FIRST-ORDER AMBISONIC CODING WITH QUATERNION-BASED INTERPOLATION OF PCA ROTATION MATRICES

6th Sept. 2019

Pierre MAHE - Orange Labs and University of La Rochelle, France

pierre.mahe@orange.com

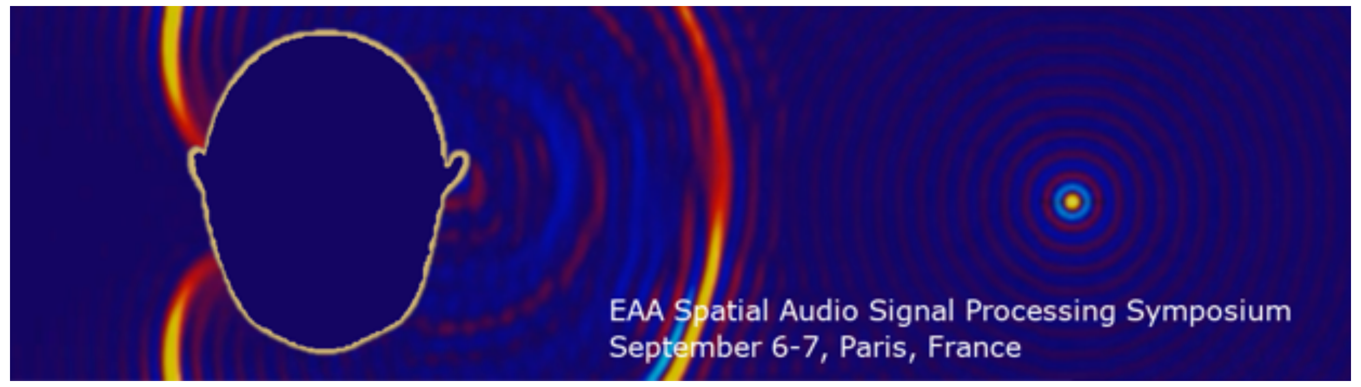
Stéphane RAGOT - Orange Labs, Lannion, France

stephane.ragot@orange.com

Sylvain MARCHAND - University of La Rochelle, France

sylvain.marchand@univ-lr.fr

orange™



EAA Spatial Audio Signal Processing Symposium
September 6-7, Paris, France

La Rochelle Université

Context and Motivations

Telephony codecs are mostly limited to mono.

Emergence of devices supporting spatial audio.

Need to compress immersive audio for telecommunication applications

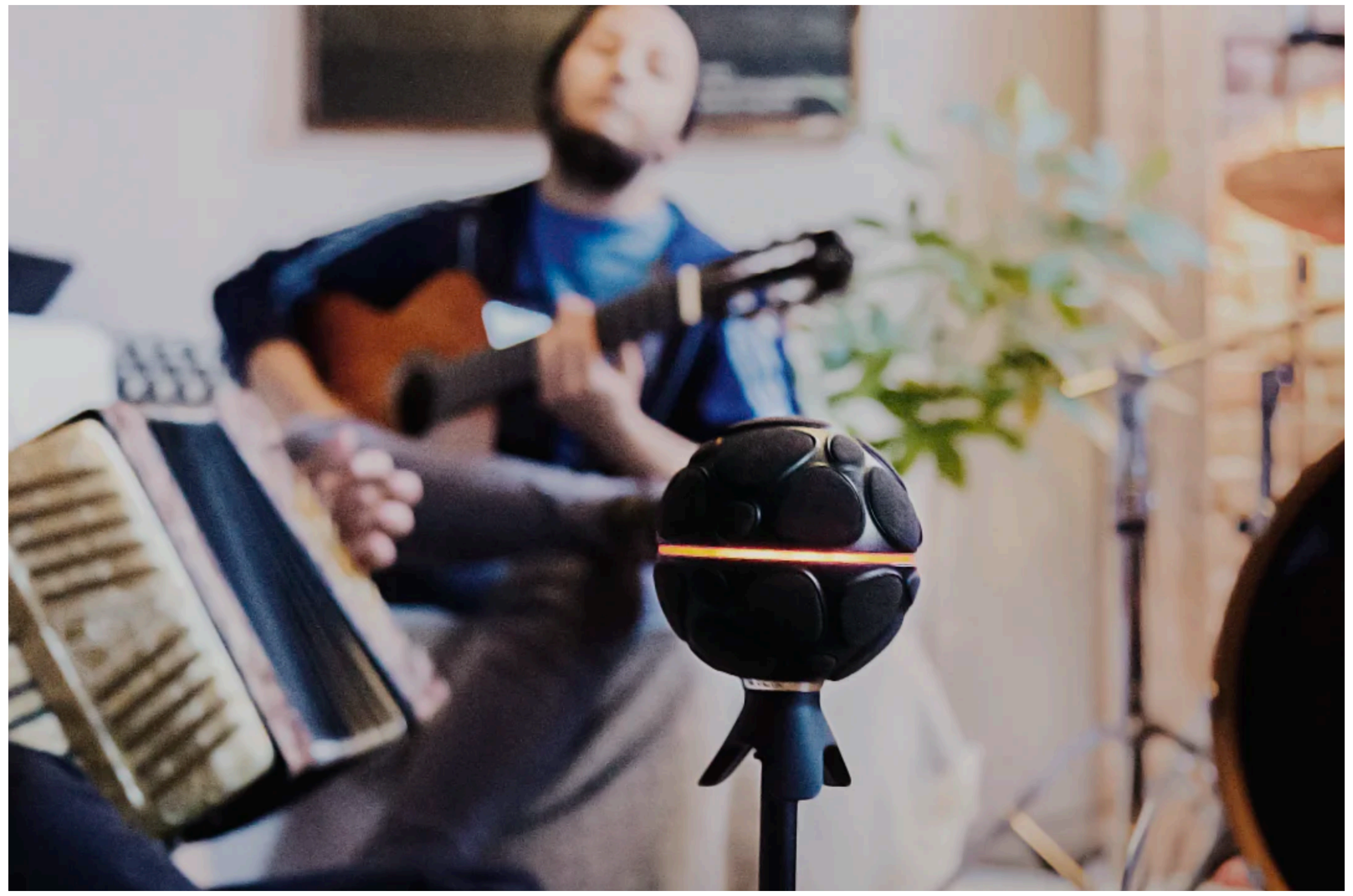
Extend existing codecs

Immersive content, for what purpose ?

— Call with ambiance sharing

— Immersive content broadcasting (360 Video, VR…)

