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

4th 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

Context and Motivations

Telephony codecs are mostly limited to mono.

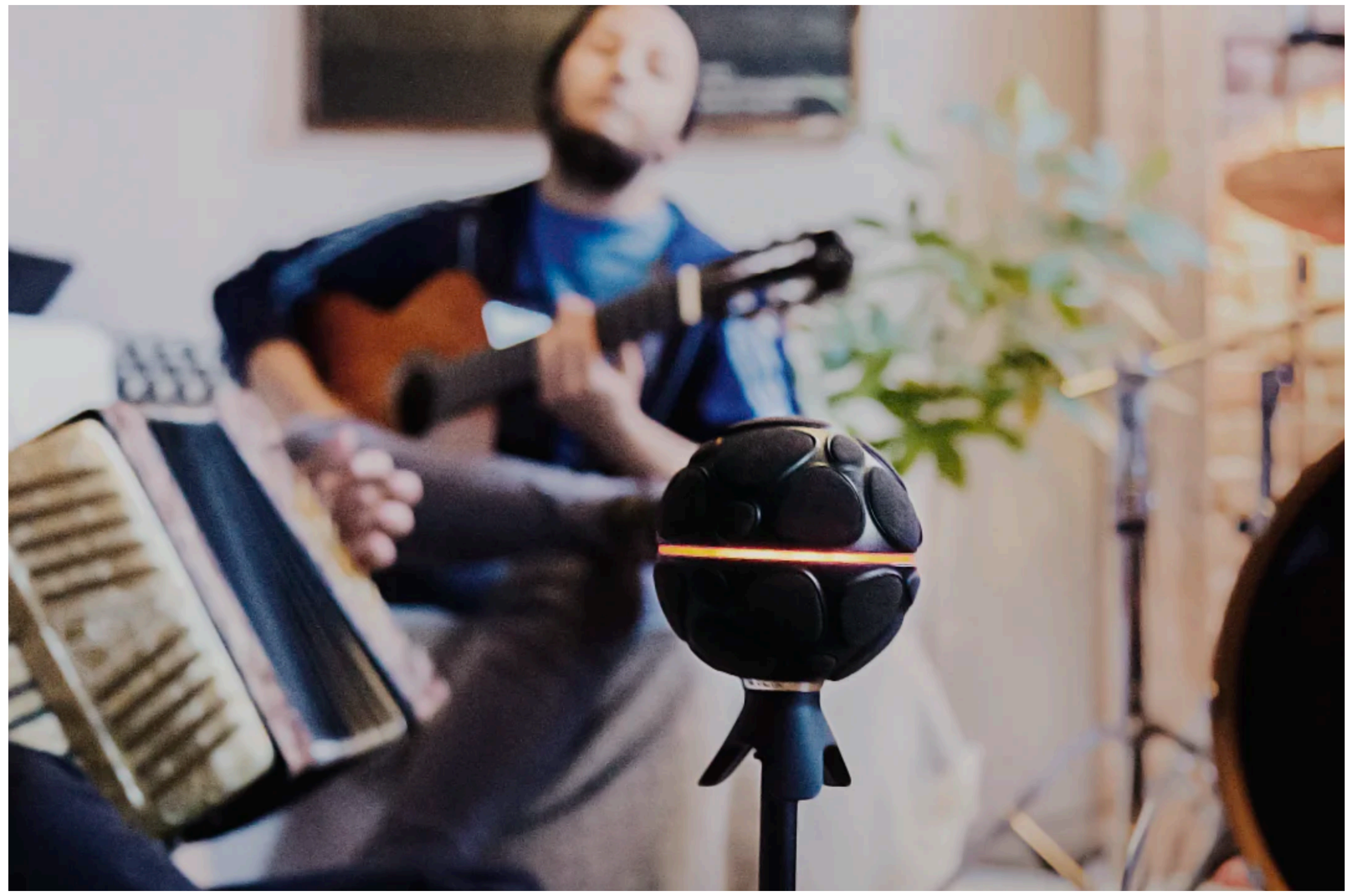
Emergence of devices supporting spatial audio.

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

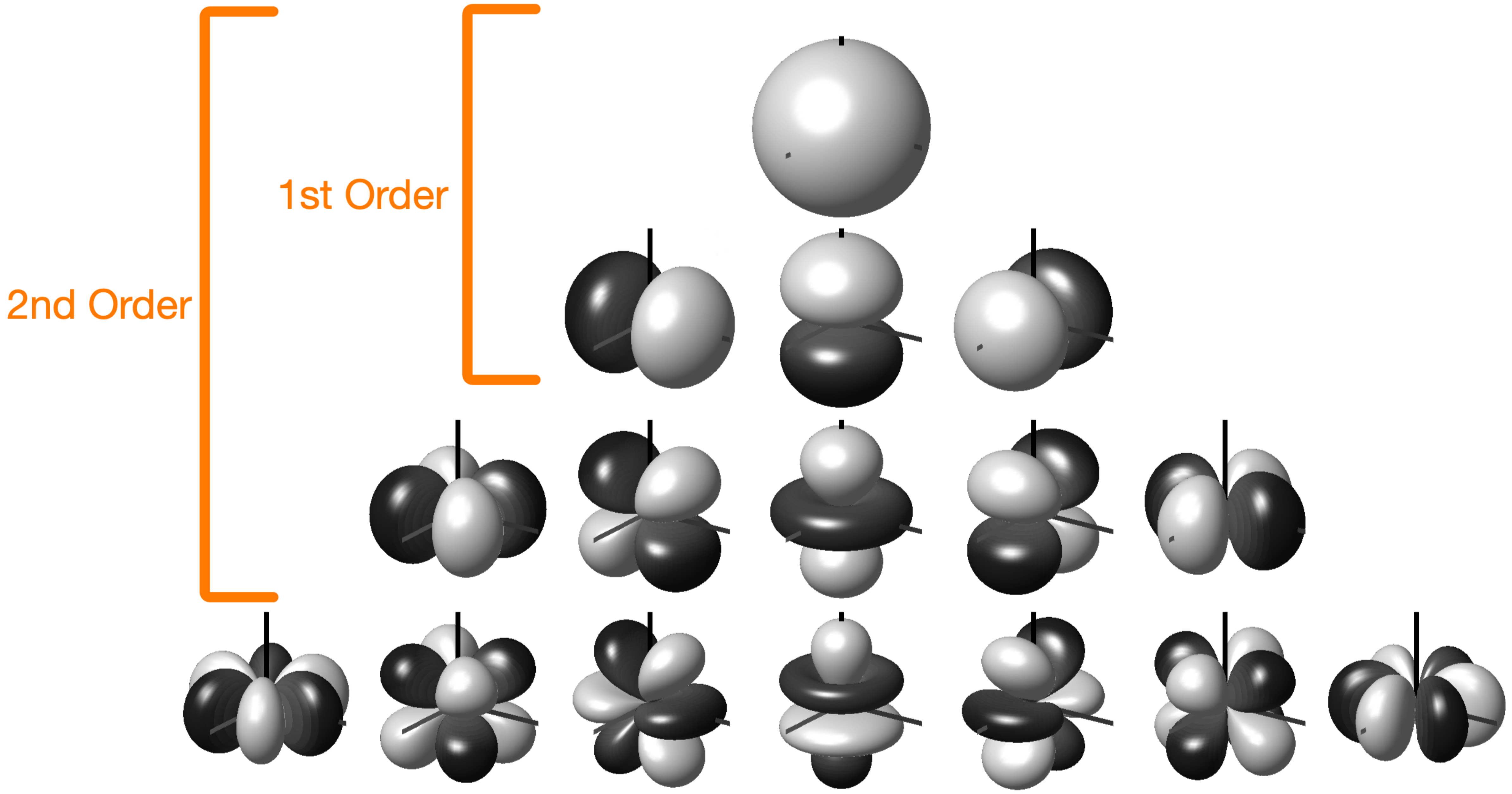


Ambisonics

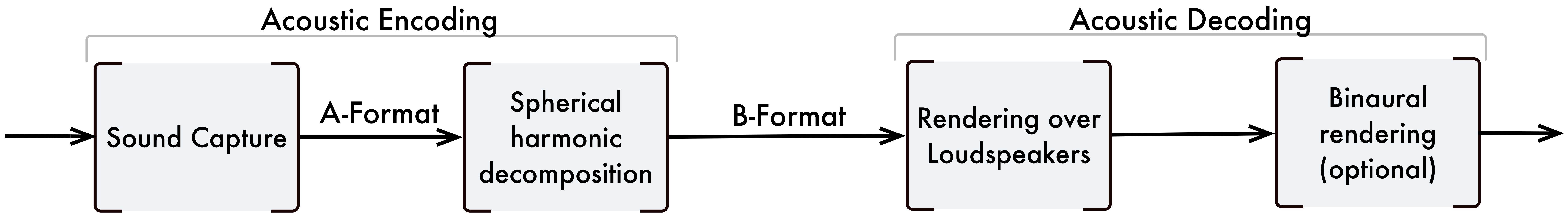
Ambisonics is a decomposition of the sound field into a spherical harmonics basis.

