FIRST-ORDER AMBISONIC CODING WITH PCA MATRIXING AND QUATERNION-BASED INTERPOLATION

Pierre Mahé
Orange Labs, Lannion, France
L3i, University of La Rochelle, France
pierre.mahe@orange.com

Stéphane Ragot
Orange Labs, Lannion, France
stephane.ragot@orange.com

Sylvain Marchand
L3i, University of La Rochelle, France
sylvain.marchand@univ-lr.fr

ABSTRACT

We present a spatial audio coding method which can extend existing speech/audio codecs, such as EVS or Opus, to represent first-order ambisonic (FOA) signals at low bit rates. The proposed method is based on principal component analysis (PCA) to decorrelate ambisonic components prior to multi-mono coding. The PCA rotation matrices are quantized in the generalized Euler angle domain; they are interpolated in quaternion domain to avoid discontinuities between successive signal blocks. We also describe an adaptive bit allocation algorithm for an optimized multi-mono coding of principal components. A subjective evaluation using the MUSHRA methodology is presented to compare the performance of the proposed method with naive multi-mono coding using a fixed bit allocation. Results show significant quality improvements at bit rates in the range of 52.8 kbit/s (4×13.2) to 97.6 kbit/s (4×24.4) using the EVS codec.

1. INTRODUCTION

Conversational applications such as telephony are typically limited to mono, with no spatial representation of the sound scene. With the emergence of new applications such as virtual reality (VR) and extended reality (XR) and the availability of devices supporting spatial audio capture and playback, there is a need to extend traditional speech/audio codecs to enable immersive communication. There are currently different spatial audio codecs developed for non-conversational applications (streaming, broadcast, etc.), including discrete coding of individual audio channels by MPEG AAC or HE-AAC, MPEG Surround [1], Dolby AC-3 or E-AC-3 [2], and more recently MPEG-H 3D Audio [3], Dolby AC-4 [4] or DTS-UHD [5]. The most recent spatial audio codecs can handle various input formats, such as multichannel audio, object-based audio, ambisonics (also called scene-based audio [3]), and multiple playback formats (e.g. mono, stereo, binaural audio, various multichannel loudspeaker setups). In this work, we investigate how spatial audio can be provided in conversational applications by extending codecs currently used in telephony or voice over IP (VoIP).

We focus in particular on reusing the Enhanced Voice Services (EVS) codec [6] which represents the state-of-art audio quality for mobile telephony applications. The EVS codec supports only mono input and output signals; a naive approach to extend this codec to spatial audio is to code each input channel by a separate instance of the mono codec – this approach is later referred to as *multi-mono coding*. Note that the Opus codec [7] supports mono, stereo, multichannel, and recently it has been extended to code ambisonic signals, using multi-mono coding or fixed channel matrixing followed by multi-stereo coding [8].

Several methods have been proposed to code ambisonic signals. One approach is to use a parametric model of the sound field, with an assumption on the number of audio sources in the scene. For First-Order Ambisonic (FOA) coding, DirAC [9, Chap.5] assumes that there is a single predominant audio source in each time-frequency tile. At the encoder side, a mono downmix signal representing the predominant source is extracted together with directional parameters – in terms of direction of arrival (DoA) and diffuseness – in each time/frequency tile. The mono signal and the parameters are coded and transmitted. At the decoder side, the sound field is reconstructed by panning the predominant source (based on the DoA, typically using VBAP); this is signal is combined with an ambiance using decorrelation and the diffuseness parameters. The DirAC method has been extended to High-Order Ambisonics (HOA) in the so-called HO-DirAC [9, Chap.6] where the sound field is divided into angular sectors. For each angular sector, one source is extracted. More recently, Compass [10] was proposed as a method inspired by DirAC. This method overcomes the limitation of the number of sources. The number of sources is estimated and sources are extracted by Principal Component Analysis (PCA) and the residual signal is downmixed to obtain an ambiance signal.

The Compass method is similar to the MPEG-H 3D Audio codec [3] which supports ambisonic signals as one type of input format. In MPEG-H 3D Audio, the input ambisonic signal is decomposed into a number of component signals; these signals represent a number of predominant sources with an ambiance signal, they are coded using a core codec derived from MPEG USAC [11]. Predominant sources may be extracted using plane wave decomposition or Principal Component Analysis (PCA). When PCA is used – noting that PCA may be equivalently implemented using a Singular Value Decomposition (SVD) – components may change dramatically between consecutive frames causing channel permutations and signal discontinuities [12] for complex sound scenes (with many audio sources, sudden changes, etc.). The MPEG-H 3D audio codec employs channel re-alignment, overlap-add and linear interpolation to mitigate these problems. Improvements to the MPEG-H 3D Audio codec were proposed in [12, 13]; the SVD decomposition is done in the MDCT domain to ensure smooth transitions across frames.

In this work, we wanted to avoid any assumption on the audio scene (e.g. presence of predominant sources, number of sources) based on existing mono and stereo codecs. The most basic approach is multi-mono coding; at low bit rates, due to the corre-

lation of ambisonic components, the signal structure is typically degraded, which creates several spatial artifacts: spatial reduction, blur, phantom source. Multi-stereo coding with fixed matrixing was proposed in [8, 14]. The principle consists in combining the input B-format channels by matrixing, which may be interpreted as beamforming in configurable directions. This approach has the advantage to transform the signal into a format which is more suited to multi-stereo coding, however the fixed matrixing is in general not optimal. An alternative approach proposed in [15] is based on predictive coding of 1st-order FOA components based on the 0th-order component, which is typically efficient when the scene has few sources. In this paper, we propose an extension of multi-mono coding approach to represent FOA signals. We make no assumption on the scene content. The proposed method relies on PCA to obtain an adaptive matrixing of FOA components in time domain. To guarantee signal continuity between components across frames, two mechanisms are implemented: principal component matching (similar to MPEG-H 3D Audio and [12, 13]) and rotation matrix interpolation in quaternion domain. We also describe an adaptive bit allocation algorithm for an optimized multi-mono coding of principal components.

