Dolby

AudioVMAF

Audio Quality Prediction with VMAF

ARIJIT BISWAS

HARALD MUNDT

155TH AES CONVENTION, NEW YORK, 26 OCTOBER 2023

Motivation & scope

Motivation

- VMAF* is a popular tool in the industry for measuring coded video quality and optimize video delivery.
- Desire to model coded audio-video quality (AVQ) with a coherent system design and optimize audio-video delivery.

Scope

- Can VMAF be leveraged to assess coded audio quality?

We aim to use deployed VMAF for coded audio quality prediction

*VMAF – Video Multi-Method Assessment Fusion: https://github.com/Netflix/vmaf

Audio quality metrics inspired from image domain

1D variant of popular 2D image distortion metrics repurposed for audio signals[1]

- Does not predict an easy-to-interpret quality score (e.g., in MOS/MUSHRA scale) — **Difficult to interpret**
- Only evaluated AAC at 8, 32, and 128 kb/s stereo. — **Not an extensive evaluation!**

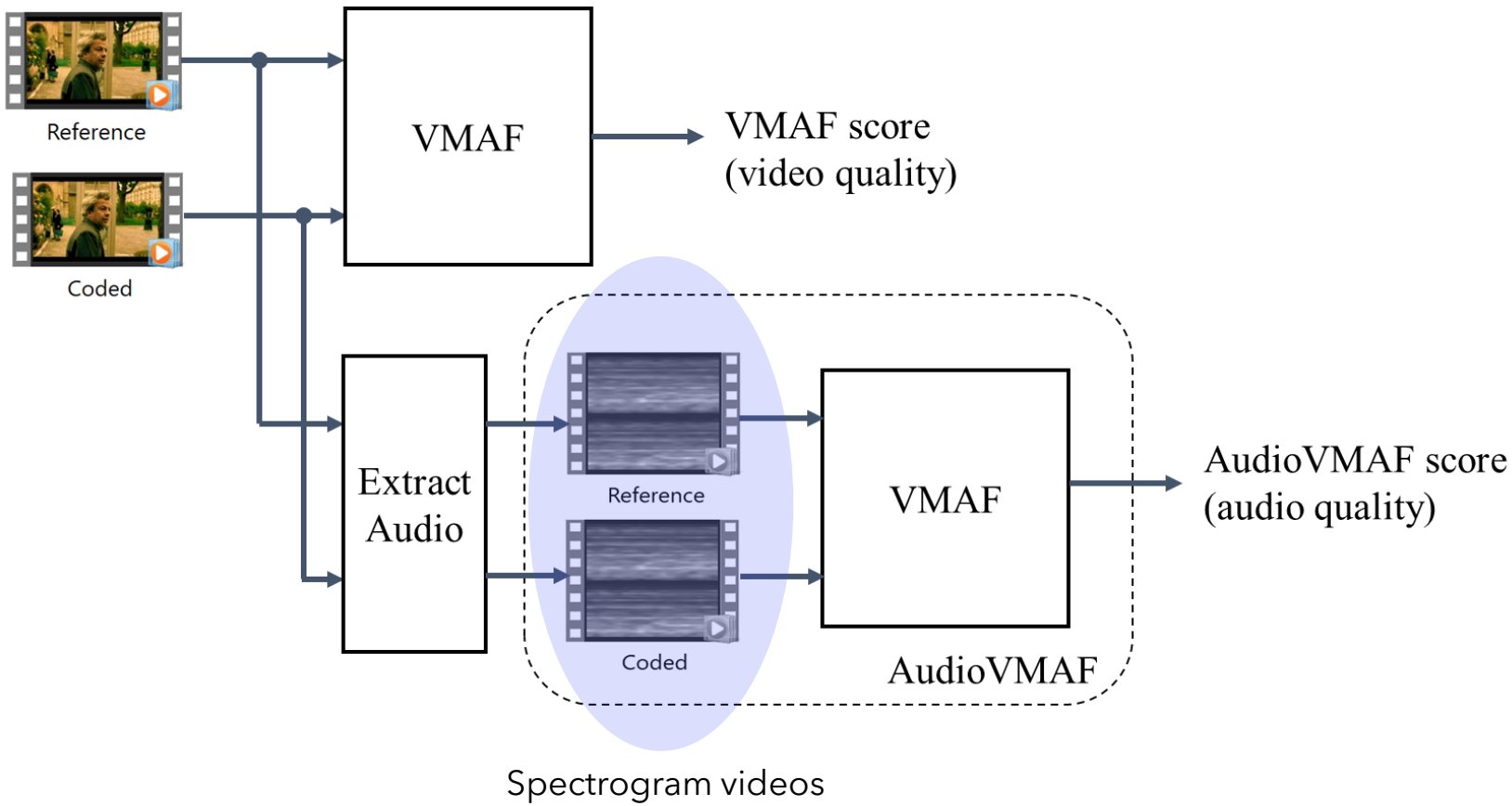
ViSQOL – distortion metric is inspired from 2D image distortion metric[2]