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

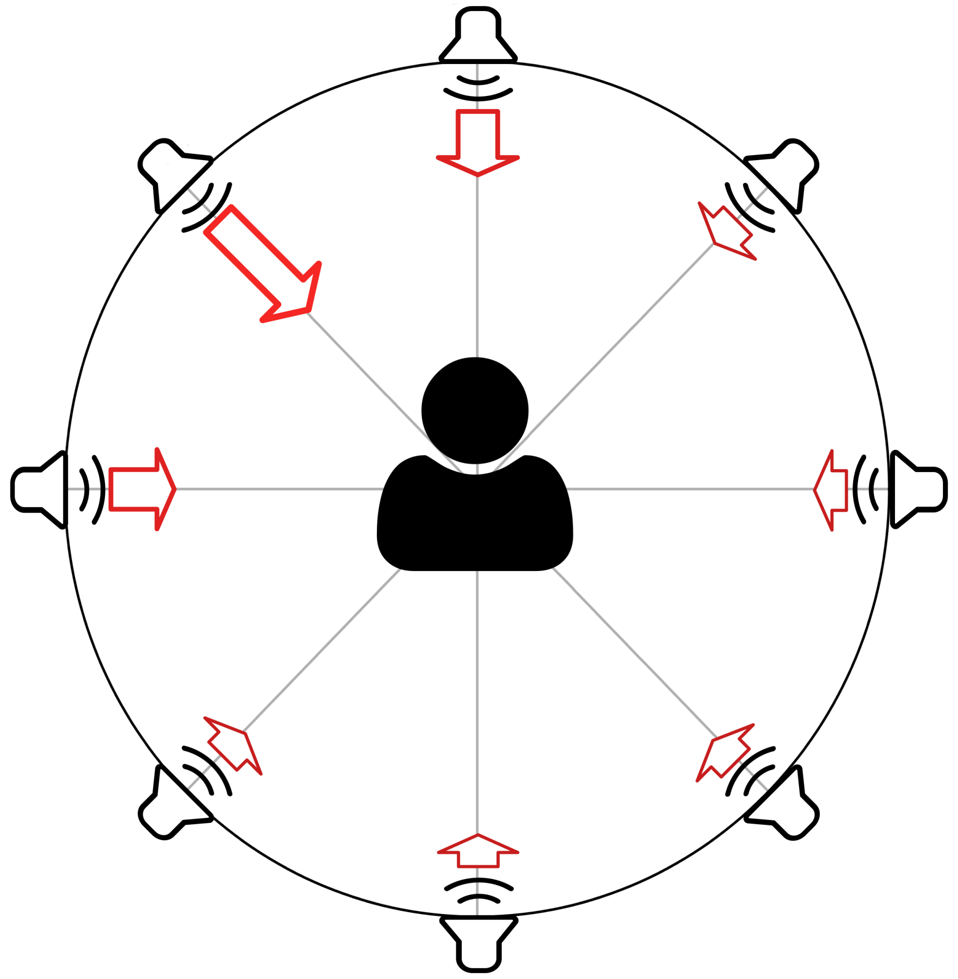
For an order N, the basis dimension is $(N+1)^2$



Ambisonics



Acoustic Encoding
Acoustic Decoding
Sound Capture
A-Format
Spherical harmonic decomposition
B-Format
Rendering over Loudspeakers
Binaural rendering (optional)



Existing approaches

Scene analysis and sources extractions

Assumptions on the scene (number of sources…)
→ If scene analysis do wrong decision, the quality are strongly impact

Fixed matrissing

Matrissing the components with fixe coefficients
No assumptions on the scene
→ The improvements are not very significative

Hybrid approaches

Used PCA or SVD to extract sources
Residual is downmix and transmitted
→ It requires important amount of metadata to garantie signal continuity between frames

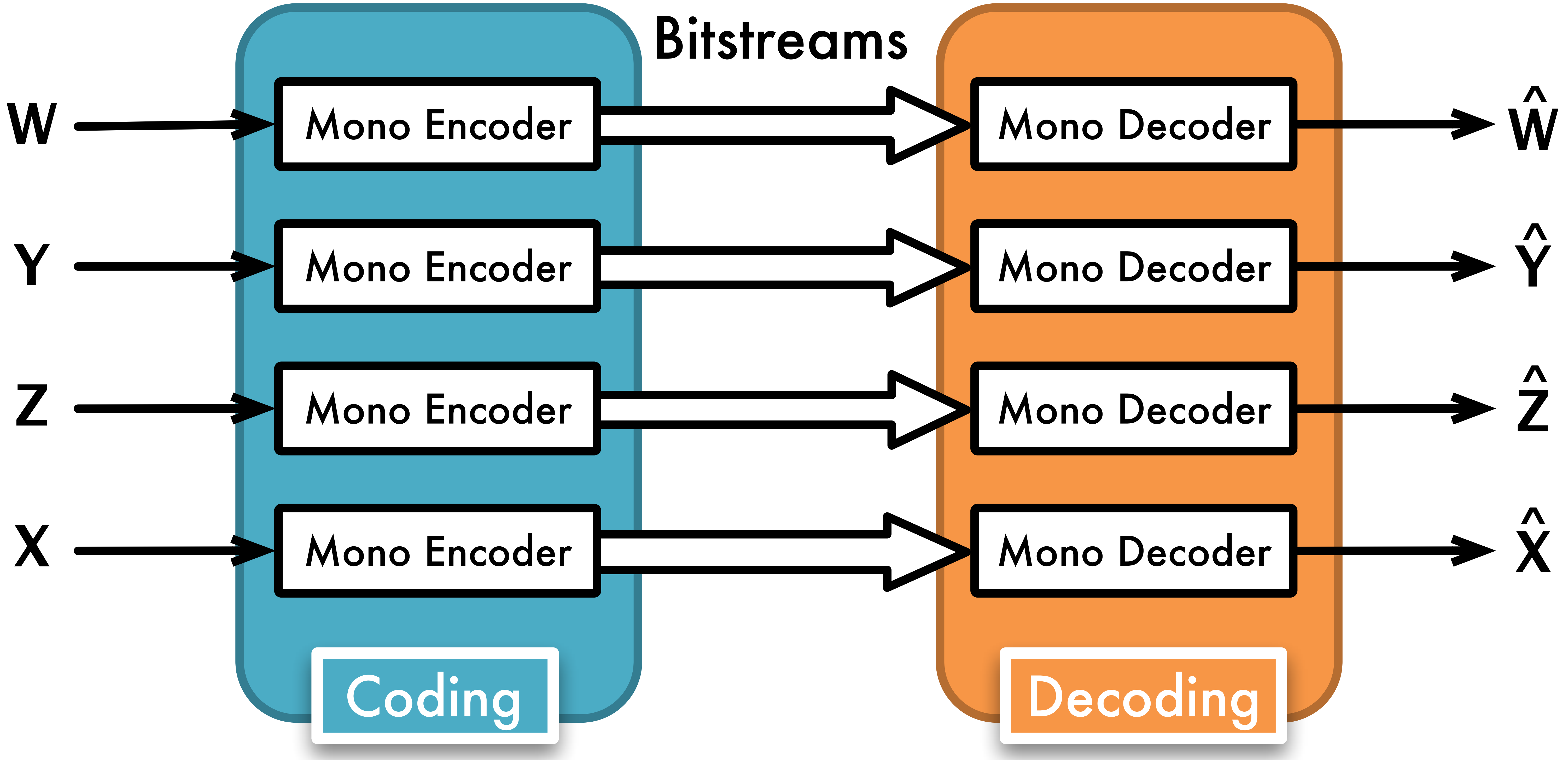
Is the naïve approach not enough ?