This paper is organized as follows. Section 2 gives a brief overview of ambisonics. Section 3 provides the relevant background on quaternions, the representation of 4D rotation matrices by double quaternions, and the interpolation of rotation matrices in quaternion domain. Section 4 describes in details the proposed coding method. Section 5 presents the results of a subjective quality evaluation comparing the proposed method and naive multi-mono coding, before concluding in Section 6.

2. AMBISONICS

Ambisonics is based on a decomposition of the sound field into an orthogonal basis of spherical harmonics. Initially limited to 1st order by Gerzon [16], the formalism was extended to high orders by Daniel [17]. The sound field may be expressed by the equation [18]:

$$p(t,f,r,\theta,\phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} p_{nm}(t,f) Y_{nm}(\theta,\phi) \qquad (1)$$

where $p(t,f,r,\theta,\phi)$ is the sound pressure at time t, frequency f, distance r, azimuth θ and elevation ϕ, $Y_{nm}(.,.)$ is the spherical harmonic function, and $p_{nm}(t,f)$ is the ambisonic coefficients of order n and degree m. In practice, the sound field representation is truncated to a finite order N. The higher N, the better the sound field representation accuracy. The so-called B-format corresponds to the ambisonic components $p_{nm}(t,f)$. For a given order N the number of ambisonic components is $(N+1)^2$ in 3D and $2N+1$ in 2D (in horizontal-only ambisonics). From the B-format, rendering is required to reproduce the sound field to the listener. For loudspeaker configurations the rendering computes the signal played by each loudspeaker. For headphone listening binaural rendering may be performed by decoding over virtual loudspeakers, convolving and combining the resulting feed signals by Head-Related Impulse Responses (HRIRs) [19] which are filters measured for each ear.

3. QUATERNIONS AND 4D ROTATION MATRICES

Quaternions (also called hypercomplex numbers or hamiltonians) were introduced in 1843 by Hamilton [20] to define a vectorial system generalizing complex numbers. A quaternion q is defined as $q = a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$, where $(a,b,c,d) \in \mathbb{R}^4$, with the following rules for $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$: $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -1$. We can also write $q = a + \mathbf{q}$, where a is the *scalar* (or real) part of q and $\mathbf{q} = b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$ is the *vector* (or imaginary) part of q. Note that $\mathbf{q}$ may be interpreted as a 3D vector by identifying $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ to three orthogonal Cartesian unit vectors. A quaternion q having zero scalar part is called a *pure quaternion*. We do not review basic operations on quaternions (e.g. addition $q_1 + q_2$, multiplication $q_1 q_2$, dot product $q_1 \cdot q_2$, conjugate $\overline{q}$ and inverse q^{-1}) – see for instance [20, 21, 22] for more details. We recall that the dot product between two quaternions $q_1 = a_1 + b_1\mathbf{i} + c_1\mathbf{j} + d_1\mathbf{k}$ and $q_2 = a_2 + b_2\mathbf{i} + c_2\mathbf{j} + d_2\mathbf{k}$ is $q_1 \cdot q_2 = a_1a_2 + b_1b_2 + c_1c_2 + d_1d_2$ which corresponds to the usual dot product of 4-dimensional vectors, and the norm $|\mathbf{q}|$ of a quaternion is: $|q| = \sqrt{q \cdot q} = \sqrt{a^2 + b^2 + c^2 + d^2}$. When $|q| = 1$, q is said to be a *unit-norm quaternion*.

Quaternions are often used as a representations of 3D rotations. We recall that the set of 3D rotations (called special orthogonal group in dimension 3 or $SO(3)$) can be mapped to the unit sphere in $\mathbb{R}^4$ under a one-to-two mapping [21, 22, 23], namely each 3D rotation matrix maps to two antipodal unit-norm quaternions: q and $-q$, therefore this mapping is not unique. A 3D rotation of angle θ and normalized axis $\mathbf{n}$ (with $|\mathbf{n}| = 1$) is represented by the unit quaternion $q = \cos(\theta/2) + \sin(\theta/2)\mathbf{n}$ (or its opposite $-q$); the rotation of a 3D vector $\mathbf{u}$ to $\mathbf{v}$ can be equivalently obtained by quaternion multiplication: $\mathbf{v} = q\mathbf{u}q^{-1} = q\mathbf{u}\overline{q}$, if $\mathbf{u}$ and $\mathbf{u}$ are interpreted as pure quaternions.

