

## INVITED PAPER

## MPEG-H 3D Audio: Immersive Audio Coding

Jürgen Herre<sup>1,\*</sup> and Schuyler R. Quackenbush<sup>2,†</sup><sup>1</sup>*International Audio Laboratories Erlangen, a joint institution of Friedrich-Alexander-Universität Erlangen-Nürnberg and Fraunhofer IIS, both Erlangen, Germany*<sup>2</sup>*Audio Research Labs,  
Scotch Plains, NJ 07076 USA**(Received 30 October 2021, Accepted for publication 28 December 2021)*

**Abstract:** The term “Immersive Audio” is frequently used to describe an audio experience that provides to the listener the sensation of being fully immersed or “present” in a sound scene. This can be achieved via different presentation modes, such as surround sound (several loudspeakers horizontally arranged around the listener), 3D audio (with loudspeakers at, above and below listener ear level) and binaural audio to headphones. This article provides an overview of the recent MPEG standard, MPEG-H 3D Audio, which is a versatile standard that supports multiple immersive sound signal formats (channels, objects, higher order ambisonics), and is now being adopted in broadcast and streaming applications.

**Keywords:** 3D audio, Audio coding, Audio compression, Audio data reduction, Immersive audio, MPEG, MPEG-H

## 1. INTRODUCTION

The MPEG-H 3D Audio specification [1,2] describes a universal audio coding and rendering environment that is designed to efficiently represent high-quality spatial/immersive audio content for storage and transmission. This is of paramount importance for many types of applications offering immersive, 3-dimensional (3D) audio, such as broadcasting or wireless streaming/download media services.

Since there is no generally accepted “one-size-fits-all” format for 3D spatial audio, it supports common loudspeaker setups including mono, stereo, surround and 3D audio (i.e., setups including loudspeakers above ear level and possibly below ear level). Furthermore, formats that are independent of loudspeaker setup like *objects* and *higher order ambisonics* are supported in the standard.

MPEG-H 3D Audio allows for flexible and optimized rendering over a wide range of reproduction conditions (i.e., various loudspeaker setups, headphones, background noise levels in various consumption environments, etc.). In addition, a considerable level of *interactivity* and *customization* is available by allowing the user to personalize content playback, including adjusting foreground/back-

ground balance, language selection and dialog enhancement (i.e., changing the dialog level).

The MPEG-H 3D Audio specification was completed in 2017 and a verification test to assess the subjective audio quality provided by the technology was completed the same year [3].

## 2. OVERVIEW

The MPEG-H 3D Audio architecture supports three important production paradigms for spatial sound: channels, objects and higher order ambisonics.

### 2.1. Channels

Traditionally, spatial sound has been delivered by producing several signals (“channels”) that drive loudspeakers positioned in a well-defined geometric setup relative to the listener (e.g., stereo, 5.1, 22.2). In this way, each channel signal is associated with a specific spatial position. The degree of spatial realism and immersion generally increases with the number of loudspeaker channels. However, one issue for channel-based content is that it expects a particular loudspeaker layout. MPEG-H 3D Audio overcomes this restriction with a conversion process that adapts the content to any loudspeaker setup.

### 2.2. Objects

A more recent approach for delivering spatial sound is to produce and deliver spatial audio content as a set of

---

\*e-mail: juergen.herre@audiolabs-erlangen.de

†e-mail: srq@audioresearchlabs.com

[doi:10.1250/ast.43.143]

“object” signals with associated metadata specifying the sound source location (and, possibly, other object properties). The locations may be time-varying to enable moving sound sources, such as a plane fly-over. These objects are then reproduced on the user loudspeaker setup (or headphones) by a rendering algorithm. This enables the user to create an interactive/personalized sound experience by adjusting the object characteristics of the rendered output [4]. For example, users could increase or decrease the level of the announcer’s commentary or actor’s dialogue relative to the other audio elements in the audio program. In contrast to the traditional channel paradigm, the object-oriented content representation is agnostic of loudspeaker layouts and permits greater spatial resolution when presented on a setup with more loudspeakers.

