットレート：
すべての音源で差分評価値 (Diff-Grade) の
95%信頼区間が-1を超えているビットレート

- 最も符号化の難しい音源は天ぷら
- 放送品質を満足するビットレート
 - MPEG-H 3DA 512kbps
 - AC-4 Level4 768kbps
- 同一ビットレートで有意差のある方式
 - 384kbps (拍手)
MPEG-H 3DA > AC-4 Level3

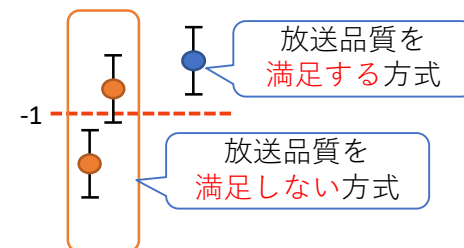
ビットレートと品質の結果 (7.1.4)



Codec

- △ MPEG-4 AAC
- MPEG-H 3DA
- AC-4 Level3
- AC-4 Level4
- ◇ Enhanced AC-3

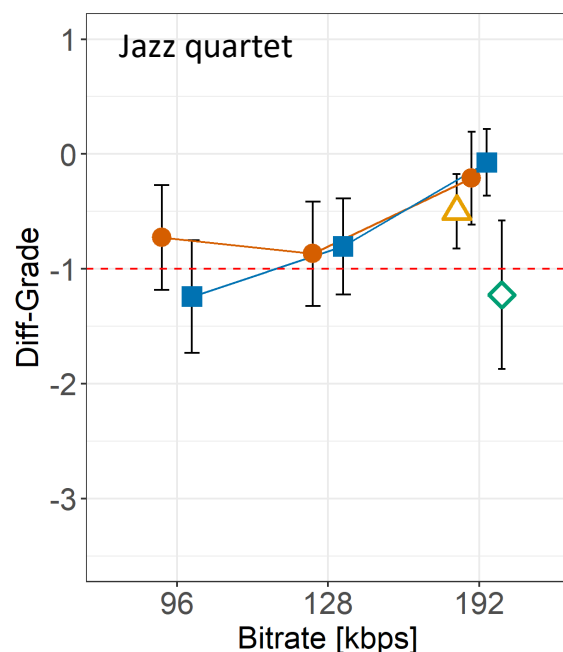
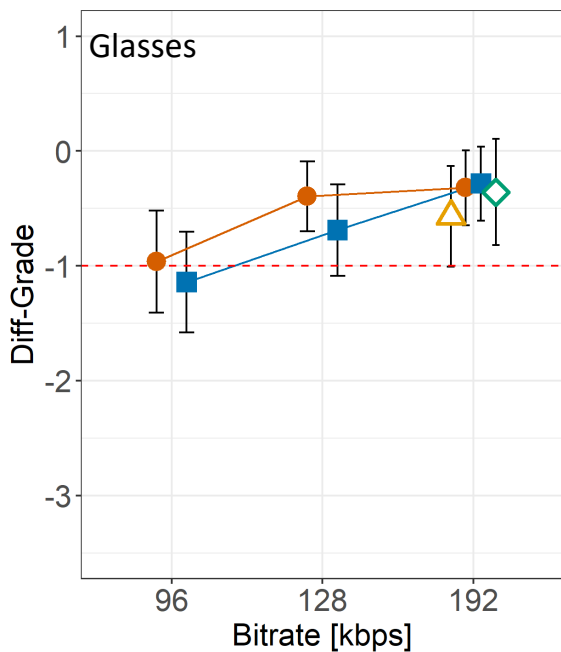
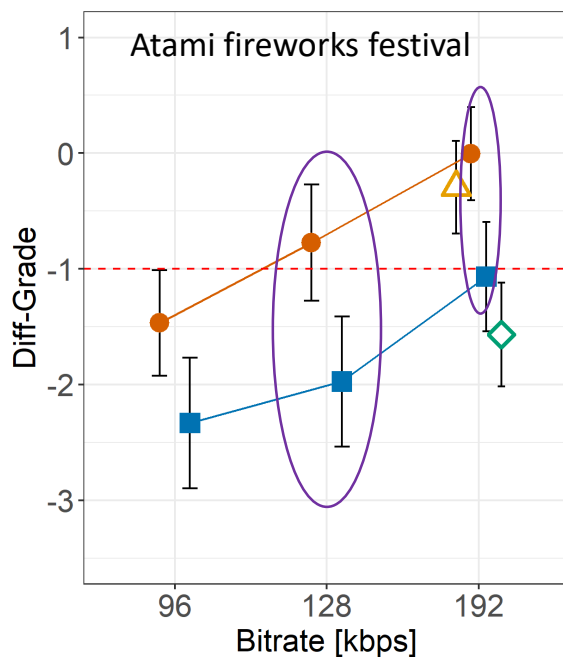
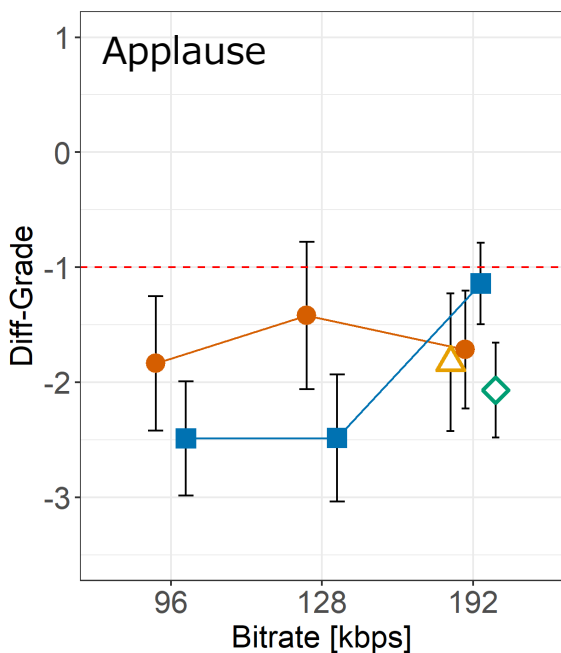
95% 信頼区間
(母集団の平均値が
存在する範囲)