— Spatialized audio conferencing

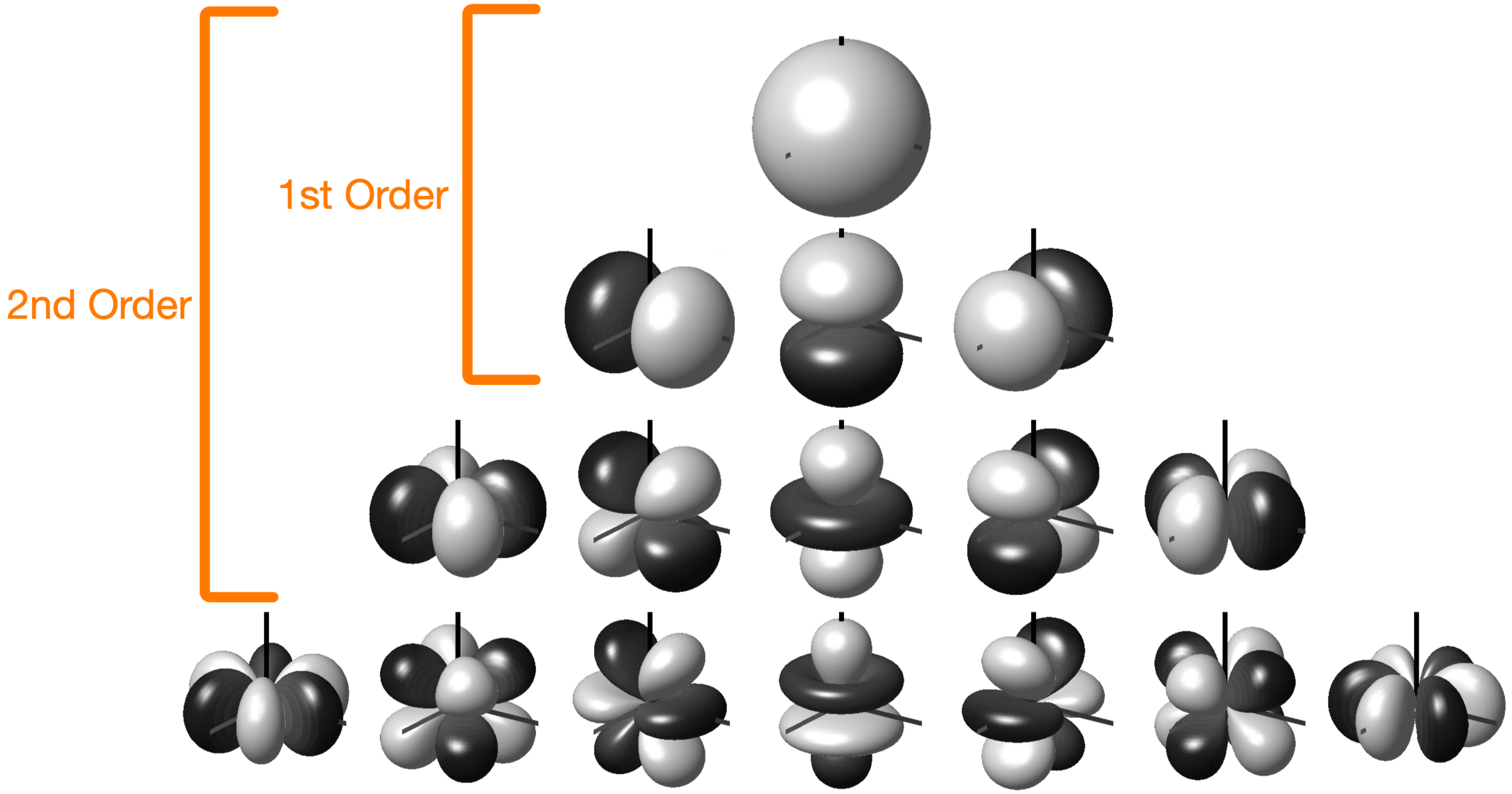


Ambisonics

Ambisonics is a decomposition of the sound field into a spherical harmonics basis.

First-order ambisonics (FOA) was invented in 1970s [Gerzon] and later extended to higher orders [Daniel].

For an order N, the basis dimension is $(\mathrm{N}+1)^2$



Is naïve approach not enough ?

Multi-mono approach

Each ambisonic components is coded independently by a mono codec.

Bitrate is uniformly distributed between components.

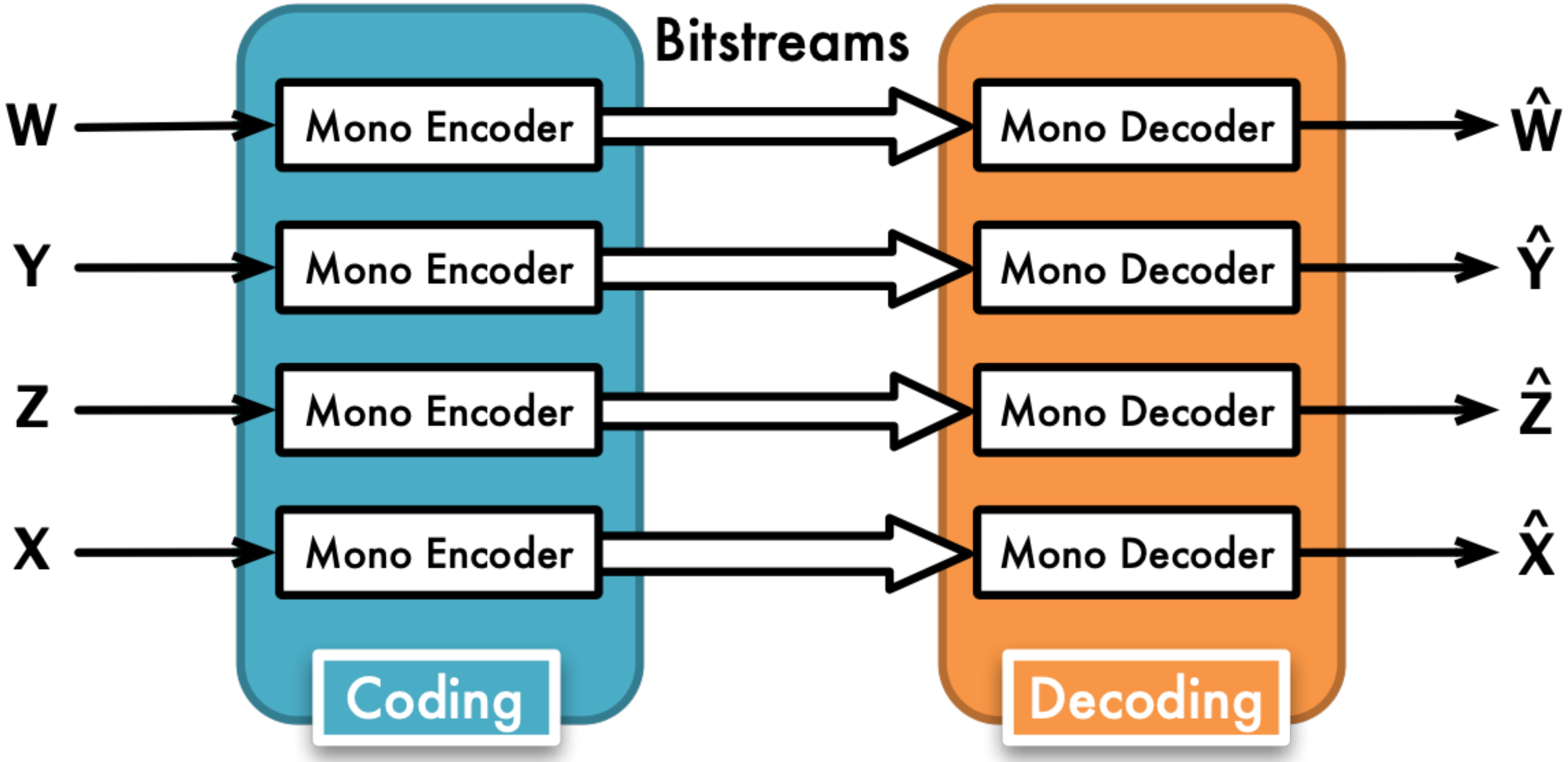
For listening tests, binaural renderering is used.

Test Results

Acceptable Quality → Bitrate higher than 4x48 kbit/s (192kbit/s)

Several artefacts :

- — Diffuse noise
- — Source positions are pushed to the front
- — Spatial blurring for percussive signal
- — Phantom sources



Existing approaches to code ambisonics

Multi-mono ou multi-stereo with fixed matrixing

Matrixing ambisonic components with fixed coefficients (e.g. for multi-stereo coding)
No assumptions on the scene

→ Improvements are not very significant

Scene analysis and source extraction [Pulkki, Politis]

Assumptions on the scene (number of sources…)

→ If wrong decisions are made in the scene analysis, quality is strongly impacted

Hybrid approaches [Herre]

Extract and transmit predominant sources (e.g; by PCA or SVD); the residual is downmixed and transmitted

