Study of Subjective and Objective Quality Assessment of Audio-Visual Signals

Xiongkuo Min, *Member, IEEE*, Guangtao Zhai, *Senior Member, IEEE*, Jiantao Zhou, *Senior Member, IEEE*, Mylène C. Q. Farias, *Senior Member, IEEE*, and Alan Conrad Bovik, *Fellow, IEEE*

***Abstract*—The topics of visual and audio quality assessment (QA) have been widely researched for decades, yet nearly all of this prior work has focused only on single-mode visual or audio signals. However, visual signals rarely are presented without accompanying audio, including heavy-bandwidth video streaming applications. Moreover, the distortions that may separately (or conjointly) afflict the visual and audio signals collectively shape user-perceived quality of experience (QoE). This motivated us to conduct a subjective study of audio and video (A/V) quality, which we then used to compare and develop A/V quality measurement models and algorithms. The new LIVE-SJTU Audio and Video Quality Assessment (A/V-QA) Database includes 336 A/V sequences that were generated from 14 original source contents by applying 24 different A/V distortion combinations on them. We then conducted a subjective A/V quality perception study on the database towards attaining a better understanding of how humans perceive the overall combined quality of A/V signals. We also designed four different families of objective A/V quality prediction models, using a multimodal fusion strategy. The different types of A/V quality models differ in both the unimodal audio and video quality prediction models comprising the direct signal measurements and in the way that the two perceptual signal modes are combined. The objective models are built using both existing state-of-the-art audio and video quality prediction models and some new prediction models, as well as quality-predictive features delivered by a deep neural network. The methods of fusing audio and video quality predictions that are considered include simple product combinations as well as learned mappings. Using the new subjective A/V database as a tool, we validated and tested all of the objective A/V quality prediction models. We will make the database publicly available to facilitate further research.**

***Index Terms*—Quality assessment, audio-visual quality, video quality, audio quality, multimodal fusion.**

I. Introduction

STREAMING media now dominate the internet, and statistics on its composition show that video and audio streaming occupy about 60% of global network traffic [1]. Video streaming services like Netflix, YouTube, Amazon Video, Facebook Watch, and Hulu consume the largest fraction, but audio streaming services like Spotify and Apple Music are also among the top tier of consumers of internet capacity. While video streaming indisputably commands more resources, audio is also a resource hog, and like video, is important to consumers. Moreover, vision and audition are the richest sources of sensory data that we use to gather information from the world around us. Furthermore, streamed video is nearly always accompanied by audio, and certainly the perceived quality of experience (QoE) when viewing streaming video is deeply affected by both perceptual video quality and perceptual audio quality, or more precisely, their conjoint quality.

In streaming applications, video and audio signals generally pass through a processing pipeline consisting of several representative stages, including content generation, processing, encoding at the server side, streaming through the network, and finally, decoding and presentation to consumers at the end-user side [2]. Various impairments may be introduced along the way to either or both of the video and audio signals, which degrade the end-user's QoE. Modern streaming media consumers are increasing savvy about audio and video (A/V) technology, and expect high QoE when viewing and listening using increasingly high-resolution and high-fidelity A/V systems, whether they be on mobile devices or in their living rooms. Thus, there is significant impetus to develop and deploy efficient and accurate audio and video quality assessment (A/V-QA) models that can be used to monitor and control end-user QoE.

End-user perceived QoE depends on a wide variety of spatio-temporal factors related to content acquisition, processing, transmission, and visual and auditory perception. Typical distortions that degrade A/V content quality include acquisition errors, compression, resizing/rate changes, and much more, including temporal factors such as rebuffering and quality switching. A wide variety of picture and video quality databases are available that contain streaming video distortions, including the LIVE Video Quality Assessment Database [3] and the EPFL-Polimi Dataset [4], which

Manuscript received October 24, 2019; revised March 1, 2020 and April 4, 2020; accepted April 7, 2020. Date of publication April 21, 2020; date of current version May 1, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61901260, Grant 61831015, Grant 61521062, and Grant 61527804, in part by the Macau Science and Technology Development Fund under Grant FDCT/022/2017/A1 and Grant FDCT/077/2018/A2, and in part by the Research Committee at the University of Macau under Grant MYRG2016-00137-FST and Grant MYRG2018-00029-FST. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiaolin Hu. *(Corresponding author: Guangtao Zhai.)*

Xiongkuo Min and Guangtao Zhai are with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: minxiongkuo@gmail.com; zhaiguangtao@sjtu.edu.cn).

Jiantao Zhou is with the State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau, and also with the Department of Computer and Information Science, University of Macau, Macau (e-mail: jtzhou@umac.mo).

Mylène C. Q. Farias is with the Department of Electrical Engineering, University of Brasilia, Brasilia 70853-060, Brazil (e-mail: mylene@ieee.org).

Alan Conrad Bovik is with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: bovik@ece.utexas.edu).

Digital Object Identifier 10.1109/TIP.2020.2988148

include compression and transmission loss distortions. Similar audio resources include the ITU-T coded-speech database [5], which includes audio encoding distortions, environmental noise, and channel degradations. More recent databases address distortions that occur over longer time spans or at the client's side. For example, the LIVE-Netflix Video QoE Database [6] models the effects of both bitrate changes and rebuffering events, while a variety of objective streaming QoE predictors are proposed in [7]. These databases have been used to create, test, and compare a large number of video quality assessment (VQA) and video QoE models [7]–[16], and also audio quality assessment (AQA) models [17]–[22]. Useful surveys of VQA and AQA studies can be found in [2], [23]–[25].

Although many studies have separately addressed video and audio quality, very few have simultaneously addressed both. This is unfortunate, since both types of sensory signals shape user-perceived QoE. Some A/V-QA studies have been conducted [26]–[31], and there are relevant surveys that summarize and critique the available A/V-QA studies [32], [33] and their limitations.

Despite the limited volume of research on the topic, A/V-QA algorithms could be of great value in practice. For example, video streaming service providers like Netflix and Amazon Prime Video benefit by adaptive streaming of both video and audio [34]. Since the goal of adaptive streaming is to provide the best overall QoE under any network conditions, it makes sense to optimize audio and video QoE simultaneously. Since bandwidth-hungry surround sound is becoming more pervasive, perceptual audio rate control has become more important, and should be a significant factor in A/V QoE optimization, especially when the available bandwidth becomes limited. These issues greatly motivate us to study these problems.

In this paper, we make a number of contributions. First, we constructed a unique new audio-visual quality resource, which we call the LIVE-SJTU A/V-QA Database. The database is comprised of 14 source A/V sequences and 336 distorted versions of them, each of which was quality-rated by 35 human subjects. We limit the study to the perception of space-time A/V distortions that are localized in time, setting aside for now the study of longer-duration temporal patterns of rebuffering events/stalls and bitrate changes. Specifically, we target streaming applications, where A/V signal compression and scaling are the main distortion sources. The videos were impaired by two types of distortions: compression, and compression after spatial downsampling. Four levels of each type of distortion were applied. The audio signals were subjected to one type of compression distortion applied at three levels of severity. All of the possible combinations of these video and audio distortion conditions constitute the overall 24 distortion conditions that were applied to each source A/V sequence.

Second, we designed and conducted a large, comprehensive subjective A/V-QA study on the new LIVE-SJTU A/V-QA Database. As described further herein, we develop an appropriate A/V subjective testing environment, and invited 35 human subjects to participate in a subjective human study of A/V quality. These subjects were each asked to record their opinions of each distorted A/V sequence. We conducted a careful statistical validation analysis of the obtained subjective ratings, including a post-study questionnaire, and we arrived at a number of interesting conclusions regarding how humans perceive and assess A/V quality under various combinations of audio and visual distortion conditions.

Third, we designed four families of objective A/V-QA models, which can be differentiated according to the strategies used to fuse the measurements made on the two signal modalities. The first family of models integrates single-mode AQA and VQA models by fusing them using simple products and weighted products. The second family of models is learned by training support vector regressors (SVRs) to fuse A/V quality scores or features. Thirdly, we developed a set of audio quality models that derive from picture quality research, by adapting a set of classical 2D visual quality models for application to 1D audio signals. By combining these visually-inspired AQA models with VQA models, we arrive at a third family of A/V-QA models. The fourth family of models is defined by first computing 2D spectral representations of the audio signals, then using a pretrained deep neural network (DNN) to learn to extract content- and distortion-aware A/V features and to predict perceptual A/V quality. We conducted extensive experiments to illuminate the absolute and relative performances of these four families of new A/V-QA models, using the new LIVE-SJTU A/V-QA Database.

The rest of the paper is organized as follows. The detailed construction of the new LIVE-SJTU A/V-QA Database and the protocol we followed when conducting the subjective A/V-QA study are described in Section II. Section III introduces and details the four families of objective A/V quality prediction models. The experimental results are laid out in Section IV. Section V gives some recommendations on practical usage and deployment of the proposed AV-QA models. Section VI concludes the paper with a number of cogent observations.

II. Subjective Audio-Visual Quality Assessment

To facilitate our work on A/V quality measurement, we first constructed the new LIVE-SJTU Audio and Video Quality Assessment (A/V-QA) Database, and then conducted a sizeable human subjective study on it. Based on the collected subjective data, we give some observations regarding the outcomes of the human study, and their implications regarding multimodal A/V perceptual fusion mechanisms.

A. Reference and Distorted Contents

1) Source Contents: We collected a set of 14 diverse source videos with corresponding soundtracks from the Consumer Digital Video Library (CDVL) [35]. All of the selected videos are of very high visual and auditory quality. The videos all have resolutions of 1920 × 1080 pixels, and are provided in raw YUV 4:2:0 format. The frame rates of the 14 videos range from 24 to 29.97 frames per second, and all of the videos are of 8 seconds duration. The corresponding audio soundtracks are stereophonic audios with two channels which are provided in raw pulse-code modulation (PCM) format with a bit depth of 16 and a sampling rate of 48 kHz. Sample

