

Perceptual Evaluation of 360 Audiovisual Quality and Machine Learning Predictions

1st Randy Frans Fela
SenseLab
FORCE Technology
Hørsholm, Denmark
rff@force.dk

2nd Nick Zacharov
SenseLab
FORCE Technology
Hørsholm, Denmark
nvz@force.dk

3rd Søren Forchhammer
Dept. Photonics Engineering
Technical University of Denmark
Kgs. Lyngby, Denmark
sofo@fotonik.dtu.dk

Abstract—In an earlier study, we gathered perceptual evaluations of the audio, video, and audiovisual quality for 360 audiovisual content. This paper investigates perceived audiovisual quality prediction based on objective quality metrics and subjective scores of 360 video and spatial audio content. Thirteen objective video quality metrics and three objective audio quality metrics were evaluated for five stimuli for each coding parameter. Four regression-based machine learning models were trained and tested here, i.e., multiple linear regression, decision tree, random forest, and support vector machine. Each model was constructed using a combination of audio and video quality metrics and two cross-validation methods (k-Fold and Leave-One-Out) were investigated and produced 312 predictive models. The results indicate that the model based on the evaluation of VMAF and AMBIQUAL is better than other combinations of audio-video quality metric. In this study, support vector machine provides higher performance using k-Fold (PCC = 0.909, SROCC = 0.914, and RMSE = 0.416). These results can provide insights for the design of multimedia quality metrics and the development of predictive models for audiovisual omnidirectional media.

Index Terms—perceptual evaluation, 360 video, spatial audio, machine learning, multimedia quality metrics, higher order ambisonics.

I. INTRODUCTION

In recent years, 360 video or omnidirectional video (ODV) has become popular and increasingly developed to be playable in a more efficient way. The features of ODV which allows users to explore a spherical image by rotating their head offers the possibility to pair this type of video with spatial audio. Several platforms, such as VLC, Youtube, and Facebook, allow users to upload 360 audiovisual content and playback through the traditional flat displays or head mounted displays (HMD). This technology raises the question about how the users perceive the quality of 360 audiovisual content and how to achieve a high-level user experience in 360 audiovisual.

As a common approach, both the affective testing and predictive measures are employed together in perceptual quality assessment of audiovisual aiming to provide the validation process of the results obtained from predictive metrics. In the 2D video, eight audiovisual quality models were identified from previous reports as comprehensively summarized in [1], [2]. In particular with ODV, the perceptual quality has been studied in [3], [4] through subjective and objective methods.

Using these models we expect to be able to produce a statistical-based perceptual quality model by utilizing techniques such as curve fitting [3], metadata-based approached [5], and by incorporating viewport information for adaptive streaming application [6]. In [7], a machine learning-based QoE model have been proposed by converting continuous scores into a dichotomous score in order to build a logistic regression model.

However, in order to obtain the overall impression of perceptual events, the presence of auditory stimuli is required. In terms of spherical projection that encourages users to look around in ODV, ambisonic spatial audio is considered a highly compatible pair for ODV. It preserves the spatial information of audio signals, allowing users to perceive sounds coming from specific directions. After its first development in the late '70s, Ambisonic, which was firstly proposed in [8] has recently gained popularity with the progress of virtual reality (VR) technology. In spatial audio quality evaluation, overall quality, as well as attributes relating to spatial qualities such as localization and timbral quality, are the central area of interest [9]. Furthermore, there is a full-reference metrics available for ambisonic as has been proposed in [10].

Although there are several perceptual evaluation studies of 360 video or spatial audio individually [3], [4], [7], [9], [11], studies in immersive contents that combine 360 video and spatial audio is relatively unexplored, thus will be the main contribution of this paper. In this domain, our earlier work has investigated the perceptual audio, video and audiovisual quality subjectively [12]. In this paper, machine learning-based models, i.e., multiple linear regression, decision tree, random forest, and support vector machine (SVM), were investigated in order to evaluate the performance of these methods in predicting perceptual quality models. According to this objective, we address the following questions:

- Based on the objective quality measures, which objective metric contributes to multimodal quality in terms of correlation?
- In terms of combined audiovisual quality metrics, which combination could produce the highest correlation with subjective audiovisual quality?
- Among the machine learning algorithms implemented, which algorithm could provide the best prediction?

TABLE I
QUALITY SPECIFICATION OF THE AUDIOVISUAL CONTENTS

Audio	Video
B-Format PCM First Order Ambisonic (FOA) AmbiX	ERP 4K (3840x1920)
48 kHz, 16 bit, 3,072 Mbps (768 kbps/channel)	29.97 fps, 8-bit depth ~30 Mbps, YUV 4:2:0

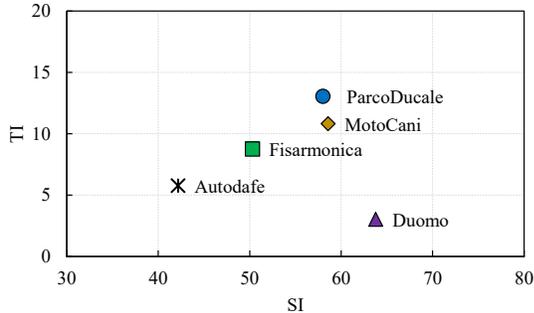


Fig. 1. Temporal and spatial indexes of testing sequences.