→ Metadata to be transmitted, issues with signal continuity between frames

Our approach

Pre-process the ambisonic components prior to multi-mono coding

Decorrelate components to avoid spatial Artifact

No assumptions on the scene

Ensure signal continuity without add extra meta-data

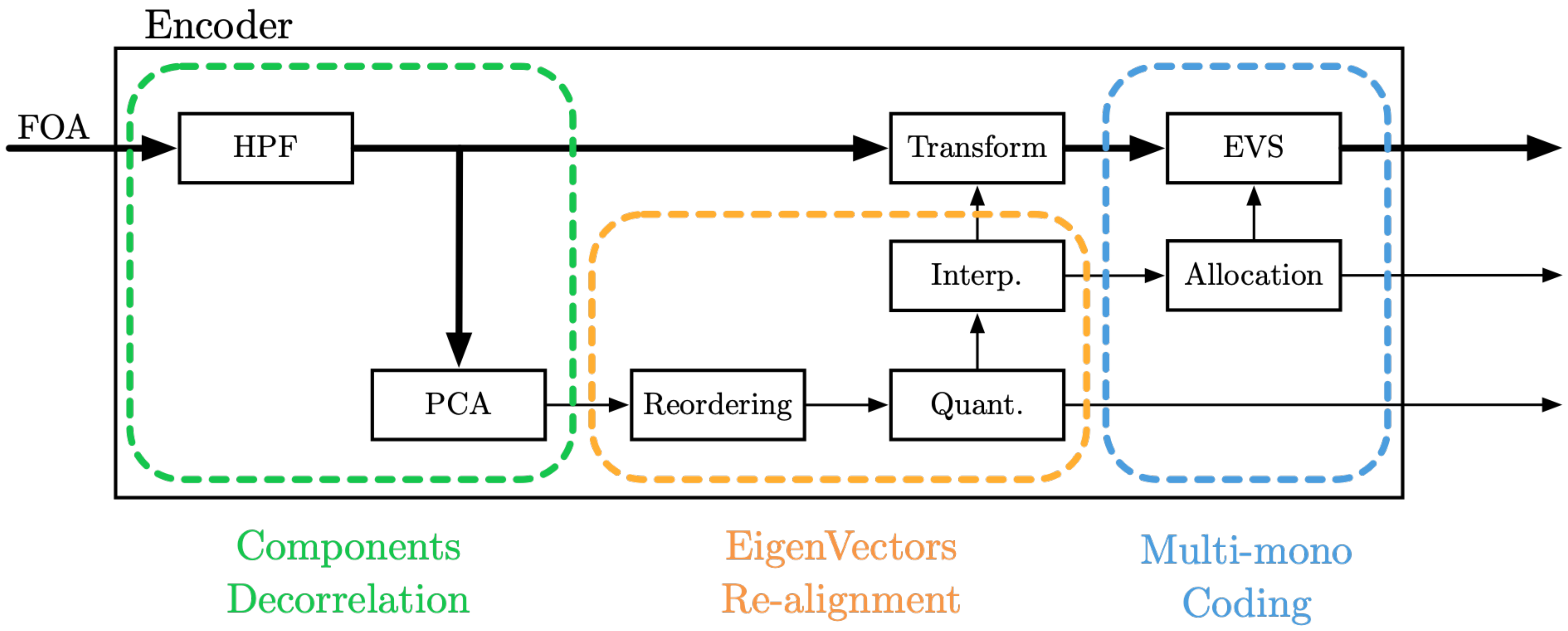
Extend existing codecs

Codec extension is composed by 3 modules

Pre-processing ambisonic which decorrelated components

Re-alignment of eigenvectors matrix

Coding of decorrelated components by core codec



Our approach

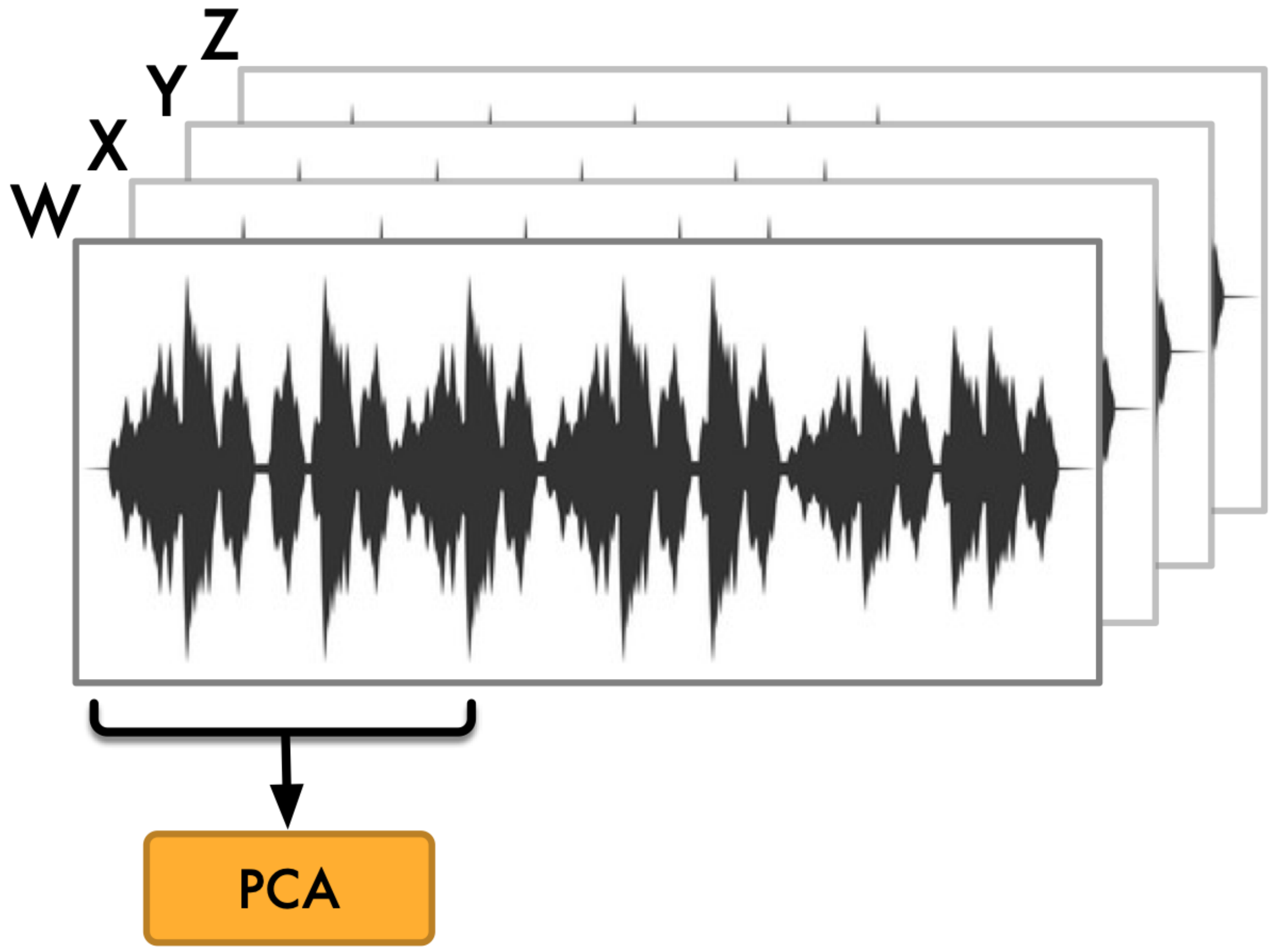
Compute PCA coefficients for the frame t

Compute the covariance matrix $\mathrm{C_{xx}}$

The matrix $\mathrm{C_{xx}}$ is factorized by Eigen decomposition

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

where V is the eigenvector matrix and $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_\mathrm{n})$



Our approach

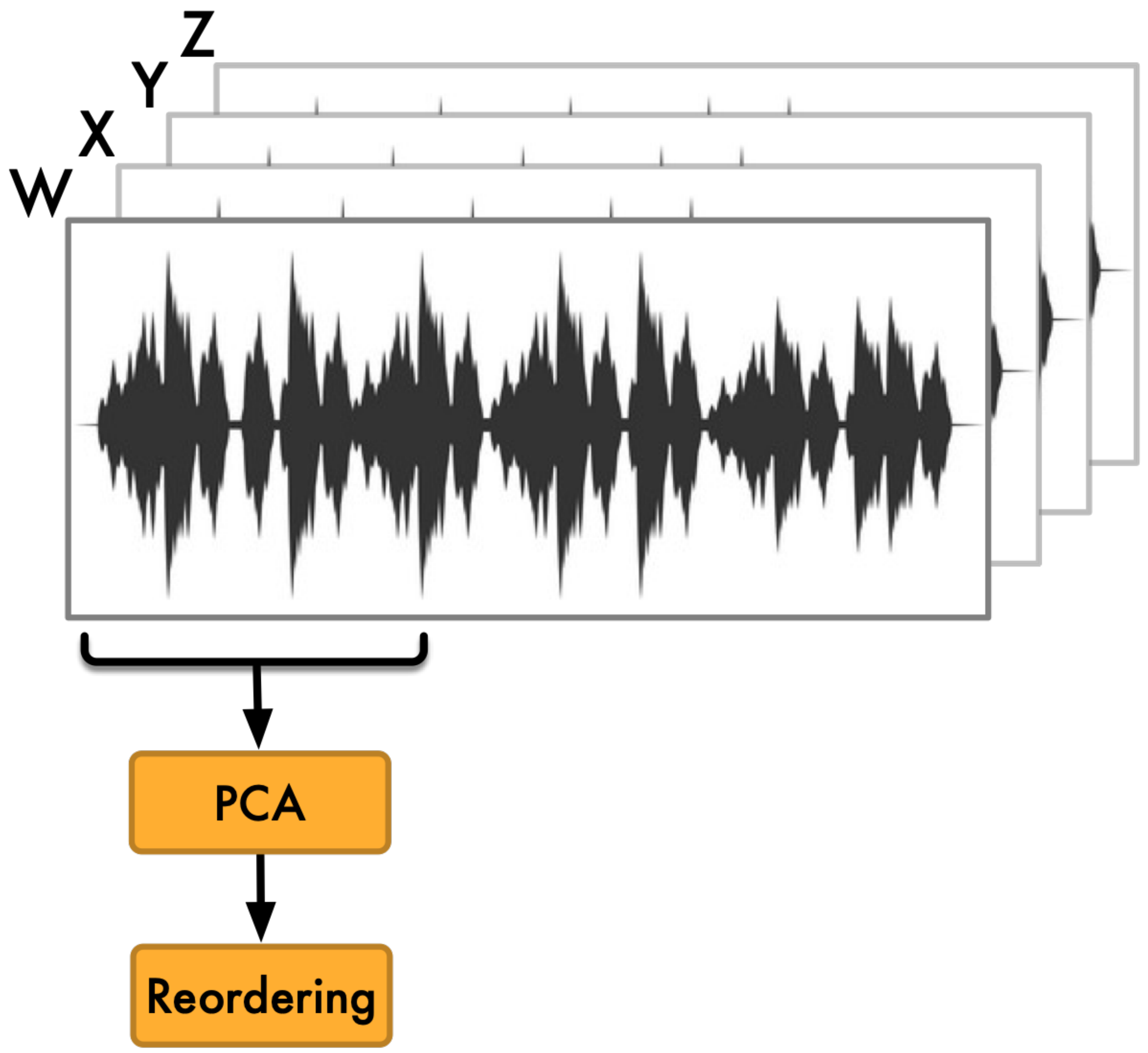
Realignment of beamformer between frames t and t-1

A permutation is found to maximize similarity between the two eigenvector bases.