放送品質を満足する所要ビットレート：
すべての音源で差分評価値 (Diff-Grade) の
95%信頼区間が-1を超えているビットレート

- 最も符号化の難しい音源はWater burst
- 放送品質を満足するビットレート
 - MPEG-H 3DA 192kbps
 - AC-4 Level3 256kbps
- 符号化しやすい音源のため参考値
- 同一ビットレートで有意差のある方式
 - 192bps (Water burst)
MPEG-H 3DA > AC-4 Level3

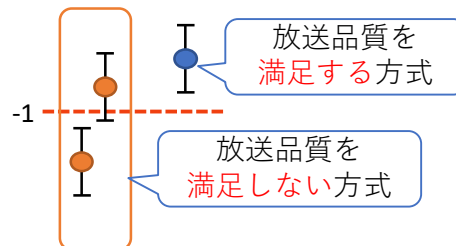
ビットレートと品質の結果 (5.1ch)



Codec

- △ MPEG-4 AAC
- MPEG-H 3DA
- AC-4 Level3
- AC-4 Level4
- ◇ Enhanced AC-3

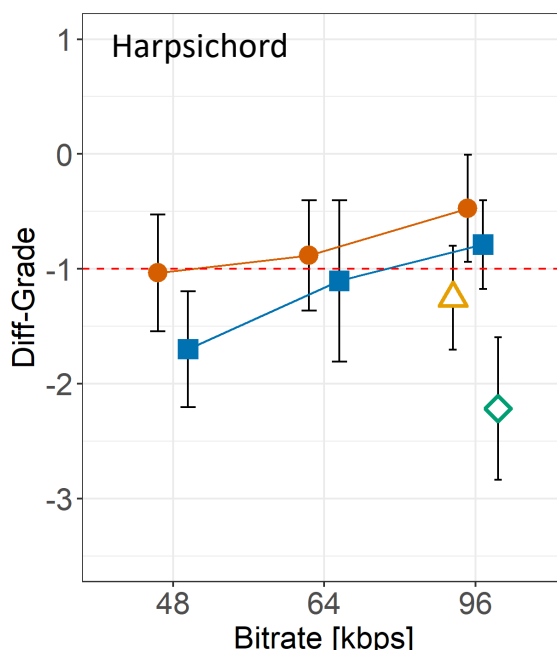
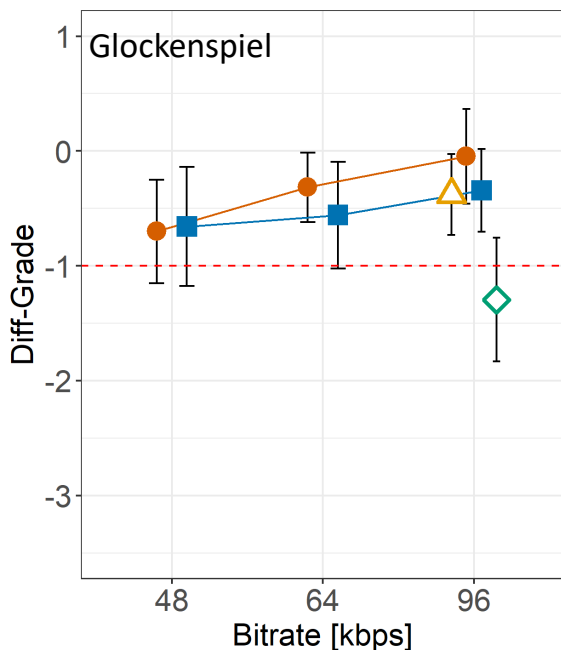
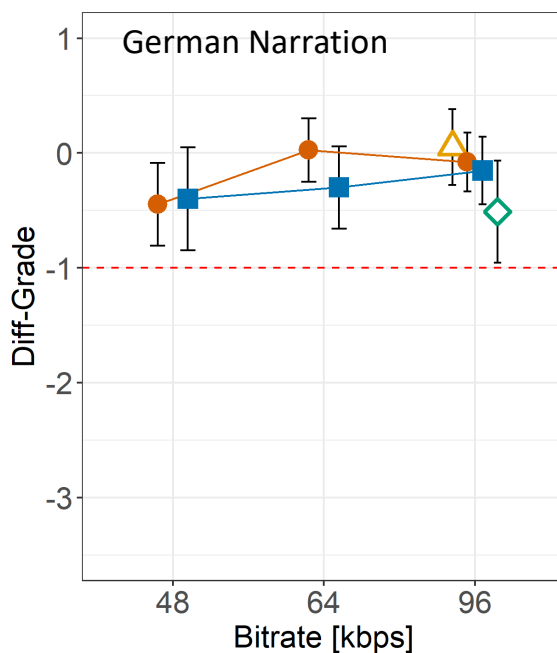
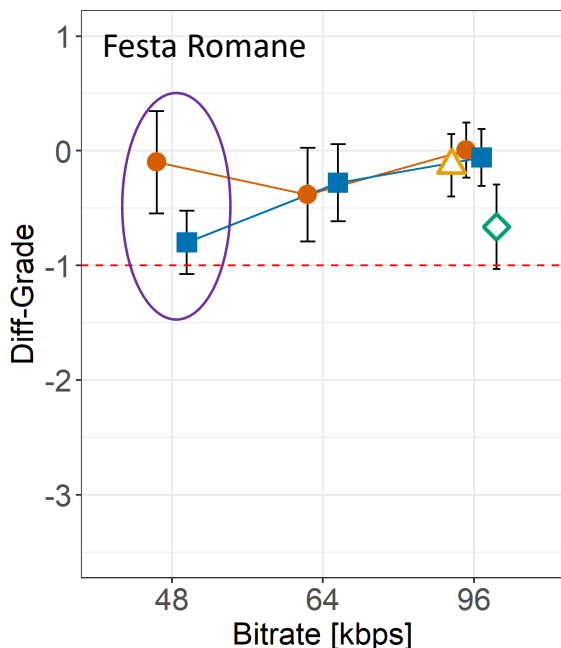
95% 信頼区間
(母集団の平均値が
存在する範囲)



