QoE Estimation of WebRTC-based Audio-visual Conversations from Facial and Speech Features

GÜLNAZIYE BINGÖL, SIMONE PORCU, ALESSANDRO FLORIS, and LUIGI ATZORI, DIEE, University of Cagliari, Italy and CNIT, University of Cagliari, Italy

The utilization of user's facial- and speech-related features for the estimation of the Quality of Experience (QoE) of multimedia services is still underinvestigated despite its potential. Currently, only the use of either facial or speech features individually has been proposed, and relevant limited experiments have been performed. To advance in this respect, in this study, we focused on WebRTC-based videoconferencing, where it is often possible to capture both the facial expressions and vocal speech characteristics of the users. First, we performed thorough statistical analysis to identify the most significant facial- and speech-related features for QoE estimation, which we extracted from the participants' audio-video data collected during a subjective assessment. Second, we trained individual QoE estimation machine learning-based models on the separated facial and speech datasets. Finally, we employed data fusion techniques to combine the facial and speech datasets into a single dataset to enhance the QoE estimation performance due to the integrated knowledge provided by the fusion of facial and speech features. The obtained results demonstrate that the data fusion technique based on the Improved Centered Kernel Alignment (ICKA) allows for reaching a mean QoE estimation accuracy of 0.93, whereas the values of 0.78 and 0.86 are reached when using only facial or speech features, respectively.

CCS Concepts: • **Information systems → Multimedia streaming;**

Additional Key Words and Phrases: Quality of Experience, WebRTC, Facial Expressions, Speech, Machine Learning, Data Fusion.

ACM Reference format:
Gülnaziye Bingöl, Simone Porcu, Alessandro Floris, and Luigi Atzori. 2024. QoE Estimation of WebRTC-based Audio-visual Conversations from Facial and Speech Features. *ACM Trans. Multimedia Comput. Commun. Appl.* 20, 5, Article 130 (January 2024), 23 pages.
https://doi.org/10.1145/3638251

1 INTRODUCTION

Internet usage has experienced a significant surge in recent years. This upward trend can be primarily attributed to the rising prominence of video streaming platforms (e.g., YouTube, Amazon Prime Video, Netflix) and videoconferencing platforms (across various domains, such

This work has been partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, "Sustainable Mobility Center" Centro Nazionale per la Mobilit Sostenibile, CNMS, CN_00000023, and by the PON "Ricerca e Innovazione" 2014-2020 (PON R&I) "Azione IV.4 Dottorati e contratti di ricerca su tematiche dell" innovazione" with D.M. 1062 on 10.08.2021.
Author's address: G. Bingöl, S. Porcu, A. Floris, L. Atzori, DIEE, University of Cagliari, Cagliari, Italy, 09123 and CNIT, University of Cagliari, Cagliari, Italy, 09123; e-mails: gulnaziye.bingol@unica.it, simone.porcu@unica.it, alessandro.floris84@unica.it, l.atzori@unica.it.

1551-6857/2024/01-ART130 $15.00
https://doi.org/10.1145/3638251

as professional meetings, online education, and social engagements), which have led video traffic to account for almost 66% of current global Internet traffic [44]. The management of such high multimedia traffic, combined with the high-quality expectations of users and the growing level of applications' interactivity and functional capabilities, has made the integration of user-centered approaches increasingly important for application and network providers. Resources in the network and at the server side need to be promptly allocated to the required multimedia sessions to assure that the target quality levels are reached to not compromise the user experience. In this regard, the assessment of the **Quality of Experience (QoE)** plays a crucial role, since it aims at quantifying *the degree of delight or annoyance of the user of an application or service* [32].

Recent studies concerning the subjective assessment of multimedia services have focused on defining personalized prediction systems to consider users' individual and personal differences [45, 50, 57]. Indeed, different users may have different reactions to the same stimuli, since the QoE depends upon many factors, mainly system-, human-, and context-related [32]. In particular, the human influence factors, classified as objective factors (e.g., demographics, cultural background) and subjective factors (user states, such as enjoyment and motivation), are the main elements of difference among users [47]. Personalized QoE models have the advantage of driving service management procedures to be better fitted to each end-user's profile with the potential to increase service success [61].