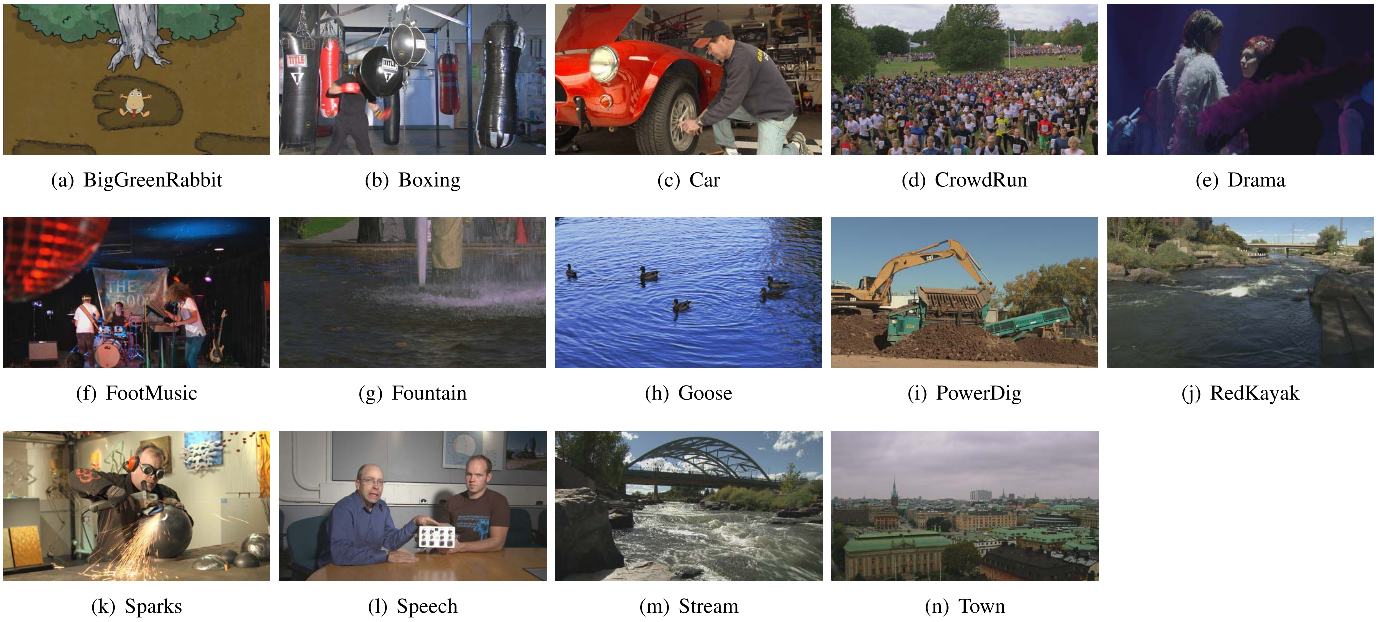


Fig. 1. Sample frames of the 14 source videos used in the LIVE-SJTU A/V-QA database.

frames of all 14 source videos are provided in Fig. 1. The video contents include normal daily activities, television show, landscape, animation, people at work, and so on. The audio contents include human speech, music, machine sounds, water, and more.

2) Distortion Sources: We generated a much larger set of distorted contents by applying quality degradation processes that occur in video streaming applications. Hence, we focused on A/V signal compression, and compression combined with scaling. Specifically for video, we modeled the following two distortion types.

- **Video compression**: We chose the high efficiency video coding (HEVC) as the video compression method, given its status as the latest ITU global video compression standard. The specific implementation of HEVC that we used is the ffmpeg x265 encoder. For each source video, 4 different compression levels were applied by selecting the constant rate factor (CRF) mode, and setting the CRF = 16, 35, 42, and 50. These quality factors were selected to generate a wide range of perceptually well-separated video compression qualities over a range containing typical operating points.
- **Video compression plus scaling**: Modern video streaming systems often spatially downsample videos prior to compressing them, then upscale them following decoding, prior to display [14]. We created distorted videos by downsampling original 1080p videos to resolution 1280 × 720 (720p), compressing these using the same compression settings as described above, then spatially upscaling them back to the original resolution of 1920 × 1080. Lanczos resampling [36] was used to upscale the reduced and decompressed videos, given its prevalence in video players and displays.

We distorted the audio content as follows.

- **Audio compression**: Audio signals are also generally compressed before being distributed to users. We chose the advanced audio coding (AAC) as the audio compression method, using the basic ffmpeg AAC encoder implementation. We again used the constant bit rate (CBR) mode, and set the bitrate at the three levels 128, 32 and 8 kbps, thereby generating three levels of perceptually well-separated audio compression distortion.

Finally, all possible combinations of the above video and audio distortions were used to generate the complete set of distorted A/V signals. In summary, 24 distortion conditions (generated from all possible combinations of 2 video distortion types, 4 video distortion levels, and 3 audio distortion levels) were applied to the 14 reference A/V sequences, yielding a total of 336 distorted A/V sequences.

B. Subjective Human Study

1) Experiment Setup: We conducted a subjective human study in LIVE to obtain data representative of how humans perceive distorted A/V quality. The A/V testing environment included a ASUS G750JX-TB71 PC equipped with a HP VH240a 23.8-inch 1920 × 1080 monitor and a Bose Companion 20 speaker which was located next to the display. We designed a user interface whereby the subjects could view/listen and rate the videos. All videos were displayed at native resolution to avoid further scaling distortions. The refresh rate of 60 Hz is larger than the frame rates of all of the the videos. Prior to the study, we carefully tested playback of all A/V sequences to eliminate any concerns regarding latency, frame drops, loss of A/V synchronization, etc.

2) Testing Methodology: We adopted a single stimulus continuous quality evaluation (SSCQE) strategy to obtain the subjective quality ratings on all of the distorted A/V sequences. After each A/V sequence was viewed, a continuous quality rating bar was presented to the subject. The quality bar was

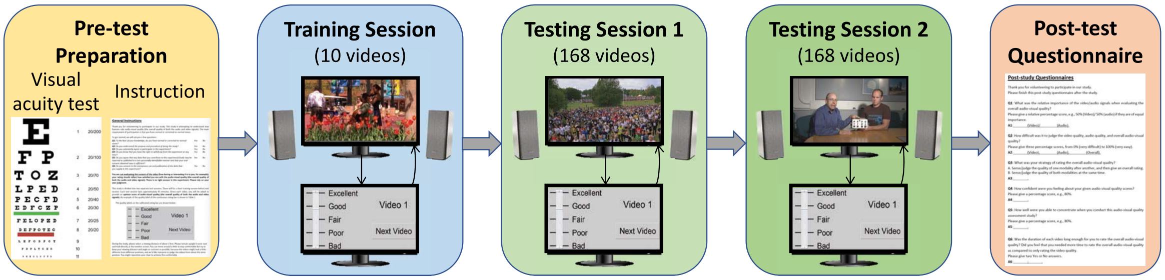


Fig. 2. Workflow of the human subjective study, which followed 5 stages: preparation, training, 2 testing sessions, and a posterior questionnaire.

labeled with five Likert adjectives: Bad, Poor, Fair, Good and Excellent, allowing subjects to smoothly drag a slider (initially centered) along the continuous quality bar to select their ratings. All subjects were instructed to *give an opinion score of the overall A/V quality they perceived.* They were seated at a distance of about 2 feet from the monitor, and this viewing distance was roughly maintained during each session.

3) Testing Procedure: The human subjective study was conducted in the LIVE subjective study room at The University of Texas at Austin. A total of 35 subjects participated in the study, most of them UT-Austin graduate students. A workflow of the human subjective study, comprised of 5 stages, is illustrated in Fig. 2. Before participating in the test, each subject read and signed a consent form which explained the human study, and participated in a Snellen visual acuity test [37]. All subjects were determined to have normal or corrected-to-normal vision. General information about the study was supplied in printed form to the subjects, along with instructions on how to participate in the A/V task. Each subject then experienced a short training session where 10 A/V sequences (not included in the actual test) were played, allowing them to become familiar with the user interface and the general range and types of distortions which may occur. The same distortion generation procedure was conducted for the training videos as for the test videos. The entire collection of 336 distorted videos was randomly and equally divided into 2 sessions. All subjects participated in both sessions, which were separated by at least 24 hours. The order in which the test videos were played was randomized and different for each subject. After participating in both testing sessions, the subjects were asked to answer a questionnaire regarding their experience. Details of the questionnaire and the results obtained from the study are given in Section II-C.

C. Subjective Data Processing and Analysis

1) MOS Calculation and Analysis: We follow the suggestions given in [38] to conduct subject rejection. Only two of the 35 subjects were detected as outliers and rejected. For the remaining 33 valid subjects, we converted the raw ratings into Z-scores, which were then linearly scaled to the range [0, 100] and averaged over subjects to obtain the final mean opinion scores (MOSs)

$$z_{ij} = \frac{r_{ij} - \mu_i}{\sigma_i}, \quad z'_{ij} = \frac{100(z_{ij} + 3)}{6}, \tag{1}$$

$$\mathrm{MOS}_j = \frac{1}{N}\sum_{i=1}^{N} z'_{ij}, \tag{2}$$

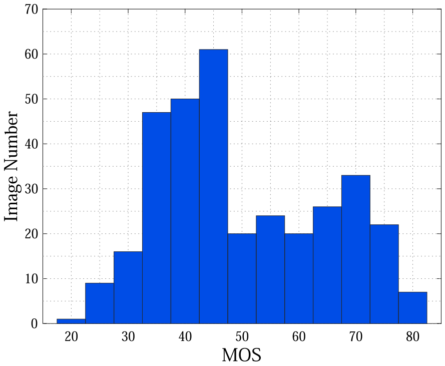


Fig. 3. Histogram of MOSs from the LIVE-SJTU A/V-QA database.

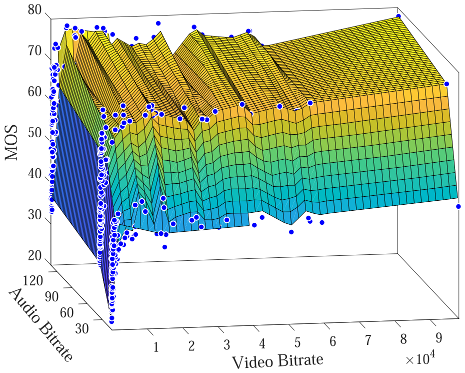
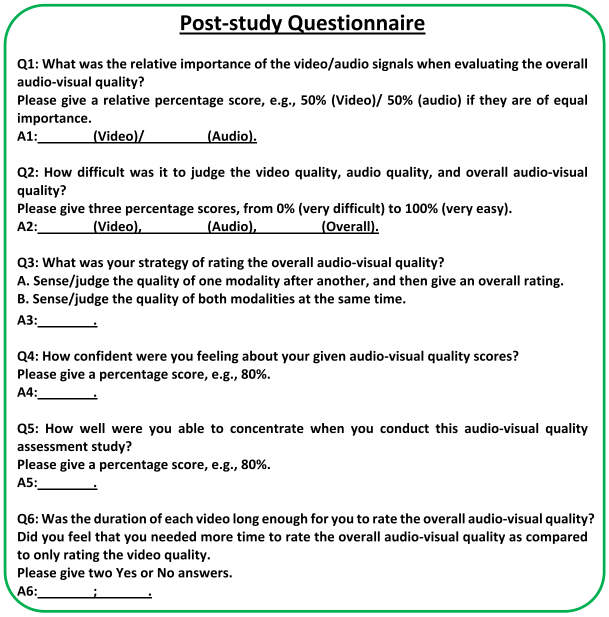


Fig. 4. Plot of video bitrate, audio bitrate, and MOS against each other. The bitrates have units of kbps. Each point corresponds to a single A/V sequence from the database.

where r_{ij} is the raw rating given by the ith subject to the jth image, μ_i is the mean rating given by subject i, σ_i is the standard deviation, and N is the total number of subjects. Fig. 3 plots the histogram of MOSs over the entire database, showing a wide range of perceptual quality scores. We also plotted video bitrate, audio bitrate, and MOS against each other in Fig. 4, where the bitrate values for both video and audio were obtained during encoding. Piecewise linear interpolation of the MOS was used to improve visibility of the trends. It may be observed that MOS generally increased with higher audio and video bitrates, but this was not always true, and the MOS trend varied non-monotonically with the combined audio-video bitrates. This further implies the need for A/V quality measures that are able to accurately predict human percepts of overall A/V quality.



