7章 次世代音声符号化方式の開発動向

杉本 岳大[†] (正会員)

キーワード：音声符号化方式，MPEG-H 3D Audio，標準化，主観評価，放送方式

1. まえがき

放送に係る音声符号化方式は，各国のデジタル放送の浸透に合わせて大きく進化してきた．日本では2000年に開始されたBSデジタル放送の音声符号化方式として，当時最新の技術であったMPEG (Moving Pictures Expert Group) -2 AAC (Advanced Audio Coding)¹⁾が採用された²⁾．MPEG-2 AACは，2003年に開始され順次カバーエリアが拡大された現行の地上デジタル放送でも採用されている²⁾．

2018年12月に放送開始予定の新4K8K衛星放送³⁾では，音声符号化方式としてMPEG-4 AAC⁴⁾が採用された．日本ではMPEG-2 AAC，MPEG-4 AACともにLC (Low Complexity) プロファイル (MPEG-4 AACではオーディオオブジェクトタイプと呼ぶ) が採用されており，基本的な符号化ツールが同じであるため，圧縮効率に違いはない．最大の相違点は，MPEG-4 AACが22.2マルチチャンネル音響 (22.2 ch 音響)⁵⁾をはじめとした，5.1サラウンドを超えるチャンネル数の音声フォーマットを送送できることにある．MPEG-4 AACでは，ダウンミックスによる音声フォーマット変換やダイアログ制御情報の伝送なども可能ではあるものの²⁾，あくまでチャンネルベース音響*1として実現可能な範囲に限られる．

ところが近年，チャンネルベース音響が中心だった放送における音響方式が拡張されつつある．ITU-R (International Telecommunication Union Radiocommunication Sector，国際電気通信連合無線通信部門) では番組制作のための先進的な音響システムに関するRec. ITU-R BS.2051-1⁵⁾が策定された．ここでは22.2 ch 音響を最上位のフォーマットとするチャンネルベース音響に加え，オブジェクトベース音響*2，

シーンベース音響*3の3つの音響方式が規定されている．既存の5.1サラウンドなどと比較すると，チャンネル配置を水平面内に限定せずに垂直方向にも拡張することを視野に入れているのが特徴である．

Rec. ITU-R BS.2051-1は番組制作に関わる規定であるが，この機運に呼応するように2010年代中頃から新しい音響方式の伝送のための音声符号化方式の開発が盛んになった．主な技術としてはMPEGで標準化されたMPEG-H 3D Audio (3DA)⁶⁾とDolby Laboratoriesが開発したAC-4⁷⁾⁸⁾が挙げられる．両技術とも，5.1サラウンドを超えるチャンネル数に対応する点，音声信号をハンドリングするためのメタデータの伝送が可能である点，ダイアログ制御やラウドネスコントロールなどのユーザフレンドリーな機能を備えている点が特徴である．本稿では，次世代の音声符号化方式の一例としてMPEG-H 3DAに着目し，標準化のプロセス，規格の概要と性能，装置の開発について紹介する．

2. MPEG-H 3DAの概要

2.1 MPEG-H 3DAの標準化

MPEG-H 3DAの標準化は，前述した先進的な音響システムの圧縮・伝送に関わる音声符号化方式の策定を目的として，2013年から開始された．2015年に規格のベースとなるMain Profileが発行され⁶⁾，その後2017年に放送用途など軽い処理負荷を想定したシステムへの実装を目的としたLCプロファイルが発行された⁹⁾．

MPEG-H 3DAに採用されている低ビットレート用音質改善ツールは多数あるが，代表的なものとしては，入力信号を人声とそれ以外の音声とに分けて各々に適した符号化アルゴリズムによって圧縮効率を向上させる手法，圧縮によってスペクトルが欠落した帯域への復号時のノイズフィリング技術 (Intelligent Gap Filling)，チャンネル間でのマスキング効果を利用したビットのダイナミックな分配法 (Multichannel Coding Tool) などがあり，どれもMPEG-4

*1 規定のチャンネル配置を対象とした音響方式．

*2 音声信号およびメタデータで記述された音情報を再生機側でレンダリングする音響方式．

[†] NHK 放送技術研究所

"Development Trend of Next-generation Audio Coding Technology" by Takehiro Sugimoto (Science and Technology Research Laboratories, NHK, Tokyo)

*3 音場全体の物理的な情報を記録し，再生条件に合わせてレンダリングする方式．高次アンビソニックス (Higher Order Ambisonics: HOA) が代表的な技術．

図1 MPEG-H 3DAデコーダの構成

AACには搭載されていない技術である。なおアルゴリズムの詳細については規格書を参照されたい⁶⁾⁹⁾。

MPEG-H 3DAは米国の次世代放送規格であるATSC 3.0に採用されたほか¹⁰⁾、2017年開始の韓国の次世代地上放送でも用いられており、次世代放送用の有力な音声符号化方式の1つとして実用化が進んでいる。

2.2 MPEG-H 3DAの構成

図1にMPEG-H 3DAデコーダの構成を示す。圧縮音声のストリームは、まずCore Decoderに入力され、復号される。復号された音声信号は、音響方式別に以下のようにレンダリングされる。