Unlike other representations Euler angles or axis-angle, quaternions are interesting to interpolate rotations. One common method is called spherical linear interpolation (slerp) and consists in the following principle [21]:

$$slerp(q_1, q_2, \gamma) = q_1(q_1^{-1}q_2)^{\gamma} \qquad (2)$$

where q_1 and q_2 are respectively the starting and ending quaternions and $0 \leq \gamma \leq 1$ is the interpolation factor. The slerp interpolation may also be formulated as [21]:

$$slerp(q_1, q_2, \gamma) = \frac{\sin((1-\gamma)\Omega)}{\sin(\Omega)} q_1 + \frac{\sin(\gamma\Omega)}{\sin(\Omega)} q_2 \qquad (3)$$

where $\Omega = \arccos(q_1 \cdot q_2)$ is the angle between the two quaternions q_1 and q_2. This interpolation boils down to interpolating along the grand circle (or geodesics) on the unit 4D sphere with a constant angular speed as a function of γ. Due to the non-unique representation of 3D rotations by antipodal unit-norm quaternions, one has to ensure that the quaternion trajectory follows the shortest path [23] on the unit sphere in $\mathbb{R}^4$ by selecting among the two possible choices $\pm q_2$ for the end point. Alternative interpolation methods, such as normalized linear interpolation (nlerp) or splines [21], are not reviewed here.

In this work, we used the representation of 4D rotation matrices as double quaternions. The multiplication of two quaternion matrices, an anti-quaternion matrix $\mathbf{Q}^\star$ and a quaternion matrix $\mathbf{P}$, associated to two quaternions $q = a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$ and $p = w + x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$, is defined as [22]:

$$\mathbf{R} = \mathbf{Q}^\star.\mathbf{P} = \mathbf{P}.\mathbf{Q}^\star \qquad (4)$$

where

$$\mathbf{Q}^{\star} = \begin{pmatrix} a & b & c & d \\ -b & a & -d & c \\ -c & d & a & -b \\ -d & -c & b & a \end{pmatrix} \tag{5}$$

and

$$\mathbf{P} = \begin{pmatrix} w & -x & -y & -z \\ x & w & -z & y \\ y & z & w & -x \\ z & -y & x & w \end{pmatrix} \tag{6}$$

It is possible [22] to show that the product $\mathbf{R}$ has the properties of a rotation matrix ($\mathbf{R}.\mathbf{R}^T = \mathbf{R}^T.\mathbf{R} = \mathbf{I}$ and $\det(\mathbf{R}) = +1$); conversely, given a 4D rotation matrix $\mathbf{R}$, it is possible [22] to factorize this matrix as in Eq. 4. This factorization (sometimes called Cayley's factorization) may be obtained using the method described in [24] (noting that a factor $1/4$ is missing in Eqs. 64 and 65 in [24]); this method is also described in [22] (together with alternative factorization approaches). The factorization relies on an intermediate matrix (called associate matrix in [24] or tetragonal transform in [22]) to determine the two quaternions q and p up to sign (i.e. $-q$ and $-p$ also form a valid solution).

Similar to 3D rotations, two 4D rotation matrices may be interpolated by interpolating separately the associated pairs of quaternions (for instance using slerp). However, it is important to keep the sign consistent between double quaternions when constraining the shortest path.

Note that there are other representations for 4D rotation matrices. In this work, we also used the generalized Euler angles defined in [25]. In general, a n-dimensional rotation matrix is characterized by $n(n-1)/2$ generalized Euler angles and for $n = 4$ we have 6 angles. We refer to [25] for details on the conversion between an n-dimensional rotation matrix and generalized Euler angles.

4. PROPOSED CODING METHOD

We describe in this section the proposed coding method. The input signal is assumed to be a first-order ambisonic signal with the ACN ordering convention (W, Y, Z, X). The $n = 4$ ambisonic components will be labeled respectively with an index $i = 1, \cdots, n$. Each ambisonic component is sampled at 32 kHz. The coding method operates on successive frames of 20 ms. The different coding and decoding steps are described in the high-level diagram shown in Figure 1. We describe below the implementation for each functional blocks.

4.1. High-pass pre-processing (HPF)

The $n = 4$ channels of the input FOA signal are separately pre-processed by a 20 Hz high-pass IIR filter from the EVS codec defined by the transfer function

$$H_{pre}(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}} \tag{7}$$

where filter coefficients (b_i, a_i) are taken from [26]. This preprocessing is used here to avoid any bias in the subsequent estimation of the covariance matrix.

4.2. Principal Component Analysis (PCA)

In each frame, the sample covariance matrix $\mathbf{C_{xx}}$ is estimated based on the 4-dimensional input $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]$:

$$\mathbf{C_{xx}} = \frac{1}{n-1} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \tag{8}$$

where $\mathbf{x}_i$ is the i-th column of $\mathbf{X}$ and corresponds to the current frame of i-th ambisonic channel. The covariance matrix $\mathbf{C_{xx}}$ is factorized by eigenvalue decomposition as:

$$\mathbf{C_{xx}} = \mathbf{V} \Lambda \mathbf{V}^T \tag{9}$$

where $\mathbf{V}$ is the eigenvector matrix (with eigenvectors as columns) and $\Lambda = diag(\lambda_1, \cdots, \lambda_n)$ is the diagonal matrix whose coefficients are the eigenvalues.

In general the matrix $\mathbf{V}$ is orthogonal and it may be either a rotation matrix ($det(\mathbf{V}) = +1$) or a reflection matrix ($det(\mathbf{V}) = -1$). In this work, we ensured that the eigenvector matrix defines a rotation matrix by inverting the sign of $\mathbf{v}_n$ if $det(\mathbf{V}) = -1$. In the following, the resulting rotation matrix in the current frame of index t will be denoted $\mathbf{V}_t$.