Post-study Questionnaire

Q1: What was the relative importance of the video/audio signals when evaluating the overall audio-visual quality?
Please give a relative percentage score, e.g., 50% (Video)/ 50% (audio) if they are of equal importance.
A1:_______(Video)/________(Audio).

Q2: How difficult was it to judge the video quality, audio quality, and overall audio-visual quality?
Please give three percentage scores, from 0% (very difficult) to 100% (very easy).
A2:_______(Video),________(Audio),________(Overall).

Q3: What was your strategy of rating the overall audio-visual quality?
A. Sense/judge the quality of one modality after another, and then give an overall rating.
B. Sense/judge the quality of both modalities at the same time.
A3:________.

Q4: How confident were you feeling about your given audio-visual quality scores?
Please give a percentage score, e.g., 80%.
A4:________.

Q5: How well were you able to concentrate when you conduct this audio-visual quality assessment study?
Please give a percentage score, e.g., 80%.
A5:________.

Q6: Was the duration of each video long enough for you to rate the overall audio-visual quality? Did you feel that you needed more time to rate the overall audio-visual quality as compared to only rating the video quality.
Please give two Yes or No answers.
A6:________;________.

Fig. 5. A list of the questions included in the post-test questionnaire.

2) Questionnaire Analysis: Post-test questionnaires can help give a better understand of how the subjects felt about the study, and how future studies might be improved. All of the questions that were asked are shown in Fig. 5, while statistical results of the replies given by all 35 subjects are given in Fig. 6. Some general observations are given as follows.

- Q1: The majority of subjects thought that the viewed video components were relatively more important to their experiences than audio components when they were rating A/V quality. The average relative importance of video and audio signals was 57%:43%.
- Q2: As compared with video quality, subjects found it harder to judge the audio quality, and even more difficult to judge the overall A/V quality.
- Q3: Different subjects relied on different internal strategies when rating overall A/V quality. A majority (60%) of subjects thought that they judged the quality of one mode first, then the other, and then gave an overall rating.
- Q4: Most subjects felt pretty confident about their ratings.
- Q5: Most subjects were able to concentrate during the test.
- Q6: Almost all of the subjects felt that 8 seconds was adequate time to be able to give an accurate overall rating. The majority thought that they would not need more time to rate the A/V quality as compared with only rating the video quality.

Although the questionnaires only broadly describe the experiences of the human subjects, they still provided some useful insights regarding the efficacy of the study and strategies going forward.

3) Subjective Audio-Visual Quality Model: Intuitively, perceived overall A/V quality is a fusion of perceived video quality and audio quality. If the A/V, video and audio qualities are described by subjective quality scores, then a model that fuses them is a subjective A/V quality model [26], [29], [30].

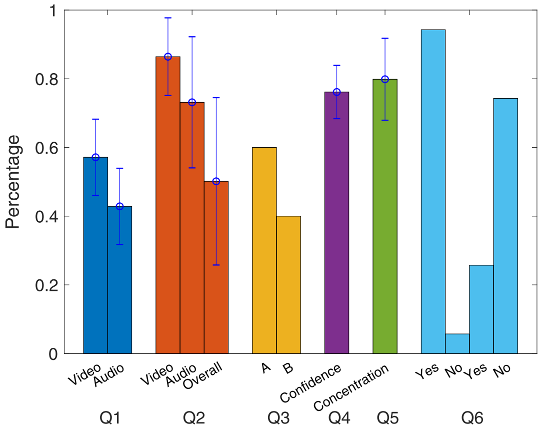


Fig. 6. Results of the post-test questionnaire. The error bar denotes one standard deviation.

TABLE I

SUBJECTIVE AV QUALITY MODELS. VALUE DENOTES THE FITTING VALUE, WHILE BOUNDS DENOTE THE 95% CONFIDENCE BOUNDS

Model	w_1, w_2		k_1, k_2	
	Value	Bounds	Value	Bounds
Model 1	0.5762	(0.5458, 0.6065)	-10.38	(-10.99, -9.761)
Model 2	0.5826	(0.5422, 0.6229)	0.8563	(0.8450, 0.8675)

We use MOS$_{av}$, MOS$_v$ and MOS$_a$ to denote the reported subjective A/V quality, video quality and audio quality, respectively. Of course, MOS$_{av}$ was obtained as the final MOS from the database. Since we did not collect single-mode ratings, we instead derived the estimated MOS (eMOS$_v$ and eMOS$_a$) as follows: for a given A/V sequence (e.g. a video coded with CRF 42 and associated audio coded with bitrate 32 kbps), let eMOS$_v$ be the value of MOS$_{av}$ corresponding to the A/V sequence having the same video distortion but the lowest (highest quality) audio distortion setting (i.e., the video coded with CRF 42 and associated audio coded with bitrate 128 kbps, which derive from the same original content). Likewise, define eMOS$_a$ to be the value of MOS$_{av}$ corresponding to the A/V sequence having the same audio distortion but the lowest video distortion setting (CRF 16). We define eMOS$_v$ and eMOS$_a$ in this way since we did not collect single-mode ratings, and since the least compressed videos and highest bitrate audios are close to pristine. Still, we recognize that there remains uncertainty in these estimates, so we only use them to generally validate our hypotheses.

We then designed two subjective A/V quality models:

- Model 1: weighted sum of single-mode estimated subjective qualities

$$\mathrm{MOS}_{av} = w_1 \cdot \mathrm{eMOS}_v + (1 - w_1) \cdot \mathrm{eMOS}_a + k_1. \quad (3)$$

- Model 2: weighted product of single-mode estimated subjective qualities

$$\mathrm{MOS}_{av} = k_2 \cdot \mathrm{eMOS}_v^{w_2} \cdot \mathrm{eMOS}_a^{1-w_2}, \quad (4)$$

where w_1 and w_2 denote linear and exponent weights, and k_1 and k_2 are biases to be fitted. We fit the above two models on the LIVE-SJTU A/V-QA Database, with results listed in Table I. It observed that the two models are in general

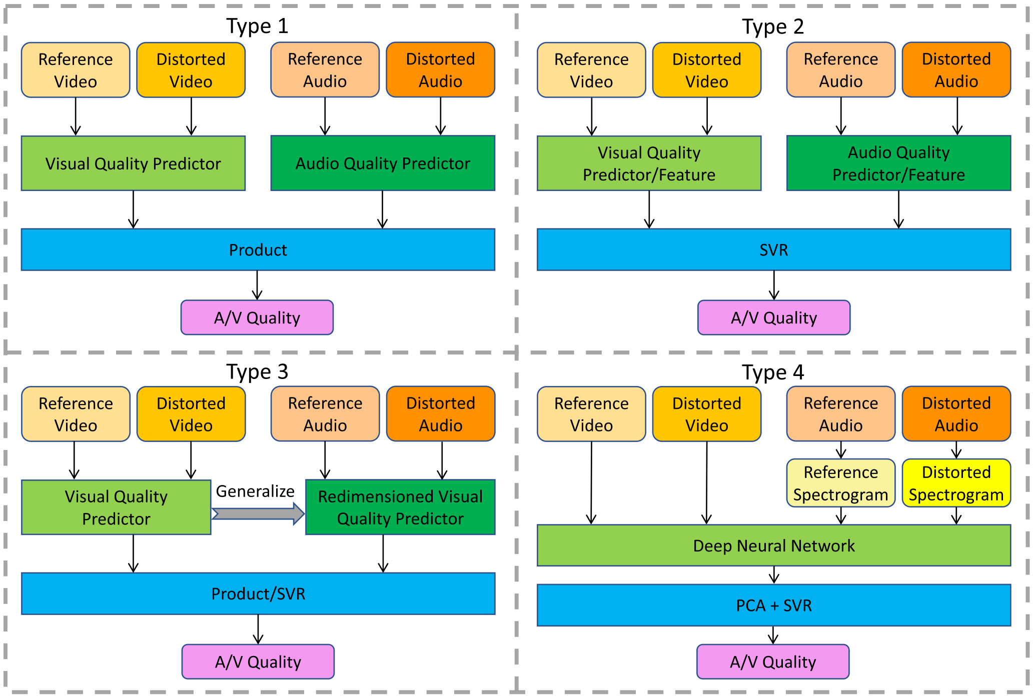


Fig. 7. Frameworks of the proposed four families of A/V quality prediction models.

agreement. In both models, the weight on the video term is larger than that on the audio term. Moreover, the relative weights are very close to the relative importance values obtained by the questionnaire.

III. Objective Audio-Visual Quality Assessment

When attempting to create models predictive of overall perceived A/V quality, it is quite reasonable to assume that a fusion of measurements on the visual and audio signals will be required. While objective VQA and AQA models have been studied for years, and many successful algorithms have been proposed, objective A/V-QA measurement remains a relatively unexplored problem. Here we attempt to better fill this gap by advancing four families of A/V-QA models, that differ in their complexities, methods of fusion, and requirements on training (if any). We will refer to these families as Types 1, 2, 3, and 4. In this paper we mainly focus on full-reference A/V-QA. The original source contents that were used to generate the distorted A/V sequences were used as the reference signals when testing the A/V-QA models, since they are of high quality and free of compression and scaling distortions. The frameworks of the four types of A/V quality models are illustrated in Fig. 7. Details of each type are given as follows.

A. Type 1: Product of Video and Audio Quality Predictors

Since numerous single-mode visual and audio quality predictors have been proposed, it is reasonable to consider using them to predict video and audio qualities first, then combining the results into a single, united A/V quality predictor by fusing them [26], [30]. The idea of such a posterior fusion strategy finds some support from the post-test questionnaire, where a majority of subjects stated that they first judged the video and audio quality separately, and then subsequently fused them. Perhaps the simplest form of bi-modal fusion is the product of AQA and VQA model responses, if both are properly scaled. As depicted at the top left of Fig. 7, we leveraged existing knowledge by deploying top-performing quality predictors for both modalities, as follows. The A/V quality prediction is

$$Q_{av} = Q_v \cdot Q_a, \tag{5}$$

where Q_v and Q_a denote the video and audio quality predictions computed on the respective components of distorted test signals. We used the following well-known video and audio quality predictors:

- Video: VMAF [14], STRRED [13], SpEED [39], VQM [40], SSIM [8], MS-SSIM [9], VIFP [10], FSIM [41], GMSD [42];
- Audio: PEAQ [17], STOI [19], VISQOL [20], log-likelihood ratio (LLR) [21], signal-to-noise ratio (SNR), and segmental SNR (segSNR) [22].