- ➢ Includes Gammatone frontend — **Designed for audio**
- ➢ Distortion metric → MOS (trained) — **Designed for coded audio quality prediction**

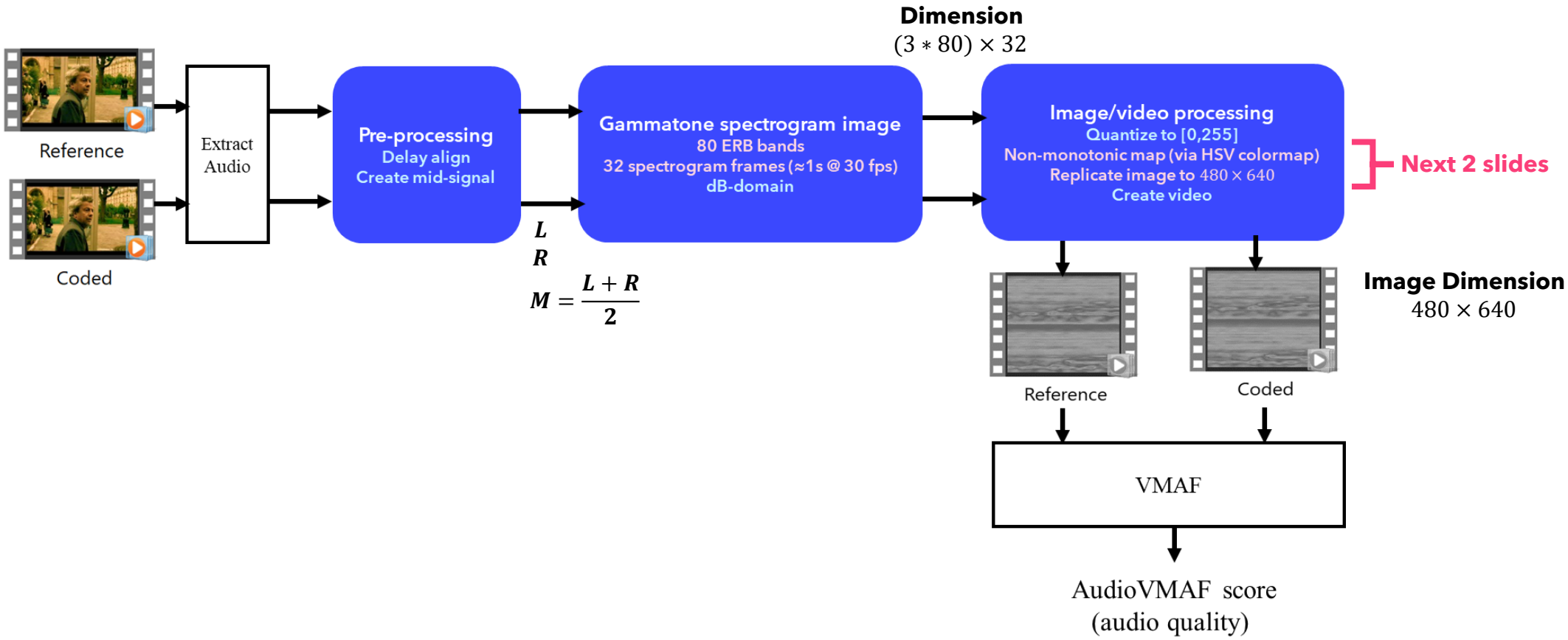
Unaware of "out-of-the-box" image/video quality metric utilized for Coded Audio Quality Prediction!