II. CONTENT AND METHODS

A. Content

The content was downloaded from Jump video dataset [13] and used with permission from the creator. Five 360 videos in equirectangular projection format containing first-order ambisonic (FOA) audio were carefully selected with the internal expectation that the selected materials would be able to elicit both the audio and visual quality responses. The specification of source materials is listed in Table I. In order to ensure that there was adequate visual information for subjective assessment, the temporal and spatial index (TI & SI) of the source video were calculated as described in [14] and the result is shown in Figure 1.

The overall workflow implemented in this study is illustrated in Figure 2. First, audio clips were extracted from test items to be processed independently (audio and video) for different encoding parameters. Each source video was encoded in FFmpeg using H.264 (libx264) to create the processed video sequences (PVS) with four quantization parameters (QP: 22, 27, 32, 37) and four resolutions (3840x1920, 2560x1280, 1920x1080 and 1280x720). Meanwhile, a low-bitrate codec (AAC-LC) and ambisonic decoding technique were employed to create processed audio excerpts (PAE) in three bitrates (64 kbps, 128 kbps, and 256 kbps) and three types of channel playback (5.0, 11.0, 22.0). Each test signal (PVS and PAE) was further processed with representative objective quality metrics for video (OQM_V) or audio (OQM_A) and delivered to the display device (head-mounted display or multichannel loudspeakers) for laboratory testing.

B. Measures

Three full-reference of OQM_A i.e., perceptual evaluation of audio quality (PEAQ) [15], ViSQOLAudio¹ [16], [17],

¹<https://github.com/google/vsqol>

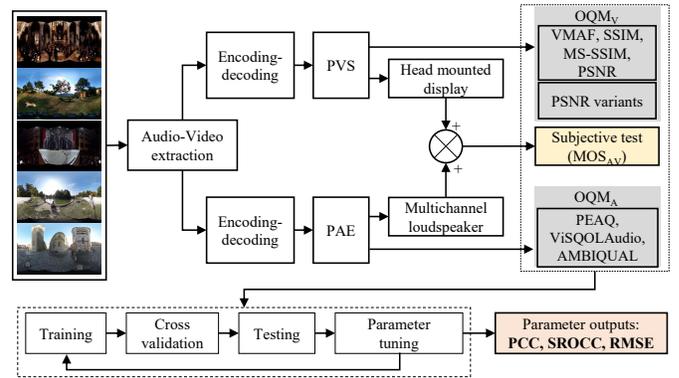


Fig. 2. Workflow of predictive models built in this study.

and AMBIQUAL [10] were computed to estimate overall listening quality. In terms of video quality metrics, nine full-reference OQM_V were computed by following the common test condition and testing procedure for 360 video described in [18]. The OQM_V was measured in the codec, cross-format (CF), and end-to-end (EE) stages, including the basic peak signal-to-noise ratio (PSNR) and its variants, i.e., weighted to spherically uniform PSNR (WS-PSNR) [19], Sphere-based PSNR with interpolation (S-PSNR-I) [20], with the nearest neighbor (S-PSNR-NN) [21], and craster parabolic projection PSNR (CPP-PSNR) [22]. In addition, a VMAF (video multimethod assessment fusion) [23], [24] source code originally developed by Netflix was also computed, generating four additional metrics including structural similarity metrics (SSIM) [25] and multi-scale SSIM [26], PSNR, and VMAF. The inclusion of VMAF in this study was motivated by the report that it is feasible to measure OQM_V of 360 video by using VMAF² without any adjustment [11].

Three subjective tests (audio, video, audiovisual) with single stimulus absolute category ratings with continuous quality scale (ACR-CQS) were conducted and participated by twenty assessors who passed a basic audiovisual screening test [14]. The user interface (UI) reliably displayed either on the projection screen (for listening session) or as a pop-up interface in a virtual environment (visual and audiovisual session). SenseLabOnline 4.0 was used to integrate the entire test setup, define the experimental design, and precisely run the tests in a double-blind random presentation order. Please note that only audiovisual quality will be discussed in this paper. The readers interested to single modality results as well as detailed experimental description are encouraged to refer to [12].