### 2.3. Higher Order Ambisonics

An alternative approach to representing spatial audio content is higher order ambisonics [5], which decomposes the three-dimensional sound field at a particular point in space into spherical harmonics, where the HOA “coefficient” signals are the weighting of the associated harmonics at given time instants. While first order ambisonics provides limited spatial resolution, higher orders provide increasingly higher resolution and better approximation of the original sound field. Similar to the object-based paradigm, HOA signals are agnostic of loudspeaker layouts and need a renderer in order to be reproduced on a target loudspeaker setup (or on headphones).

### 2.4. Decoder Architecture

Figure 1 shows a high-level overview of an MPEG-H 3D Audio decoder. The main part of the incoming MPEG-H 3D Audio bitstream is decoded by the *Core Decoder*

which reproduces the encoded waveforms that represent either channel signals, object signals or HOA coefficient signals. These waveforms are then processed in dedicated processing chains for each of these signal types. A *Format Converter* processes the channel signals to convert them to the target rendering loudspeaker layout. Object signals and HOA components are rendered to the target layout by the *Object Renderer* and the *HOA Renderer*, respectively. All three contributions are summed together in a *Mixer* to generate the final loudspeaker feed signals. Optionally, binaural output can be produced.

The most important blocks of the system are discussed in more detail in the following sections.

### 2.5. Audio Signal Compression

The core part of the codec compresses and represents the waveforms of the channel, object and HOA signals. To this end, the MPEG-H 3D Audio codec is based closely on the technology of the previously developed MPEG Unified Speech and Audio Coding (USAC) [6,7] system which is the state-of-the-art MPEG codec for compression of mono, stereo and multi-channel audio signals. It provides the best audio coding quality for both audio and speech signals at rates down to 8 kbit/s per channel by combining elements from perceptual audio coding and speech coding. Both codec parts are tightly integrated and can be adaptively selected depending on the nature of the input signal at each processing frame, in this way choosing the more efficient of the two technologies.

The codec bitstream has been designed to enable seamless transitions between encoded streams, for example with different coding bitrates, which is particularly advantageous for adaptive streaming over IP networks. This means that MPEG-H 3D Audio is fully compatible with MPEG Dynamic Adaptive Streaming (DASH) [8].

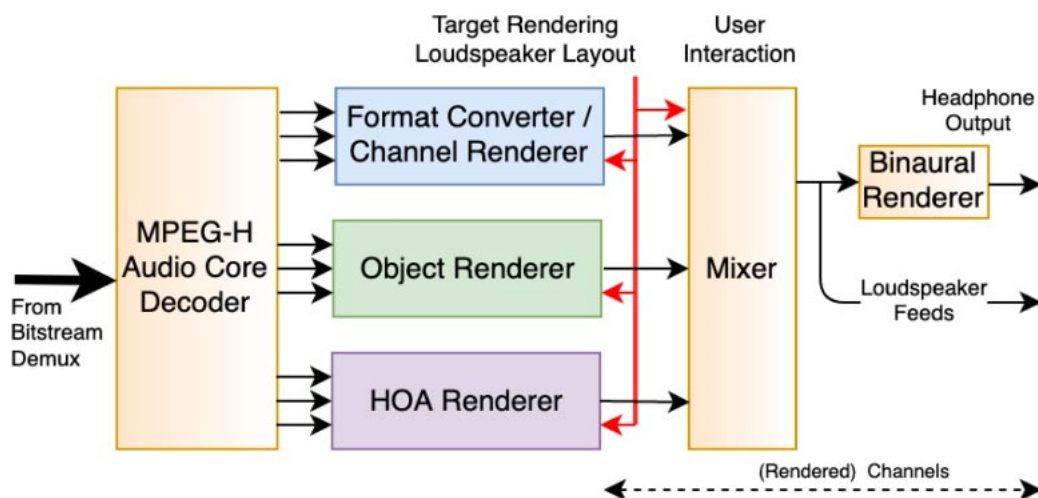


Fig. 1 Top-level architecture of MPEG-H 3D Audio decoder.

### 3. RENDERING

#### 3.1. Channel Rendering: Format Conversion

In order to allow reproduction of encoded channel-based content on any available loudspeaker setup connected to the MPEG-H 3D Audio decoder, the format converter maps the encoded channel signals to the target speaker layout. As an example, the decoder may detect a 5.1 surround reproduction loudspeaker setup while the content has been encoded in 22.2 channel format. Thus, an appropriate high-quality downmix has to be performed in order to enable the best possible listening experience, given the available speaker layout. To provide the best possible user experience, the format converter considers a number of specific aspects of signal downmix, as described below.