Any video quality measure can, in principle, be combined with any audio quality measure, if they are appropriately scaled or normalized. However, since the ranges of the video and audio predictors may differ, it is proper to normalize them prior to forming the product

$$Q_{av} = \hat{Q}_v \cdot \hat{Q}_a, \tag{6}$$

where either (to match a desired decreasing or increasing trend)

$$\hat{Q}_a = \frac{Q_a - Q_{a_{min}}}{Q_{a_{max}} - Q_{a_{min}}}, \quad \text{or } \hat{Q}_a = 1 - \frac{Q_a - Q_{a_{min}}}{Q_{a_{max}} - Q_{a_{min}}}, \tag{7}$$

where $Q_{a_{max}}$ and $Q_{a_{min}}$ bound the known range of Q_a, which may need to be determined empirically. The normalized video

TABLE II
VIDEO AND AUDIO QUALITY MEASURES AND THE CORRESPONDING DECOMPOSED FEATURES USED IN TYPE 2 MODELS

Category	Measure	#Feature	Decomposed features
Video	VMAF [14]	6	4 scales of VIF, detail loss, motion
	STRRED [13]	6	Full and single number versions of SRRED, TRRED, STREED
	SpEED [39]	6	Full and single number versions of spatial, temporal, spatial-temporal SpEED
	VQM [40]	7	4 spatial gradient features, 2 chrominance features, 1 contrast and motion feature
	SSIM [8]	2	Luminance similarity, contrast and structural similarity
	MS-SSIM [9]	6	Luminance similarity, 5 scales of contrast and structural similarity
	VIFP [10]	4	4 scales of VIFP features
	FSIM [41]	3	Phase congruency, gradient magnitude, and chrominance similarity
	GMSD [42]	2	Mean and standard deviation of gradient magnitude similarity
Audio	PEAQ [17]	11	11 model output variables before the neural Network
	STOI [19]	1	The complete algorithm
	VISQOL [20]	3	Narrowband, wideband, fullband versions of VISQOL
	LLR [21]	1	The complete algorithm
	SNR [21]	1	The complete algorithm
	segSNR [22]	1	The complete algorithm

score $\hat{Q}_v$ is also obtained using one of the forms in (7). Naturally, both $\hat{Q}_a$ and $\hat{Q}_v$ are defined with the same sign of trend. After normalization, the overall product quality score will monotonically increase (or decrease, as desired) with ground-truth A/V quality.

Since the video and audio modalities have different importances, a weighted product

$$Q_{av} = \hat{Q}_v^w \cdot \hat{Q}_a^{1-w}, \tag{8}$$

may instead be employed, where $0 \leqslant w \leqslant 1$. The optimal weight depends on the particular unimodal quality predictors used, as well as the particular application, which may be characterized by more or less severe distortions, for example. In any case, the (weighted) product has virtues of simplicity, efficiency, and easy interpretability, and as we shall see, it performs reasonably well. Of course, other simple fusion schemes might also be considered such as linear regression, linear regression plus product, harmonic mean, etc.

B. Type 2: Fusion of Video and Audio Quality Predictors by SVR

We can also make use of available data to derive a trained regressor to integrate quality predictions derived from single-mode quality models. An efficient way is to deploy an SVR [43] to learn the quality fusion

$$Q_{av} = \mathrm{SVR}(Q_v, Q_a), \tag{9}$$

where Q_v and Q_a have the same definitions as in (5). In this case, the SVR is trained on the predicted single-mode quality scores and the subjective ground-truth A/V quality labels.

This may be improved further, by instead using quality-aware feature vectors $\mathbf{f}_v$ and $\mathbf{f}_a$ which may be independently derived, or may be components (features) of existing VQA and AQA models. Then use these features to train the SVR

$$Q_{av} = \mathrm{SVR}(\mathbf{f}_v, \mathbf{f}_a). \tag{10}$$

The video and audio quality quality-aware feature vectors that we use here were drawn from top-performing AQA and VQA models, and are summarized in Table II. Other basic machine learning based regression techniques might also to be considered, such as random forests.

C. Type 3: A/V-QA Models Defined Using 1D and 2D Visual Quality Predictors

Visual and audio quality assessment have both been widely researched for decades, yet work in the two areas has been largely mutually isolated. But the neurosensory apparatus of the visual and audio modalities bear important similarities, and it is reasonable to consider whether suitably redimensioned VQA models might be adapted for audio quality prediction. Moreover, VQA models have largely been designed on the basis of perceptual concepts that are shared with audio perception. For example, visual masking, including luminance and contrast masking [44], [45], is implemented by the most successful video quality models. Likewise, auditory masking principles are well-understood, and simultaneous masking is embodied by VQA models like SSIM and MS-SSIM [8], [9]. Intensity masking also holds for audio perception [46], [47]. Indeed, SSIM has already been shown to be quite effective for audio quality prediction [48], [49]. Furthermore, most VQA models utilize multi-scale modeling, which is fundamental to both visual [9], [50] and audio signal processing [51]. Some of the most successful VQA models rely on natural scene statistics models to characterize quality, including VIF [10], STRRED [13], VMAF [14], SpEED [39], etc. Similar signal statistics have been observed to be fundamental to audio signals with similar implications for audio perception [52], [53]. This potentially includes the prediction of audio quality, and which is a contribution we make here. It turns out that natural statistics-based features make very good audio quality predictors, for the same reasons as they do for pictures and videos. Other of the employed VQA models, like GMSM and GMSD, utilize image gradients [42], similar to the temporal derivatives used in AQA [17]. These concepts greatly motivated us to generalize 2D spatial visual quality models for application to the 1D audio quality prediction. We accomplish this by reducing all 2D processes in a given visual quality model into 1D processes, then use the dimension-reduced models to predict audio quality. These generalized measures

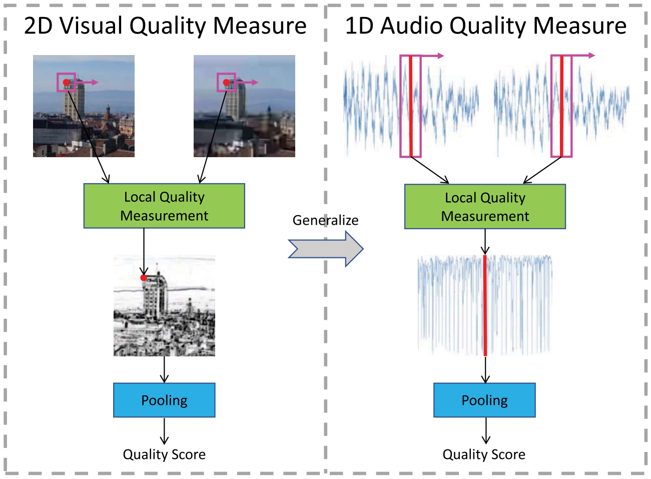

2D Visual Quality Measure
1D Audio Quality Measure
Local Quality Measurement
Generalize
Local Quality Measurement
Pooling
Pooling
Quality Score
Quality Score

Fig. 8. Workflows of 2D VQA model (left) and redimensioned 1D AQA model adapted for 1D audio quality measurement (right). Local fidelity/quality is measured by extracting and comparing local features using local moving windows in both 2D and 1D.

are then integrated with visual quality predictors using the fusion methods employed in the Type 1 and Type 2 models.

Workflows of this kind of A/V quality prediction process are illustrated in Fig. 8. Generally, data is captured and analyzed along one dimension rather than two. For example, a 1D SSIM is easily defined using 1D data windows on both the reference and test signals. In this general manner, we generalized 1D instances of the popular video frame/picture quality predictors SSIM, MS-SSIM, VIFP, GMSM, and GMSD, which we then apply to the audio signal components. Denote these redimensioned 1D video frame quality models as SSIM_{1D}, $\text{MS-SSIM}_{\text{1D}}$, VIFP_{1D}, GMSM_{1D}, and GMSD_{1D}. Then define the 1D-2D hybrid VQA type of A/V-QA model as

$$Q_{av} = Q_{\text{2D}}^{w} \cdot Q_{\text{1D}}^{1-w}, \tag{11}$$

where Q_{2D} can be any of SSIM, MS-SSIM, VIFP, GMSM, GMSD (at least), and Q_{1D} is a redimensioned 1D version of the same model, applied to the audio component.

Compared with the Type 1 and 2 models, the Type 3 models utilize the same methodology to predict the qualities of the video and audio components, and thus are more consistent. We specifically named these models as ***Audio-Visual SSIM (AVSSIM)***, ***Audio-Visual MS-SSIM (AVMSSSIM)***, ***Audio-Visual Information Fidelity in Pixel domain (AVIFP)***, ***Audio-Visual GMSM (AVGMSM)***, and ***Audio-Visual GMSD (AVGMSD)***. Of course, the redimensioned 1D VQA models can be also combined with other models than their original 2D versions, using the methods of fusion in (8), (9) and (10). Specifically, since VMAF (whose basis is VIF) utilizes an SVR, we combine it with VIFP_{1D} via the feature based SVR fusion used in the Type 2 models, and refer to the model as ***Audio and Video Multimethod Assessment Fusion (AVMAF)***. In the experimental section, we will show that AVSSIM, AVMSSSIM, AVIFP, AVGMSM, and AVGMSD all performed pretty well, while AVMAF was excellent.

D. Type 4: Deep Neural Families of A/V Quality Predictors

All of the above families of A/V-QA models are based on hand-crafted features. Given the successes of deep neural networks (DNNs) on wide swathes of visual problems, we also employ them for A/V quality prediction. Our framework for DNN based A/V quality prediction is illustrated Fig. 9. Specifically, we use a DNN pretrained on ImageNet [54] as a feature extractor, by extracting deep features from the final layers of the DNN, then feeding them to an SVR trained to predict the overall A/V quality. Since the resolution of video frames is generally much larger than the input dimension of available pretrained DNNs, we randomly cropped N patches whose resolutions fit the input of the DNN from each video frame. The frame patches were then fed into the DNN, to extract patch features from its final layers. The patch features of all N patches were then averaged to produce the video frame features.

Each 1D audio signal was first transformed into a 2D representation, by calculating the spectrum of each audio segment. Similar to [55], [56], we developed the spectrogram, which only includes spectral magnitudes. The short-time Fourier transform (STFT) was applied to calculate the spectrogram, which was fed to the same pretrained DNN, while audio features were extracted from the same layer of the video DNN. For both reference and distorted signals, the same procedures were followed, and the reference and distorted video/audio features were extracted. The reference and distorted video/audio features and their differences were then fused using an SVR. Principal component analysis (PCA) was applied to these features to reduce the feature dimension, prior to feeding them to the DNN.

We used the pretrained ResNet-50 model [57] as an exemplar DNN. We removed the last fully connected layer to extract content-aware features, whose dimension is 2048. The input dimension of ResNet-50 is 224 × 224, thus we cropped image patches of this size, and the audio spectrogram was also calculated as this size. To calculate spectrograms of this size from the audio, 20 milliseconds windows having 75% overlap at a step of 5 milliseconds were used to apply the STFT to the $224 \times 5 = 1120$ milliseconds of the audio segment nearest to the corresponding video frame. Then, 224 frequency points uniformly distributed on the mel scale were sampled and converted to the hertz scale. As illustrated in Fig. 10, the 224 sampled frequencies span the human audible frequency range of 20 Hz to 20 kHz. The conversion function between mel scale and hertz scale is

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right), \tag{12}$$

where m and f are mel scale and hertz scale frequencies.