[1]X. Min, et al., "Study of Subjective and Objective Quality Assessment of Audio-Visual Signals," *IEEE Trans. on Image Processing*, 2020.
[2]M. Chinen, et al., "ViSQOL v3: An Open Source Production Ready Objective Speech and Audio Metric," *QoMEX*, 2020.

Video and audio quality prediction with VMAF

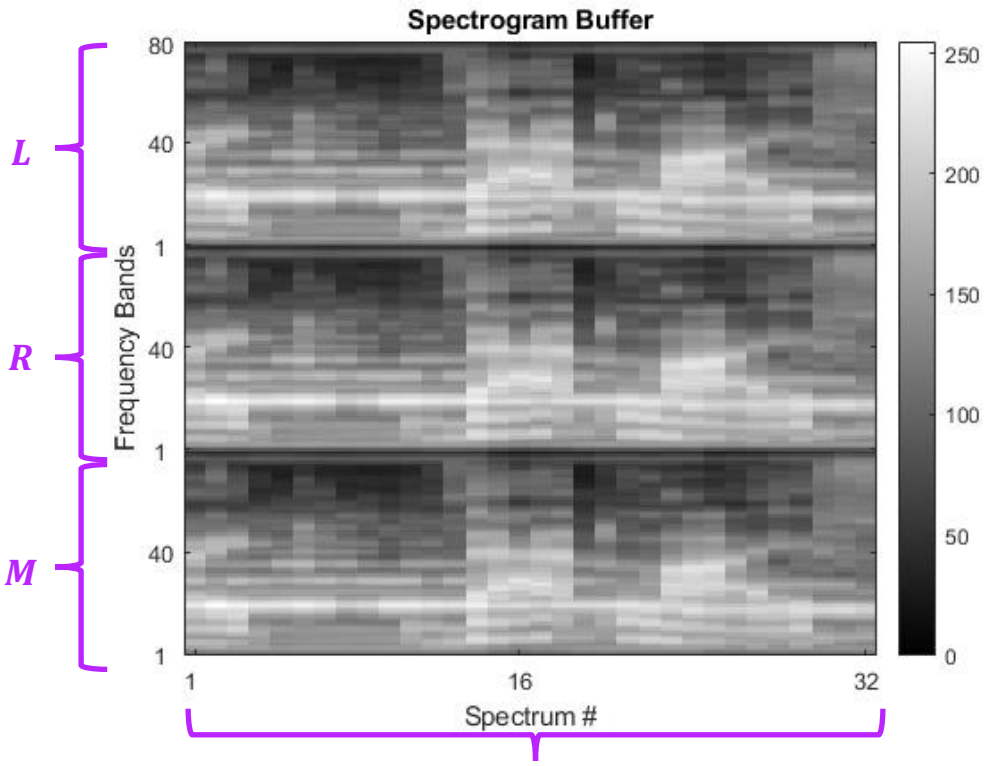


Creation of spectrogram videos



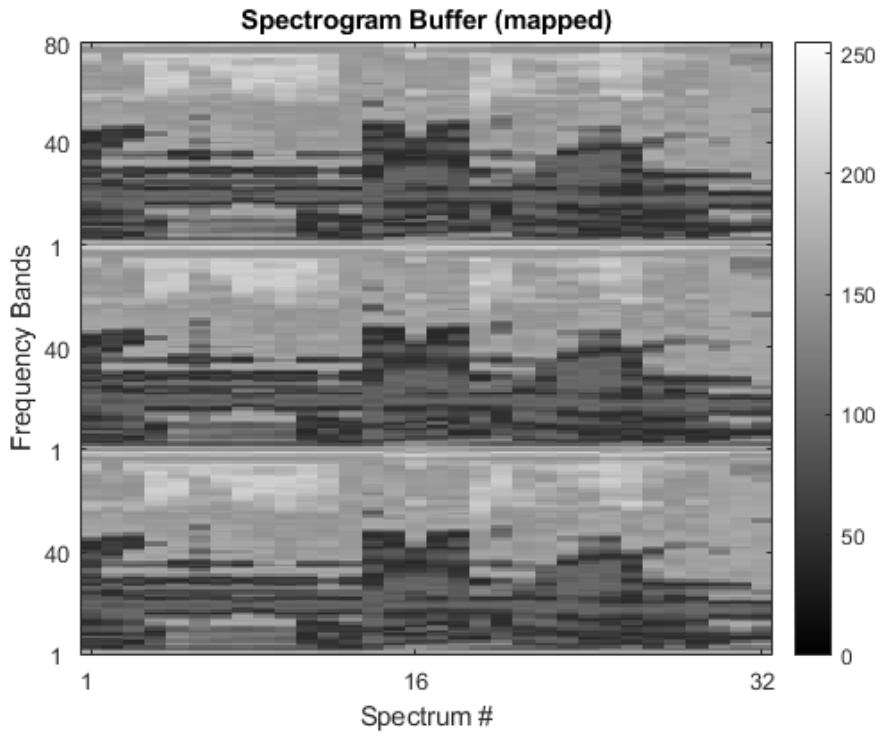