##### 3.1.1. Optimized downmix for all setups

The format converter is capable of automatically generating optimized downmix matrices for all target loudspeaker setups including non-standard positions (as appear frequently in users' homes) to map the transmitted channel configuration to the output loudspeaker layout.

This is achieved by an iterative search through an internal look-up table of tuned mapping rules for each input channel that happens once during the initialization phase of the format converter module. Each rule describes the mapping of one particular input channel to one or more output loudspeaker channels, possibly accompanied by a specific equalization curve that is to be applied when this particular rule has been selected. Even asymmetric loudspeaker setups can be accommodated.

##### 3.1.2. Advanced active downmix

When the optimized downmix coefficients have been found, they can be used as part of an advanced downmix process that is designed to avoid downmix artefacts, such as signal cancellations or comb-filtering. Such artifacts could occur when using static gains to linearly combining input signals that may exhibit significant correlations. In practice, such effects are quite common for current 3D audio content, since it is frequently produced from available 2D (surround) content by populating the missing channel signals with delayed and filtered copies of the 2D signals.

The MPEG-H 3D Audio active downmix process adapts to the properties of the input signal by measuring the correlations between the incoming signals and aligning the phases of these signals, if required. Furthermore, a frequency selective energy compensation ensures that energy is preserved after the downmix step and thus avoids timbral colorations.

#### 3.2. Object Rendering

In MPEG-H 3D Audio, objects consist of an encoded monophonic waveform and associated metadata describing

how to render these objects to a specific spatial location. An associated object gain also can be transmitted and both position and gain can be defined as dynamic trajectories (see MPEG-H 3D Audio metadata overview [9]).

The MPEG-H 3D Audio object renderer is based on the Vector Base Amplitude Panning (VBAP [10]) algorithm and renders the transmitted audio objects to any given output (target) loudspeaker setup. It processes one decoded audio stream per audio object, the associated decoded object metadata (e.g., time-varying position data and gains), and the geometry of the target rendering setup.

VBAP positions sound sources by triangulating the 3D surface surrounding the listener. MPEG-H 3D Audio provides an automatic triangulation algorithm that supports both standard and arbitrary target loudspeaker setups. Object locations that are not covered well by the target loudspeaker setup (e.g., locations below the horizontal speaker plane) are supported by the addition of imaginary speakers to provide, for any loudspeaker setup, complete 3D triangle meshes to the VBAP algorithm [2].

In accordance with the VBAP approach, the MPEG-H 3D Audio renderer searches for the loudspeaker triangle that encloses each object and builds an associated vector base. For this loudspeaker triangle, panning gain values are computed that preserve overall signal energy. The computed gains can be linearly interpolated between the panning values computed at sequential time stamps. Finally, the contributions of each object are summed to form the final renderer output signals.

#### 3.3. HOA Decoding and Rendering

Higher Order Ambisonics (HOA) describes the audio scene as a three-dimensional acoustic sound field that is represented as a truncated expansion of the wave field into spherical harmonics [5]. This completely determines the acoustic quantities within a certain source free region around the listener's position up to an upper frequency limit beyond which spatial aliasing limits the expansion's accuracy. The time-varying coefficients of the spherical harmonics expansion are called HOA coefficient signals and carry the information of the wave field that is to be described for transmission or reproduction.

Generally speaking, for a complete 3D expansion of order  $n$ ,  $(n + 1)^2$  coefficient signals have to be carried and many of those signals exhibit—depending on the nature of the sound field—considerable correlations (and thus redundancy) between them. Hence, both redundancy and irrelevance reduction have to be considered to arrive at a bitrate-efficient and yet high-quality representation. To this end, the MPEG-H 3D Audio encoder includes a dedicated tool set for HOA coding which decomposes the input HOA signals into a different, bitrate-efficient, internal representation that is used for rate-reduced transmission in the

MPEG-H 3D Audio bitstream. Conversely, this process is inverted in the MPEG-H 3D Audio decoder. The underlying algorithmic concepts will be described in the following.