C. Implementation and Evaluation

The machine learning models were implemented in the R programming language with the `caret` package [27], [28]. Note that this is a basic benchmarking study and four machine learning algorithms were selected based on the common practice in perceptual quality studies. For each model, there are three variable inputs which consist of OQM_A ($n = 3$), OQM_V

²<https://github.com/Netflix/vmaf>

TABLE II
COMPUTATIONAL SETUP FOR MACHINE LEARNING PREDICTION

Settings	LM	DT	RF	SVM
returnResamp	All	All	All	All
search	random	random	random	random
method	lm	rpart	rf	svmRadialSigma
tuning	minsplrit: 20 maxdepth: 30	-	ntree: 500 tuneLength: 15	-

($n = 13$), and each combination gives $n = 39$. In contrast, the variable output is only a mean opinion score of perceived audiovisual quality (MOS_{AV}). This form is motivated by the basic audiovisual quality model, which was originally expressed in linear form, as shown in (1) [2],

$$MOS_{AV} = \alpha_1 + \beta_1 Q_A + \gamma_1 Q_V + \zeta_1 Q_A Q_V \quad (1)$$

where in this case, the symbol Q could be the quality obtained from the objective metrics of audio or video, respectively, and MOS_{AV} denotes the audiovisual quality score obtained from the subjective evaluation. The relevant settings for the machine learning model are presented in Table II.

In order to maintain the accuracy of the model prediction due to the relatively small dataset, we split the data into 80:20 ratio for the training set and test set. Two types of cross validation (CV) were carried out in this study i.e., k-Fold ($n = 10$ splits) and leave-one-out (LOOCV) in order to investigate how the results from content-based split will differ to random split based on k-Fold. In LOOCV, each class of content was treated as a test set. The remaining class ($N = 5-1$) was implemented as a training set consecutively so that the prediction accuracy was the average of the overall results.

1) *Decision Trees*: The decision tree in R worked based on Gini impurity, which measures the proportion of the incorrect labels of the randomly selected elements from the set according to the label distribution in the subset [29]. We used the `rpart` library for the decision tree model with the tuning parameters of a minimum split “*minsplrit*” and maximum depth “*maxdepth*” adjusted by default to 20 and 30, respectively. The “*minsplrit*” represents the minimum number of data points needed to attempt a split before it is enforced to build a terminal node, whereas the “*maxdepth*” is the maximum number of internal nodes built between the root and the terminal nodes.

2) *Random forest*: Random forest [30] is considered more effective than decision tree when working with large datasets and can retain consistency when missing data exists. The `rf` method in the `caret` package was used to perform the random forest work in R. The default tuning parameter is set to “*mtry*” = 7 and “*ntree*” = 500. The parameter “*mtry*” specifies the number of randomly sampled variables as candidates at each split, while “*ntree*” specifies the number of trees to grow. Here, we configured the `tuneLength` parameter, which allowed the system to adjust the algorithm automatically. It indicated the number of different values to be tested for each adjustment parameter, for example, “*mtry*” as a random forest. Assuming

the `tuneLength` = 5, which means to try five different “*mtry*” values and find the best “*mtry*” value based on these five.

3) *Support Vector Machine*: The support vector machine (SVM)³ is a machine learning technique which aims to construct a hyperplane in the an n dimensional space, where n represents the number of features, to find a decision boundary of two or more classes. The objective of SVM is to find the hyperplane that could maximize the distance margin to the closest vectors called support vectors. Hence, in the case where the goal is to predict the label of the class is called classification problem. SVM can also be used for regression problem where the goal is to predict the appropriate hyperplane position relative to the support vectors. By maximizing the distance margin between support vectors to the hyperplane could approximate the actual function represented by data points. Support vector machines use a hypothetical space of linear functions in a higher-dimensional feature space which are trained using an optimization theory learning algorithm, which uses learning biases derived from statistical learning theory. In this study, we used support vector regression with a radial basis function (RBF) for kernel parameters as reported in a similar study showing a greater accuracy than a linear kernel [7]. The method chosen is `svmRadialSigma` as available in library `kernlab` [31]. The radial basis function can be expressed in (2) as follows [32],

$$K(x_1, x_2) = \exp(-\sigma \|x_1 - x_2\|^2) \quad (2)$$

In the RBF kernel, the value depends on the distance from the origin or a certain point. Finally, the distance information of the vector in the original space can be used to determine the dot product (similarity) of x_1 and x_2 . The tuning parameters are regularization parameter c and kernel parameter σ . Parameter c controls error by adjusting the margin distance. In the case where the value of c increases and σ decreases, the model is overfitting. In a random search, `tuneLength` parameter is the total number of (c, σ) pairs to evaluate.

III. RESULTS

The audio-video quality prediction metrics were computed in all possible combinations between OQM_V and OQM_A as independent variables and audiovisual subjective scores (MOS_{AV}) as a dependent variable. In this section, we aim to answer three questions addressed earlier in Section I. The

³This technique is originally named as SVM. Later, a term of support vector regression (SVR) is frequently used in certain community to distinguish the application in regression problem.

TABLE III
PEARSON'S CORRELATION COEFFICIENT (PCC) OF ALL METRIC COMBINATIONS FOR 1) K-FOLD (LEFT) AND 2) LOOCV (RIGHT).