Multi-mono approach

Each ambisonic component is coded independently by a mono codec.

Bitrate is uniformly distributed between components.

For listening tests, binaural rendering is used.



Is the naïve approach not enough ?

Test conditions

MUSHRA test

Mono coding by 3GPP EVS

4 evaluated bitrates : 4x13.2, 4x16.4, 4x24.4, 4x48 kbit/s

7 participants: expert or experienced listeners

Results

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

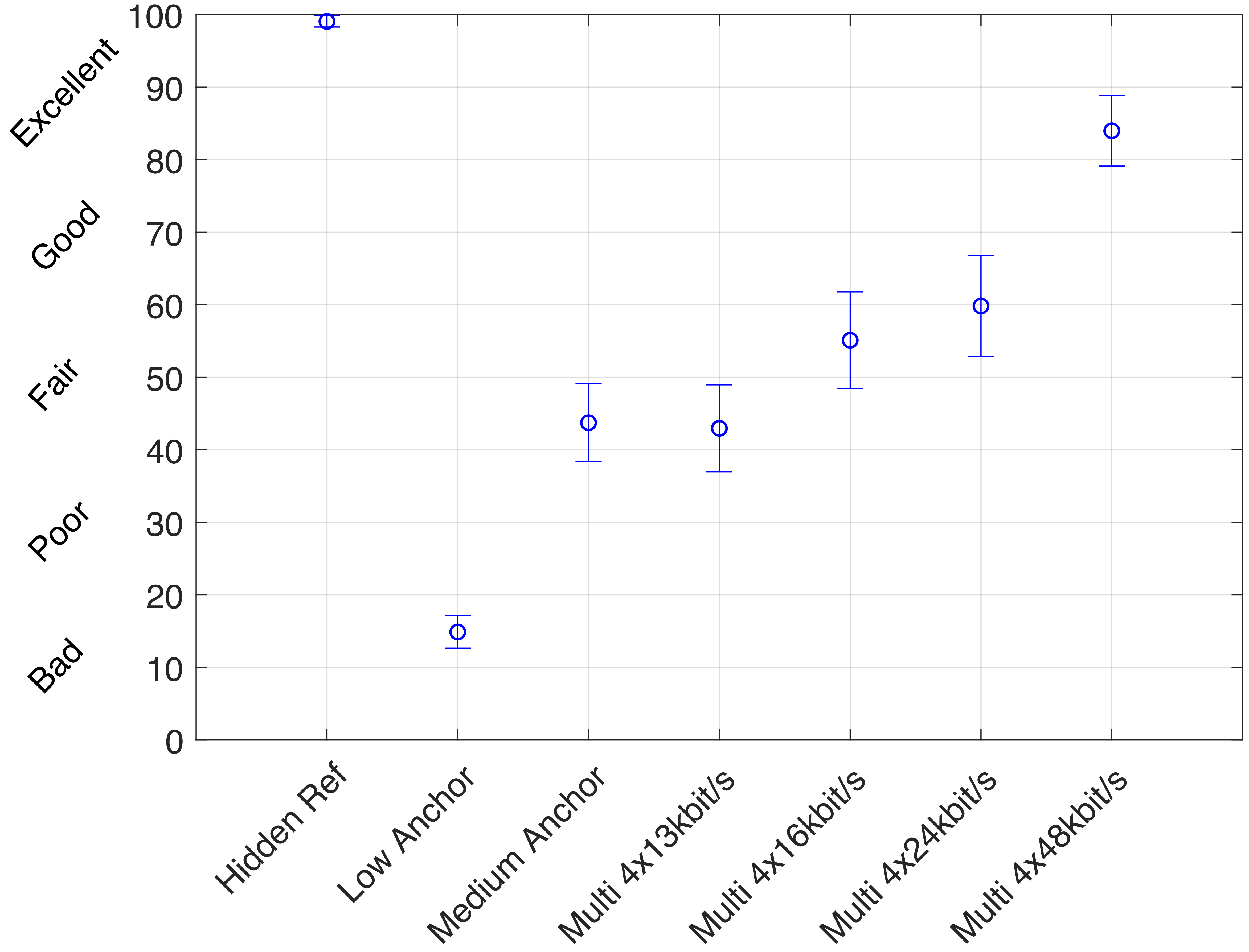
Several artefacts :

— Diffuse noise

— Source positions are pushed to the front

— Spatial blurring for percussive sounds

— Phantom sources



Our approach

Pre-process the ambisonic components

- Decorrelate components to avoid spatial Artifact
- No assumptions on the scene
- Garantie signal continuity without add extra meta-data
- Extend existing codecs

Our approach

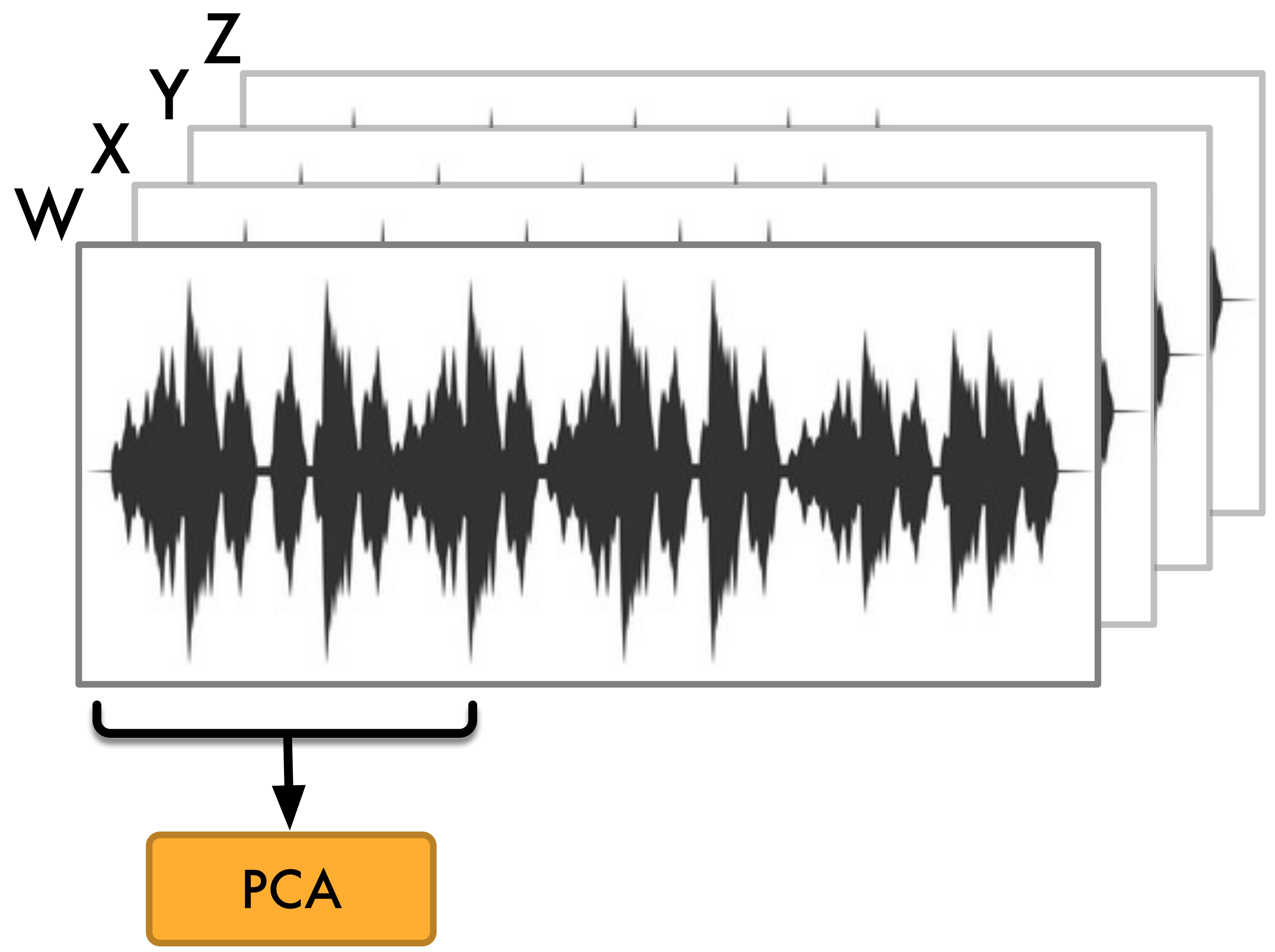
Compute PCA coefficients for frame *t*

Compute the covariance matrix C_{xx}

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

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

where $\mathbf{V}$ is the eigenvector matrix and $\Lambda = diag(\lambda_1, \cdots, \lambda_n)$