Non-monotonic mapping (via HSV colormap*)

Dimension: 240×32



32 spectrogram frames (≈1s audio at 30 fps)

Dimension: 240×32



Effect of non-monotonic mapping

*HSV (Hue, Saturation, Value) Colormap Array. https://www.mathworks.com/help/matlab/ref/hsv.html/

Replication of audio spectrograms

Image dimension: 480 × 640

240 × 32

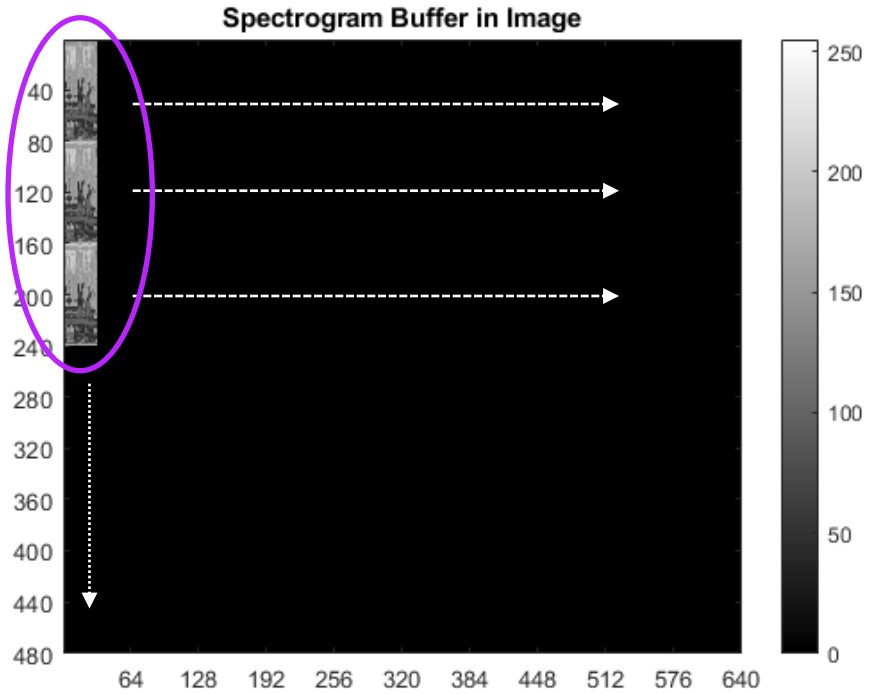
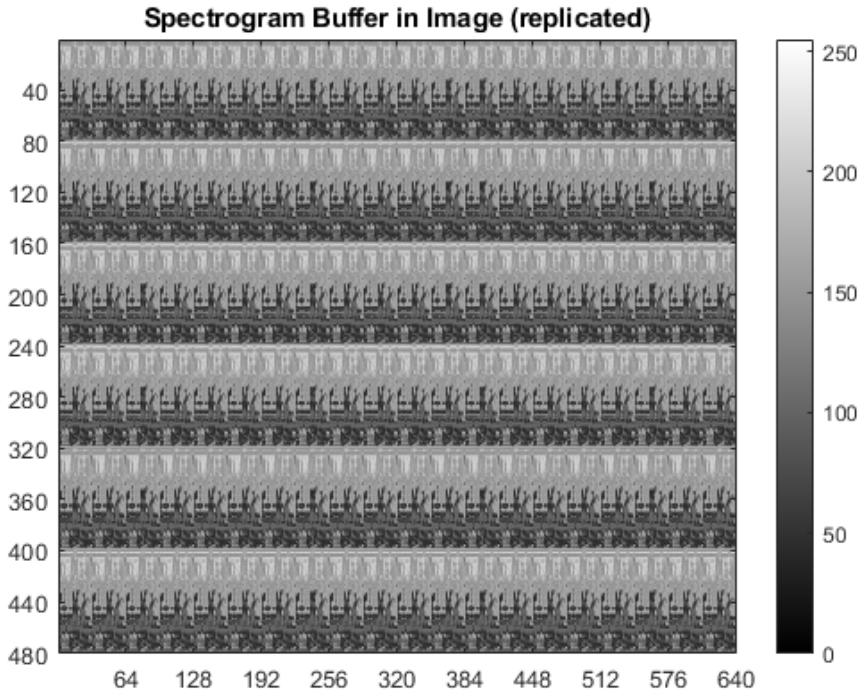