4.3. Re-alignment of eigenvectors

From frame to frame eigenvectors might change significantly. These modifications might create discontinuities in the signal, which can degrade audio quality. To improve signal continuity between successive frames, a signed permutation is applied to the eigenvector matrix $\mathbf{V}_t$ in the current frame of index t based on the eigenvector matrix $\mathbf{V}_{t-1}$ in the previous frame.
The signed permutation is obtained in two steps:

1. A permutation is found by matching eigenvectors in frames t and $t-1$ according to the axes (not directions) of each basis vector. This problem can be seen as an assignment problem where the goal is to find the closest eigenvector in frame t for each eigenvector in the previous frame $t-1$. To solve this assignment problem, we used the Hungarian algorithm to find the optimal solution in a similar way to [12]. After this step, eigenvectors are permuted. This allows to maximize similarity between the two basis.

   The similarity being defined as:

   $$\mathbf{J}_t = tr(|\mathbf{V}_t.\mathbf{V}_{t-1}^T|) \tag{10}$$

   where $tr(|.|)$ is the trace of the matrix $|\mathbf{V}_t.\mathbf{V}_{t-1}^T|$ whose coefficients are the absolute values.

   After applying the optimal permutation to $\mathbf{V}_t$ we obtain a new matrix of eigenvectors $\tilde{\mathbf{V}}_t$.

2. The direction of each eigenvector is determined based on the permuted eigenvector matrix in the current frame $\tilde{\mathbf{V}}_t$ for frame t and the rotation matrix $\mathbf{V}_{t-1}$ in the previous frame $t-1$

   $$\mathbf{\Gamma}_t = \tilde{\mathbf{V}}_t.\mathbf{V}_{t-1}^T \tag{11}$$

   A negative diagonal value in $\mathbf{\Gamma}_t$ indicates a direction inversion between two frames. The sign of the respective columns of $\tilde{\mathbf{V}}_t$ is inverted to compensate for this change of direction.

Note that the resulting matrix $\tilde{\mathbf{V}}_t$ in the current frame will be saved for the next frame processing to become $\mathbf{V}_{t-1}$.

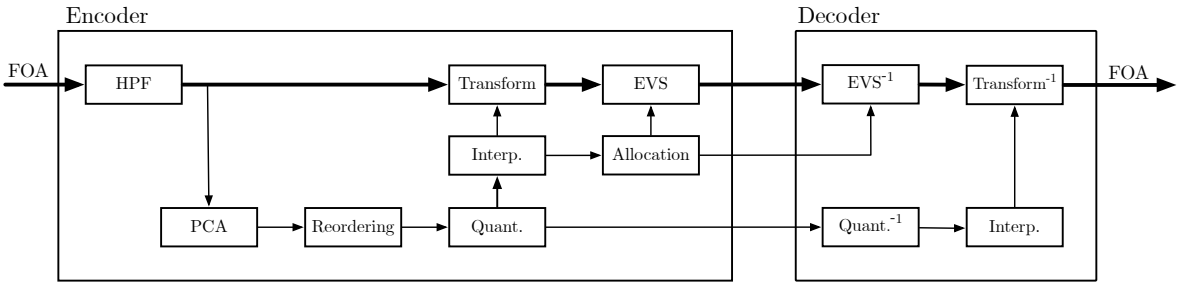


Figure 1: *Overview of proposed coding method.*

4.4. Quantization of the 4D PCA rotation matrix

In general a n-dimensional rotation matrix has $n(n-1)/2$ degrees of freedom. In [27] 2D and 3D rotation matrices were quantized with an angle representation: one angle in 2D, three Euler angles in 3D. In this work we used a similar idea by quantizing 6 generalized Euler angles. The 6 angles were obtained from the 4D rotation matrix [25] and coded by scalar quantization with a budget of respectively 8 and 9 bits for angles defined over a support of length π and 2π, with an overall budget of 51 bits per frame for 6 angles.

4.5. Interpolation of quantized rotation matrices by subframe

The rotation matrices are interpolated by subframes to smooth variations of PCA rotation matrices and avoid signal discontinuities. The rotation matrix representation is not suitable for interpolation. Instead, we converted 4D rotation matrices to double quaternions as explained in Section 3. The current frame of length $L = 640$ samples is divided into K sub-frames. We used $K = 128$ which gives a subframe length of $L/K = 10$ samples (0.3125 ms). For each subframe of index $1 \leq k \leq K$ in the current frame, the left (q_{t-1}, q_t) and right quaternions (p_{t-1}, p_t) are interpolated by the slerp algorithm with an interpolation factor given by $\gamma = k/K$ as defined in Eq. 3. The interpolated double quaternions are converted back to a 4D matrix using Eqs. 4, 5, 6.

4.6. PCA matrixing

The pre-processed FOA signal is transformed into 4 principal components by applyinng the interpolated rotation matrix in each subframe.