#### 3.3.1. Representation of direct sound components

In a first step, a decomposition of the sound field into direct and diffuse sound components is carried out to reduce redundancy in the HOA coefficient signals. Strong sound events that originate from a distinct direction ('direct' components) introduce highly correlated components into many HOA signals, as can be seen from a spherical harmonics expansion of plane waves [11]. In order to identify such redundancies, the encoder performs an analysis of the input signal to detect the presence of significant direct sounds (called "predominant sounds" in the context of MPEG-H 3D Audio) and transmits them separately as parametrically coded plane waves plus associated directional metadata. Then, the direct sound contributions are subtracted from the HOA coefficients of the remaining ambient sound field components. In this way, a considerable reduction of redundancy can be achieved for input with significant direct sound contributions.

#### 3.3.2. Representation of ambient sound components

In a second step, the remaining ambient sound components are processed. Since localization accuracy is typically not of high perceptual importance for such non-directional sound field components, their HOA representation order can be reduced without perceptual penalty in order to increase coding efficiency. Nonetheless, this representation still may include high correlations between the HOA coefficient signals. Not only is this disadvantageous from a redundancy point of view, but individual perceptual coding of these coefficient signals can also lead to undesirable spatial unmasking of the quantization noise contributions introduced by each of the coefficient signal coders. As a solution to this challenge, the ambience HOA representation is first transformed into a different spatial domain using a spherical Fourier transform. The resulting virtual loudspeaker signals exhibit less correlation and are then used as input to a multi-channel codec core for bitstream encoding and transmission. The number of transmitted virtual loudspeaker signals can be chosen according to the available bitrate and perceptual quality considerations.

In the MPEG-H 3D Audio decoder, the reverse processing takes place as compared to the encoding process: The parametrically represented predominant sound components are re-synthesized as HOA contributions. Then, the transmitted virtual loudspeaker channels representing the ambient sound field are mapped back to the HOA domain. Finally, both contributions are added and rendered to the reproduction setup using a generic HOA

rendering matrix that is adapted to the target loudspeaker layout.

## 4. ADDITIONAL TOOLS & FUNCTIONS

Besides the core components described so far, the MPEG-H 3D Audio decoder also includes a number of additional tools and functions that enhance its performance or applicability in specific situations. These are discussed below.

### 4.1. Dynamic Range Control and Loudness Control

In order to enable optimum playback in each consumption environment, the MPEG-H 3D Audio decoder also includes a *Dynamic Range Control (DRC)* feature that can be applied to the final output signals, and also to individual intermediate sound components, such as objects. The underlying DRC technology is that from the MPEG-D specification [12] and allows to control the dynamic range of the playback in consideration of the background noise conditions of the playback environment (e.g., living room, car, airplane, ...), effectively providing better subjective user experience in very noisy environments by increasing the level of the most quiet portions of the audio program such that they are more easily heard. Furthermore, a loudness normalization function avoids jumps in loudness when the user switches between different programs.

### 4.2. Binaural Rendering

Besides reproduction on loudspeakers, MPEG-H 3D Audio also supports binaural reproduction of spatial sound on headphones. This allows convincing consumption of immersive audio on common mobile devices, such as mobile phones, handhelds, and portable music players, where headphones would be used.

### 4.3. Content-Based Interactivity

The use of audio objects with rich metadata in audio productions opens up the possibility of *user interaction* with the content. To this end, all sound components (channels, objects and HOA components) that are embedded in the MPEG-H 3D Audio bitstream can be selected by the user during playback and adjusted in level offering the possibility of personalized playback. A simple adjustment might be increasing/decreasing the level of the commentary signal relative to the other audio elements according to user preference, and in this way enhancing the intelligibility of the dialog. The extent of possible user interaction is under full control of the producer by means of embedding control metadata during content creation.

## 5. PERFORMANCE

As part of the standardization process, MPEG conducts a verification test of all standardized technology. The

Verification Test for MPEG-H 3D Audio [3] assessed the performance of a subset of the standardized technology, the Low Complexity Profile. It consisted of four subjective listening tests, that each evaluated audio quality and compression performance at an operating point representing a distinct use case.

The test material consisted of 36 audio items selected to represent typical and critical audio material. The material was either channel-based signals, both channel-based and object signals, or scene-based signals, as HOA of a designated order, possibly also including object signals.