Image dimension: 480 × 640



After replication

USAC Verification Listening Test 3* (stereo high-rates)

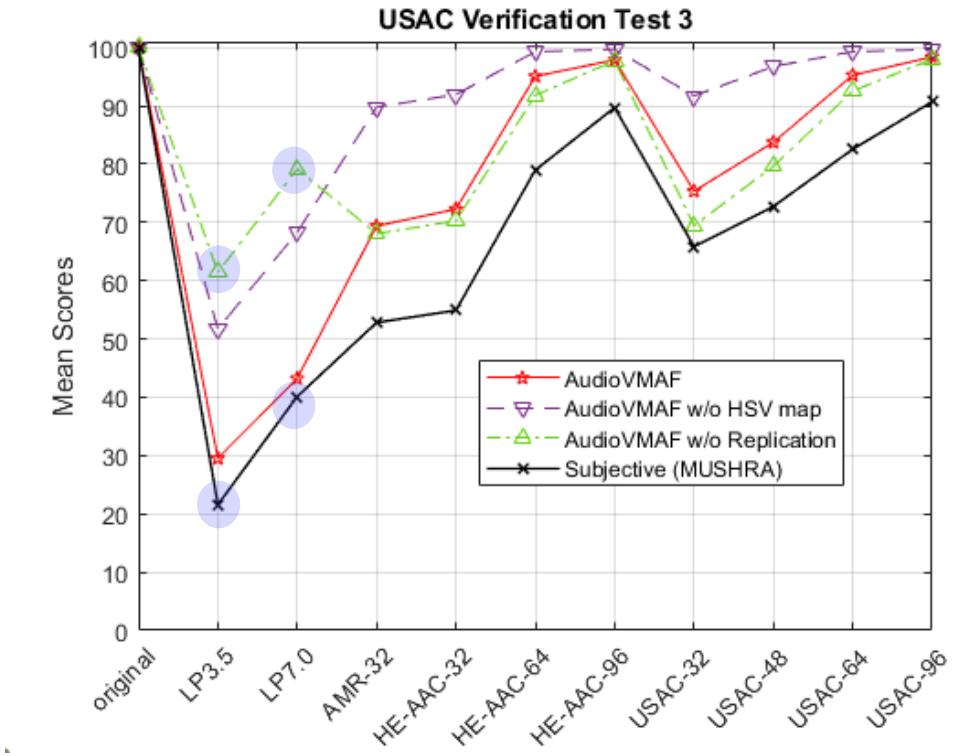
	w/ anchors		w/o anchors		
	R_p	R_s	R_p	R_s	
ViSQOL-v3	0.823	0.904	0.769	0.852	Dedicated audio quality metric
$SSIM_{1D}$	0.263	0.417	0.702	0.803	
$MS\text{-}SSIM_{1D}$	0.460	0.559	0.752	0.814	
$VIFP_{1D}$	0.389	0.517	0.332	0.581	1D variant of 2D image distortion metrics
$GMSM_{1D}$	0.115	0.239	0.678	0.807	
$GMSD_{1D}$	0.116	0.248	0.738	0.797	
AudioVMAF	**0.909**	**0.938**	**0.818**	**0.898**	