4.7. Adaptive bit rate allocation to multi-mono EVS coding

After PCA matrixing the $n = 4$ channels could have been coded using a fixed bit allocation, as in naive multi-mono coding. Depending on the input signals, the signals after PCA matrixing may vary significantly in importance and it was found experimentally that an adaptive bit allocation is necessary to optimize quality. We use a greedy bit allocation algorithm which aims at maximizing the following score:

$$S_(b_1, \cdots, b_n) = \sum_{i=1}^{n} Q(b_i).E_i^{\beta} \tag{12}$$

where b_i and E_i are respectively the bit allocation and the energy of the i^{th} channel in the current frame and $Q(b_i)$ is a quality score reflecting the estimated quality of the codec for a bit rate corresponding to b_i bits per 20 ms frame. This optimization is subject to the constraint $b_1 + \cdots + b_n \leq B$ where B is the budget allocated for multi-mono coding. The term $Q(b_i)$ is defined here in Table 1 to reflect the fact that the EVS codec quality does not increase [28] linearly with increasing bit rate and it was found experimentally the theoretical rate-distortion function that would apply for a usual source model (e.g. Gaussian source) would not be suitable. The scores $Q(b_i)$ correspond here to average MOS (Mean Opinion Score) values found during the performance characterization of the EVS codec by 3GPP [28]. Note that if another core codec than EVS was used, the values $Q(b_i)$ could be adjusted accordingly; for instance, a quality evaluation of Opus can be found in [29]. To guarantee maximum compatibility with the EVS codec, the bit allocation to individual audio channels is restricted to the set of EVS bit rates: 9.6, 13.2, 16.4, 24.4, 32, 48, 64, 96, 128 kbit/s. Note that a minimum bit allocation of 9.6 kbit/s was defined to ensure a super-wideband coded bandwidth. The energy term E_i^{β} is raised to the power β, where $0 \leq \beta \leq 1$ is defined as an extra tuning parameter: when β is close to 1, channels with more energy will dominate the bit allocation and when $\beta = 0$ channels would receive an equal allocation. The value of $\beta = 0.5$ was selected experimentally. The bit allocation $b_1, \cdots, b_n$ selected in the current frame is coded and transmitted to the decoder.

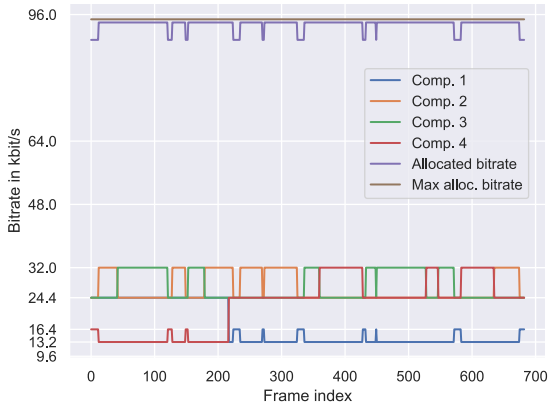


Figure 2: *Bit rate allocation example at* 4×24.4 *kbit/s.*

Figure 2 shows an example of bit rate allocation to components (denoted Comp. 1, 2, 3, 4) in successive 20 ms frames for one test item. The maximum allocated bit rate (94.75 kbit/s) corresponds to the difference between 4×24.4 kbit/s and the bit rate used for side information (2.85 kbit/s). The overall allocated bit

Table 1: *Bit allocation parameters.*

b_i	192	264	328	488	640	960	1280	1920	2560
rate kbit/s	9.6	13.2	16.4	24.4	32	48	64	98	128
$Q(b_i)$	3.62	3.79	4.25	4.60	4.53	4.82	4.83	4.85	4.87

rate (corresponding to $b_1 + b_2 + b_3 + b_4$) does not always use the available budget and in this case padding bits have to be added to the bitstream.

4.8. Multi-mono EVS coding

The bit rate to code the meta-data (rotation matrices, bit allocation) is subtracted to the target bit rate and the remaining bit rate is used for the adaptive bit allocation to multi-mono coding. The transformed FOA channels are coded by separate instances of the EVS codec and the associated bitstreams are transmitted to the decoding part. In this work we used the fixed-point implementation of EVS (V14.2.0) with discontinuous transmission disabled.

A bitstream structure example is shown in Figure 3 for the bit rate of $4 \times 24.4 = 97.6$ kbit/s; the bitstream is divided in several sections: bit allocation (6 bits), quantized generalized Euler angles (51 bits), four coded channels ($b_1 + b_2 + b_3 + b_4$ bits) and padding (when needed).

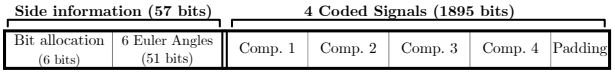


Figure 3: *Bitstream structure example at* $4 \times 24.4 kbit/s$.

4.9. Decoding part

The decoding part is similar to the encoding part. The bit allocation is demultiplexed and the 4 coded channels are decoded by separate instances of the EVS decoder. In addition, the generalized Euler angles are decoded and converted into a 4D rotation matrix. The interpolation in quaternion domain done in each subframe is performed. The signal in the current frame divided in K subframes is then transformed by inverse PCA matrixing to retrieve the reconstructed ambisonic components.

5. EVALUATION

5.1. Test methodology

We conducted subjective tests according to the MUSHRA methodology [30] to compare the perceptual performance of naive multi-mono coding and the proposed coding method. For each item, subjects were asked to evaluate the quality of conditions with an integer grading scale ranging of 0 to 100. This interval is divided in 5 sections of 20 points: bad (0-20) to excellent (80-100).

The test conditions included three specific items: the hidden reference (FOA) and two anchors. Traditional MUSHRA tests for mono signals typically use a low anchor (reference processed by a 3.5kHz low-pass filter) and a medium anchor (reference processed by a 7kHz low-pass filter). For MUSHRA tests with stereo, it is suggested to use "reduced stereo image" as degradations in anchors [30]. There is no clear recommendation on spatial alterations for MUSHRA tests with ambisonic signals. In this work we used anchors with some spatial deformation. We use the following spatial reduction defined by:

$$FOA = \begin{pmatrix} W \\ \alpha X \\ \alpha Y \\ \alpha Z \end{pmatrix}, \qquad \alpha \in [0, 1] \tag{13}$$

with $\alpha = 0.65$ and $\alpha = 0.8$ for the low and medium anchors, respectively.