A total of 6 groups (from the reference and distorted video and audio sequences, and the feature differences between the reference and distorted signals) of 2048-dimensional features were extracted. To reduce the feature dimension, we applied PCA to all $2048 \times 6 = 12288$ features before SVR feature fusion. The feature dimension was reduced to 25, which is a typical feature dimension for many SVR based quality predictors. The method illustrated in Fig. 9 is a frame based

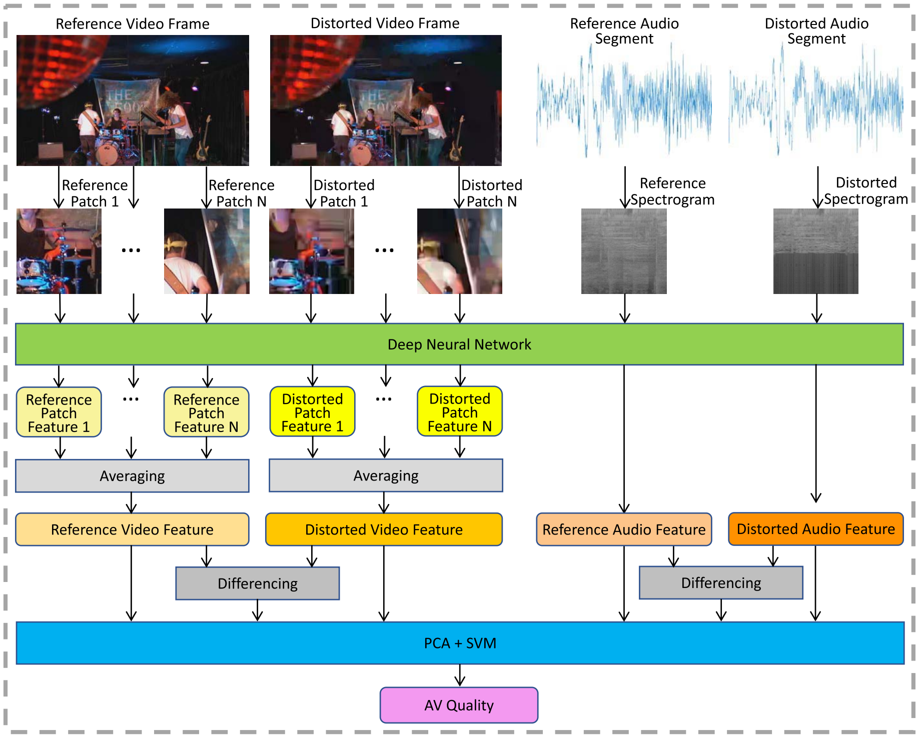


Fig. 9. Framework of DNN based A/V quality model. For the video component, randomly crop N patches whose sizes fit the input of the DNN from the video frame, which are then fed into the pretrained DNN. Patch features are extracted from the final layers, then the the features from all N patches are averaged to produce video frame features. For the audio component, the spectrogram is computed from each audio segment, then fed into the same DNN, then audio features are extracted from the final layers. The extracted A/V features and their differences are then fused by the trained SVR. PCA is applied before the fusion to reduce the feature dimension.

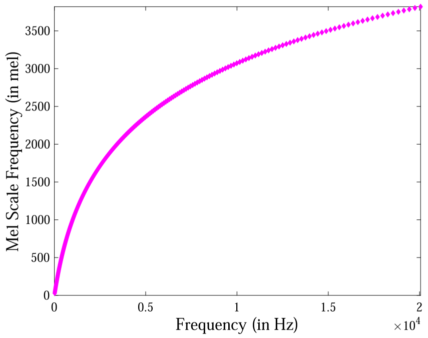


Fig. 10. 224 frequency points uniformly distributed in mel scale are sampled and converted into hertz scale to calculate the spetrogram.

method, hence it is applied to the video frames and its corresponding audio segments to predict the frame quality. The overall video quality is calculated as the average of the frame quality predictions. To reduce computation, a frame skip of 10 is employed, meaning that one frame and its concurrent audio segment are extracted every 10 frames to compute the frame quality.

IV. Experimental Results

We relied upon the LIVE-SJTU A/V-QA Database to test and compare the 4 families of A/V quality predictors in the preceding. These experiments also served to validate the utility of the A/V-QA database.

A. Experimental Setting

The four families of A/V-QA models that were introduced in Section III were tested. When applying image (frame) quality algorithms to videos, the computed frame quality predictions were averaged over all frames to produce the final video quality predictions (viz., average pooling). The above unimodal models will serve as component audio and video quality predictors in the Type 1, 2, and 3 A/V-QA models to be evaluated.

To evaluate the various quality predictors, we followed the recommendations given in [58], and used a five-parameter logistic function to fit the quality scores:

$$Q' = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + e^{\beta_2 (Q - \beta_3)}} \right) + \beta_4 Q + \beta_5, \qquad (13)$$

where Q and Q' are the objective and best-fitting quality scores, and the parameters $\{\beta_i | i = 1, 2, ..., 5\}$ were determined via curve fitting during the evaluation. The consistency between the ground-truth subjective ratings and the fitted quality scores is measured to evaluate the quality model. We used the Spearman rank-order correlation coefficient (SRCC) to measure the prediction monotonicity of the models and the Pearson linear correlation coefficient (PLCC) to measure the prediction linearity. For both SRCC and PLCC, larger values denote better performance.

Exemplars from all four families of A/V quality models described in Section III were evaluated. Among them, some models involve training, while others do not. For fair compari-

TABLE III
PERFORMANCES OF TYPE 1 (TOP HALF) AND TYPE 2 (BOTTOM HALF) A/V QUALITY MODELS. THE TOP 3 MODELS OF EACH SUB-TYPE ARE IN BOLD

Criteria	Video Model	Product						Weighted Product					
		PEAQ	STOI	VISQOL	LLR	SNR	segSNR	PEAQ	STOI	VISQOL	LLR	SNR	segSNR
SRCC	VMAF	0.5920	**0.9287**	0.8450	0.7640	0.8164	0.8356	0.7536	**0.9334**	0.8436	0.7596	0.8871	**0.9324**
	STRRED	0.3782	0.8396	0.7019	0.5962	0.6804	0.7050	0.7274	0.8981	0.8125	0.7526	0.8605	0.8859
	SpEED	0.3628	0.8305	0.6881	0.5770	0.6691	0.6947	0.7303	0.8978	0.8134	0.7555	0.8657	0.8924
	VQM	0.5949	**0.9091**	0.8145	0.7548	0.8224	0.8460	0.7287	0.9052	0.8183	0.7428	0.8644	0.9037
	SSIM	0.2917	0.7937	0.6100	0.4898	0.6113	0.6399	0.7261	0.9192	0.8172	0.7546	0.8749	0.9117
	MS-SSIM	0.2818	0.7799	0.5971	0.4698	0.6064	0.6362	0.7119	0.9092	0.8082	0.7480	0.8665	0.9039
	VIFP	0.5791	**0.9122**	0.8275	0.7150	0.8138	0.8285	0.7405	0.9146	0.8276	0.7348	0.8800	0.9238
	FSIM	0.2634	0.7486	0.5715	0.4199	0.5942	0.6272	0.7462	**0.9429**	0.8427	0.7718	0.8709	0.9016
	GMSD	0.4882	0.8808	0.7975	0.6915	0.7441	0.7713	0.7165	0.8964	0.7985	0.7245	0.8470	0.8939
PLCC	VMAF	0.6639	**0.9534**	0.8646	0.7606	0.8392	0.8369	0.7581	**0.9499**	0.8606	0.7568	0.9112	0.9404
	STRRED	0.5981	0.8425	0.7446	0.6136	0.6638	0.6596	0.7268	0.9302	0.8379	0.7520	0.8847	0.8991
	SpEED	0.5987	0.8281	0.7349	0.6000	0.6491	0.6542	0.7289	0.9309	0.8389	0.7532	0.8912	0.9054
	VQM	0.6593	**0.9324**	0.8383	0.7483	0.8472	0.8536	0.7309	0.9311	0.8395	0.7397	0.8951	0.9188
	SSIM	0.5697	0.7952	0.7042	0.5385	0.5922	0.5890	0.7263	**0.9471**	0.8448	0.7555	0.8996	0.9204
	M-SSSIM	0.5652	0.7764	0.6965	0.5215	0.5839	0.5775	0.7185	0.9394	0.8358	0.7485	0.8914	0.9124
	VIFP	0.6390	**0.9263**	0.8312	0.7005	0.8390	0.8354	0.7176	0.9243	0.8385	0.7052	0.8924	0.9267
	FSIM	0.5244	0.7393	0.6644	0.4743	0.5610	0.5589	0.7461	**0.9596**	0.8628	0.7707	0.8973	0.9121
	GMSD	0.6446	0.9210	0.8299	0.7005	0.7694	0.7732	0.7205	0.9231	0.8249	0.7249	0.8786	0.9095

Criteria	Video Model	SVM (quality score)						SVM (quality feature)					
		PEAQ	STOI	VISQOL	LLR	SNR	segSNR	PEAQ	STOI	VISQOL	LLR	SNR	segSNR
SRCC	VMAF	0.8223	**0.9471**	0.8308	0.7626	0.9105	0.9273	0.9102	0.9507	**0.9590**	0.7319	0.9112	0.9317
	STRRED	0.7924	0.9014	0.7854	0.7181	0.8645	0.8827	0.8805	0.9270	0.9369	0.7585	0.8810	0.9040
	SpEED	0.7804	0.8935	0.7810	0.7108	0.8530	0.8726	0.8666	0.9271	0.9295	0.7558	0.8831	0.9038
	VQM	0.7856	0.9230	0.8091	0.7451	0.8766	0.8953	0.8872	0.9442	0.9522	0.7571	0.9131	0.9325
	SSIM	0.8163	0.9291	0.8064	0.7270	0.8876	0.9053	0.8517	0.9386	0.9464	0.7561	0.9051	0.9187
	MS-SSIM	0.8007	0.9175	0.7994	0.7074	0.8822	0.8978	0.8620	0.9159	0.9200	0.7377	0.8972	0.9165
	VIFP	0.8001	0.9287	0.8113	0.7103	0.9068	0.9165	0.8674	0.9552	**0.9572**	0.7683	0.9222	0.9415
	FSIM	0.8358	**0.9550**	0.8352	0.7664	0.9191	**0.9346**	0.8670	0.9543	**0.9574**	0.7674	0.9261	0.9415
	GMSD	0.7850	0.9092	0.7905	0.7161	0.8662	0.8834	0.8105	0.9015	0.9106	0.7055	0.8574	0.8774
PLCC	VMAF	0.8233	**0.9629**	0.8507	0.7564	0.9257	0.9374	0.9277	0.9664	**0.9752**	0.7368	0.9247	0.9384
	STRRED	0.7955	0.9179	0.8078	0.7263	0.8802	0.8868	0.8994	0.9456	0.9555	0.7533	0.8988	0.9149
	SpEED	0.7834	0.9084	0.7990	0.7185	0.8696	0.8774	0.8855	0.9446	0.9490	0.7488	0.9017	0.9153
	VQM	0.7867	0.9455	0.8348	0.7394	0.9026	0.9137	0.9164	0.9624	0.9685	0.7577	0.9275	0.9389
	SSIM	0.8208	**0.9490**	0.8349	0.7306	0.9053	0.9147	0.8670	0.9586	0.9641	0.7529	0.9212	0.9281
	MS-SSIM	0.8070	0.9391	0.8272	0.7137	0.9000	0.9078	0.8748	0.9445	0.9456	0.7406	0.9196	0.9303
	VIFP	0.7942	0.9446	0.8298	0.7044	0.9201	0.9251	0.8875	0.9679	**0.9715**	0.7613	0.9335	0.9463
	FSIM	0.8352	**0.9690**	0.8574	0.7615	0.9324	0.9426	0.8759	0.9683	**0.9713**	0.7620	0.9365	0.9458
	GMSD	0.7901	0.9327	0.8178	0.7129	0.8875	0.8985	0.8495	0.9282	0.9338	0.7086	0.8829	0.8949

son of all models, we randomly split the LIVE-SJTU A/V-QA Database into a training set of 80% of the A/V sequences and a testing subset with the remaining 20% of the A/V sequences. All of the distorted A/V sequences arising from a same original content were placed into the same subset to ensure a complete content separation between training and testing data. The training based models were trained on the training subset, and tested on the testing subset. Models that were not trained were tested on the same (20%) test subset. This process was repeated over 1,000 random train-test divisions when evaluating the Type 1, 2, and 3 models, as shown in Tables III and IV. For Type 4 models (Table V), only 100 of the 1,000 random splits were used because of the implied large computation. The tables report the mean SRCC and PLCC achieved by each model on the LIVE-SJTU A/V-QA Database over all train-test splits. For the weighted product models of Types 1, 2, and 3, we varied the weight from 0 to 1 using a step increment of 0.05, found the weight that generated the highest SRCC on the training set, then tested the model with the optimal weight on the test set.