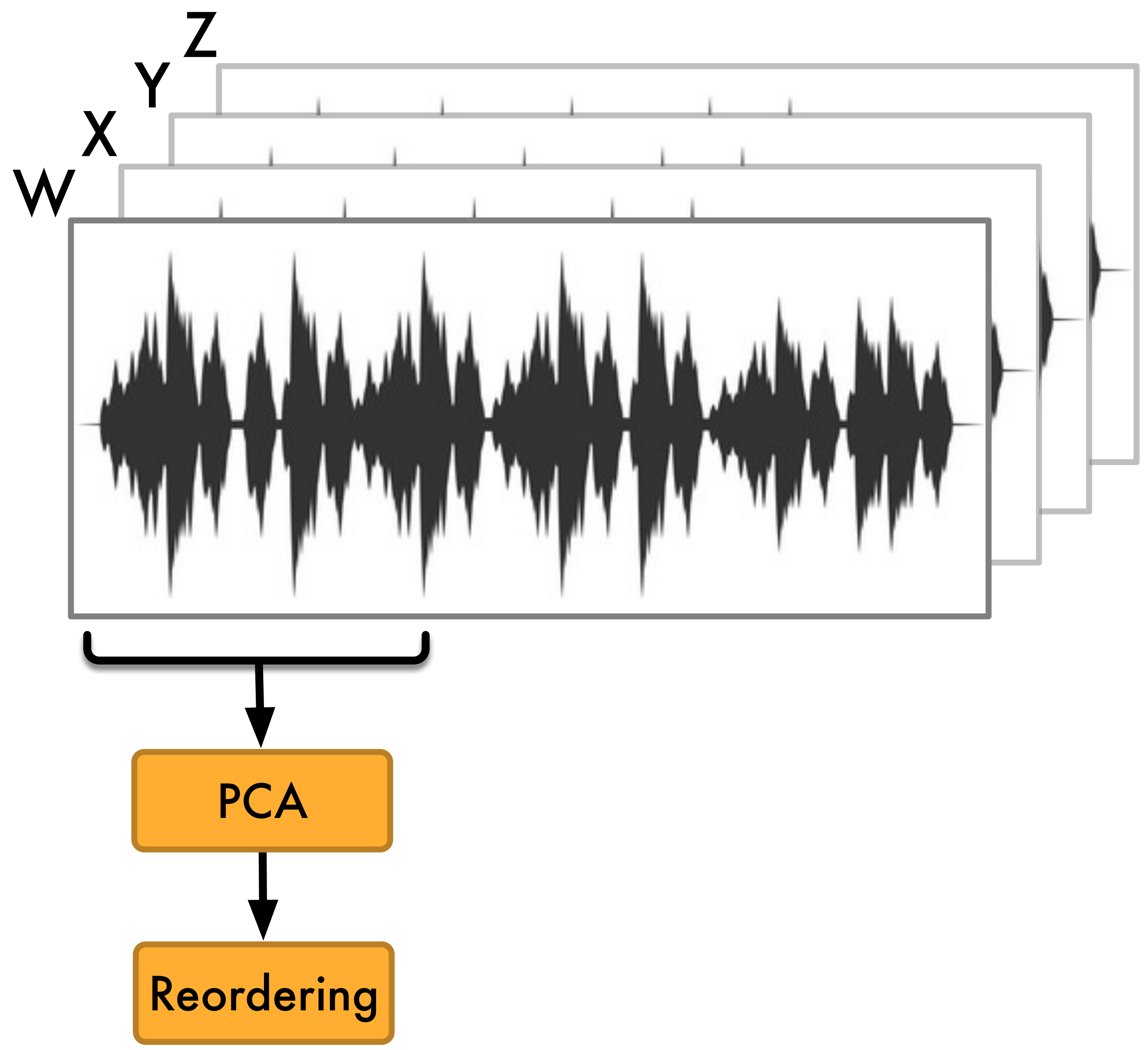


Before PCA

After PCA

…

Our approach



Reordering of eigenvectors 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.

Our approach

Quaternion Interpolation

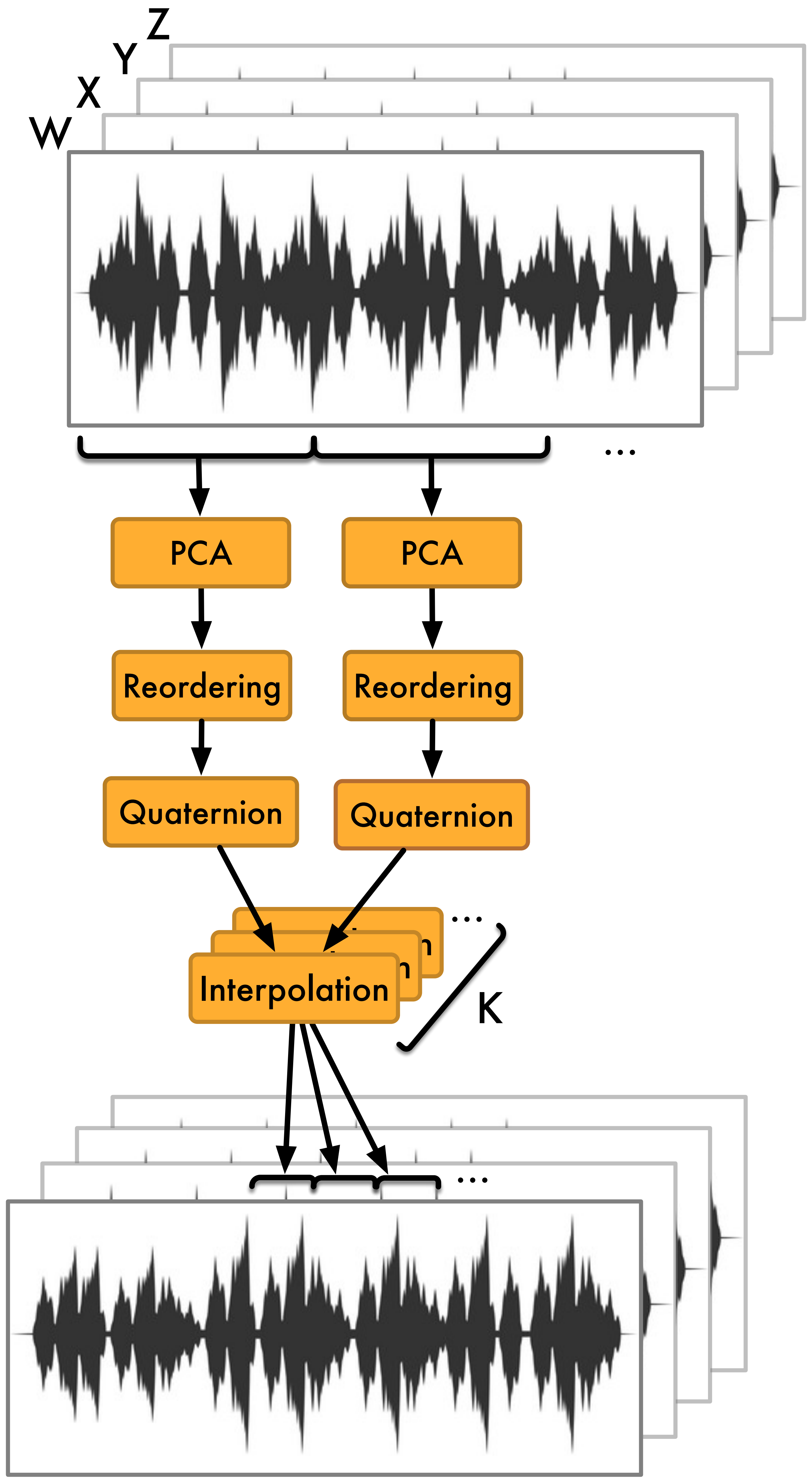
Each frame is subdivided into K subframes.