Table 2: *List of MUSHRA conditions.*

Short name	**Description**
HREF	FOA hidden reference
LOW_ANCHOR	3.5 kHz LP-filtered and spatially-reduced FOA ($\alpha = 0.65$)
MED_ANCHOR	7 kHz LP-filtered and spatially-reduced FOA ($\alpha = 0.8$)
MULTI52	FOA coded by multimono EVS at 4×13.2 kbit/s
MULTI65	FOA coded by multimono EVS at 4×16.4 kbit/s
MULTI97	FOA coded by multimono EVS at 4×24.4 kbit/s
PCA52	FOA coded by proposed method at 52.8 kbit/s
PCA65	FOA coded by proposed method at 65.6 kbit/s
PCA97	FOA coded by proposed method at 97.6 kbit/s

5.2. Experimental setup

The subjects were placed in an audio listening room. The room complied with ITU-R recommendations [31] in terms of acoustics properties, reverberation time and background noise. All subjects conducted the listening test with the same hardware, which included an external sound-card (Focusrite Scarlett 6i6) and high-quality headphones (Sennheiser HD 650) and a test interface running on a MacBook pro computer.

The test items consisted of 10 challenging ambisonic items: 4 voice items, 4 music items and 2 ambient scenes. Six items were real recordings made by ambisonic microphones (EigenMike or SoundField SPS200), others were synthetic items. A description of these items can be found in Appendix 7. Each item was about 10 s long. All FOA items were binauralized with Resonance Audio renderer [32] using generic HRTFs corresponding to the KU100 manikin. Original FOA signals were normalized in loudness using the following procedure: the rerefence FOA signals were binauralized using the Resonance audio renderer; the resulting binaural signal was normalized to -23 dBLUFS according to ITU-R BS.1770 [33], and after verifying that there was no saturation in binaural signals, the obtained scale factor was re-applied to the respective coded FOA signal.
In total 11 listeners participated in the test; all of them are expert or experienced listeners without hearing impairments. Each item was codec at three bit rates for multi-mono coding: 52.8, 65.6, 97.6 kbit/s which corresponds to a fixed allocation of 13.2, 16.4 and 24.4 kbit/s per channel. For the proposed coding method, as explained in Section 4, the bit rate was dynamically distributed between channels; however the target (maximum) bitrate was set to the same bit rate as multi-mono coding for a fair comparison. All test conditions are summarized in Table 2.

5.3. Test results

The subjective evaluation results, including the mean and 95% confidence intervals, are presented in Figure 4. They show that the proposed coding method improves quality when compared with multi-mono coding at the same bit rate. It is particularly noticeable that multi-mono coding at 65.6 kbit/s is equivalent to the proposed coding method at 52.8 kbit/s.

These quality improvements are largely due to the suppression of spatial artifacts. These artifacts are present at every bit rate in multi-mono coding and they can be classified into three categories: diffuse blur, spatial centering, phantom source. These three types of artifacts in multi-mono coding result from the degraded structure between ambisonic components. With the proposed coding method, these artifacts are mostly removed because the structure is less important after PCA matrixing. This explanation was supported by the feedback from some subjects, after they conducted the subjective test.

It is possible to analyze MUSHRA scores for different item categories. Figure 5 shows the scores for recorded and synthetic scenes. As can be seen, the proposed coding method brings significant improvements for synthetic items. The result for each item is the same whatever the position of sources or the number of sources. Two assumptions may explain this result. The first one is that the sources do not interact together, and the PCA can decorrelate each source, consequently the proposed method avoided spatial artifacts. The second one is that spatialization is more pronounced in synthetic items (wider displacement, more localized source). This spatialization puts the emphasis for the listener on spatial artifacts.

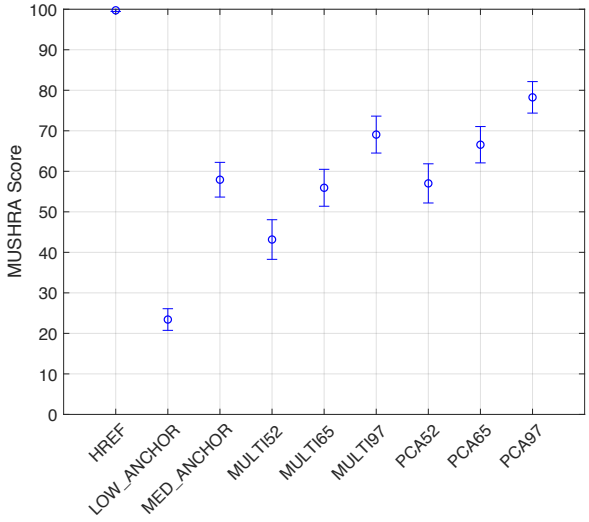


Figure 4: *MUSHRA mean scores with 95% confidence intervals.*

6. CONCLUSION

This paper presented a spatial extension of a mono speech/audio codec. The proposing coding method operates in time domain to avoid extra delay and allow maximum compatibility with existing codecs (e.g. EVS or Opus) which are used as a black box. The ambisonic components are transformed by adaptive matrixing depending on the audio scene.