放送品質を満足する所要ビットレート：
すべての音源で差分評価値 (Diff-Grade) の
95%信頼区間が-1を超えているビットレート

- 最も符号化の難しい音源は拍手
- 放送品質を満足するビットレート
 - 見だせなかった
 - 符号化が難しい音源
- 同一ビットレートで有意差のある方式
 - 128kbps (花火)
MPEG-H 3DA > AC-4 Level3
 - 192kbps (花火)
MPEG-H 3DA > AC-4 Level3

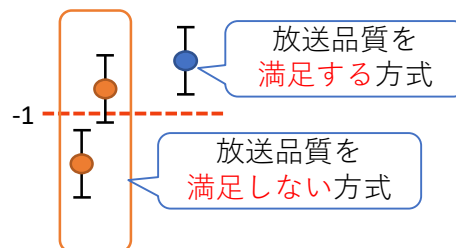
ビットレートと品質の結果（ステレオ）



Codec

- △ MPEG-4 AAC
- MPEG-H 3DA
- AC-4 Level3
- AC-4 Level4
- ◇ Enhanced AC-3

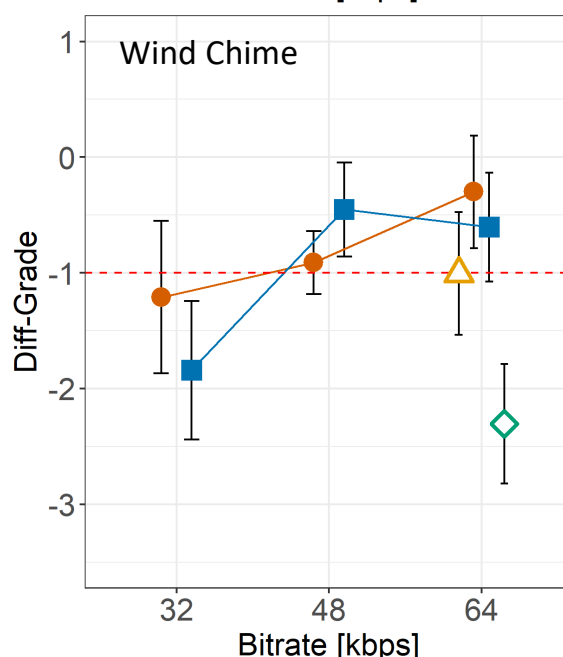
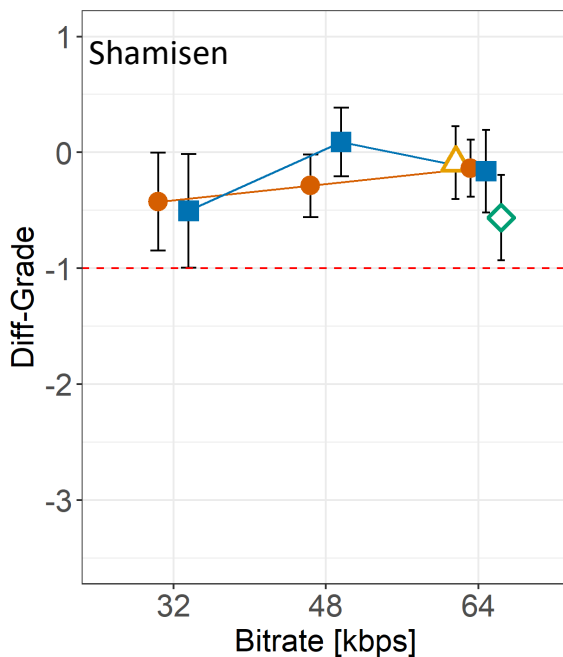
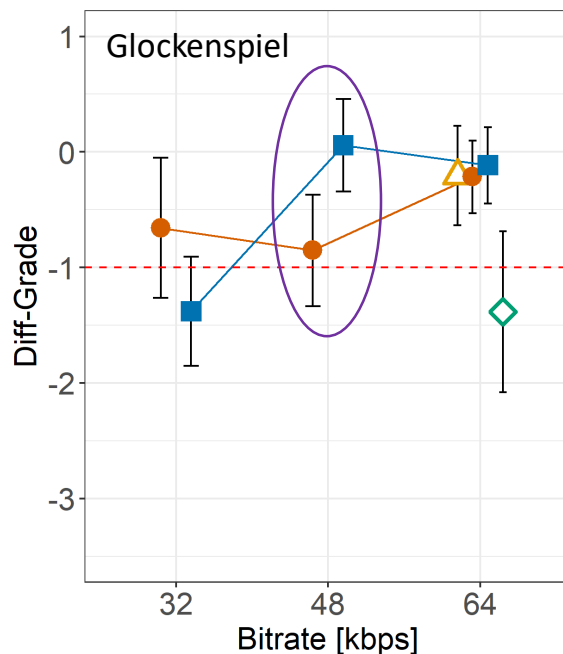
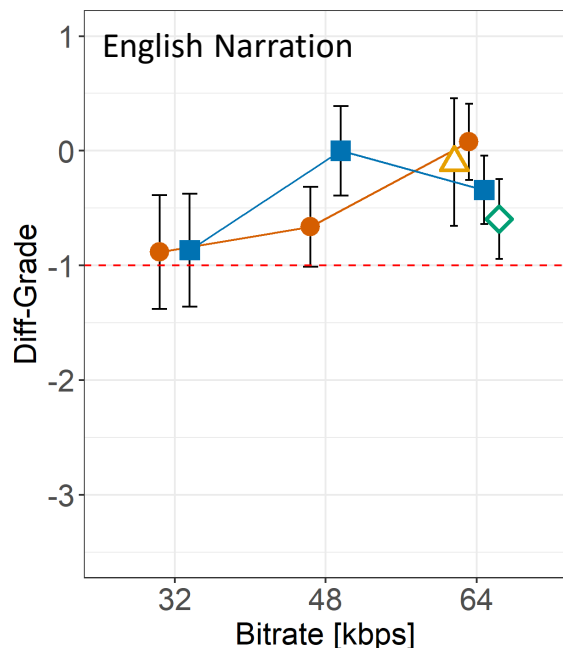
95% 信頼区間
(母集団の平均値が
存在する範囲)