The similarity being defined as :

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

The Hungarian algorithm was used to find the optimal combination.



Matrixing as beaminforming

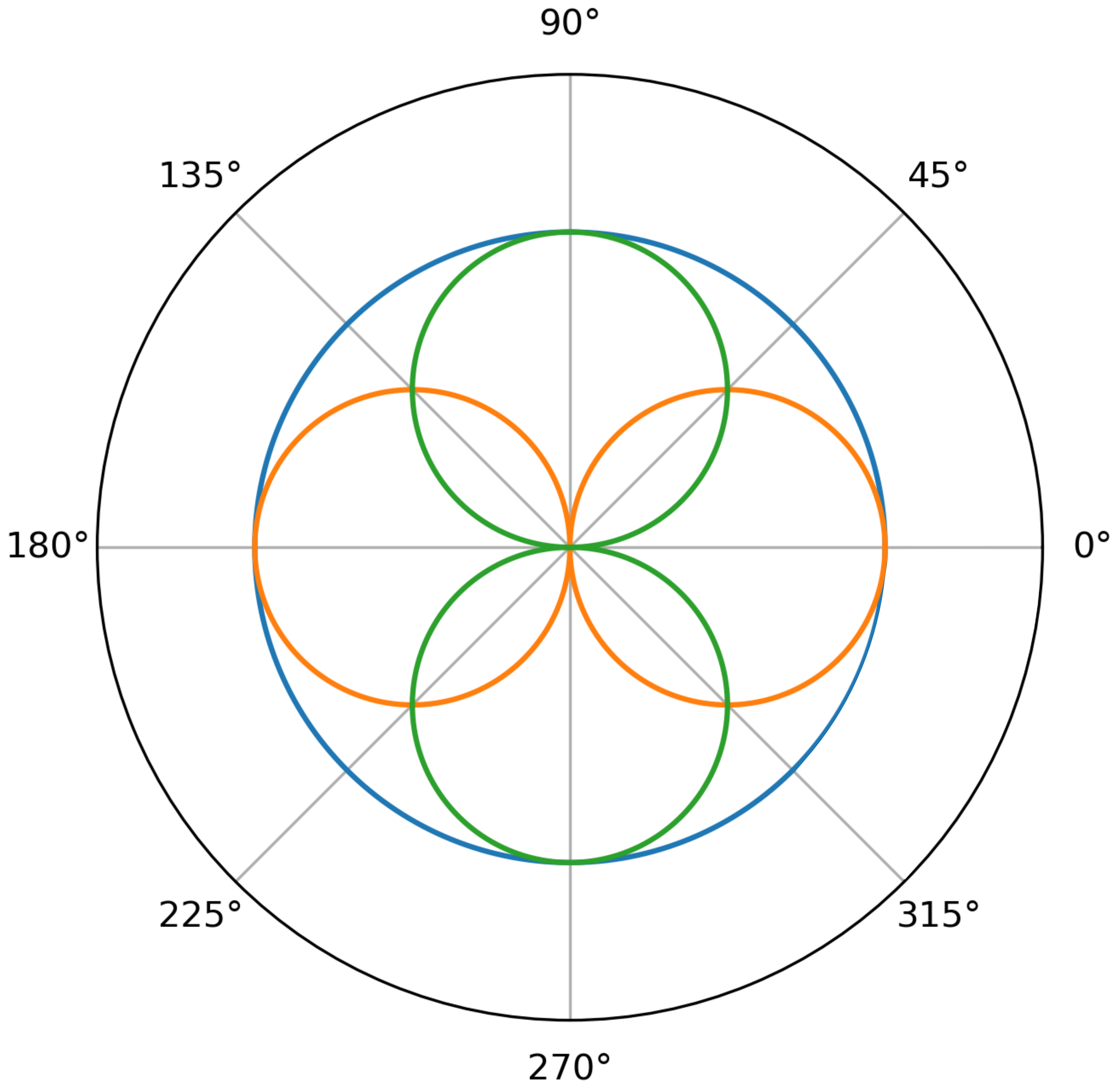
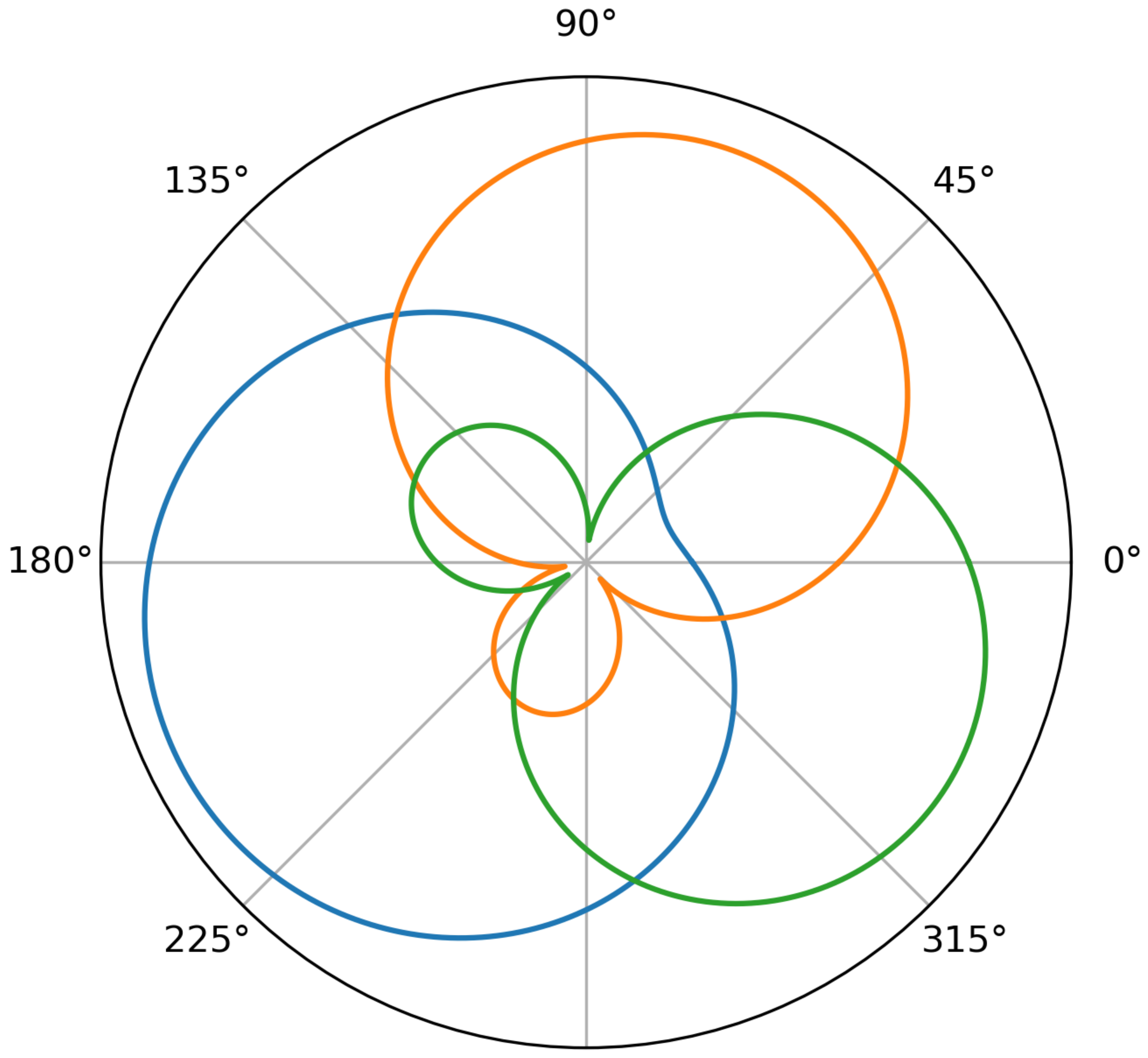
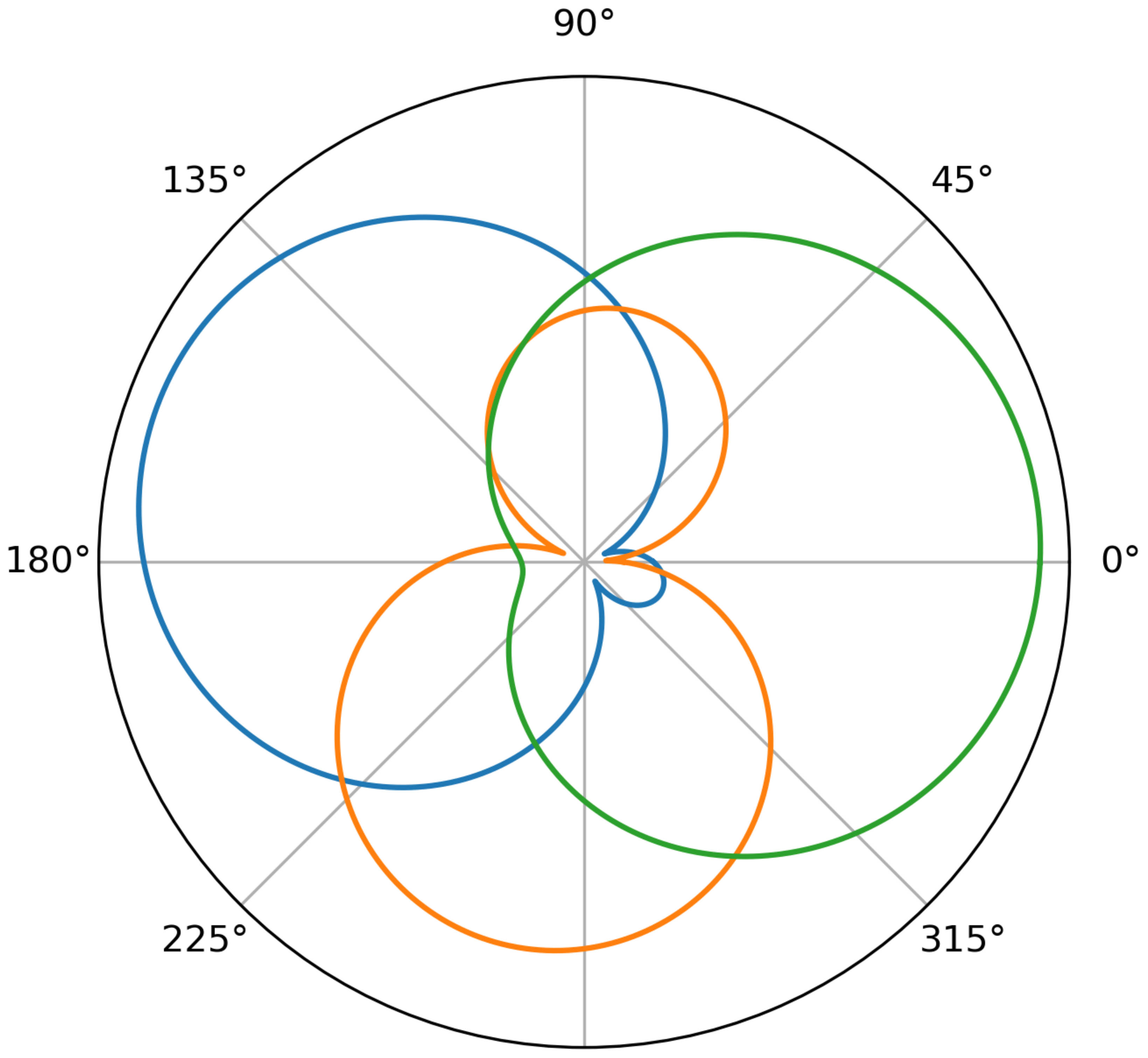