Pearson's correlation coefficient

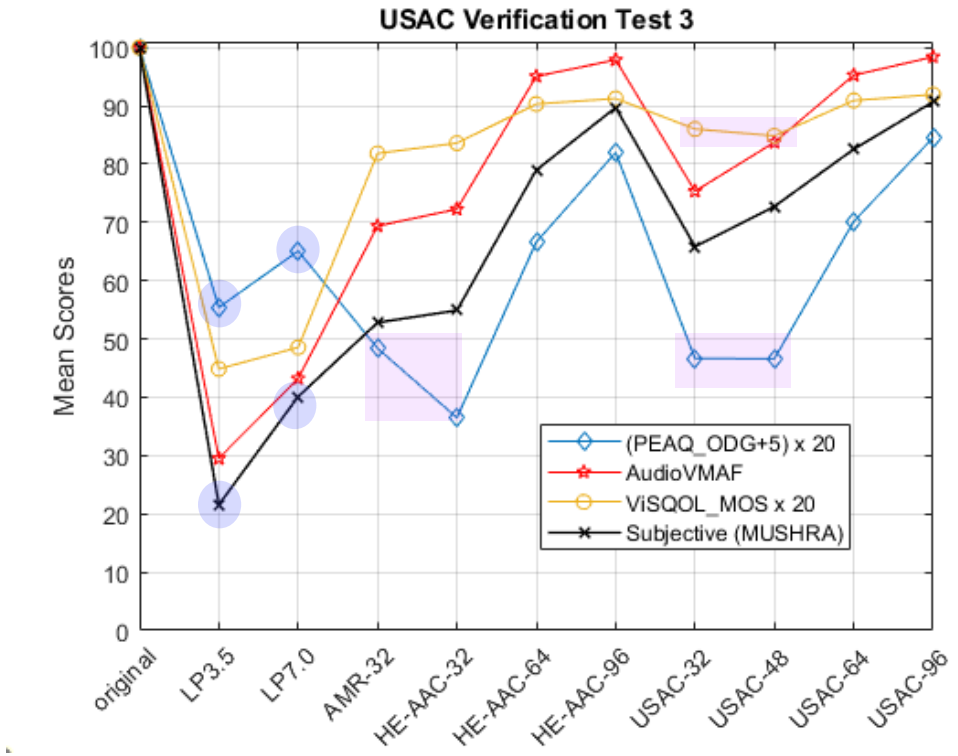
Spearman's Rank correlation coefficient

*Results for tests 1 & 2 are reported in our paper. Codecs included in all three MUSHRA tests were AMR-WB+, HE-AAC, and USAC at various bitrates.
Note: ViSQOL-v3 and 1D variant of 2D image distortion metrics evaluates the mid-signal: $M = \frac{1}{2}(L + R)$

With/without replication and non-monotonic mapping

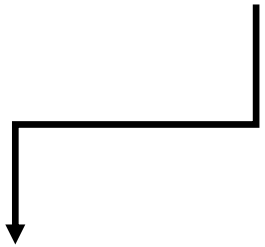


AudioVMAF versus ViSQOL and PEAQ*



*https://www.mmsp.ece.mcgill.ca/MMSP/Documents/Software/

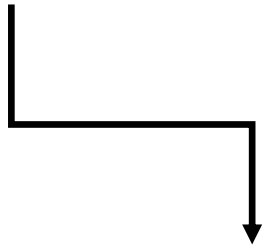
Summary



- VMAF can be used to predict coded audio quality
- Proposed preprocessing to deployed VMAF
- New angle for audio quality prediction

  → Joint audio-video quality (AVQ) measure using a coherent system design

Next steps



- Better understand VMAF (design & training data)
- Extend towards multichannel/immersive
- Joint AVQ modeling

—

THANK YOU

—

APPENDIX

Effect of non-monotonic mapping (via HSV colormap)