OQM _v	LM1			DT1			RF1			SVM1			LM2			DT2			RF2			SVM2			
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	
PSNR	0.704	0.653	0.731	0.675	0.655	0.738	0.629	0.671	0.726	0.731	0.700	0.802	0.733	0.768	0.810	0.569	0.583	0.759	0.699	0.607	0.830	0.702	0.775	0.864	0.92
WS.PSNR	0.720	0.685	0.749	0.660	0.671	0.699	0.682	0.696	0.751	0.713	0.717	0.818	0.734	0.765	0.812	0.631	0.578	0.780	0.693	0.649	0.841	0.701	0.776	0.865	0.88
EE.WS.PSNR	0.733	0.727	0.771	0.669	0.627	0.788	0.749	0.701	0.809	0.728	0.741	0.756	0.776	0.789	0.848	0.588	0.616	0.773	0.750	0.701	0.876	0.759	0.812	0.896	0.84
EE.SPSNR.NN	0.734	0.727	0.772	0.676	0.632	0.765	0.748	0.695	0.812	0.724	0.747	0.761	0.776	0.789	0.848	0.572	0.616	0.790	0.755	0.701	0.877	0.760	0.813	0.896	0.80
EE.SPSNR.I	0.735	0.728	0.773	0.684	0.627	0.772	0.758	0.696	0.810	0.724	0.747	0.763	0.776	0.789	0.848	0.572	0.616	0.790	0.752	0.701	0.876	0.759	0.813	0.896	0.76
EE.CPP.PSNR	0.734	0.728	0.772	0.669	0.630	0.798	0.754	0.695	0.806	0.729	0.746	0.759	0.776	0.789	0.848	0.588	0.611	0.779	0.748	0.693	0.873	0.760	0.812	0.896	0.72
CF.SPSNR.NN	0.727	0.778	0.778	0.766	0.695	0.841	0.798	0.787	0.839	0.742	0.786	0.829	0.778	0.791	0.849	0.690	0.656	0.789	0.749	0.695	0.882	0.779	0.808	0.892	0.68
CF.SPSNR.I	0.734	0.727	0.772	0.676	0.627	0.772	0.755	0.700	0.811	0.724	0.748	0.761	0.776	0.789	0.848	0.572	0.616	0.790	0.756	0.705	0.878	0.760	0.813	0.896	0.64
CF.CPP.PSNR	0.733	0.727	0.772	0.681	0.613	0.772	0.743	0.686	0.807	0.723	0.747	0.761	0.775	0.788	0.847	0.596	0.611	0.790	0.753	0.702	0.875	0.760	0.812	0.896	0.60
PSNR.vm	0.722	0.702	0.758	0.725	0.574	0.758	0.724	0.657	0.794	0.731	0.725	0.799	0.776	0.792	0.848	0.581	0.640	0.781	0.752	0.701	0.872	0.755	0.814	0.896	
SSIM	0.720	0.677	0.766	0.743	0.601	0.798	0.734	0.705	0.832	0.807	0.736	0.835	0.772	0.758	0.800	0.639	0.645	0.724	0.744	0.703	0.905	0.767	0.794	0.889	
MS.SSIM	0.715	0.671	0.752	0.773	0.682	0.849	0.786	0.737	0.828	0.813	0.764	0.823	0.753	0.790	0.821	0.668	0.638	0.714	0.745	0.720	0.900	0.763	0.826	0.904	
VMAF	0.800	0.816	0.848	0.729	0.795	0.869	0.804	0.770	0.882	0.825	0.831	0.909	0.779	0.796	0.851	0.676	0.652	0.794	0.758	0.736	0.899	0.774	0.816	0.905	

TABLE IV
SPEARMAN'S RANK ORDER CORRELATION COEFFICIENT (SROCC) OF ALL METRIC COMBINATIONS FOR 1) K-FOLD (LEFT) AND 2) LOOCV (RIGHT).