Fig. 3 plot of the regular pattern of spherical harmonics

Interpolation



t - *1*



t

Quaternion-Based Interpolation

Quaternions generalize complex numbers

$$q = a + b.i + c.j + d.k \text{ where } i^2 = j^2 = k^2 = ijk = -1$$

They are often used to represent 3D rotations. A 3D rotation of angle θ and axis n can be represented by a unit quaternion

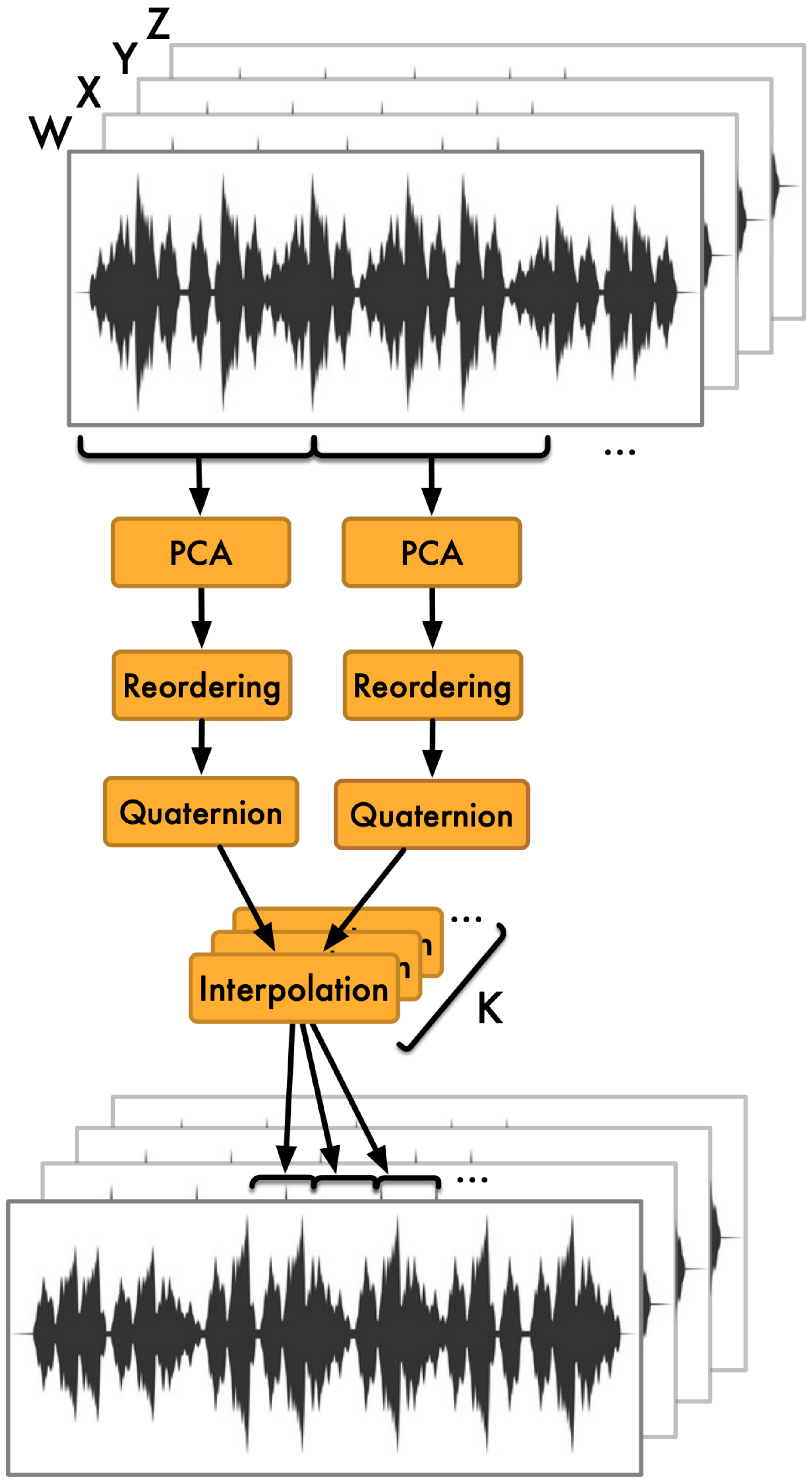
$$q = \cos(\theta/2) + \sin(\theta/2).n$$

The 4D rotation matrix V_t can be decomposed into a pair of quaternions by the Cayley's factorization [Perez-Gracia]