- (1) チャンネルベース音響の信号は、Format Converterに入力され、再生環境のスピーカ数やスピーカ配置に合わせてダウンミックスおよびレンダリングされる。
- (2) オブジェクトベース音響の信号は、音響オブジェクトの動作や位置などを記述するメタデータとともにObject Rendererに入力され、再生環境のスピーカ数やスピーカ配置に合わせてレンダリングされる。
- (3) シーンベース音響の信号(MPEG-H 3DAではHOA)は、HOAの次数に基づきHOA Rendererによって再生環境のスピーカ数やスピーカ配置に合わせてレンダリングされる。

レンダリングステージに続くMixerでは、必要に応じてレンダリング済みの音声信号同士を重ね合わせる。一例を挙げるなら、チャンネルベース音響の背景音に対して複数のダイアログの音響オブジェクトを重ね合わせて再生する場合である。

スピーカで再生する場合(図1の左側のパス)、Mixerからの出力はDynamic Range Controlに直接入力される。Dynamic Range Controlでは聞き取りやすさ向上のための

表1 LCプロファイルの各レベルとその制約条件

レベル	サンプリング 周波数 [kHz]	取り扱い可能な最大音声信号数		最大音声 フォーマット
		伝送	同時復号	
1	48	10	5	ステレオ
2	48	18	9	7.1
3	48	32	16	7.1.4
4	48	56	28	22.2
5	96	56	28	22.2

ダイナミックレンジ圧縮などを施し、スピーカ再生用の音声信号を出力する。一方、再生デバイスがヘッドホンやイヤホンの場合には(図1の右側のパス)、Binaural RendererでMixerからの出力にバイノーラル室内インパルス応答(Binaural Room Impulse Response: BRIR)を畳み込むことで、擬似的に3次元音響を再現可能なバイノーラル信号を生成する。その後Dynamic Range Controlの処理を適用して、ヘッドホン用の音声信号を出力する。

2.3 LCプロファイルのレベル

3次元音響の空間表現能力や音響オブジェクトを用いた機能性は、原理的に取り扱える音声信号の数が多いほど向上する。しかしながら、実用的な符号化方式の規格としては受信機や音響機器への実用的な実装条件を提示する必要があるため、MPEG-H 3DAではLCプロファイル内に処理負荷量に応じて5つのレベルを設けている。

表1にLCプロファイルの各レベルとその制約条件を示す。取り扱い可能な最大音声信号数における伝送と同時復号の数値が異なるのは、多言語放送におけるダイアログの差し替えやホーム&アウェイ^{*4}においては、伝送された音響オブジェクトを必ずしもすべて同時に再生する必要がないことによる。

3. MPEG-H 3DAによる圧縮符号化音の品質評価

本節では、MPEG-H 3DAで圧縮符号化した22.2ch音声信号の音質を主観評価し、放送品質を満たすビットレートを導出した検討¹¹⁾を紹介する。

3.1 主観評価の手続き

実際の主観評価では、符号化劣化を検出しやすい評価音源を3種選んだが、本稿ではその中から典型的な例として拍手を取り上げる。評価するビットレートは22.2chトータルで288, 512, 704, 880, 1200 kbpsの5種とした。評定者は音声符号化技術の非専門家12名、年齢は20～40代であり、実験後にポストスクリーニングで2名を除外した。またMPEG-4 AAC(LCプロファイル)による圧縮符号化音も合わせて評価し、MPEG-H 3DAと比較した。

*4 スポーツ番組などで、制作側から提供された各チームまたは各選手の応援に特化した複数のダイアログや歓声を、視聴者が好みに応じて受信機側で自由に選択できるサービス。

評価手法は、Rec. ITU-R BS.1116-3¹²⁾に規定されている隠れ基準付き三刺激二重盲検法を採用した。評定者は非圧縮符号化音を基準音として、圧縮符号化音の劣化度合いを、表2に示す5段階の劣化尺度に沿って小数点以下一桁の精度で評価した。主観評価の各試行ではR, A, Bの3種の評価音が提示される。Rが基準音であることは既知とし、A, Bのどちらか一方には基準音と同じ非圧縮符号化音、もう一方には圧縮符号化音が試行毎にランダムに割り当てられる。評定者はA, B両方の基本音声品質を評価し、基準音と同一と判断した評価音（隠れ基準音）には5.0を、圧縮符号化音と判断した評価音には4.9以下を与えることとした。聴取回数に制限はなく、自由に評価音を切り替えながら試聴可能とした。評価は一人ずつ実施し、聴取位置は22.2 ch 音響配置の中心とした。

3.2 主観評価の結果

図2に拍手を音源とした実験結果を、95%信頼区間と平均値で示した。実験結果は、圧縮符号化音の評点から隠れ基準音の評点を差し引いた主観評価値 (Diff-Grade) で示した。なおRec. ITU-R BS.1548-4¹³⁾によると、デジタル放送における放送品質の要求条件は、符号化劣化を検知しやすい音源の主観評価値の平均値が-1.0以上であることと規定されている。実験結果から、MPEG-H 3DAは512 kbps以上で放送品質を満たすことがわかる。一方MPEG-4 AACは、放送品質を満たすためには1200 kbps以上のビットレートが必要であり、MPEG-H 3DAはMPEG-4 AACと比較して圧縮効率が大幅に向上していることが確認できた。