To build personalized models, extensive subjective studies are required that consider human influencing factors, which, however, are affected by relevant limitations: (i) the need to ask for explicit feedback from the user; (ii) the rating process may be influenced by the rating scale, which may not reflect well the user's internal perception of quality; (iii) it is time-consuming and money-consuming; (iv) it is not suitable for real-time management systems [37]. Accordingly, alternative unobtrusive (i.e., that do not require user feedback) methods to assess the individual QoE have emerged, which rely on the physiological bases of perceptual and cognitive processes, such as electroencephalography, [35] heart rate, and electrodermal activity measurements [15, 18]. Among these techniques, facial expressions and vocal speech characteristics have gathered particular attention in recent years, because they naturally convey human emotions driven by the user's emotional state [27, 33]. Although literature studies are limited in this regard, the design of QoE estimation models based on facial- and speech-related features is promising. However, to the best of the authors' knowledge, none of the state-of-the-art studies has ever integrated facial and speech features in a QoE estimation model. Indeed, facial features have been commonly used to estimate video quality [2, 40], whereas speech features have been used to estimate speech quality [6, 57]. Therefore, further research is needed to confirm the suitability and applicability of facial and speech features to other multimedia services. This new approach could be added to the traditional network-based approaches to monitor the QoE when deploying multimedia services [10].

In this study, we focused on videoconferencing tools based on WebRTC technology, which have been widely used for audio-visual conversation over the Internet in the past years due to the COVID-19 pandemic. Also, audio-visual calls naturally require the user to show his face while speaking, which enables capturing both the facial expressions and vocal speech characteristics. In Reference [8], we published the results of a subjective test concerning the conversational QoE of a two-party WebRTC-based audio-visual telemeeting service disturbed by combinations of network impairments (i.e., delay, jitter, and packet loss). The face and speech of the test participants were recorded during the test, and they were asked to rate the overall perceived QoE at the end of each conversation.

The following are the major contributions of this article:

— We extracted facial and speech features from the participants' audio-video data collected during the subjective assessment described in Reference [8], and we performed thorough statistical analysis to identify the most significant facial- and speech-related features for QoE estimation.
— We trained ML-based algorithms with the most significant facial features to evaluate the QoE estimation performance of facial features only. Note that preliminary results are published in Reference [7].
— We trained ML-based algorithms with the most significant speech features to evaluate the QoE estimation performance of speech features only.
— We employed data fusion techniques to combine the facial and speech datasets into a single dataset to train an ML model for enhanced QoE estimation performance due to the integrated knowledge provided by the fusion of facial and speech features.

The article is structured as follows: Section 2 presents related work. Section 3 describes the methodology followed. Section 4 and Section 5 present the proposed QoE estimation models based on facial and speech features, respectively. In Section 6, we exploit data fusion techniques to propose a QoE estimation model based on integrated facial and speech features. Section 7 concludes the article.

2 RELATED WORK

Up to this point, QoE management has been approached from various, sometimes complementary, angles [49]. It involves the utilization of different control points distributed throughout the delivery process. QoE-centric application management has primarily concentrated on control and adjustments made at the end-user level and within application hosting or cloud services. This perspective is often examined from the viewpoint of application providers aiming to enhance the quality of **Over-The-Top (OTT)** applications and services. For instance, applications like adaptive video streaming over HTTP dynamically adjust to changing network conditions to ensure a consistently high QoE. Conversely, network providers mainly depend on performance and traffic-monitoring solutions integrated into their access and core networks to gain insights into the issues experienced by end-users. QoE-oriented network management strategies are therefore centered on the network provider's standpoint and involve control mechanisms such as optimized network resource allocation and efficiency, especially in wireless systems [28]. These strategies also encompass admission control and QoE-driven routing, among others.

A key component in QoE management is the assessment, which should be able to predict the level of quality as perceived by the end-user. The outcome is the input to the major procedures mentioned above. The literature presents many studies concerning the assessment of the QoE for various multimedia services, such as video streaming [12, 13, 48], VoIP [26], Web browsing [5], and virtual reality applications [53]. However, studies on WebRTC-based applications are ongoing and still limited in the literature, in particular, those involving subjective assessments of interactive conversations. Vučić et al. focused on the multiparty telemeeting scenario and investigated the impact of several factors on the QoE, i.e., packet loss and Google Congestion Control [54], bandwidth limitation and video resolution [56], video bitrate and frame rate [55]. Major findings of these studies indicate that: The impact of packet loss on overall QoE was found to differ greatly among participants, and as long as the audio quality remained satisfactory, most participants provided high-quality scores; higher video resolutions contribute to better video quality but it requires higher processing system capabilities and may lead to congestion under limited bandwidth conditions. De Moor et al. [14] combined different network and application distortions into four technical conditions: no