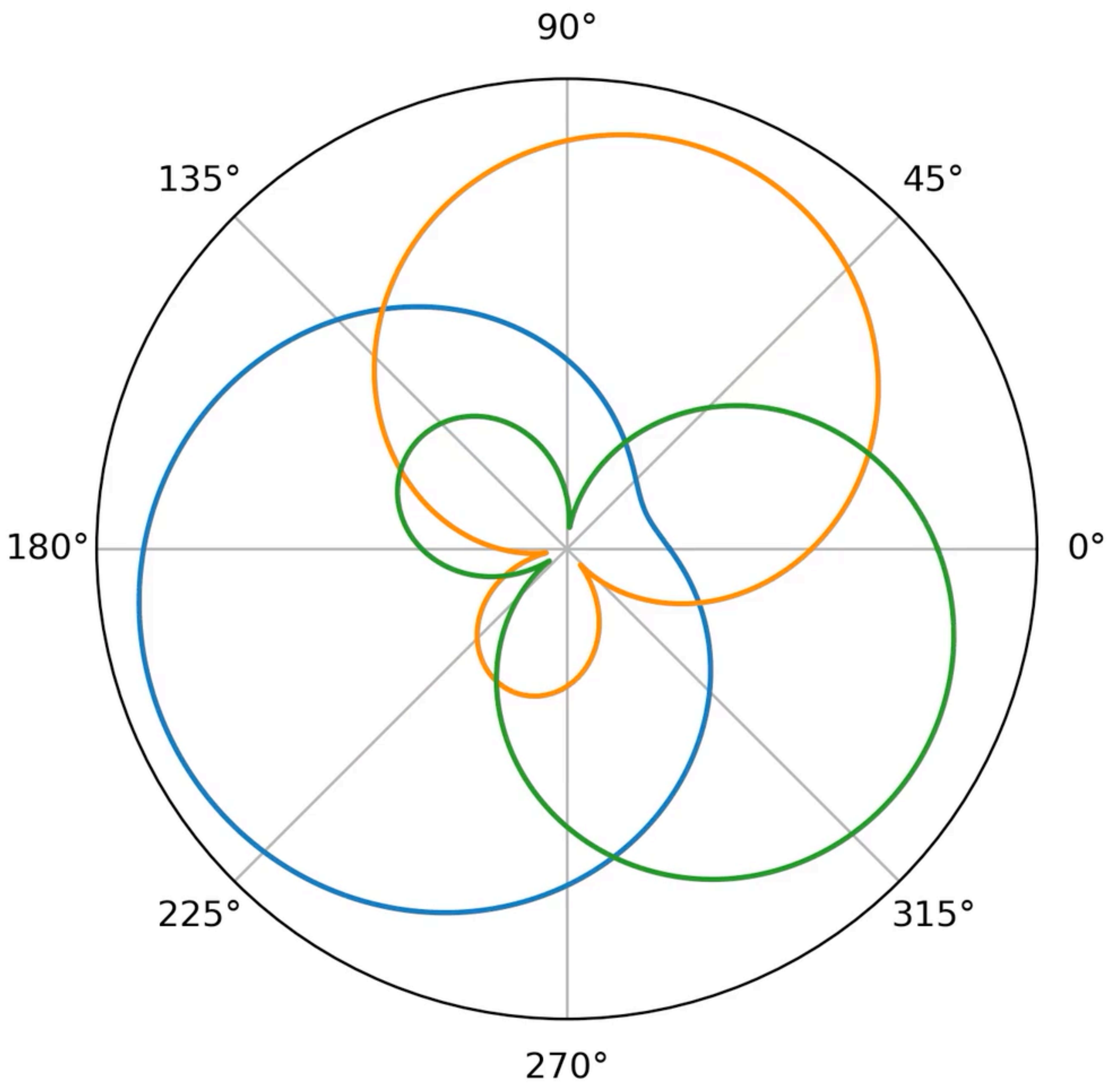
Quaternions in frames t and $t-1$ are interpolated by spherical linear interpolation (slerp)

$$slerp(q_L, q_{L'}, \gamma) = q_1(q_1^{-1}q_2)^{\gamma}$$

Where $\gamma = \frac{k}{K}$ and k is the subframe index.



Interpolation



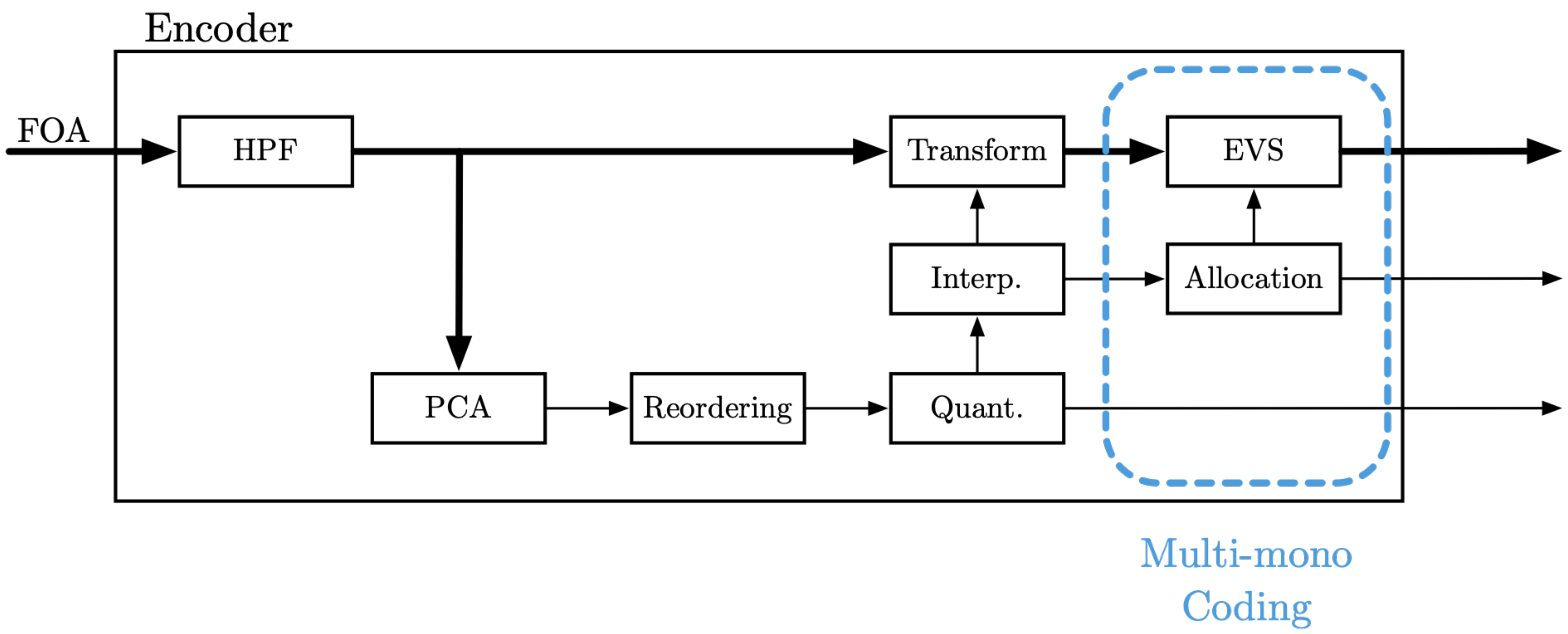
32 interpolations peer 20 ms frame
(subframes of 0.625 ms)

Adaptive bitrate allocation

A global bitrate are define, it must be divide among the components.

After decorrelation, each components don't have the same importance.

An adaptive bit allocation is necessary to optimize quality



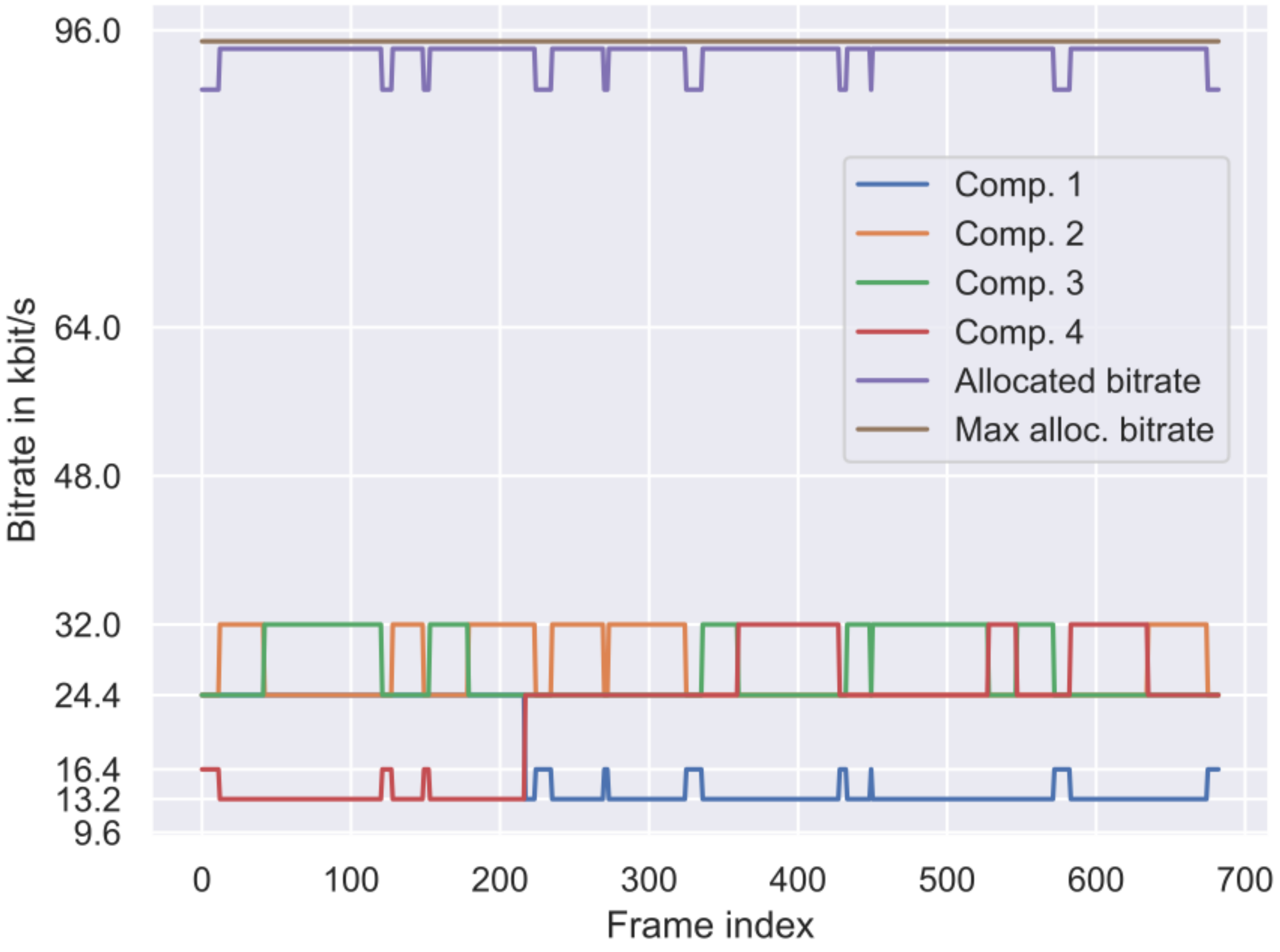
Adaptive bitrate allocation between components