放送品質を満足する所要ビットレート：
すべての音源で差分評価値（Diff-Grade）の
95%信頼区間が-1を超えているビットレート

- 最も符号化の難しい音源は Harpsichord
- 放送品質を満足するビットレート
 - MPEG-H 3DA 96kbps
- 同一ビットレートで有意差のある方式
 - 48kbps (Festa Romane)
MPEG-H 3DA > AC-4 Level3

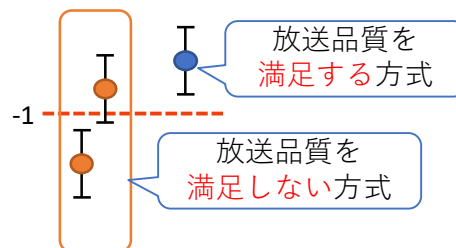
ビットレートと品質の結果 (モノ)



Codec

- △ MPEG-4 AAC
- MPEG-H 3DA
- AC-4 Level3
- AC-4 Level4
- ◇ Enhanced AC-3

95% 信頼区間
(母集団の平均値が
存在する範囲)



放送品質を満足する所要ビットレート：
すべての音源で差分評価値 (Diff-Grade) の
95%信頼区間が-1を超えているビットレート

- 最も符号化の難しい音源は
Wind Chime
- 放送品質を満足するビットレート
 - AC-4 48kbps
 - MPEG-H 3DA 64kbps
- 同一ビットレートで有意差のある方式
 - 48kbps (Glockenspiel)
AC-4 Level3 > MPEG-H 3DA