The subjective tests were conducted at seven test labs. The first three tests (Tests 1, 2 and 3) were conducted in high-quality listening rooms that were calibrated to conform to the criteria set forth in BS.1116 [13] and also calibrated to be perceptually similar to each other. The fourth test (Test 4) was conducted in acoustically isolated sound booths. Altogether, the tests results were based on 190 subjects and nearly ten thousand subjective scores.

**Test 1: Ultra HD Broadcast:** This use case assumed an 8k video broadcast and a mix of immersive 22.2 channel and 11.1 channel (as 7.1+4H) audio presentation formats. Since the video would be expected to have a high bit rate, the audio coding bitrate was proportional at 768 kb/s, the highest bit rate amongst the four tests. This test used the “Triple Stimulus Hidden Reference” subjective test methodology (Rec. ITU-R BS.1116) [13]. The subjective results showed that the MPEG-H 3D Audio Low Complexity profile operating at 768 kb/s with highly immersive audio content achieved a quality of “Excellent” on the BS.1116 quality scale. It had a mean score (as a BS.1116 “diff score”) of  $-0.31$  with a 95% confidence interval of  $\pm 0.04$  which is well above the  $-1.0$  limit (4.0 out of 5.0 in “absolute scores”) recommended in Rec. ITU-R BS.1548-4 [14] for “High-quality emission” for broadcast applications.

**Test 2: HD Broadcast:** This use case assumed a broadcast program with HD video and immersive audio with 11.1 channel (as 7.1+4H) or 7.1 channel (as 5.1+2H) channel loudspeaker layouts. All audio was coded at three bit rates: 512 kb/s, 384 kb/s and 256 kb/s.

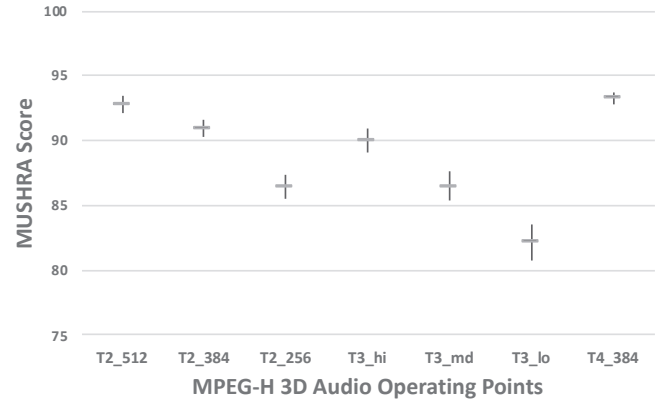
**Test 3: High Efficiency Broadcast:** As in Test 2, all audio was coded at three bit rates, but this use case assumed a need for greater compression, the specific bit rates depended on the number of channels in the audio material. Bit rates ranged from 256 kb/s (5.1+2H) to 48 kb/s (stereo).

**Test 4: Mobile:** This use case assumed that content consumption would be “on the go” (via a mobile device). It used the coded immersive audio content from Test 2, at the 384 kb/s rate, and rendered for headphone presentation using the MPEG-H 3D Audio binauralization engine.

Tests 2, 3 and 4 all used the “Method for the subjective

**Table 1** MUSHRA descriptors and associated score range.

Descriptor	Score
Excellent	80–100
Good	60–80
Fair	40–60
Poor	20–40
Bad	0–20



**Fig. 2** Plot of subjective audio quality of MPEG-H 3D Audio for Test 2, Test 3 and Test 4 (for greater visibility, only the MUSHRA scale above 75 points is shown here).

assessment of intermediate quality level of coding systems” (MUSHRA) [14]. In a MUSHRA test, the correspondence of subjective quality (indicated by descriptor) and the range of subjective score is shown in Table 1.

The test results are given in Fig. 2, where the vertical axis is subjective score (for greater visibility of results, its low end is 75 rather than 0), and the horizontal axis shows the MPEG-H 3D Audio operating points tested. In the specific operating point names, the prefix T2\_, T3\_, and T4\_ indicates configurations tested in Test 2, Test 3 and Test 4, respectively, and the numerical suffix indicates the bit rate, in kb/s. A suffix of hi, md, lo, indicates the high, medium and low bitrates for audio signals in Test 3. The plot shows mean scores as a horizontal tick and 95% confidence intervals on the mean as a vertical stroke.

The subjective results show that, for all bit rates in each of Test 2, Test 3 and Test 4, MPEG-H 3D Audio Low Complexity profile achieved a quality of “Excellent” on the MUSHRA subjective quality scale.