OQM _v	LM1			DT1			RF1			SVM1			LM2			DT2			RF2			SVM2			
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	
PSNR	0.709	0.661	0.727	0.702	0.654	0.740	0.624	0.684	0.725	0.732	0.693	0.779	0.724	0.757	0.821	0.510	0.562	0.770	0.682	0.607	0.838	0.692	0.769	0.862	0.92
WS.PSNR	0.721	0.671	0.748	0.674	0.665	0.735	0.698	0.731	0.755	0.725	0.712	0.791	0.726	0.756	0.819	0.613	0.559	0.792	0.683	0.639	0.841	0.691	0.769	0.864	0.88
EE.WS.PSNR	0.700	0.688	0.764	0.685	0.634	0.814	0.712	0.695	0.819	0.736	0.702	0.766	0.757	0.786	0.851	0.547	0.613	0.758	0.722	0.679	0.880	0.748	0.801	0.898	0.84
EE.SPSNR.NN	0.700	0.691	0.759	0.680	0.661	0.768	0.721	0.685	0.818	0.725	0.710	0.768	0.759	0.786	0.850	0.529	0.613	0.793	0.729	0.679	0.878	0.746	0.800	0.897	0.80
EE.SPSNR.I	0.701	0.690	0.760	0.691	0.634	0.774	0.726	0.684	0.818	0.726	0.711	0.769	0.759	0.786	0.850	0.529	0.613	0.793	0.730	0.686	0.880	0.747	0.801	0.897	0.76
EE.CPP.PSNR	0.704	0.689	0.762	0.685	0.634	0.818	0.730	0.677	0.822	0.734	0.703	0.766	0.759	0.786	0.851	0.547	0.604	0.765	0.727	0.672	0.875	0.749	0.801	0.898	0.72
CF.SPSNR.NN	0.719	0.752	0.788	0.786	0.701	0.852	0.806	0.788	0.845	0.772	0.762	0.810	0.759	0.780	0.850	0.673	0.655	0.791	0.738	0.673	0.888	0.775	0.797	0.892	0.68
CF.SPSNR.I	0.700	0.694	0.762	0.680	0.634	0.774	0.727	0.679	0.818	0.726	0.710	0.768	0.759	0.786	0.850	0.529	0.613	0.793	0.734	0.683	0.879	0.746	0.800	0.897	0.64
CF.CPP.PSNR	0.706	0.691	0.762	0.679	0.609	0.774	0.702	0.680	0.815	0.732	0.707	0.766	0.758	0.783	0.849	0.552	0.604	0.793	0.729	0.681	0.880	0.747	0.798	0.897	0.60
PSNR.vm	0.699	0.667	0.743	0.724	0.572	0.724	0.701	0.647	0.795	0.727	0.691	0.779	0.760	0.786	0.854	0.531	0.633	0.784	0.732	0.675	0.876	0.745	0.800	0.897	
SSIM	0.719	0.688	0.785	0.764	0.650	0.784	0.750	0.702	0.838	0.807	0.724	0.843	0.745	0.804	0.858	0.613	0.630	0.683	0.726	0.674	0.908	0.766	0.797	0.900	
MS.SSIM	0.708	0.681	0.758	0.765	0.693	0.833	0.790	0.728	0.830	0.849	0.757	0.824	0.756	0.814	0.863	0.635	0.613	0.674	0.737	0.701	0.910	0.767	0.835	0.914	
VMAF	0.787	0.797	0.864	0.711	0.783	0.864	0.828	0.743	0.895	0.850	0.809	0.914	0.769	0.797	0.861	0.653	0.650	0.804	0.739	0.726	0.901	0.773	0.808	0.911	

TABLE V
ROOT MEAN SQUARED ERROR (RMSE) OF ALL METRIC COMBINATIONS FOR 1) K-FOLD (LEFT) AND 2) LOOCV (RIGHT).

OQM _v	LM1			DT1			RF1			SVM1			LM2			DT2			RF2			SVM2			
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	
PSNR	0.714	0.755	0.680	0.733	0.749	0.697	0.770	0.734	0.689	0.683	0.708	0.593	0.682	0.655	0.589	0.817	0.804	0.647	0.710	0.802	0.557	0.707	0.641	0.520	0.40
WS.PSNR	0.698	0.728	0.661	0.769	0.738	0.734	0.724	0.711	0.656	0.707	0.694	0.571	0.680	0.655	0.586	0.769	0.809	0.620	0.718	0.760	0.537	0.705	0.638	0.515	0.45
EE.WS.PSNR	0.675	0.683	0.631	0.761	0.782	0.621	0.657	0.707	0.584	0.690	0.666	0.650	0.627	0.627	0.530	0.795	0.784	0.628	0.661	0.708	0.478	0.641	0.592	0.461	0.50
EE.SPSNR.NN	0.674	0.683	0.630	0.753	0.790	0.657	0.657	0.712	0.580	0.691	0.658	0.644	0.628	0.626	0.529	0.817	0.784	0.609	0.658	0.708	0.477	0.641	0.592	0.461	0.55
EE.SPSNR.I	0.674	0.682	0.629	0.745	0.782	0.648	0.646	0.711	0.583	0.692	0.658	0.641	0.627	0.626	0.529	0.817	0.784	0.609	0.660	0.709	0.480	0.641	0.592	0.461	0.60
EE.CPP.PSNR	0.674	0.682	0.630	0.761	0.781	0.605	0.651	0.713	0.589	0.689	0.660	0.647	0.627	0.627	0.529	0.795	0.790	0.621	0.663	0.718	0.484	0.633	0.592	0.461	0.65
CF.SPSNR.NN	0.679	0.628	0.622	0.649	0.723	0.543	0.598	0.610	0.540	0.676	0.613	0.556	0.628	0.604	0.528	0.714	0.754	0.610	0.663	0.722	0.467	0.623	0.579	0.463	0.70
CF.SPSNR.I	0.674	0.683	0.629	0.753	0.782	0.648	0.649	0.708	0.583	0.691	0.658	0.645	0.628	0.626	0.529	0.817	0.784	0.609	0.656	0.705	0.475	0.641	0.591	0.461	0.75
CF.CPP.PSNR	0.675	0.682	0.630	0.743	0.802	0.649	0.664	0.721	0.588	0.692	0.659	0.645	0.628	0.628	0.530	0.793	0.790	0.610	0.659	0.708	0.481	0.640	0.592	0.461	0.80
PSNR.vm	0.687	0.707	0.646	0.700	0.834	0.660	0.684	0.752	0.608	0.688	0.681	0.595	0.628	0.624	0.530	0.806	0.766	0.620	0.659	0.704	0.486	0.639	0.590	0.461	0.85
SSIM	0.700	0.737	0.649	0.676	0.817	0.686	0.689	0.706	0.578	0.594	0.672	0.545	0.685	0.652	0.607	0.765	0.764	0.683	0.669	0.711	0.428	0.641	0.613	0.470	
MS.SSIM	0.702	0.744	0.663	0.640	0.730	0.543	0.615	0.672	0.572	0.601	0.642	0.565	0.662	0.623	0.581	0.738	0.762	0.692	0.667	0.691	0.435	0.639	0.570	0.439	
VMAF	0.595	0.578	0.526	0.694	0.605	0.490	0.588	0.633	0.466	0.565	0.555	0.426	0.620	0.609	0.519	0.731	0.752	0.605	0.649	0.675	0.434	0.622	0.571	0.432	