For each frame, a PCA allows to find a new basis where the components are decorrelated. To guarantee smooth transition between consecutive frames, rotation matrices are interpolated in quaternion domain. Subjective test results show that the PCA matrixing together with the adaptive bit allocation give significant improvements over naive multi-mono coding for bit rates from 4×13.2 to 4×24.4 kbit/s. For future work, it will be interesting to characterize the spatial artifacts for various coding methods and to apply the proposed method to Opus.

7. ACKNOWLEDGMENTS

The authors would like to thank all participants in the audio test. They also thank Jérôme Daniel for helpful discussions on test items and spatial reduction for MUSHRA anchors.

8. REFERENCES

[1] J. Herre et al., "MPEG Surround—The ISO/MPEG Standard for Efficient and Compatible Multichannel Audio Coding," *J. Audio Eng. Soc.*, vol. 56, no. 11, pp. 932–955, 2008.

[2] ATSC Standard, Doc. A/52:2018, "Digital Audio Compression (AC-3, E-AC-3)," January 2018.

[3] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H audio—the new standard for universal spatial/3D audio coding," *J. Audio Eng. Soc.*, vol. 62, no. 12, pp. 821–830, 2015.

[4] ETSI TS 103 190 V1.1.1, "Digital Audio Compression (AC-4) Standard," April 2014.

[5] ETSI TS 103 491 V1.1.1, "DTS-UHD Audio Format; Delivery of Channels, Objects and Ambisonic Sound Fields," April 2017.

[6] S. Bruhn et al., "Standardization of the new 3GPP EVS codec," in *Proc. ICASSP*, 2015, pp. 5703–5707.

[7] J.-M. Valin, K. Vos, and T. Terriberry, "Definition of the Opus Audio Codec," RFC 6716, September 2012.

[8] J. Skoglund, "Ambisonics in an Ogg Opus Container," RFC 8486, October 2018.

[9] V. Pulkki, A. Politis, M.-V. Laitinen, J. Vilkamo, and J. Ahonen, "First-order directional audio coding (DirAC)," in *Parametric Time-Frequency Domain Spatial Audio*, V. Pulkki, S. Delikaris-Manias, and A. Politis, Eds., chapter 5. John Wiley & Sons, 2018.

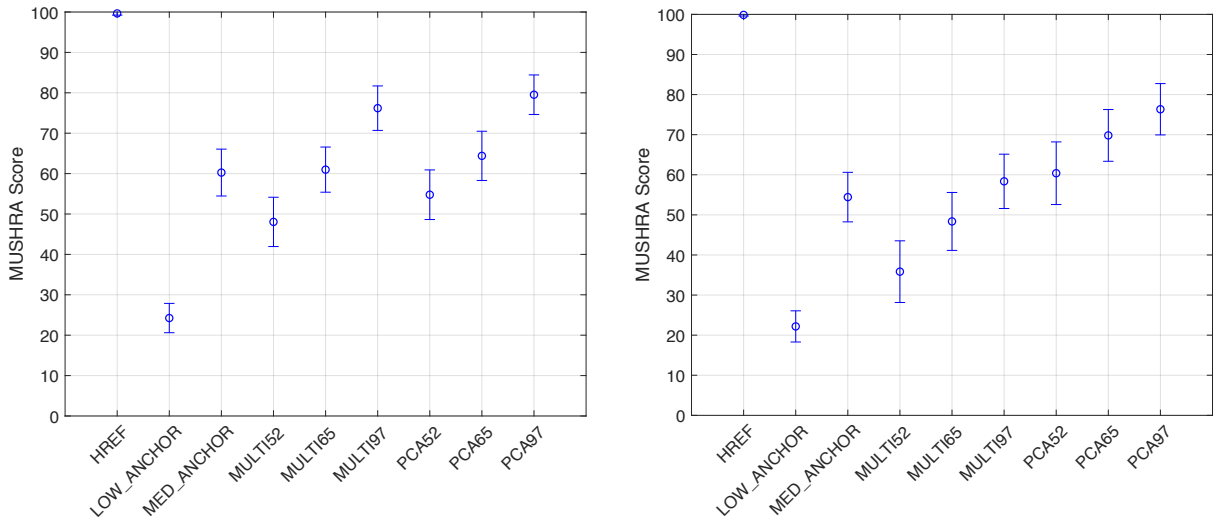
[10] A. Politis, S. Tervo, and V. Pulkki, "Compass: Coding and multidirectional parameterization of ambisonic sound scenes," in *Proc. ICASSP*, 2018, pp. 6802–6806.

[11] M. Neuendorf et al., "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.*, vol. 61, no. 12, pp. 956–977, 2013.

[12] S. Zamani, T. Nanjundaswamy, and K. Rose, "Frequency domain singular value decomposition for efficient spatial audio coding," in *Proc. WASPAA*, 2017, pp. 126–130.

[13] S. Zamani and K. Rose, "Spatial Audio Coding with Backward-Adaptive Singular Value Decomposition," in *145th AES Convention*, 2018.

[14] 3GPP Tdoc S4-171342, "Encoding First-Order Ambisonics with HE-AAC," Source: Dolby.



(a) Recorded scenes

(b) Synthetic scenes

Figure 5: *MUSHRA mean scores with 95% confidence intervals for different types of scenes.*

[15] D. McGrath et al., "Immersive Audio Coding for Virtual Reality Using a Metadata-assisted Extension of the 3GPP EVS Codec," in *Proc. ICASSP*, May 2019.