distortions; distorted audio (CPU usage limited to 20%); distorted video (packet loss ratio of 20%); and distorted audio and video (delay of 500 ms and jitter of 300 ms). The lowest QoE was perceived when both audio and video were distorted by delay and jitter. In Reference [21], García et al. considered seven test conditions: no network impairments, packet loss impairment (15%, 30%, and 45%), and jitter impairment (25 ms, 50 ms, and 75 ms). It is found that the impact of these impairments differs for different video genres and that the video QoE tends to be poor when the PLR and jitter overcome 20% and 25 ms, respectively. Tsiaras et al. [51] evaluated the impact of single network distortions on the QoE of WebRTC voice calls, namely, delay (150 to 1,600 ms), jitter (0 to 400 ms), and packet loss (5% to 40%). They calibrate the proposed **Deterministic QoE model (DQX)** using the collected subjective results, which outperformed the performance of the traditional ITU E-model [25] in terms of **Mean Opinion Score (MOS)** estimation. The achieved results highlight the greater robustness of WebRTC applications to network impairments (compared to traditional VoIP applications to which the E-model is the reference QoE model) and to network delay in particular.

A limited number of literature studies explored the relationship between the QoE and facial/speech-related human emotional responses. With regard to facial expressions, these can be classified as the six+one basic emotions (i.e., anger, fear, disgust, happiness, surprise, sadness, plus the neutral emotion) [16] or can be analyzed in terms of the Facial **Action Units (AUs)** defined by the **Facial Action Coding System (FACS)** [17]. The AUs reflect the activation and intensity of facial muscles and can also be related to basic emotions. In Reference [41], facial expressions were investigated to predict the quality perception of video meeting participants. A score was given after each presentation by all participants except the presenter, and it was found that the happier the speaker was, the happier and less neutral the audience was. Also, the presentations that triggered wide swings in "fear" and "happiness" among the participants are correlated with a higher rating. In Reference [2], test participants were asked to watch videos subjected to quality (video resolution) and network (limited bandwidth) impairments. Then, selected facial-, video-, and network-related features were used to build various ML-based QoE estimation models. The **Pearson correlation coefficient (PCC)** computed between the subjective MOS and estimated MOS achieved 0.79 when the Random Forest bagging-based algorithm was used. However, the QoE model was trained on the MOS (and not the individual QoE) and the number of video sequences (8) and testers (14) were limited. In Reference [40], we have already considered facial expressions and gaze direction to build an ML-based QoE estimation model for video services impaired by buffering- and blurring-related distortions. The model achieved a mean QoE estimation accuracy of 0.878 and 0.939 when training the **k-nearest neighbors (k-NN)** algorithm with facial features only and with a combination of facial features and distortion-related features, respectively. Also, the achieved PCC between subjective MOS and estimated MOS was 0.989. In Reference [7], we have trained ML models with selected facial features and a **Fully Convolutional Network (FCN)** with the full facial expression features dataset. The aim was to let the FCN filters process the facial input matrices and to find the most relevant information to estimate the QoE. However, the ML model trained on selected facial features achieved greater QoE estimation accuracy than the FCN accuracy.

Concerning speech, several speech features are indicative of different emotions, such as the vocal tract features and the prosodic features [29, 38]. Vocal tract features produce different sound units in different emotions and are represented by the **MFCCs (Mel frequency cepstral coefficients)** derived from the cepstral domain. Prosodic features make human speech natural, including duration, intonation and intensity. In addition, they are represented by acoustic features, such as pitch frequency features, duration- and energy-related features. All of these speech features are commonly referred to as **Low-Level Descriptors (LLD)**. In Reference [6], test participants were asked to rate the QoE perceived when speech communication was impaired by different network distortions, i.e., delay, bandwidth, and loss rate. Then, three types of speech features (acoustic,