## 6. CONCLUSIONS

MPEG-H 3D Audio is an audio coding standard that provides state-of-the-art signal compression with rich metadata. The standard was developed for highly immersive content and supports the coding and transmission of

audio as audio channels, audio objects, or Higher Order Ambisonics. It can support up to 64 loudspeaker channels and 128 core codec channels and offers additional tools such as loudness normalization, dynamic range control and can adapt audio content for presentation on any loudspeaker layout or can produce binauralized output for headphone presentation. The rich metadata enables new aspects of user interactivity, such as dialog enhancement. The 3D Audio Low Complexity Profile has excellent performance at bitrates that are compatible with broadcast and streaming delivery service rates.

## REFERENCES

- [1] ISO/IEC 23008-3:2019, (Information technology — High efficiency coding and media delivery in heterogeneous environments — Part 3: 3D Audio (2019).
- [2] J. Herre, J. Hilpert, A. Kuntz and J. Plogsties, “MPEG-H audio: The new standard for coding of immersive spatial audio,” *IEEE J. Sel. Top. Signal Process.*, **9**, 770–779 (2015).
- [3] N16584, “MPEG-H 3D Audio verification test report.” Available <https://mpeg.chiariglione.org/standards/mpeg-h/3d-audio/mpeg-h-3d-audio-verification-test-report> (accessed 24 January 2022).
- [4] R. Bleidt, D. Sen, A. Niedermeier, B. Czelhan, S. Füg, S. Disch, J. Herre, J. Hilpert, M. Neuendorf, H. Fuchs, J. Issing, A. Murtaza, A. Kuntz, M. Kratschmer, F. Küch, R. Füg, B. Schubert, S. Dick, G. Fuchs, F. Schuh, E. Burdiel, N. Peters and M. Kim, “Development of the MPEG-H TV audio system for ATSC 3.0,” *IEEE Trans. Broadcast.*, **63**, 202–236 (2017).
- [5] F. Zotter and M. Frank, *Ambisonics, A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement and Virtual Reality* (Springer, Cham, Switzerland, 2019).
- [6] ISO/IEC 23003-3:2012 Information technology — MPEG audio technologies — Part 3: Unified speech and audio coding (2012).
- [7] M. Neuendorf, M. Multus, N. Rettelbach, G. Fuchs, J. Robilliard, J. Lecomte, S. Wilde, S. Bayer, S. Disch, C. Helmrich, R. Lefebvre, P. Gournay, B. Bessette, J. Lapiere, K. Kjörling, H. Purnhagen, L. Villemoes, W. Oomen, E. Schuijers, K. Kikuri, T. Chinen, T. Norimatsu, C. Seng, E. Oh, M. Kim, S. Quackenbush and B. Grill, “The ISO/MPEG unified speech and audio coding standard — consistent high quality for all content types and at all bit rates,” *J. Audio Eng. Soc.*, **61**, 956–977 (2013).
- [8] ISO/IEC 23009-1:2019, Information technology — Dynamic adaptive streaming over HTTP (DASH) — Part 1: Media presentation description and segment formats (2019).
- [9] S. Füg, A. Hölzer, C. Borß, C. Ertel, M. Kratschmer and J. Plogsties, “Design, coding and processing of metadata for object-based interactive audio,” *137th AES Convention*, Los Angeles (2014).
- [10] V. Pulkki, “Virtual sound source positioning using vector base amplitude panning,” *J. Audio Eng. Soc.*, **45**, 456–466 (1997).
- [11] B. Rafaely, “Plane-wave decomposition of the sound field on a sphere by spherical convolution,” *J. Acoust. Soc. Am.*, **116**, 2149 (2004), DOI:10.1121/1.1792643.
- [12] ISO/IEC 23003-4:2015 Information technology — MPEG audio technologies — Part 4: Dynamic Range Control (2015).
- [13] ITU-R Recommendation BS.1116-3 (02/2015), “Methods for the subjective assessment of small impairments in audio systems” (2015).
- [14] ITU-R Recommendation BS.1534-3 (10/2015), “Method for the subjective assessment of intermediate quality level of coding systems,” also known as “MULTi Stimulus test with Hidden Reference and Anchor (MUSHRA)” (2015).