correlation coefficients (PCC and SROCC) and prediction errors RMSE are computed as evaluation performance in this study between the predicted MOS and actual MOS from the test set. In this section, all description in this section refers to Tables III, IV and V. The column number in tables is represented as OQM_A as follows: 1) PEAQ, 2) ViSQOLAudio, and 3) AMBIQUAL.

A. Perceptual Metrics

In OQM_A , AMBIQUAL significantly outperforms other metrics regardless of the algorithms and the type of cross-

validation used, proving its robustness (range of PCC: 0.731–0.909, SROCC: 0.725–0.915, RMSE: 0.697–0.426). In contrast, PEAQ and ViSQOL results vary depending on the algorithm and validation method. In a linear correlation, ViSQOL produces comparable or higher performance over PEAQ in a linear model (PCC: 0.653–0.816) and support vector machine (PCC: 0.700–0.831) for which the highest scores were produced in k-Fold CV. In monotonic correlation, ViSQOL yields better performance in LOOCV for all algorithms but random forest, ranging from 0.562 (DT2) to 0.835

(SVM2).

Meanwhile, VMAF generally performs the best scores over OQM_V (range of PCC: 0.729–0.909, SROCC: 0.633–0.915, RMSE: 0.752–0.426). SSIM metrics perform relatively lower than PSNR metrics specifically in linear model (LM1 & LM2). Nevertheless, in particular paired with OQM_A and mainly for LOOCV, several metrics i.e., SSIM (PCC-RF2: 0.905, RMSE-RF2: 0.428), MS-SSIM (SROCC-RF2 & SVM2: 0.910 & 0.914) and CF-SPSNR-NN could present a slightly comparable (CF-SPSNR-NN) or higher score (SSIM & MS-SSIM) than VMAF.

B. Machine Learning Predictions

The results indicated that regardless of the cross-validation method used, VMAF or AMBIQUAL can improve the prediction performance both in terms of linear and monotonic correlation, thus reducing the prediction error. Furthermore, the VMAF–AMBIQUAL pair can generally achieve the highest performance of all machine learning prediction models (PCC ≥ 0.794 , SROCC ≥ 0.804 , RMSE: 0.605–0.426.). However, it is observed that a number of OQM_V , including CF-SPSNR-NN, SSIM, and MS-SSIM paired with AMBIQUAL produces slightly better performance (SSIM PCC: 0.905, SROCC: 0.908 & MS-SSIM PCC: 0.900, SROCC: 0.910 both in RF2) or close to VMAF–AMBIQUAL (MS-SSIM PCC: 0.904, SRCC: 0.914 in SVM2). The root-mean-squared error (RMSE) values are determined as small as 0.426 and 0.428 respectively in SVM1: VMAF–AMBIQUAL and RF2: SSIM–AMBIQUAL. According to the comparison of the machine learning models, the results imply that overall, the support vector machine could produce the highest performance with PCC up to 0.909 (VMAF–AMBIQUAL) and SROCC up to 0.914 (MS-SSIM–AMBIQUAL). Although some related studies have found that random forest-based prediction models could better be suited among their respective models [33], it can be said that a limitation of random forest is that it cannot be extrapolated and the prediction is resulted only from the average of previous data observed in a training set. Therefore, in the regression problem, the prediction range of the random forest is bound by the highest and lowest labels in the training data. It can be problematic when the range or data distribution vary for the training and test sets. However, most studies reported that random forest can perform closely, the same, or better than SVM but repeatedly perform better than the remaining models such as multiple linear regression and decision tree.