lexical, and discourse) were extracted from the recorded speech and used to train different ML algorithms. The **Support Vector Machine (SVM)** algorithm trained with a combination of the three kinds of speech features achieved a QoE estimation accuracy of 0.68. Moreover, results have shown that training with acoustic features only achieved greater QoE estimation accuracy than training with lexical and discourse features. However, the aggregation of all features slightly improves the estimation performance. The reason can be that, unlike acoustic features, lexical features (language-related information) and discourse features (only word repetition was considered) lack speech-related information (e.g., pitch and tone), which many studies have identified as indicative of human emotions and thus may be more suitable for estimating the user-perceived quality. For example, speech produced in a state of fear, anger, or joy becomes loud and fast, with a higher and broader range in pitch, whereas emotions such as sadness or tiredness generate slow and low-pitched speech [57]. Afshari et al. conducted a similar study in Reference [1], but they also considered emotional behaviors in addition to vocal and lexical speech features. The mean QoE estimation accuracy training the SVM using the fusion of speech- and behavior-related features achieved 0.828, outperforming the utilization of speech features only (0.721). The SpeechQoE model is proposed in Reference [57], a personalized QoE assessment model that converts speech-based cues into a QoE score. The model is built with a **Convolutional Neural Network (CNN)** classifier aimed at extracting, from the input time-frequency domain speech spectrogram, explicit and implicit features that are identified as the most beneficial for QoE classification. Test participants were asked to rate the QoE perceived when the VoIP communication was impaired by packet loss rate, latency, and background noise. SpeechQoE achieved a mean QoE estimation accuracy of 91.4%, with at least 90% accuracy for each single ACR class. In Reference [39], we investigated the quality perceived by employees when conducting remote working activities through implicit emotional responses estimated from the speech of video calls. The **Analysis of Variance (ANOVA)** results indicated significant changes in speech features (i.e., a combination of MFCCs, Chroma, and Mel features) when remote employees perceived different quality levels. In particular, the ANOVA tests between MFCC and ACR scores, between Chroma features and ACR scores, and between Mel features and ACR scores produced a p-value < 0.001 across the full ACR quality scale. For MFCC and Chroma features, the pairwise comparison showed an adjusted p-value < 0.001 only for the most extreme ACR scores of 1 and 5, while ACR scores 2, 3, and 4 exhibited no significant differences in means among themselves. For Mel features, the pairwise comparisons revealed an adjusted p-value < 0.001 for all ACR scores except for the pair 3–4, where the means were not significantly different from each other. However, it is worth noting that this p-value (0.004) was very close to the threshold. These findings indicate that by gathering a sufficient amount of data, Mel features can discern how individuals perceive the quality of a Web call, while MFCC and Chroma features can aid in distinguishing whether the perceived quality is positive, neutral, or negative. Consequently, these features can be effectively employed to train an ML-based model for estimating the perceived quality of remote working. Moreover, we have found that good-quality perceptions are related to neutral and positive emotion polarity, whereas low-quality perceptions are related to negative emotions.

These literature studies demonstrate the potential of facial- and speech-related features to describe human emotions driving the user's QoE perceived when utilizing multimedia services, which could be used to improve network and client-side QoE Management in streaming services [30]. In this article, we first investigate the performance of ML-based QoE estimation models trained on separated datasets of selected facial and speech features. Then, we employed a data fusion technique to combine the trained separated models into a single ML model for enhanced QoE estimation performance due to the integrated knowledge provided by the facial and speech features.