B. Evaluation of Type 1 *Models*

1) Performance Evaluation: We tested two variants of Type 1: Type 1(a) (product) and Type 2(b) (weighted product) models. A total of 9 (video models) × 6 (audio models) × 2 (product forms) = 108 models were tested. To normalize the component quality models, the following empirically determined normalization functions were used: $Q'_{\mathrm{VMAF}} = Q_{\mathrm{VMAF}}/100$, $Q'_{\mathrm{VQM}} = 1 - Q_{\mathrm{VQM}}/1.01$, $Q'_{\mathrm{STRRED}} = 1 - Q_{\mathrm{STRRED}}/1500$, $Q'_{\mathrm{SpEED}} = 1 - Q_{\mathrm{SpEED}}/4600$, $Q'_{\mathrm{GMSD}} = 1 - Q_{\mathrm{GMSD}}/0.25$, $Q'_{\mathrm{PEAQ}} = 1 + Q_{\mathrm{PEAQ}}/3.5$, $Q'_{\mathrm{LLR}} = 1 - (Q_{\mathrm{LLR}} - 1.1)/(1.5 - 1.1)$, $Q'_{\mathrm{SNR}} = Q_{\mathrm{SNR}}/35$, $Q'_{\mathrm{segSNR}} = (Q_{\mathrm{segSNR}} + 1)/(30 + 1)$. Since SSIM, MS-SSIM, VIFP, FSIM, STOI and VISQOL are already bounded on [0, 1], no further normalization was required. The performances of the tested Type 1 models are summarized in the top half of Table III. Among the simple product based models, the models defined as products between VQA algorithms VMAF, VQM, and VIFP and the AQA algorithms STOI, VISQOL, SNR, and segSNR yielded relatively good performances. Among the weighted product models, the A/V-QA performance differences obtained

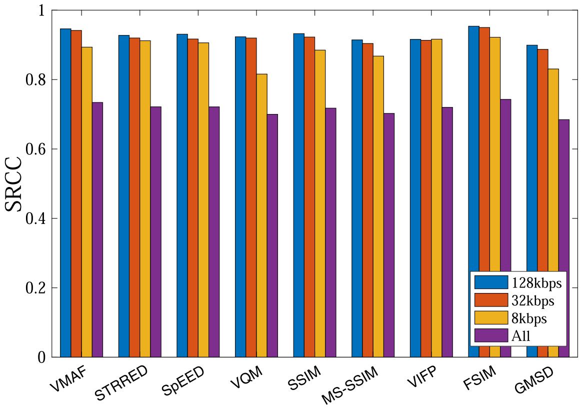


Fig. 11. Performances of VQA models on overall A/V quality prediction.

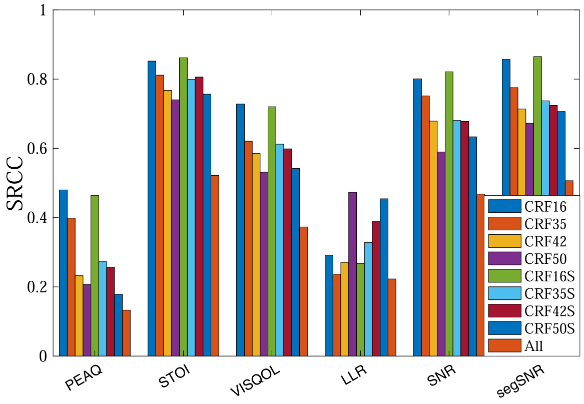


Fig. 12. Performances of AQA models on overall A/V quality prediction.

using different VQA components narrowed, while the choice of AQA component had increased impact. More particularly, choosing STOI, SNR, or segSNR was advantageous. Moreover, most Type 1(b) models were better than the Type 1(a) models, hence, relatively weighting the AQA and VQA components was advantageous.

2) Analysis of Single-Mode Quality Models: From the above performance analysis, it may be observed that the effectiveness of the fused A/V quality models depends heavily on the performances of each of the single-mode component models. Thus it is interesting to analyze the individual effectiveness of each of the single-mode quality prediction components. Hence, we evaluated the component VQA models on all videos having the same audio distortion condition, e.g. 128 kbps, 32 kbps or 8 kbps of audio compression. We also evaluated their effectiveness on the entire database. We evaluated the component AQA models in a similar way. The distortion conditions include CRF16, CRF35, CRF42, CRF50, CRF16S, CRF35S, CRF42S, CRF50S, or the overall database. Here the suffix 'S' indicates compression plus scaling distortion. All of these distortion conditions were described earlier, in Section II-A.

The model performances are illustrated in Figs. 11 and 12, from which we make some useful observations. Most of the VQA models performed at very similar levels, and all of them were able to predict A/V quality effectively when the audio distortion level was fixed. Among the AQA models, STOI, VISQOL, SNR, and segSNR were more effective, but their effectiveness for A/V quality prediction when the video distortion level was fixed was worse than that of the video models just described. When testing on the entire database (nothing held fixed), none of the single-mode measures were effective enough, but the video models were still better at predicting A/V quality than the audio models (SRCC of about 0.7 vs. 0.5). This is likely true in part because the video modality tend to dominate perceived A/V quality. Further, current VQA models may be more well-developed than AQA models.

C. Evaluation of Type 2 Models

1) Performance Evaluation: Two sub-types of this family of models were tested: Type 2(a) quality score driven, Type 2(b) quality feature driven SVR fusion. A total of 9 (video models) × 6 (audio models) × 2 (SVR forms) = 108 models were tested. For the Type 2 models, the normalization process was left to the SVR. The performances of the Type 2 models are summarized in the bottom half of Table III. The performances of the Type 2(a) models showed some similarities to the Type 1(b) models: STOI, SNR, and segSNR yielded advantageous performances, while the differences in performance allowed by the different VQA models was small. For Type 2(b) models, the performance differences between different VQA models were also not large, but all of the AQA models (except for LLR) were able to predict A/V quality effectively when combined with VQA models.

2) Influences of SVR and Feature Decomposition: It is also interesting to study whether replacing the weighted product with SVR fusion is more effective, and also whether decomposing the constituent AQA and/or VQA model into quality-aware features during SVR fusion can contribute to A/V quality prediction. The first and second questions can be answered by comparing the performances of the Type 1(b) models against Type 2(a) models, and Type 2(a) models against Type 2(b) models, respectively. We first calculated the performance improvement (in percentage) afforded by each combination model, then averaged it over all models of each modality and over all evaluation criteria, finally arriving at a measurement of the performance improvement obtained by each single-mode quality model. The amount of improvement obtained by each unimodal model is illustrated in Fig. 13.

The performance improvement gained by all models by replacing the weighted product with an SVR was limited, except for PEAQ, which suggests that the weighted product is generally an effective fusion device. To make use of the power of the learner, it may be more efficient to decompose the component VQA and AQA models into their constituent features, as appropriate. The efficacy of such a feature decomposition is made evident by comparing the Type 2(a) and Type 2(b) models. A degree of improvement was obtained in almost all cases. Among models which are not easy to be decomposed, e.g. STOI, LLR, SNR, and segSNR, the improvement was smaller. For models like PEAQ and VISQOL, the improvements were larger.

TABLE IV

PERFORMANCES OF TYPE 3 A/V QUALITY MODELS. THE TOP 3 MODELS OF EACH SUB-TYPE ARE IN BOLD

Criteria	Video Model	Weighted Product					SVM (quality score)					SVM (quality feature)			
		SSIM_{1D}	MS-SSIM_{1D}	VIFP_{1D}	GMSM_{1D}	GMSD_{1D}	SSIM_{1D}	MS-SSIM_{1D}	VIFP_{1D}	GMSM_{1D}	GMSD_{1D}	SSIM_{1D}	MS-SSIM_{1D}	VIFP_{1D}	GMSM_{1D}
SRCC	VMAF	0.9239	0.9285	0.9059	0.9439	**0.9463**	0.9408	0.9459	0.9447	0.9463	0.9483	0.9496	0.9555	**0.9603**	0.9482
	STRRED	0.8815	0.8928	0.8810	0.9035	0.9048	0.8955	0.8985	0.8954	0.9029	0.9050	0.9275	0.9287	0.9284	0.9268
	SpEED	0.8860	0.8881	0.8819	0.9033	0.9098	0.8869	0.8905	0.8882	0.8952	0.8980	0.9311	0.9292	0.9269	0.9355
	VQM	0.9006	0.9052	0.8906	0.9160	0.9166	0.9119	0.9191	0.9173	0.9159	0.9203	0.9450	0.9481	0.9446	0.9507
	SSIM	0.9077	0.8983	0.9002	0.9244	0.9301	0.9218	0.9252	0.9257	0.9318	0.9359	0.9371	0.9361	0.9367	0.9414
	MS-SSIM	0.9030	0.8925	0.8943	0.9204	0.9228	0.9138	0.9168	0.9158	0.9247	0.9301	0.9210	0.9216	0.9199	0.9232
	VIFP	0.9103	0.9096	0.8822	0.9386	0.9384	0.9262	0.9282	0.9265	0.9382	0.9427	0.9589	**0.9608**	0.9519	0.9589
	FSIM	0.9196	0.9306	0.9102	**0.9465**	**0.9541**	0.9467	**0.9511**	0.9497	**0.9530**	**0.9552**	0.9555	**0.9596**	0.9540	0.9567
	GMSD	0.8858	0.8895	0.8756	0.9109	0.9151	0.9012	0.9072	0.9061	0.9124	0.9183	0.8987	0.8911	0.9014	0.9065
PLCC	VMAF	0.9464	0.9528	0.9372	0.9534	0.9558	0.9566	0.9623	0.9600	0.9588	0.9594	0.9636	0.9678	**0.9722**	0.9593
	STRRED	0.9110	0.9203	0.9133	0.9247	0.9205	0.9155	0.9162	0.9116	0.9177	0.9198	0.9439	0.9490	0.9473	0.9407
	SpEED	0.9122	0.9173	0.9125	0.9251	0.9294	0.9079	0.9082	0.9046	0.9089	0.9118	0.9451	0.9469	0.9433	0.9461
	VQM	0.9289	0.9348	0.9252	0.9308	0.9315	0.9374	0.9441	0.9401	0.9372	0.9392	0.9595	0.9645	0.9617	0.9613
	SSIM	0.9353	0.9296	0.9311	0.9429	0.9430	0.9419	0.9453	0.9443	0.9459	0.9487	0.9547	0.9578	0.9563	0.9565
	MS-SSIM	0.9306	0.9211	0.9275	0.9382	0.9360	0.9351	0.9375	0.9353	0.9389	0.9421	0.9438	0.9475	0.9442	0.9443
	VIFP	0.9258	0.9266	0.9010	0.9458	0.9424	0.9409	0.9421	0.9396	0.9488	0.9507	0.9673	**0.9707**	0.9645	0.9665
	FSIM	0.9485	**0.9597**	0.9487	**0.9626**	**0.9662**	0.9609	**0.9661**	0.9636	**0.9645**	**0.9660**	0.9652	**0.9711**	0.9673	0.9659
	GMSD	0.9141	0.9186	0.9114	0.9275	0.9300	0.9263	0.9315	0.9282	0.9313	0.9334	0.9264	0.9241	0.9286	0.9259