The eigenvector matrix $\mathbf{V}_t$ is 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)

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

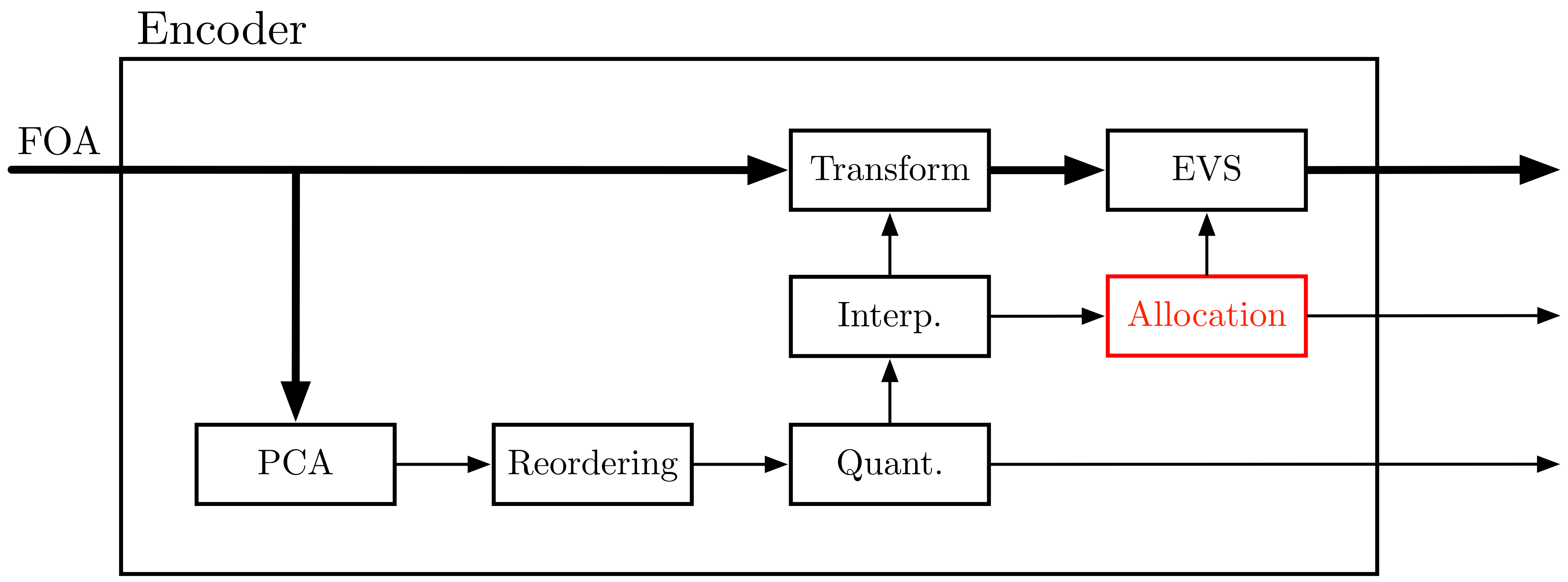
Where $\gamma = \dfrac{k}{K}$ and k is the subframe index.

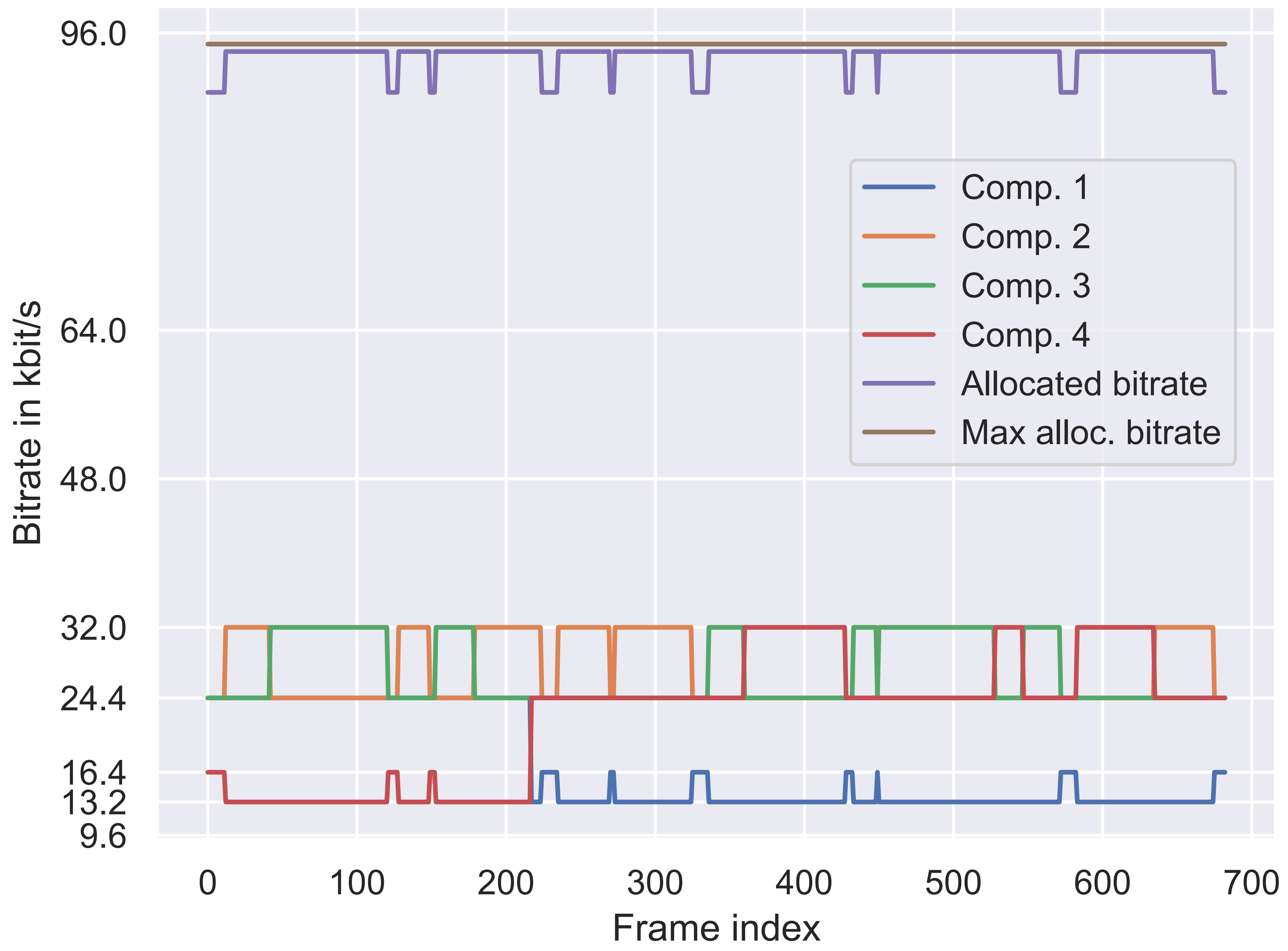


Adaptive bitrate allocation between components

An adaptive bit allocation is necessary to optimize quality.