In cross-validation method, it is shown that the type of CV could diversify the distribution of prediction results yielded among various pairs of metrics and machine learning approaches. It is noticeable that k-Fold could produce wider range than LOOCV, for instance by comparing SVM1 (PCC: 0.802–0.909, SROCC: 0.779–0.914 & RMSE: 0.593–0.426) and SVM2 (PCC: 0.864–0.905, SROCC: 0.852–0.914 & RMSE: 0.520–0.432). A k-Fold also produces more consistent results in terms of different pairs of audiovisual metrics and machine learning approaches, depicted by the VMAF–AMBIQUAL that outperforms in all other models

whereas the SSIM-based results can be found as superior in LOOCV. In addition, it should be noted that the cross-validation method also depends on the data size and type of data. This is a bias-variance trade-off when choosing a cross-validation method for building a model. In LOOCV, since each training set contains only $n-1$ examples, the estimation of the test error has a lower bias. It can produce a higher variance, which means that it will use the training set for each iteration. Since there is a large overlap between the training sets, the effect is a larger variance, which means that the average of the test estimates of the test error will have a more significant variance. In contrast, the overlap between the training sets in k-Fold CV is relatively small, so the correlation of the test error estimates is small. As a result, the average test error value will not have as large a variance as LOOCV. At the end of this results, based on Tables III, IV, and V, k-Fold shows better results than LOOCV.

IV. CONCLUSIONS

Perceptual evaluation of audiovisual quality was carried out through subjective experiment for 360 video over the head-mounted display and low-bitrate ambisonic over loudspeaker playback. The perceptual subjective results, as well as perceptual quality models based on subjective data, have been described earlier in [12]. This study evaluates a combination of objective quality metrics computed for ambisonic audio and 360 video to build the perceptual audiovisual quality models by utilizing machine learning prediction. A test methodology, as well as the test results, have been described, and the main conclusions can be summarized as follows:

- In general, VMAF and AMBIQUAL, respectively, provide consistent performance better than the other video/audio quality metrics in terms of the highest correlation coefficient hence lowering prediction error (PCC ≥ 0.794 , SROCC ≥ 0.804 , RMSE: 0.605–0.426.).
- The combination of audio-video metrics analyzed for 360 audiovisual contents demonstrates that VMAF–AMBIQUAL outperforms other combinations and presents a good agreement in any condition and tested machine learning algorithms (PCC: 0.794–0.909, SROCC: 0.804–0.914, RMSE: 0.605–0.426.).
- Corresponding to cross-validation techniques, there is a slightly different performance between k-Fold and leave-one-out cross-validation for all machine learning algorithms but decision tree, where noticeable differences were identified (e.g., SVM1: PCC 0.802–0.909, SROCC 0.779–0.914 & RMSE 0.593–0.426; SVM2: PCC 0.864–0.905, SROCC 0.852–0.914 & RMSE 0.520–0.432).

V. FUTURE WORKS

At the time of the writing of this paper, it could be stated that a comprehensive 360 audiovisual quality database remains scarce. As expected, our subjective assessment as described in [12] showed that the MOS score did not fully span the scale range. Therefore, the availability of a 360 audiovisual

database, which we are currently developing, is highly critical in order to collect better quality data from retraining and building the models.

As the results shows in Tables III, IV, and V, combination of certain audio-video quality metrics reveals a promising correlation with subjective data. It is therefore exposing a new interest to consider a number of potential objective metrics used in various application to be included. Although the metric is typically application-dependent (speech, music, annotation)⁴, there are at least 28 audio quality metrics exist as has been identified and proposed in [34] and could be selectively applied to advance this study.

The data augmentation by means increasing the features through perceptual measures, feature extraction, or by incorporating annotated data recorded from physiological sensors for learning will allow us to explore the more optimal model based on the combinatorial features and/or deep learning algorithm, which is expected to provide more accurate results.

ACKNOWLEDGMENT

The authors wish to thank Prof. Angelo Farina for the permission in using the source materials and to the authors in [10] for providing the Matlab code of AMBIQUAL. This research is funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No.765911 RealVision.

REFERENCES

- [1] Z. Akhtar and T. H. Falk, "Audio-visual multimedia quality assessment: A comprehensive survey," *IEEE Access*, vol. 5, pp. 21 090–21 117, 2017.
- [2] J. You, U. Reiter, M. M. Hannuksela, M. Gabbouj, and A. Perkis, "Perceptual-based quality assessment for audio-visual services: A survey," *Signal Processing: Image Communication*, vol. 25, no. 7, pp. 482–501, 2010.
- [3] H. T. Tran, C. T. Pham, N. P. Ngoc, A. T. Pham, and T. C. Thang, "A study on quality metrics for 360 video communications," *IEICE Trans. Inf. Syst.*, vol. 101, no. 1, pp. 28–36, 2018.
- [4] Z. Fei, F. Wang, J. Wang, and X. Xie, "QoE evaluation methods for 360-degree VR video transmission," *IEEE J. Sel. Topics Signal Process.*, 2019.
- [5] S. Fremerey, S. Göring, R. R. Rao, R. Huang, and A. Raake, "Subjective test dataset and meta-data-based models for 360° streaming video quality," in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSp)*. IEEE, 2020, pp. 1–6.
- [6] S. Xie, Y. Xu, Q. Shen, Z. Ma, and W. Zhang, "Modeling the perceptual quality of viewport adaptive omnidirectional video streaming," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.
- [7] M. S. Anwar, J. Wang, W. Khan, A. Ullah, S. Ahmad, and Z. Fei, "Subjective QoE of 360-Degree Virtual Reality Videos and Machine Learning Predictions," *IEEE Access*, vol. 8, pp. 148 084–148 099, 2020.
- [8] M. A. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Eng. Soc.*, vol. 21, no. 1, pp. 2–10, 1973.
- [9] T. Rudzki, I. Gomez-Lanzaco, J. Stubbs, J. Skoglund, D. T. Murphy, and G. Kearney, "Auditory localization in low-bitrate compressed ambisonic scenes," *Appl. Sci.*, vol. 9, no. 13, p. 2618, 2019.
- [10] M. Narbutt, J. Skoglund, A. Allen, M. Chinen, D. Barry, and A. Hines, "AMBIQUAL: Towards a quality metric for headphone rendered compressed ambisonic spatial audio," *Appl. Sci.*, vol. 10, no. 9, p. 3188, 2020.
- [11] M. Orduna, C. Díaz, L. Muñoz, P. Pérez, I. Benito, and N. García, "Video Multimethod Assessment Fusion (VMAF) on 360VR contents," *IEEE Trans. Consum. Electron.*, vol. 66, no. 1, pp. 22–31, 2019.