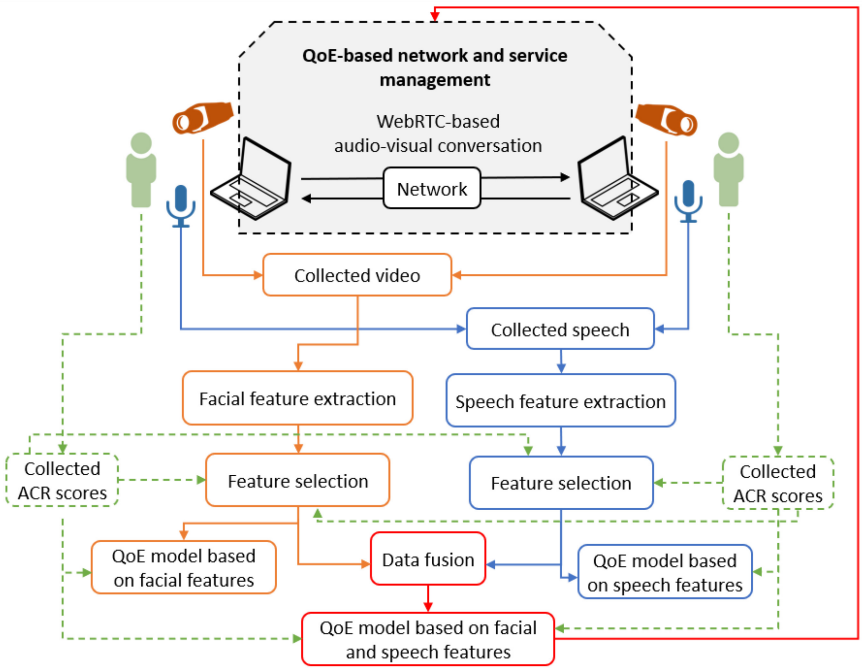


Fig. 1. The proposed methodology. The dashed lines indicate the collection of ACR scores, which are only needed for the training phase.

3 METHODOLOGY

This study aims to investigate whether the QoE of users utilizing a WebRTC-based audio-visual application can be estimated through features extracted from the user's facial expressions and speech. The rationale behind this study stands on the fact that the perceived QoE is driven by the user's emotions, which are conveyed by facial expressions and speech vocal characteristics.

To this aim, we designed and conducted a subjective quality assessment during which participants had to talk with a conversation partner using a WebRTC-based application. The network was impaired by a combination of different degradation factors to impair the quality of the audio-visual calls. During the talk, the faces and the speech of the participants were recorded to collect a data stream for extracting facial expressions and speech features. At the end of the talk, the participants were asked to rate the perceived QoE. The proposed methodology is illustrated in Figure 1. The dashed lines indicate the collection of the ACR scores (the subjective ratings provided by the users; see Section 3.1), which are only needed during the training phase as the ground truth for identifying the most suitable facial and speech features for estimating the QoE and for training the ML-based models using the selected significant features. We trained separate QoE models using facial or speech features. In addition, we relied on a data fusion technique to train a QoE model based on both facial and speech features.

The developed QoE model can be used during the operational phase for estimating the QoE based on the observed user's facial and speech features. The estimated QoE can be used to drive network and/or service management actions, in particular, when QoE degradation is detected. Note that such an approach does not preclude the utilization of additional network- and service quality-related parameters, which may complement the information that is possible to obtain with our approach. In particular, the QoE estimation provided by our approach can be correlated with

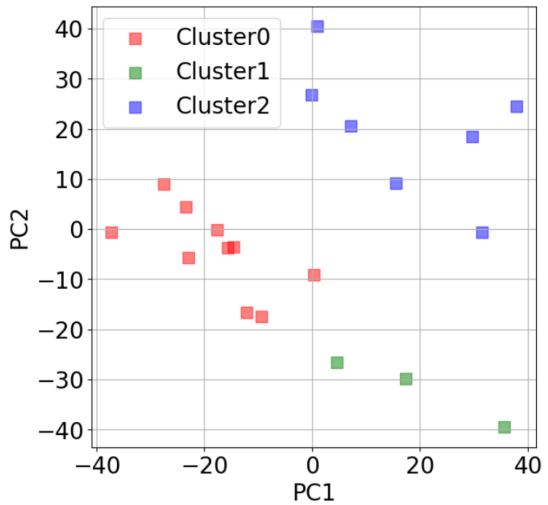


Fig. 2. The three clusters identified by clustering on the 100 psychoacoustic sharpness features extracted from the recorded speech of the 20 participants. The PCA was used to reduce feature dimensionality to the two principal components (PC1 and PC2).

network and service performance indicators to derive a root cause analysis and develop approaches to predict in advance possible quality issues and to better allocate resources to the deployed services.

The rest of this section summarizes the conducted subjective experiment and the results published in Reference [8], which are used as the basis for analysis presented in the following.

3.1 Subjective Experiment

The subjective experiment aimed to assess the conversational quality of WebRTC-based audio-visual conversations under controlled network conditions. In particular, we considered three network parameters: delay, jitter, and **packet loss rate (PLR)**. The test environment included two desks, placed in two different rooms, where the test participants were provided with a laptop to make the WebRTC-based video calls using the Google Chrome browser. We used an access point to create a dedicated wireless network communication between the two laptops to avoid undesirable network distortion provided by the Internet. Moreover, the PyNetem tool (Python network emulator) was used to impose the desired network conditions on the communication.