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





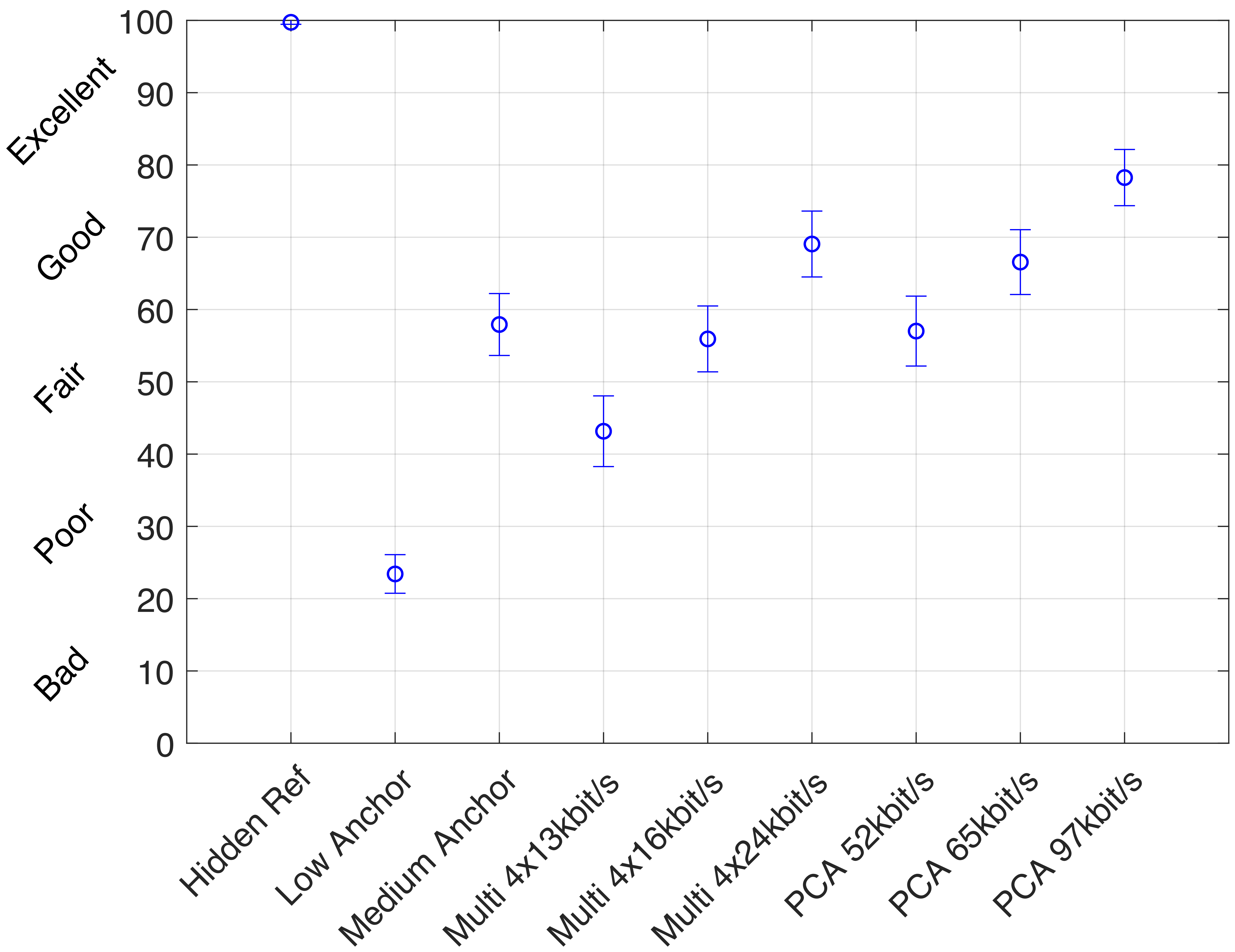
Results

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.*

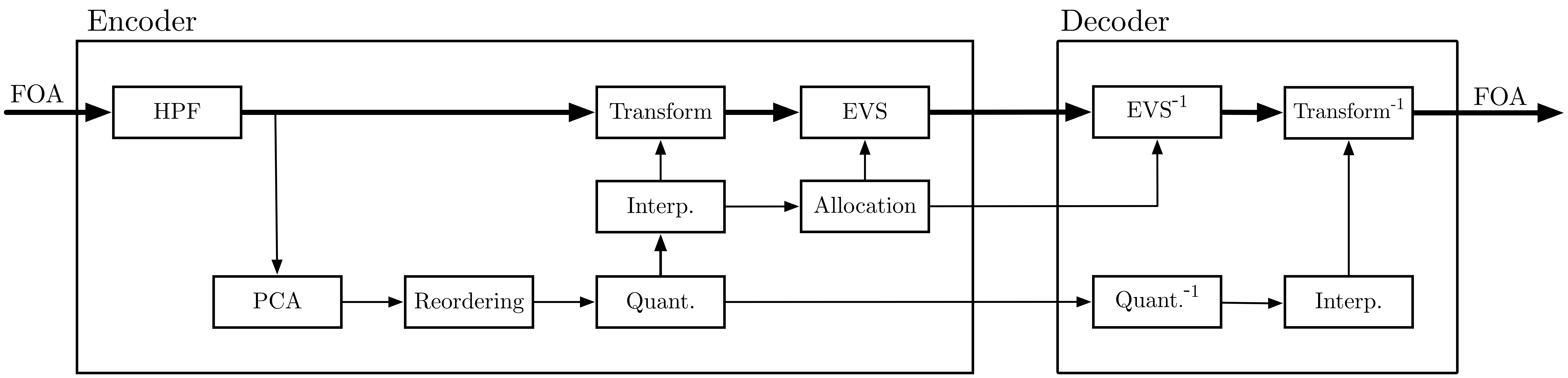
[Politis] A. Politis, S. Tervo, and V. Pulkki, "Compass: Coding and multidirectional parameterization of ambisonic sound scenes," in *Proc. ICASSP*, 2018

[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, normalized by ITU-R.

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

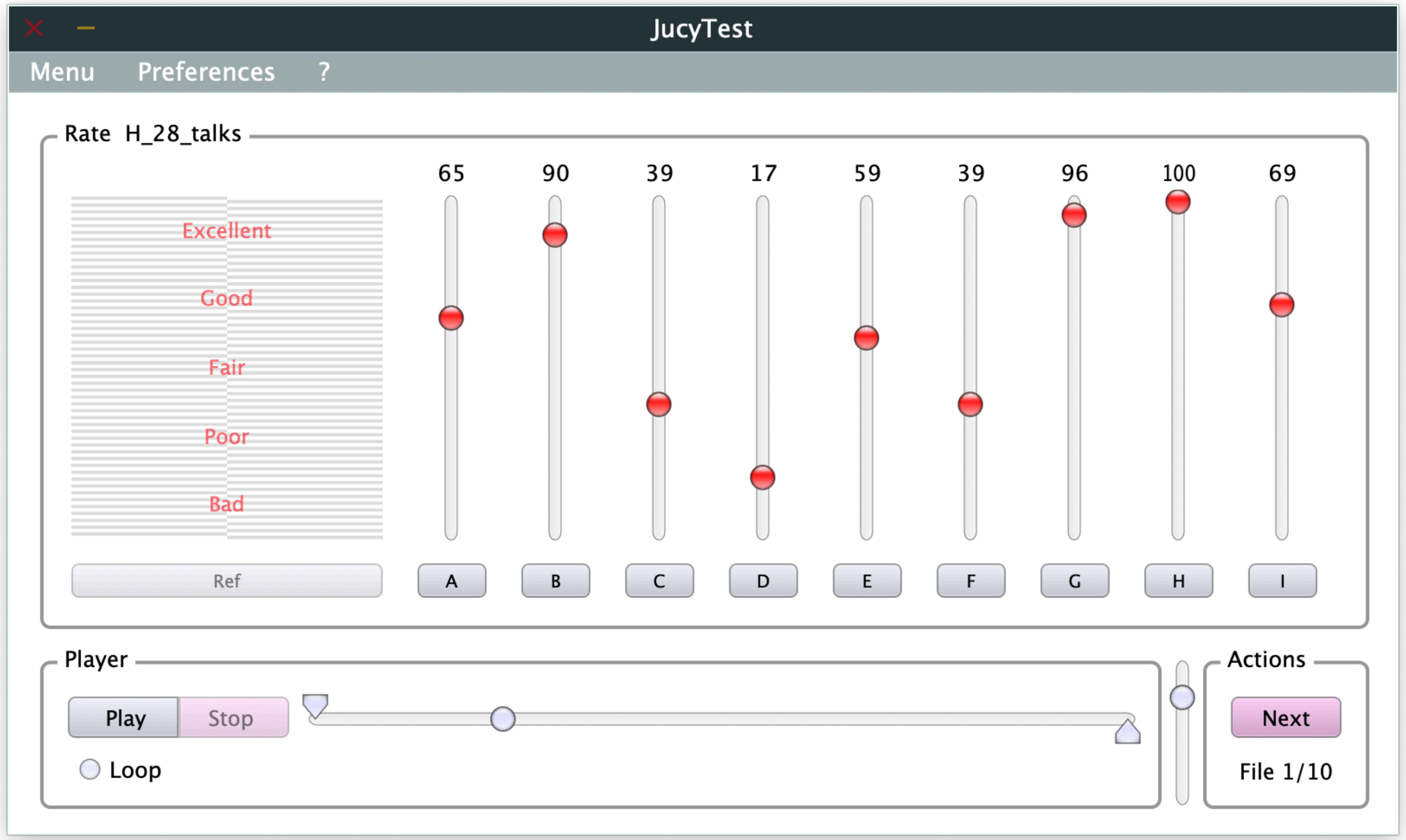
This interval is divided in 5 sections from bad (0-20) to excellent (80-100).