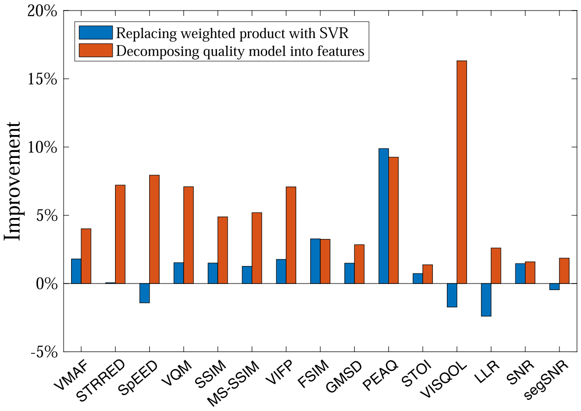


Fig. 13. Performance improvements introduced by replacing the weighted product with a SVR (by comparing the performances of Type 1(b) and Type 2(a) models), and decomposing quality models into features during SVR fusion (by comparing the performances of Type 2(a) and Type 2(b) models).

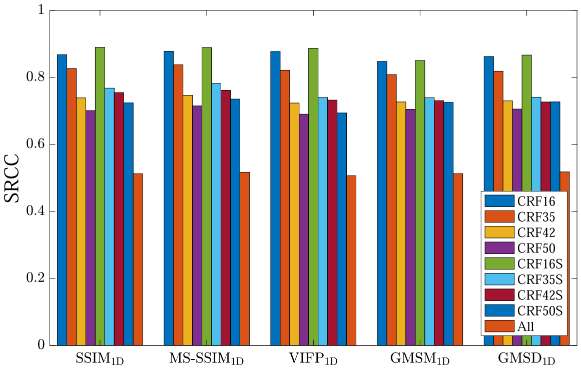


Fig. 14. Performances of the redimensioned VQA models repurposed as AQA models on overall A/V quality prediction.

D. Evaluation of Type 3 Models

1) Performance Evaluation: Type 3 models replace the component AQA models with redimensioned and repurposed VQA models. Here three fusion variants were tested: Type 3(a) weighted product, Type 3(b) quality score/SVR based, and Type 3(c) quality feature/SVR based. A total of 9 (video models) × (5 + 5 + 4) (audio models and fusion forms) = 126 models were tested. In the Type 3(c) models, GMSD_{1D} was taken as a feature of GMSM_{1D}. Normalization for all VQA models was conducted in the same way as for Type 1 models. Among the redimensioned VQA models being used as AQA models, only GMSD_{1D} requires normalization: $Q'_{\text{GMSD}_{1D}} = 1 - Q_{\text{GMSD}_{1D}}/0.4$. The performances of the Type 3 models are summarized in Table IV. Most of these models achieved SRCC or PLCC performances better than 0.9, while the remaining models obtained performances very close to 0.9. The performance differences of using different VQA models were not large, though several models yielded slightly better performances. For example, VMAF, FSIM, or VIFP fused with MS-SSIM_{1D}, GMSM_{1D}, GMSD_{1D} all performed better. Among the fusion functions, Type 3(c) was better than Type 3(b), while Type 3(b) was better than Type 3(a), although the performance differences were not large. Specifically, AVSSIM, AVMSSSIM, AVIFP, AVGMSM, and AVGMSD were able to achieve state-of-the-art performances, while AVMAF is one of the best-performing models overall.

2) Comparison of Redimensioned and Repurposed 1D VQA Models Against True AQA Models: The models in Tables III and IV were evaluated under the same settings, thus the performances in these two tables are directly comparable. Comparing the Type 3(a), Type 3(b), and Type 3(c) models against the Type 1(b), Type 2(a), and Type 2(b) models respectively, highlights the successes of the Type 3 models. The average performances of the Type 3 models were noticeably better than those of Type 2 and Type 1 models. Among the best performing models of each sub-type, the best-performing Type 3 models were slightly better. Similar to Section IV-B.2

TABLE V
PERFORMANCES OF TYPE 4 A/V QUALITY MODELS HAVING DIFFERENT SETTINGS. THE BEST ONE IS IN BOLD

Final PCA	2×2048 features		4×2048 features		6×2048 features	
Dimension	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
5	0.9186	0.9334	0.5710	0.6188	0.6179	0.6452
25	0.9490	0.9636	0.9345	0.9472	**0.9536**	**0.9627**
125	0.9298	0.9518	0.9156	0.9262	0.9334	0.9481

and Fig. 12, we also evaluated the redimensioned single-mode AQA models by fixing the video distortion conditions. The resulting performance values are shown in Fig. 14. By comparing Fig. 14 with Fig. 12, it may be observed that all of the redimensioned 1D VQA models that were repurposed as AQA models performed at a level comparable to the best existing AQA models. This result is both remarkable and provocative, given the relative simplicity and easy availability of the VQA models used.

E. Evaluation of Type 4 *Models*

1) Performance Evaluation: The family of Type 4 models are all frame based, meaning that feature extraction and fusion are conducted at a frame level. During SVR training, we labeled each frame with the MOS of the corresponding A/V sequence, and used each frame as a training instance to enrich the training data. During testing, we used the model to predict single frame qualities, which were then averaged to predict the video quality. Performances of the Type 4 models are listed in Table V, where the best performance (in bold) denotes the model using the final settings. We used only the first 100 of the same 1,000 random splits that were used for the Type 1, 2, and 3 models when training and testing the Type 4 models. By comparing Table V with Tables III and IV, it may be observed that the Type 4 model was comparable to the best-performing Type 1, 2, and 3 models. Note that we only used the pretrained DNN to extract content-aware quality features. The DNN was not fine-tuned or retrained. Given significantly more subjective data, it may be worthwhile to embark upon a larger scale study using end-to-end or retrained DNNs.

2) Influences of PCA Dimension and Feature Settings: The key settings of the Type 4 model include the PCA dimension and the feature setting. The first one controls how many dimensions of features are used following PCA dimension reduction. We tested it three settings: 5, 25, and 125. The feature setting dictates how many groups of features are input to the PCA module. As described in Section III-D, four groups of 2048-dimensional features were extracted from the reference and distorted video and audio sequences by the DNN. Two groups of features were derived by calculating feature differences between the reference and distorted signals. These six groups of features are all used by the Type 4 model.

We also test the Type 4 model under the two additional settings: first, using the two groups of features that include only feature differences, and second, using the four groups of features that include only the raw DNN features extracted from the reference and distorted signals. The performances of the Type 4 models under all settings are summarized in Table V, from which it may be observed that a moderately large feature dimension is helpful to the model, as is preprocessing of the raw features.

V. RECOMMENDATIONS ON PRACTICAL A/V-QA MODEL USAGE AND DEPLOYMENT

From the evaluation results given above, we may observe that regardless of which family of models is used, some of the member A/V-QA models achieved accurate performances (SRCC of 0.9+). This suggests that predicting overall A/V quality is a problem that may be successfully addressed, provided that accurate predictions of the corresponding video and audio components can be obtained. It should be possible to augment existing video quality prediction models that are deployed in practical systems, by adding suitable audio quality prediction models and multimodal quality fusion modules.

When deploying the proposed AV-QA models in practice, one may need to choose the component AQA/VQA models and the fusion schemes, that is, choosing one from all of the proposed 4 families and hundreds of A/V-QA models. For the component AQA models, we suggest using one of the redimensioned VQA models which are repurposed as AQA models, since any of the SSIM_{1D}, $\text{MS-SSIM}_{\text{1D}}$, VIFP_{1D}, GMSM_{1D}, and GMSD_{1D} can be efficiently combined with the current VQA models and achieve the state-of-the-art performances. With regards to the component VQA models, from the experimental results we can observe that all of the tested VQA models are pretty effective, and the performance differences of using different VQA models are not large.

Among the fusion methods, product fusion has the advantages of simplicity and easy interpretability, and it performed well. A weight for the product improves performance, and the final A/V-QA model is generally stable within certain ranges of the weight. To achieve the best performance, it can be tuned to a use case. Among the weighted product based models, ***AVSSIM***, ***AVMSSSIM***, ***AVIFP***, ***AVGMSM***, and ***AVGMSD*** are recommended, since the qualities of both modalities are estimated using the same methodology. If performance is a critical criterion, then quality feature based SVR fusion is advisable. Specifically, ***AVMAF*** is a good choice, since VMAF is a top VQA model, and they perform well together. AVMAF is one of the best-performing A/V-QA models. The DNN based A/V-QA approach is worth further study, although the current model is somewhat heavy and does not yet give a performance advantage. Likely, larger A/V subjective datasets are needed.

VI. CONCLUSION

We conducted an in-depth exploration of the problem of assessing the quality of A/V signals. Specifically, we constructed a sizable and unique resource: the LIVE-SJTU A/V-QA Database, which includes several hundred A/V sequences processed by distortions representative of those encountered in the streaming space. A subjective A/V-QA study was then conducted to obtain ground-truth quality ratings of all of the distorted A/V sequences included in the database. The collected subjective rating data suggest that while the video modality is generally more important in forming subjective impressions than the audio modality, audio quality is an important contributor to overall QoE.

We also designed four families of objective A/V quality prediction models which fuse single-mode quality predictors or quality-aware features. All four families of proposed A/V-QA models delivered promising results on the LIVE-SJTU A/V-QA Database. The fact that we were able to obtain good prediction performances using fusion models ranging from very simple to somewhat sophisticated, suggests that existing video quality prediction systems for streaming control might be easily and effectively augmented by fusion with audio quality modules.

REFERENCES

[1] (Oct. 2018). *The Global Internet Phenomena Report*. [Online]. Available: https://www.sandvine.com/hubfs/downloads/phenomena/2018-phenomena-report.pdf

[2] A. C. Bovik, "Automatic prediction of perceptual image and video quality," *Proc. IEEE*, vol. 101, no. 9, pp. 2008–2024, Jul. 2013.

[3] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.

[4] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, and T. Ebrahimi, "A H. 264/AVC video database for the evaluation of quality metrics," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 2430–2433.