Twenty people (11 females and 9 males) participated in the test. They were between the ages of 23–36 years (mean 28.7, standard deviation 4.24). All participants are of European-Italian origin. We based on the five-level Fitzpatrick skin scale to identify the participants' skin tones [22]. Among the male participants, four were identified with a Type II skin tone (light-colored but darker than fair) and five with a Type III skin tone (golden honey or olive). Among the female participants, five were identified with a Type II skin tone, five with a Type III skin tone, and one with a Type IV skin tone (moderate brown). Moreover, we have used the openSMILE software to extract a total of 100 psychoacoustic sharpness features from the recorded speech of the 20 participants. We have applied the K-means algorithm to these features, which has identified 3 separate clusters, i.e., the participants can be classified into three separate groups based on their voice sharpness. These clusters are shown in Figure 2 using the **Principal Component Analysis (PCA)** for reducing feature dimensionality from 100 to 2, where the 2 principal components (PC1 and PC2) include 75% of the total features' variance. Finally, we have computed **Analysis of Variance (ANOVA)** between the

Table 1. Test Conditions (TCs) and MOS with 95% CI

TC	Delay (ms)	Jitter (ms)	PLR (%)	MOS	CI 95%
1	0	0	0	3.89	±0.404
2	500	0	0	3.67	±0.511
3	1,000	0	0	3.72	±0.338
4	500	500	0	3.44	±0.414
5	1,000	500	0	3.44	±0.316
6	0	0	15	3.22	±0.329
7	500	0	15	3.06	±0.392
8	1,000	0	15	2.72	±0.207
9	500	500	15	2.33	±0.344
10	1,000	500	15	2.44	±0.492
11	0	0	30	2.56	±0.316
12	500	0	30	2.11	±0.404
13	1,000	0	30	2.11	±0.459
14	500	500	30	1.89	±0.404
15	1,000	500	30	1.44	±0.352

100 features and the 3 identified clusters, and we have found that the 3 most significant features (p-value < 0.001), i.e., those suitable to distinguish among different speech sharpness groups, are *pcm_fftMag_psySharpness_sma_de_maxSegLen*, *pcm_fftMag_psySharpness_sma_de_segLenStddev*, and *pcm_fftMag_psySharpness_sma_de_peakMeanRel*.

The test participants were divided into pairs and were asked to play the "Who am I?" celebrity name-guessing conversational task during the video calls. This game consists in guessing the celebrity chosen by the partner by asking, in turn, yes/no questions to the conversation partner. Each pair of participants made 15 two-minute-long video calls under different network conditions. We recorded videos of the participants' faces and audio of the participants' speech during the test sessions. At the end of each conversation, the participants were asked to rate the perceived QoE using the five-level (Bad, Poor, Fair, Good, and Excellent) single discrete **Absolute Category Rating (ACR)** scale, according to the ITU-R Rec. P.800 [24]. Table 1 reports the 15 **test conditions (TCs)** created from the combination of the three considered network parameters we used to impair WebRTC communications. We chose those settings after extensive preliminary tests to be sure that the QoE of the test participants would have been impacted to different extents during the conversations. These are in line also with the other literature studies focused on QoE assessment for WebRTC applications, which have considered values of delay, packet loss, and jitter comparable to those considered in our study [14, 21, 51]. The selected values of delay, jitter, and packet loss may look worse than expected in normal two-way real-time multimedia communications. For instance, one-way delays greater than 400 ms or high values of PLR (e.g., 5%) are considered unacceptable for traditional VoIP services [52]. However, the utilization of error concealment techniques, such as packet retransmissions and **forward error correction (FEC)**, makes WebRTC-based applications more robust to network distortions. In particular, our WebRTC implementation included the Opus in-band FEC to protect the audio streaming, while the media streaming through the **RTCP (Real-time Transport Control Protocol)** protocol was protected by a combination of **redundant audio data (RAD)**, FEC, and retransmission mechanisms. In addition, the Google Chrome browser used for the experiment implemented the **Google Congestion Control (GCC)** algorithm, which adapts the media sending rate to the link capacity [9]. For these reasons, high values for the network impairment parameters have been selected, as shown in Table 1, to provide different levels of quality to the test participants.

Table 1 also shows the corresponding MOS with the 95% **confidence interval (CI)**. Note that the MOS is the average of the single ACR scores provided by 18 subjects, since 2 out of the 20 participants were identified as outliers. As expected, the MOS decreases as the impact of the network impairments increases. In particular, the PLR had the most negative influence on the QoE when introduced both alone and in combination with other network impairments. Indeed, almost sufficient to poor quality was perceived by participants when the packet loss impaired the communication. Delay and jitter also impacted negatively on the QoE, but participants still perceived sufficient quality when these impairments were even added simultaneously. Naturally, the greatest QoE was perceived when no impairments were added to the network.