[16] M.A. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Eng. Soc.*, vol. 21, no. 1, pp. 2–10, 1973.

[17] J. Daniel, *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*, Ph.D. thesis, Université Paris 6, 2000.

[18] B. Rafaely, *Fundamentals of spherical array processing*, Springer, 2015.

[19] H. Møller, M.F. Sørensen, D. Hammershøi, and C.B. Jensen, "Head-related transfer functions of human subjects," *J. Audio Eng. Soc.*, vol. 43, no. 5, pp. 300–321, 1995.

[20] W.R. Hamilton, "On a new species of imaginary quantities connected with a theory of quaternions," *Proc. R. Ir. Acad.*, vol. 2, pp. 424–434, 1843.

[21] A.J. Hanson, *Visualizing Quaternions*, Morgan Kaufmann Publishers, 2006.

[22] P. De Casteljau, *Les quaternions*, Dunod, 1987.

[23] K. Shoemake, "Animating rotation with quaternion curves," *ACM SIGGRAPH Computer Graphics*, vol. 19, no. 3, pp. 245—-254, 1985.

[24] A. Perez-Gracia and F. Thomas, "On Cayley's factorization of 4D rotations and applications," *Advances in Applied Clifford Algebras*, vol. 27, no. 1, pp. 523–538, 2017.

[25] D.K. Hoffman, R.C. Raffenetti, and K. Ruedenberg, "Generalization of Euler Angles to N-Dimensional Orthogonal Matrices," *Journal of Mathematical Physics*, vol. 13, no. 4, pp. 528–533, 1972.

[26] 3GPP TS 26.445, "Codec for Enhanced Voice Services (EVS); Detailed algorithmic description," 2019.

[27] M. Briand, *Études d'algorithmes d'extraction des informations de spatialisation sonore : application aux formats multicanaux*, Ph.D. thesis, INPG Grenoble, 2007.

[28] 3GPP TS 26.952, "Codec for Enhanced Voice Services (EVS); Performance Characterization," 2019.

[29] A. Rämö and H. Toukomaa, "Voice quality characterization of IETF Opus codec," in *Proc. Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[30] ITU-R Rec. BS.1534-3, "Method for the subjective assessment of intermediate quality level of coding systems," 2015.

[31] ITU-R BS.1116-3, "Methods for the subjective assessment of small impairments in audio systems including Multichannel Sound Systems," 2015.

[32] "Resonance audio : Rich, immersive, audio," `https://resonance-audio.github.io/resonance-audio`.

[33] ITU-R Rec. BS.1170-4, "Algorithms to measure audio programme loudness and true-peak audio level," 2015.

[34] ISO/IEC JTC1/SC29/WG11/N13633, "Submission and Evaluation Procedures for 3D Audio," 2013.

[35] "Ambisonia," `www.ambisonia.com`.

[36] E. Roncière, "Compte-rendu de captation (Rapport projet BiLi) – Fiction CNSMDP : 2 femmes pour un fantôme de René de Obaldia," `www.bili-project.org`.

[37] M. Kronlachner, "ambix v0.2.8 – ambisonic plug-in suite," `www.matthiaskronlachner.com/?p=2015`.

APPENDIX: DESCRIPTION OF MUSHRA ITEMS

Synthetic items were generated in Orange Labs, recorded items were captured and mixed by Orange Labs or done jointly with partners (see details below). HOA items were truncated to FOA for testing.

- **Drums (synthetic)** Drum tracks were generated with a MIDI sequencer. HOA mixing, spatialization and near-field simulation were done with an internal Matlab library. This item is a shorterned version of the H_02_Drums1 test item [34] contributed by Orange Labs to MPEG-H standardization.
- **Modern Music (synthetic)** Single source (jazz music) crossing left to right with near-field effect. The mono source is an except of "The Present" in Laurent de Wilde's album "Time for Change". Spatial rendering was done with an internal Matlab library. This item is a shorterned version of the H_03_Modern test item [34] contributed by Orange Labs to MPEG-H standardization.
- **Opera (recorded)** Italian Opera with female singer, harpsichord and strings in concert hall from the Ambisonia database [35]. Players were in front of a SoundField SPS200 microphone.
- **Orchestra (recorded)** Orchestra in concert hall. An EigenMike microphone was placed in middle of the orchestra and spot microphones were also used for ambisonic mixing.
- **Theater (recorded)** Theater play with 3 moving talkers (in French), one in near-field and two in far-field. Recorded with an EigenMike microphone during the Bili Project [36].
- **Kids Playground (synthetic)** Two talkers (female and male in English) mixed with an ambisonic recording of kids playground. The mix was done with Reaper, the mono speech signals were spatialized with the Ambix framework [37], the ambisonic kids playground signal came from the Ambisonia database [35].
- **Little Prince (recorded)** Man reading excerpt of "Le Petit Prince" (in French) at a fixed (left, top) position in a living room. Recorded with an EigenMike microphone.
- **Talks (recorded)** People at different positions talking (in French) in a large room. Recorded with an EigenMike microphone during the Bili Project [36].
- **Nature (synthetic)** Artificial scene with bumblebee sound moving in space, mixed with an ambisonic brook noise and a bird singing at fixed position. The mix was done with Reaper, the mono signals of bumblebee and bird were spatialized with the Ambix framework [37], the bird noise came from the Ambisonia database [35].
- **Applause (recorded)** Applause at the end of a concert. A SoundField SPS200 microphone was placed in the middle of the crowd. The recording came from the Ambisonia database [35].