表2 5段階劣化尺度

5.0	違いを検知できない
4.0	違いを検知できるが気にならない
3.0	違いがやや気になる
2.0	違いが気になる
1.0	違いが非常に気になる

図2 拍手による実験結果

4. MPEG-H 3DAを用いたリアルタイム符号化・復号装置の開発

NHKでは地上波によるスーパーハイビジョン放送の実現を目指し、地上放送技術の高度化をメーカーや大学と連携して進めている。2018年度は実験試験局を使用した都市部での大規模な実証実験を実施し¹⁴⁾、これまでに検討してきた伝送方式の有効性を検証・評価して課題抽出を行う予定になっている。本節では、実証実験に向けて開発したMPEG-H 3DAを用いたリアルタイム符号化・復号装置¹⁵⁾の仕様と構成について紹介する。

4.1 装置の仕様

近年のCPUの性能向上のおかげで、汎用PCでも音声信号のリアルタイム符号化・復号処理が充分可能な条件が整ってきた。そこでファームウェア更新の自由度と機器整備の容易さを勘案して、Mac mini (3.0 GHz Core i7, 16 GB RAM) を符号化・復号装置のプラットフォームとして用いることとした。表3に符号化・復号装置の入力音声フォーマットと選択可能なビットレートを示す。今回はLCプロファイル、レベル4の範囲で3種の音声フォーマットを選択し、チャンネルベース音響に限定して実装した。

開発した装置における遅延量は符号化で400 ms、復号で30 ms程度であった。一般に8K映像は符号化処理のみで1 s程度を想定するため、本装置は地上放送の高度化に向けた実証実験用の機器として充分なリアルタイム性を実現したと言える。

4.2 伝送デモンストレーション

2018年5月のNHK技研公開で、8K映像と22.2 ch 音声信号を組み合わせ、暫定的な次世代地上放送方式の伝送デモンストレーションを実施した。展示の様子を図3に示す。

表3 符号化・復号装置の仕様

入力音声フォーマット	ビットレート [kbps]
ステレオ	48, 64, 80, 96, 128, 192
5.1サラウンド	128, 160, 192, 256, 320
22.2 ch	512, 640, 768, 1000, 1400

図3 技研公開2018での展示風景

音声信号のビットレートは512 kbpsとし、復号した22.2 ch 音声信号はバイノーラルヘッドホンで再生した。

展示ではおよそ一週間、毎日8時間以上連続して運用したが、常に動作は安定しており、品質にも問題は発生しなかった。これによりMPEG-H 3DAは、22.2 ch 音響を高い圧縮効率でリアルタイムに伝送するための規格として実用可能であると結論づけた。

5. むすび

本稿ではMPEG-H 3DAを例にとり、次世代の音声符号化方式の開発動向を紹介した。次世代の音声符号化方式は高い圧縮効率による品質の向上のみならず、個人適応やアクセシビリティの改善など機能面からも視聴者の聴取体験充実に貢献できる技術である。これらの技術が将来、日本国内の放送方式に応用され、放送サービスの高度化が推進されることを期待したい。

謝辞 本稿で紹介した符号化・復号装置の開発におけるフラウンホーファー集積回路研究所の協力を感謝します。

(2018年8月6日受付)

〔文 献〕

- 1) ISO/IEC 13818-7:2006 (2006)
- 2) ARIB 標準規格STD-B32 3.9版 (2016)
- 3) 総務省 4K放送・8K放送情報サイト, http://www.soumu.go.jp/menu_seisaku/ictseisaku/housou_suishin/4k8k_suishin.html
- 4) ISO/IEC 14496-3:2009 (2009)
- 5) Rec. ITU-R BS.2051-1 (2017)
- 6) ISO/IEC 23008-3:2015 (2015)
- 7) ETSI TS 103 190-1 V1.3.1 (2018)
- 8) ETSI TS 103 190-2 V1.2.1 (2018)
- 9) ISO/IEC 23008-3:2015/Amd.3:2017 (2017)
- 10) ATSC Standard: A/342 Part 3 (2017)
- 11) 杉本, 小森, 音講論 (秋), pp.419-420 (2016)
- 12) Rec. ITU-R BS.1116-3 (2015)
- 13) Rec. ITU-R BS.1548-4 (2013)
- 14) 岡野ほか, 映像学技報, 42, pp.47-50 (2018)
- 15) 杉本ほか, 音講論 (秋), pp.197-198 (2018)

杉本 岳大 2001年、NHK入局。2004年より、放送技術研究所にて、放送システムにおける音声符号化方式、3次元音響システムにおける信号処理、音響トランスデューサの研究開発に従事。また、MPEG・ARIBにおいて、音声符号化方式の標準化活動に従事。博士(工学)。正会員。