An adaptive bit allocation is necessary to optimize quality.

The audio quality was modeled by the MOS score and weighted by the energy component

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

- Where $\mathrm{b_i}$ and $\mathrm{E_i}$ are respectively the bit allocation and the energy of the $\mathrm{i^{th}}$ channel in the current frame and $\mathrm{Q(b_i)}$ is a quality score.



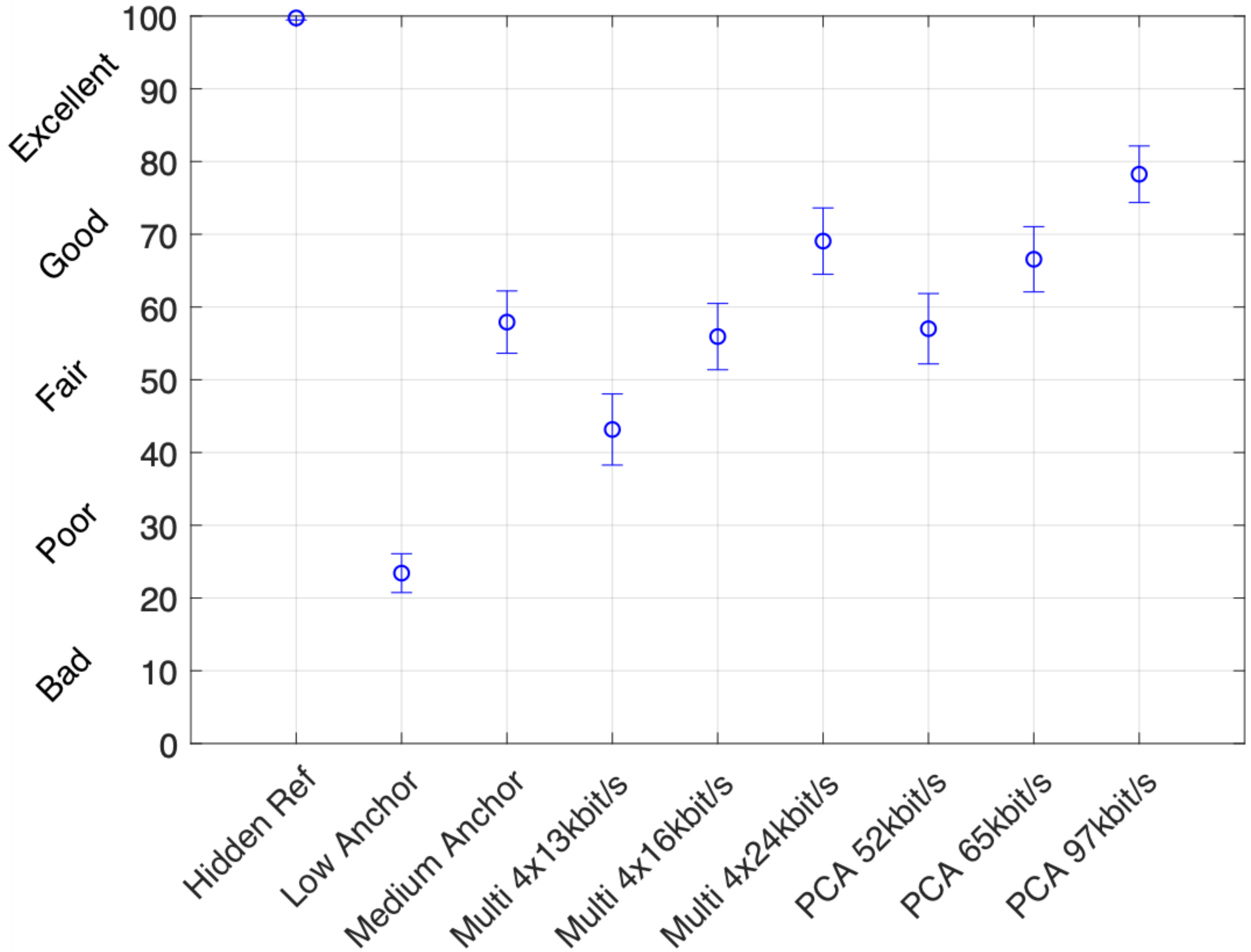
Results

Test conditions

- MUSHRA test
- 3 evaluated bitrates for each method (Naive and PCA)
- 11 participants: expert or experienced listeners

Test conclusions

- At the same bitrate, our approach is better than multimono
- Most spatial artefacts are removed



Conclusion

- Our approach proposed a spatial extension to existing codecs to handle FOA.
- To avoid spatial artifacts, the ambisonic components are decorrelated by PCA.
- The signal continuity is guaranteed by PCA matrix interpolation in (double) quaternion domain.
- Subjective test results showed significant improvements over naive multi-mono coding.

Thanks for your attention

Any questions ?

Pierre MAHE - Orange Labs and University of La Rochelle, France

pierre.mahe@orange.com

References

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

[Daniel] J. Daniel, Représentation de champs acoustiques, applica- tion à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia, Ph.D. thesis, Uni- versité Paris 6, 2000.

[Pulkki] V.Pulkki, A.Politis, M.-V.Laitinen, J.Vilkamo and J.Ahonen, “First-order directional audio coding (DirAC),” in Para- metric Time-Frequency Domain Spatial Audio, chapter 5. 2018.

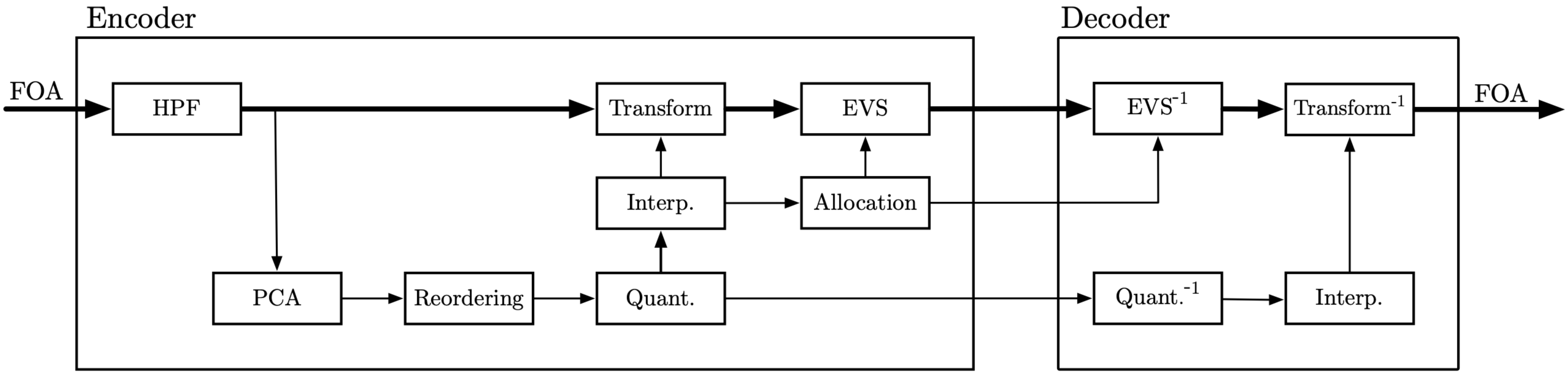
[Politis] A. Politis, S. Tervo, and V. Pulkki, “Compass: Coding and multidirectional parameterization of ambisonic sound scenes,” in Proc. ICASSP, 2018

[Herre] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, “MPEG-H audio - the new standard for universal spatial/3D audio coding,” Audio Eng. Soc., 2015.