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

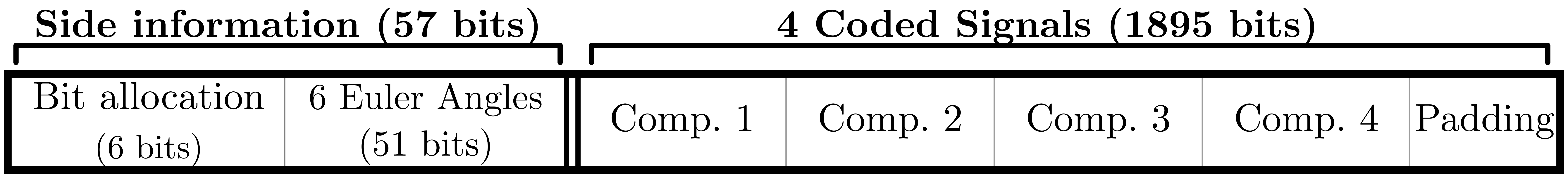
Anchor spatial reduction :

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

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



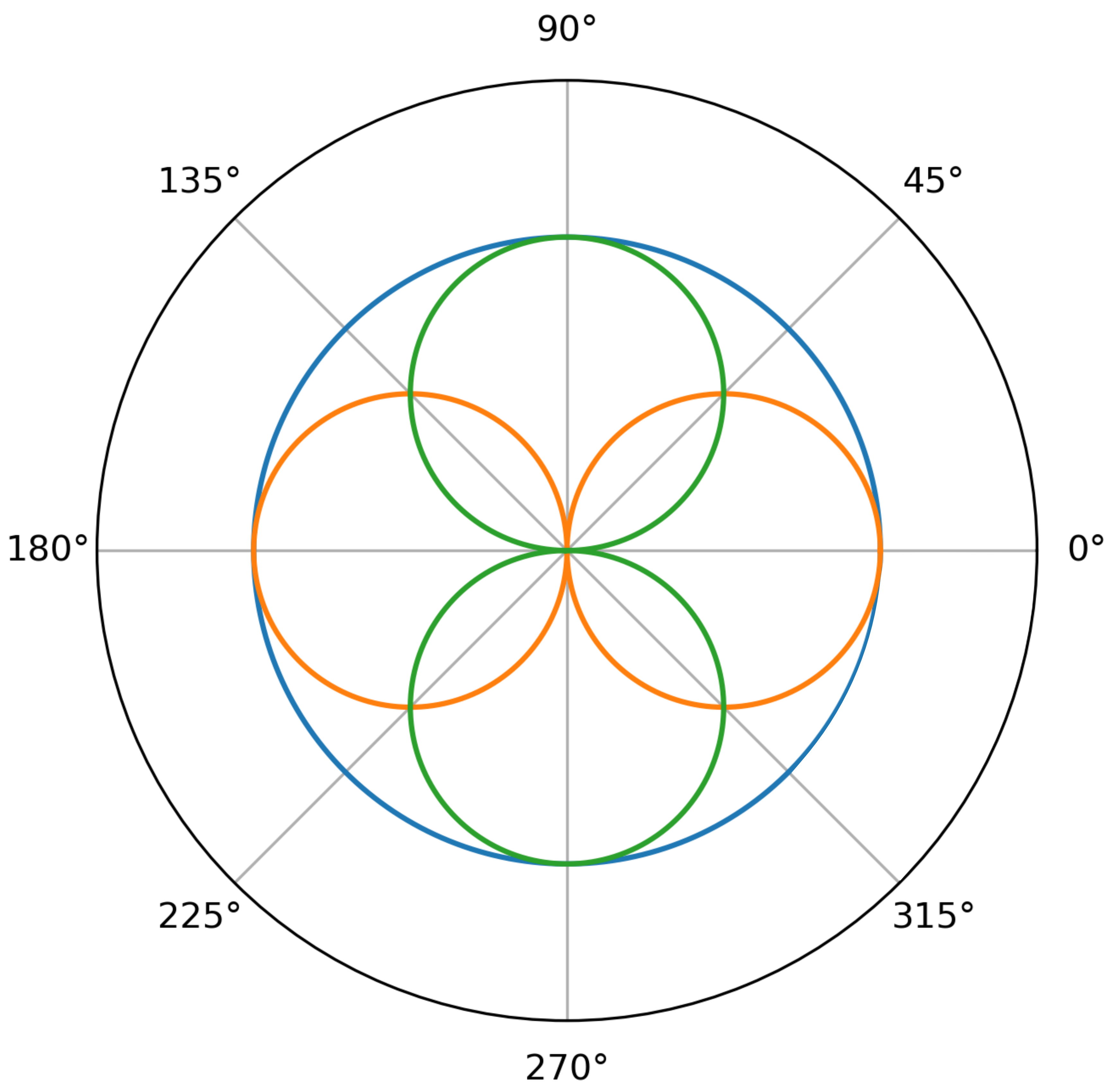
Bitstream structure