[5] *ITU-T Coded-Speech Database*, International Telecommunication Union, document ITU-T Rec. P 23, 1998.

[6] C. G. Bampis, Z. Li, A. K. Moorthy, I. Katsavounidis, A. Aaron, and A. C. Bovik, "Study of temporal effects on subjective video quality of experience," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5217–5231, Nov. 2017.

[7] C. G. Bampis, Z. Li, I. Katsavounidis, and A. C. Bovik, "Recurrent and dynamic models for predicting streaming video quality of experience," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3316–3331, Jul. 2018.

[8] 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, Apr. 2004.

[9] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, 2003, pp. 1398–1402.

[10] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[11] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[12] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most-apparent-distortion model for video quality assessment," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 2505–2508.

[13] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 684–694, Apr. 2013.

[14] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," The Netflix Tech Blog, Tech. Rep., 2016.

[15] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, and C. W. Chen, "Blind quality assessment based on pseudo-reference image," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2049–2062, Aug. 2018.

[16] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind image quality estimation via distortion aggravation," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 508–517, Jun. 2018.

[17] T. Thiede *et al.*, "PEAQ-the ITU standard for objective measurement of perceived audio quality," *J. Audio Eng. Soc.*, vol. 48, nos. 1–2, pp. 3–29, Feb. 2000.

[18] *Perceptual Objective Listening Quality Assessment (POLQA)*, document ITU-T Rec. P 863, International Telecommunication Union, 2011.

[19] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Feb. 2011.

[20] A. Hines, E. Gillen, D. Kelly, J. Skoglund, A. Kokaram, and N. Harte, "ViSQOLAudio: An objective audio quality metric for low bitrate codecs," *J. Acoust. Soc. Amer.*, vol. 137, no. 6, pp. EL449–EL455, Jun. 2015.

[21] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[22] J. H. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. Int. Conf. Spoken Lang. Process.*, vol. 7, 1998, pp. 2819–2822.

[23] W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey," *J. Vis. Commun. Image Represent.*, vol. 22, no. 4, pp. 297–312, May 2011.

[24] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 165–182, Jun. 2011.

[25] A. W. Rix, J. G. Beerends, D.-S. Kim, P. Kroon, and O. Ghitza, "Objective assessment of speech and audio quality—Technology and applications," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 1890–1901, Nov. 2006.

[26] D. S. Hands, "A basic multimedia quality model," *IEEE Trans. Multimedia*, vol. 6, no. 6, pp. 806–816, Dec. 2004.

[27] S. Winkler and C. Faller, "Perceived audiovisual quality of low-bitrate multimedia content," *IEEE Trans. Multimedia*, vol. 8, no. 5, pp. 973–980, Oct. 2006.

[28] M. H. Pinson *et al.*, "The influence of subjects and environment on audiovisual subjective tests: An international study," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 640–651, Oct. 2012.

[29] H. B. Martinez and M. C. Farias, "Full-reference audio-visual video quality metric," *J. Electron. Imag.*, vol. 23, no. 6, Sep. 2014, Art. no. 061108.

[30] H. A. B. Martinez and M. C. Q. Farias, "Combining audio and video metrics to assess audio-visual quality," *Multimedia Tools Appl.*, vol. 77, no. 18, pp. 23993–24012, Sep. 2018.

[31] H. B. Martinez and M. C. Farias, "A no-reference audio-visual video quality metric," in *Proc. Eur. Signal Process. Conf.*, 2014, pp. 2125–2129.

[32] Z. Akhtar and T. H. Falk, "Audio-visual multimedia quality assessment: A comprehensive survey," *IEEE Access*, vol. 5, pp. 21090–21117, 2017.

[33] J. You, U. Reiter, M. M. Hannuksela, M. Gabbouj, and A. Perkis, "Perceptual-based quality assessment for audio-visual services: A survey," *Signal Process., Image Commun.*, vol. 25, no. 7, pp. 482–501, Aug. 2010.

[34] *Engineering a Studio Quality Experience With High-Quality Audio at Netflix*. Accessed: Sep. 24, 2019. [Online]. Available: https://medium.com/netflixtechblog/engineering-a-studio-quality-experience-with-high-qualityaudio-at-netflix-eaa0b6145f32

[35] *The Consumer Digital Video Library*. Accessed: Apr. 20, 2019. [Online]. Available: https://cdvl.org/

[36] K. Turkowski, "Filters for common resampling tasks," in *Graphics Gems*. New York, NY, USA: Academic, 1990, pp. 147–165.

[37] *Snellen Chart*. Accessed: Mar. 2, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Snellen_chart

[38] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document Rec. ITU-R BT.500-13, Jan. 2012.

[39] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, "SpEED-QA: Spatial efficient entropic differencing for image and video quality," *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1333–1337, Sep. 2017.

[40] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Sep. 2004.

[41] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[42] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2014.

[43] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, Apr. 2011.

[44] D. Kahneman, "Method, findings, and theory in studies of visual masking," *Psychol. Bull.*, vol. 70, no. 6, pp. 404–425, 1968.

[45] G. E. Legge and J. M. Foley, "Contrast masking in human vision," *J. Opt. Soc. Amer.*, vol. 70, no. 12, pp. 1458–1471, Dec. 1980.

[46] G. A. Miller, "Sensitivity to changes in the intensity of white noise and its relation to masking and loudness," *J. Acoust. Soc. Amer.*, vol. 19, no. 4, pp. 609–619, Jul. 1947.

[47] D. D. Greenwood, "Auditory masking and the critical band," *J. Acoust. Soc. Amer.*, vol. 33, no. 4, pp. 484–502, Apr. 1961.

[48] S. Kandadai, J. Hardin, and C. D. Creusere, "Audio quality assessment using the mean structural similarity measure," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 221–224.

[49] 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, Dec. 2017.

[50] B. M. H. Romeny, *Front-End Vision and Multi-Scale Image Analysis: Multi-scale Computer Vision Theory and Applications, Written in Mathematica*, vol. 27. Dordrecht, The Netherlands: Springer, 2008.

[51] S. Dieleman and B. Schrauwen, "Multiscale approaches to music audio feature learning," in *Proc. Int. Soc. Music Inform. Retr. Conf.*, 2013, pp. 116–121.

[52] O. Schwartz and E. P. Simoncelli, "Natural signal statistics and sensory gain control," *Nature Neurosci.*, vol. 4, no. 8, pp. 819–825, Aug. 2001.

[53] H. Attias and C. E. Schreiner, "Temporal low-order statistics of natural sounds," in *Proc. Adv. Neural Inf. Process. Syst.*, 1997, pp. 27–33.

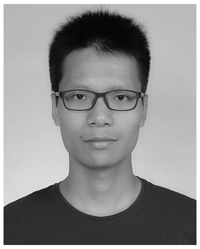
[54] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[55] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 131–135.

[56] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1096–1104.

[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[58] Video Quality Experts Group, "Final report from the video quality experts group on the validation of objective models of video quality assessment," in *Proc. VQEG Meeting*, Ottawa, ON, Canada, 2000.

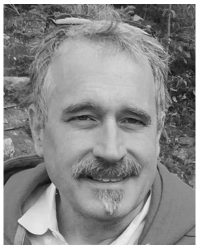


Xiongkuo Min (Member, IEEE) received the B.E. degree from Wuhan University, Wuhan, China, in 2013, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2018. From 2016 to 2017, he was a Visiting Student with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. He is currently a Postdoctoral Fellow with Shanghai Jiao Tong University. His research interests include visual quality assessment, visual attention modeling, and perceptual signal processing. He received the Best Student Paper Award from the IEEE ICME 2016.

Guangtao Zhai (Senior Member, IEEE) received the B.E. and M.E. degrees from Shandong University, Shandong, China, in 2001 and 2004, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2009, where he is currently a Research Professor with the Institute of Image Communication and Information Processing. From 2008 to 2009, he was a Visiting Student with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, where he was a Postdoctoral Fellow from 2010 to 2012. From 2012 to 2013, he was a Humboldt Research Fellow with the Institute of Multimedia Communication and Signal Processing, Friedrich Alexander University of Erlangen-Nuremberg, Germany. He received the Award of National Excellent Ph.D. Thesis from the Ministry of Education of China in 2012. His research interests include multimedia signal processing and perceptual signal processing.

Jiantao Zhou (Senior Member, IEEE) received the B.Eng. degree from the Department of Electronic Engineering, Dalian University of Technology, in 2002, the M.Phil. degree from the Department of Radio Engineering, Southeast University, in 2005, and the Ph.D. degree from the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, in 2009. He held various research positions with the University of Illinois at Urbana-Champaign, The Hong Kong University of Science and Technology, and the McMaster University. He is currently an Associate Professor with the Department of Computer and Information Science, Faculty of Science and Technology, University of Macau. His research interests include multimedia security and forensics, multimedia signal processing, artificial intelligence and big data. He holds four granted U.S. patents and two granted Chinese patents. He has coauthored two articles that received the Best Paper Award at the IEEE Pacific-Rim Conference on Multimedia in 2007 and the Best Student Paper Award at the IEEE International Conference on Multimedia and Expo in 2016. He is an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING.

Mylène C. Q. Farias (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the Federal University of Pernambuco (UFPE), Brazil, in 1995, the M.Sc. degree in electrical engineering from the State University of Campinas (UNICAMP), Brazil, in 1998, and the Ph.D. degree in electrical and computer engineering from the University of California Santa Barbara (UCSB), USA, in 2004, for her work in no-reference video quality metrics. She has worked as a Research Engineer at CPqD (Brazil) in video quality assessment and validation of video quality metrics. She has also worked as an intern for Philips Research Laboratories, The Netherlands in video quality assessment of sharpness algorithms, and for Intel Corporation, Phoenix, USA, in developing no-reference video quality metrics. She is currently an Associate Professor with the Department of Electrical Engineering, University of Brasilia (UnB). She has published over 120 scientific articles. Her current interests include quality of experience and bio-inspired processing of images and videos 2D, 3D, HDR, and 360. She is a member of the IEEE Signal Processing Society, ACM, and SPIE. She served as a TPC Co-Chair for the IEEE QoMEX 2020 and EI Image Quality and System Performance 2019-2020. She serves as an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS and a regular reviewer for several international conferences and journals.



Alan Conrad Bovik (Fellow, IEEE) is the Cockrell Family Regents Endowed Chair Professor with The University of Texas at Austin. His research interests include image processing, digital photography, digital television, digital streaming video, and visual perception. For his work in these areas he has been the recipient of the 2019 Progress Medal from The Royal Photographic Society, the 2019 IEEE Fourier Award, the 2017 Edwin H. Land Medal from the Optical Society of America, the 2015 Primetime Emmy Award for Outstanding Achievement in Engineering Development from the Television Academy, and the Norbert Wiener Society Award and the Karl Friedrich Gauss Education Award from the IEEE Signal Processing Society. A perennial Web of Science Group Highly-Cited Researcher, he has also received about ten best journal paper awards, including the 2016 IEEE Signal Processing Society Sustained Impact Award. His books include The Essential Guides to Image and Video Processing. He co-founded and was longest-serving Editor-in-Chief of the IEEE TRANSACTIONS ON IMAGE PROCESSING and also created/Chaired the IEEE International Conference on Image Processing which was first held in Austin, Texas, 1994.