- [12] R. F. Fela, N. Zacharov, and S. Forchhammer, "Towards a perceived audiovisual quality model for immersive content," in *Proceedings of the 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020, pp. 2472–2484.
- [13] "Index of /Public/Jump-Videos," <http://pcfarina.eng.unipr.it/Public/Jump-Videos/>, 2019.
- [14] ITU-T Rec. P.910, *Subjective Video Quality Assessment Methods for Multimedia Applications*, International Telecommunication Union Std., 2008.
- [15] ITU-R Rec. BS.1387-1, *Method for Objective Measurements of Perceived Audio Quality*, International Telecommunication Union Std., 2001.
- [16] C. Sloan, N. Harte, D. Kelly, A. C. Kokaram, and A. Hines, "Objective assessment of perceptual audio quality using ViSQOLAudio," *IEEE Trans. Broadcast.*, vol. 63, no. 4, pp. 693–705, 2017.
- [17] M. Chinen, F. S. Lim, J. Skoglund, N. Gureev, F. O’Gorman, and A. Hines, "Visqol v3: An open source production ready objective speech and audio metric," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2020, pp. 1–6.
- [18] E. Alshina, J. Boyce, A. Abbas, and Y. Ye, "JVET Common Test Conditions and Evaluation Procedures for 360 Degree Video," *JVET document, JVET-G1030*, 2017.
- [19] Y. Sun, A. Lu, and L. Yu, "Weighted-to-spherically-uniform quality evaluation for omnidirectional video," *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1408–1412, 2017.
- [20] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes," in *2015 IEEE International Symposium on Mixed and Augmented Reality*. IEEE, 2015, pp. 31–36.
- [21] H. Lin, C. Huang, C. Li *et al.*, "AHG8: inter Digital's Projection Format Conversion Tool Joint Video Exploration Team of ITU-T SG16 WP3 and ISO." IEC JTC1/SC29/WG11, JVET-D0021. In: 5th Meeting, Geneva, 2017.
- [22] V. Zakharchenko, K. P. Choi, and J. H. Park, "Quality metric for spherical panoramic video," in *Optics and Photonics for Information Processing X*, vol. 9970. International Society for Optics and Photonics, 2016, p. 99700C.
- [23] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," 2017. [Online]. Available: <https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652>
- [24] Z. Li, C. Bampis, J. Novak, A. Aaron, K. Swanson, M. Anush, and J. D. Cock, "VMAF: The journey continues," 2018. [Online]. Available: <https://netflixtechblog.com/vmaf-the-journey-continues-44b51ee9ed12>
- [25] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.
- [26] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. IEEE, 2003, pp. 1398–1402.
- [27] M. Kuhn *et al.*, "Building predictive models in r using the caret package," *J Stat Softw.*, vol. 28, no. 5, pp. 1–26, 2008.
- [28] M. Kuhn, J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, R. C. Team *et al.*, "Package 'caret'," *The R Journal*, p. 223, 2020.
- [29] T. M. Therneau, E. J. Atkinson *et al.*, "An introduction to recursive partitioning using the rpart routines," Technical report Mayo Foundation, Tech. Rep., 1997.
- [30] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [31] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "kernlab-an s4 package for kernel methods in r," *Journal of statistical software*, vol. 11, no. 9, pp. 1–20, 2004.
- [32] A. Karatzoglou, D. Meyer, and K. Hornik, "Support vector machines in r," *Journal of statistical software*, vol. 15, no. 9, pp. 1–28, 2006.
- [33] E. Demirebilek and J.-C. Gregoire, "Machine learning-based parametric audiovisual quality prediction models for real-time communications," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, no. 2, pp. 1–25, 2017.
- [34] M. Torcoli, T. Kastner, and J. Herre, "Objective measures of perceptual audio quality reviewed: An evaluation of their application domain dependence," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1530–1541, 2021.

⁴<https://www.amtoolbox.org/models.php>