[Perez-Gracia] 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.

Further Slides

Codec Diagram



MUSHRA Test

MUSHRA stands for MUltiple Stimuli with Hidden Reference and Anchor [ITU-R BS 1534-3]

For each item, subjects evaluated the quality with a scale ranging of 0 to 100.

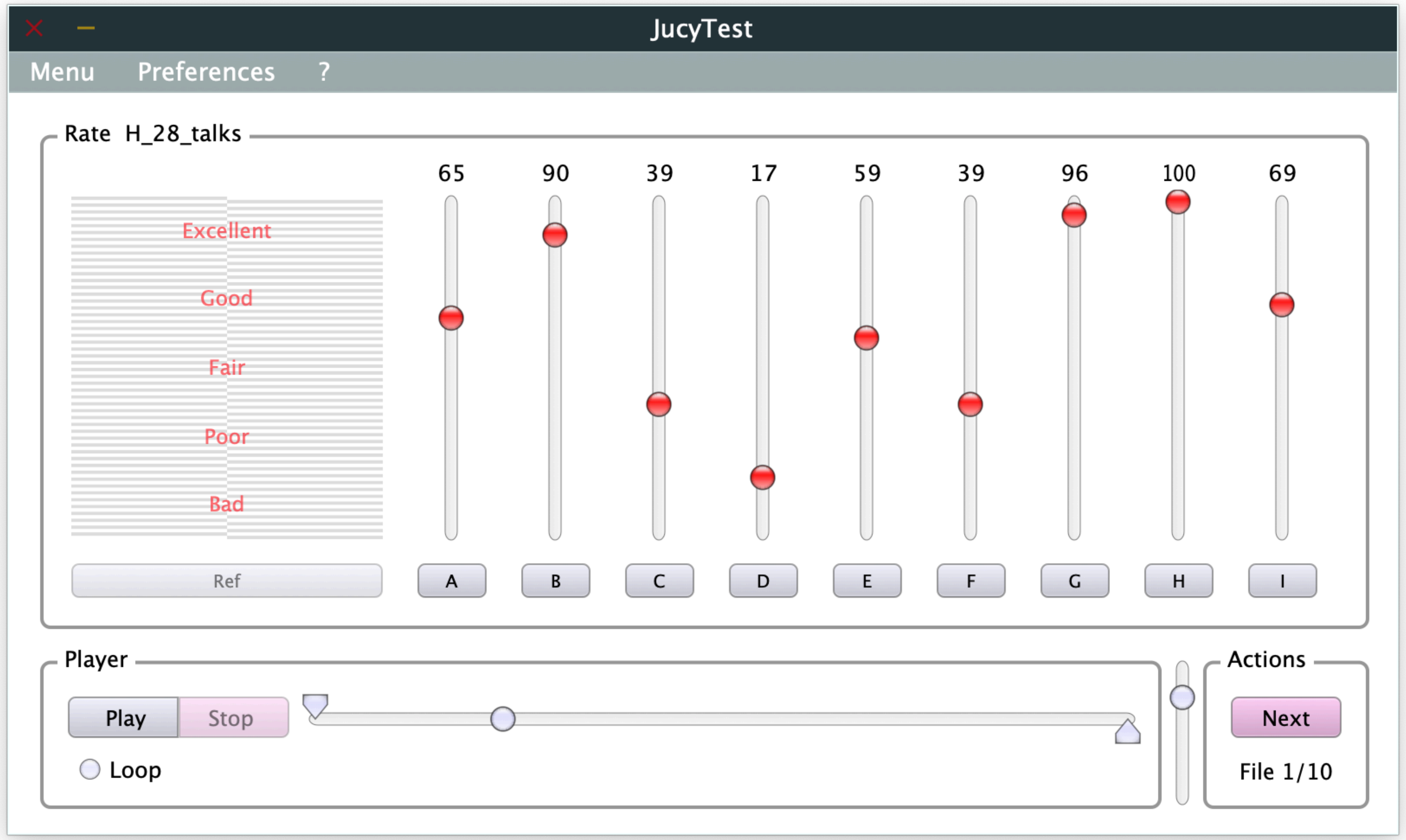
This interval is divided in 5 sections from bad (0-20) to excellent (80-100).

Three specific items: the hidden reference (FOA) and two anchors.

Anchor spatial reduction :

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

with α = 0.65 and α = 0.8 for the low and medium anchors.



Rate H_28_talks

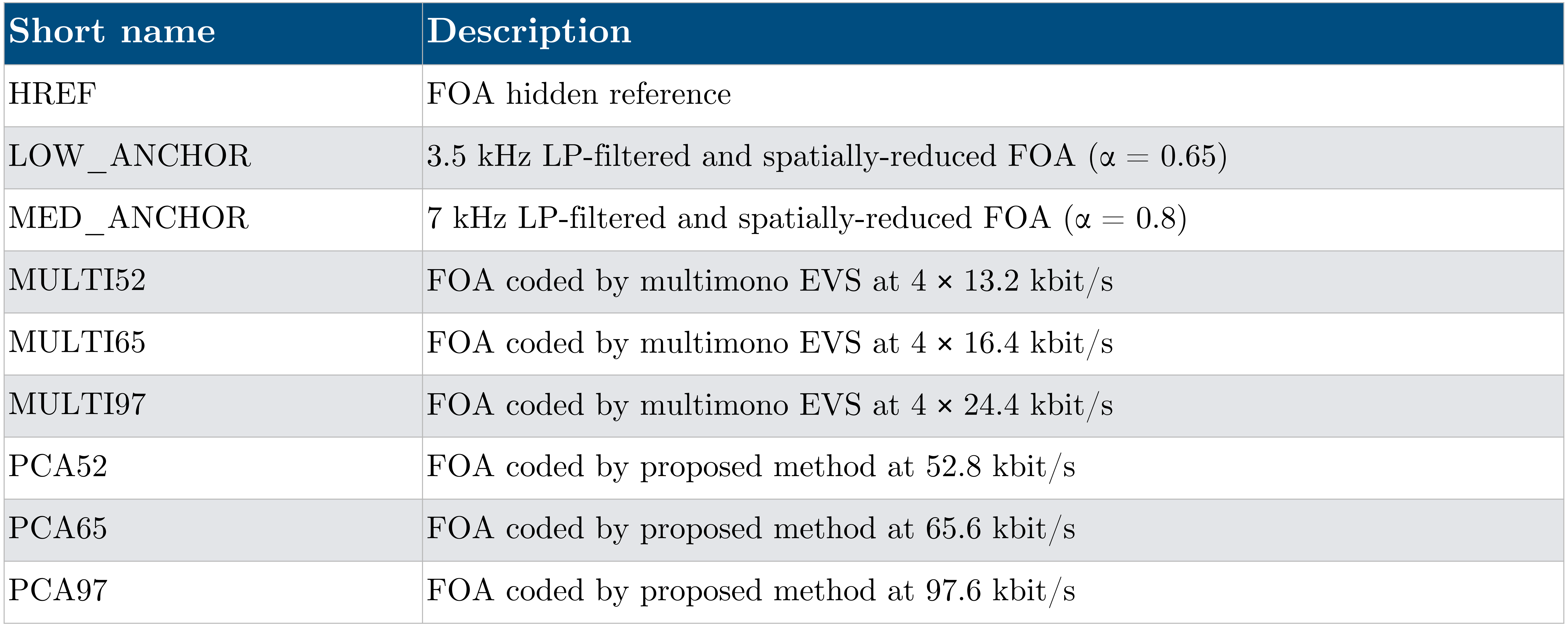
A	B	C	D	E	F	G	H	I
65	90	39	17	59	39	96	100	69

File 1/10

Fig. MUSHRA Test Interface

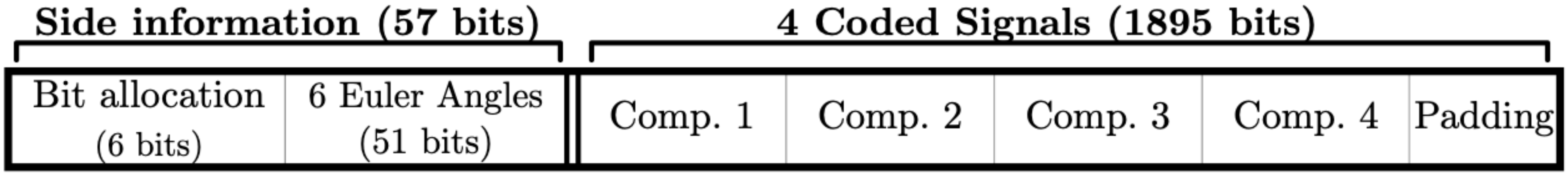
MUSHRA Test

Table 1. 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

Bitstream structure



Side information (57 bits)		4 Coded Signals (1895 bits)				
Bit allocation (6 bits)	6 Euler Angles (51 bits)	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Padding

Fig. 5 Bitstream structure example at 4 x 24.4kbit/s