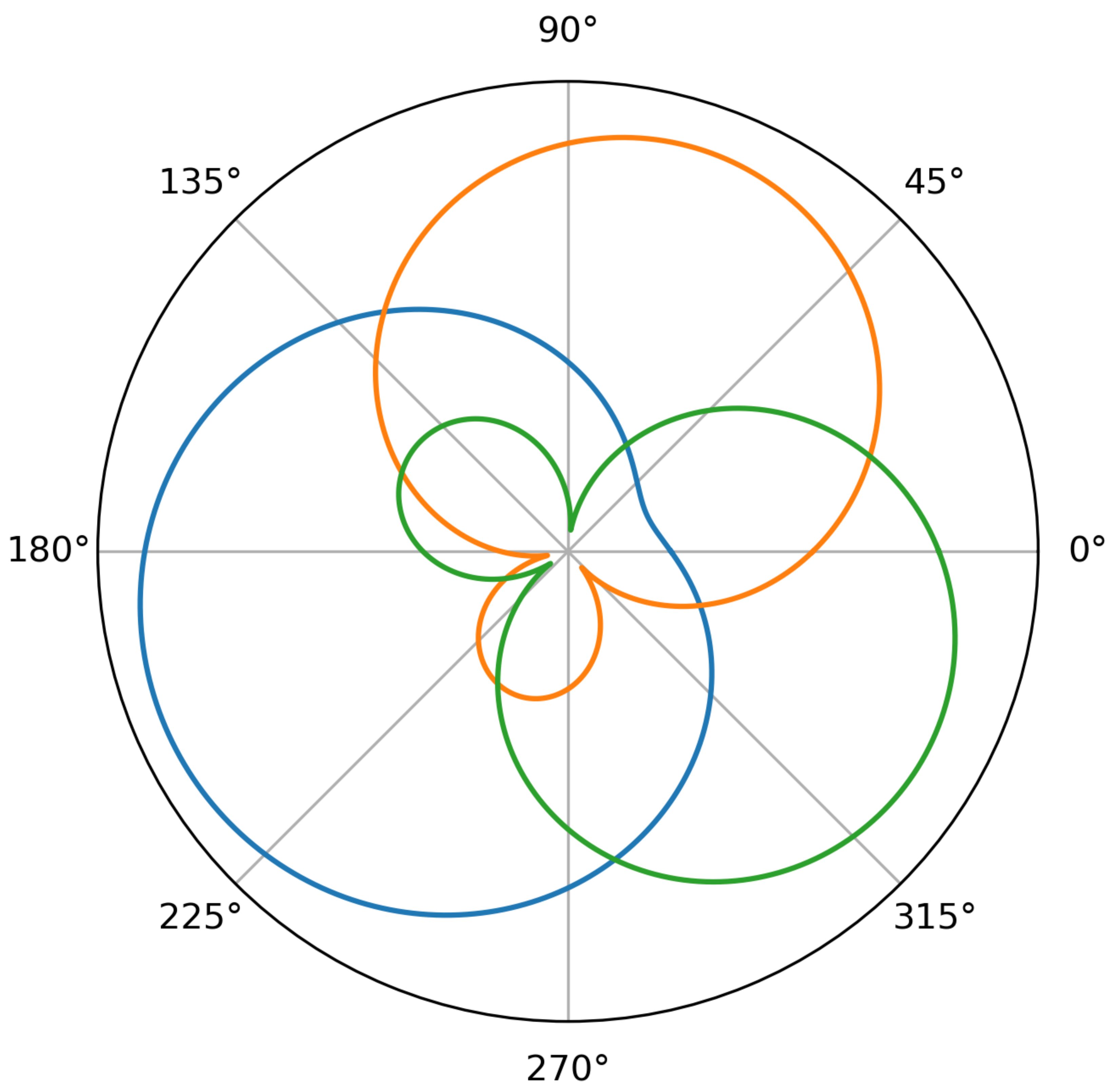
MUSHRA Test

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

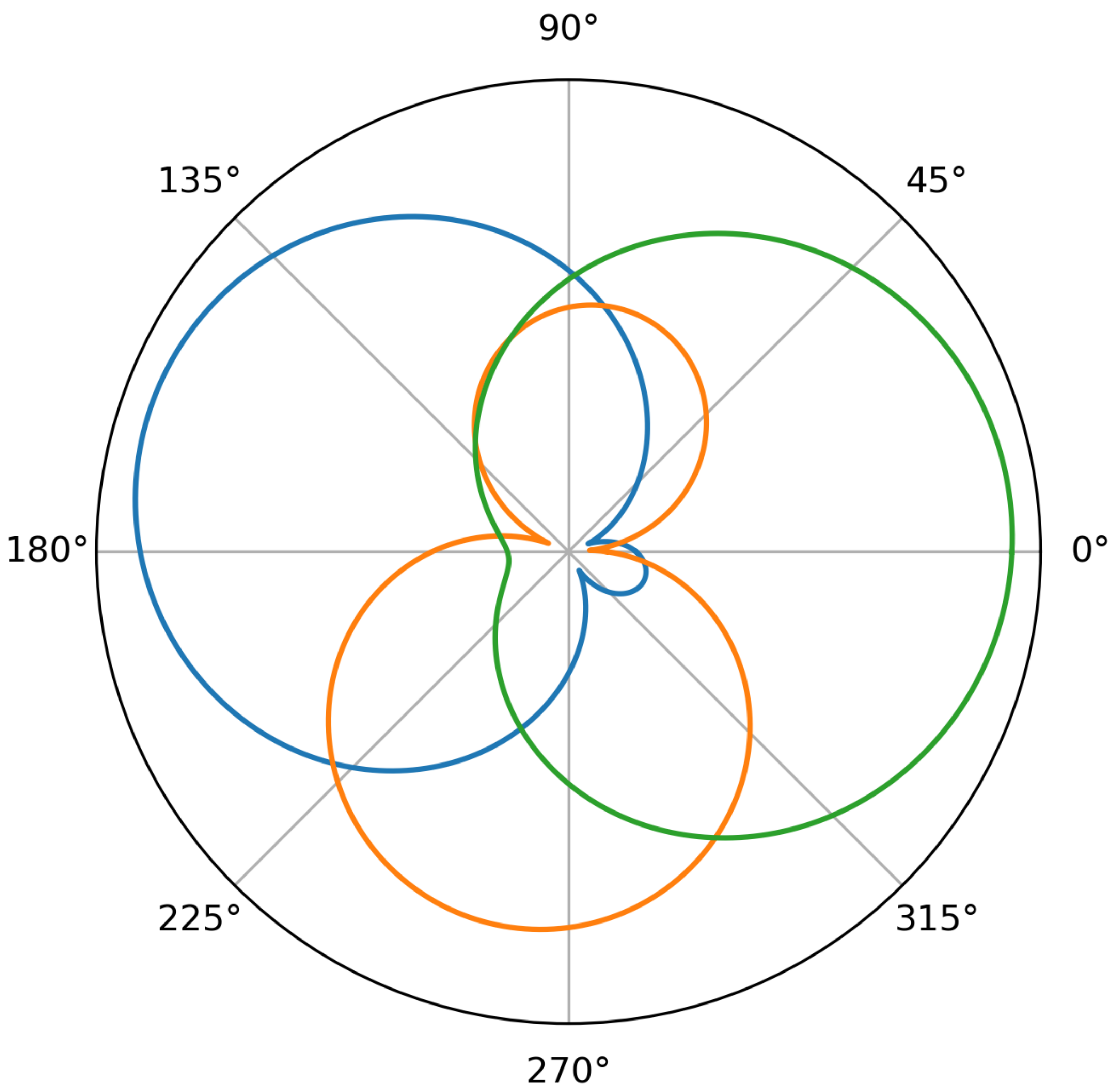
Matrixing as beaminforming



Interpolation

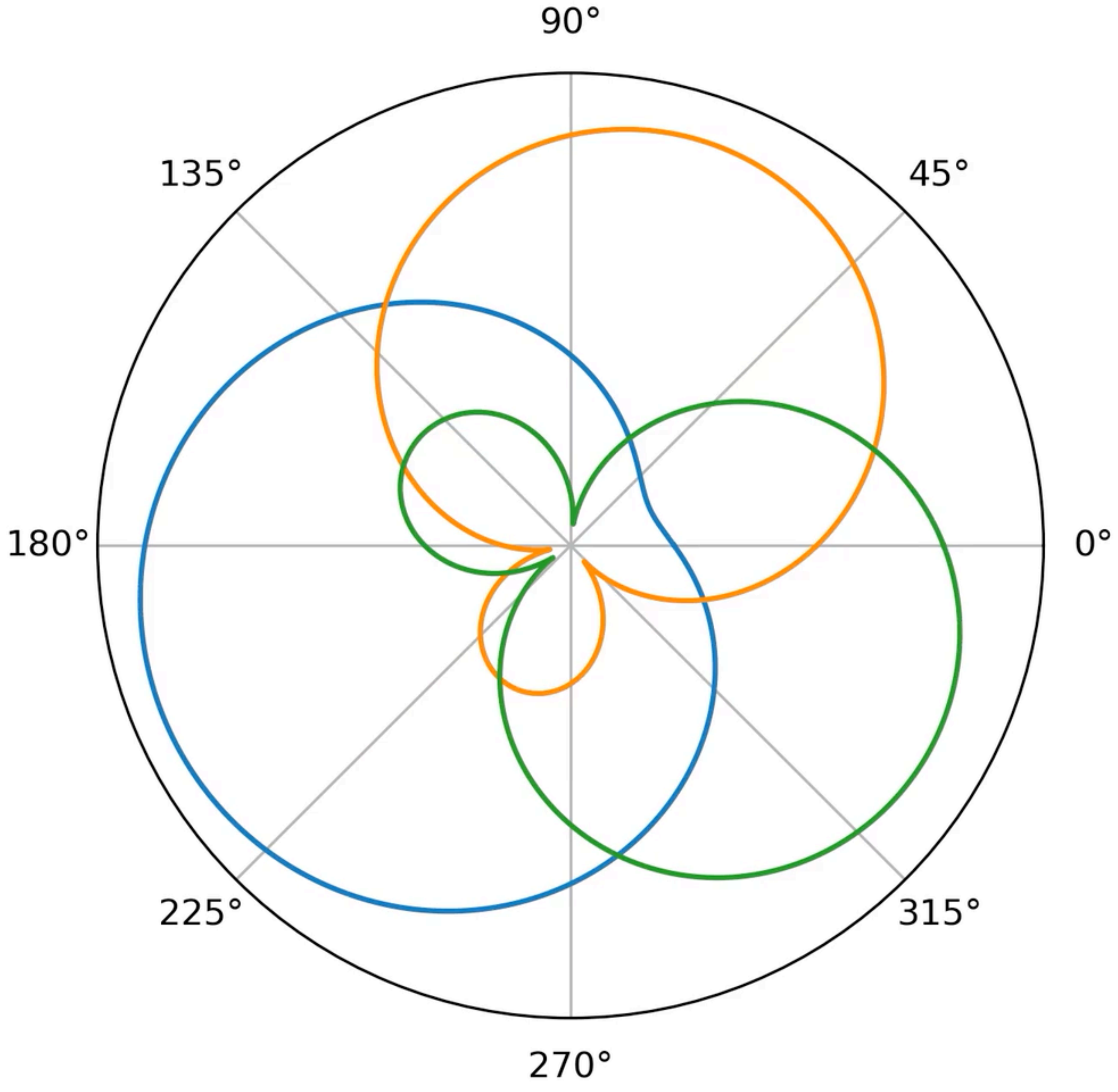


t-1



t

Interpolation



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