4 QOE ESTIMATION BASED ON THE FACIAL FEATURES

In this section, we first describe how we applied statistical analysis to facial expression features (Section 4.1). Then, we present the performance of the ML-based QoE models trained on selected facial features (Section 4.2).

4.1 Statistical Analysis of Facial Expression Features

As we mentioned in Section 3, we recorded the faces of the participants while they were having audio-visual conversations. We recorded a total of 270 two-minute-long videos at full HD quality and 30 fps. Then, we used the OpenFace toolkit [3, 4, 60] to extract facial expression features and gaze direction features from the recorded face images. Concerning facial expressions, the features were extracted in terms of the facial **Action Units (AUs)** defined by the **Facial Action Coding System (FACS)**[1] [17]. The AUs are 35: 18 AU_c detect the activation of a specific muscle, whereas 17 AU_r detect the muscle activation intensity (from 1 to 5). Concerning the **gaze direction (GD)**, the features were extracted in terms of 6 arrays of coordinates indicating where the gaze was directed. From each video, we extracted up to 3,600 values (i.e., a feature value per frame) for each of the 41 (6 GD + 35 AUs) facial features. Note that in OpenFace, we set a value of 98% for the face-tracking confidence, which means that the features' values were extracted only for the frames where a face was identified with a confidence higher than 98%. The number of features' values then varies depending on the number of frames the OpenFace toolkit managed to track of the participant's face.

The next step concerns the computation of statistics to reduce the data dimensionality and obtain a single value for each of the 41 features. To this, we considered three statistics:

(S1) We computed the frequency of activation F_{AU_c} for each AU_c, the intensity of activation I_{AU_r} for each AU_r, and the variance of the GD V_{GD} for each gaze feature, following the equations provided in Reference [40].
(S2) We computed the mean value μ of each of the 41 facial expression features over the (up to) 3,600 values extracted for each test.
(S3) We computed the standard deviation σ of each of the 41 facial expression features over the (up to) 3,600 values extracted for each test.

We obtained a total of 11,070 features' statistics (15 TCs $\times$ 18 participants $\times$ 41 features) for S_1, S_2, and S_3. Then, for each statistic, we computed the one-way ANOVA between the features' statistics (computed for different TCs) and the corresponding ACR scores. Table 2 shows the features' statistics that achieved a p-value < 0.05 for S_1, S_2, and S_3. The results show that S_1, S_2, and S_3 include, respectively, 9, 6, and 3 features' statistics that have means statistically significantly different for different ACR scores. Thus, we can consider these features' statistics as statistically significantly relevant for building QoE estimation models. It is interesting to note that S_1 and S_2

[1] https://www.cs.cmu.edu/~face/facs.htm

Table 2. ANOVA Results: Features' Statistics that Achieved a P-Value < 0.05 for S_1, S_2, and S_3

S1	S2	S3
F_{AU06_c}, F_{AU12_c}, F_{AU14_c}, μ_{AU12_c}, μ_{AU26_c}, μ_{AU45_c}, σ_{AU04_c}, σ_{AU26_c}, σ_{AU45_c} F_{AU25_c}, F_{AU26_c}, I_{AU05_r}, μ_{AU06_r}, μ_{AU12_r}, μ_{AU26_r} I_{AU06_r}, I_{AU12_r}, I_{AU26_r}		

Table 3. Number of ACR Scores (and Corresponding Facial Features) before and after Data Augmentation

ACR	Collected samples	Augmented samples
1	37	105
2	65	121
3	103	103
4	44	101
5	21	99
Total	**270**	**529**

statistics regard both AU activation (AU_c) and AU intensity (AU_r), whereas S_3 statistics are only significant for AU intensity. Thus, there are specific facial muscle movements that are found to be correlated with QoE perception. In particular, AU26 (jaw drop) is present in all statistics, AU06 (cheek raiser) and AU12 (lip corner puller) are present in both S_1 and S_2 statistics, and AU45 (blink) is present in both S_2 and S_3 statistics. However, none of the significant features' statistics concerns the eye gaze, which can be excluded as a piece of relevant information for estimating the QoE for this specific study.

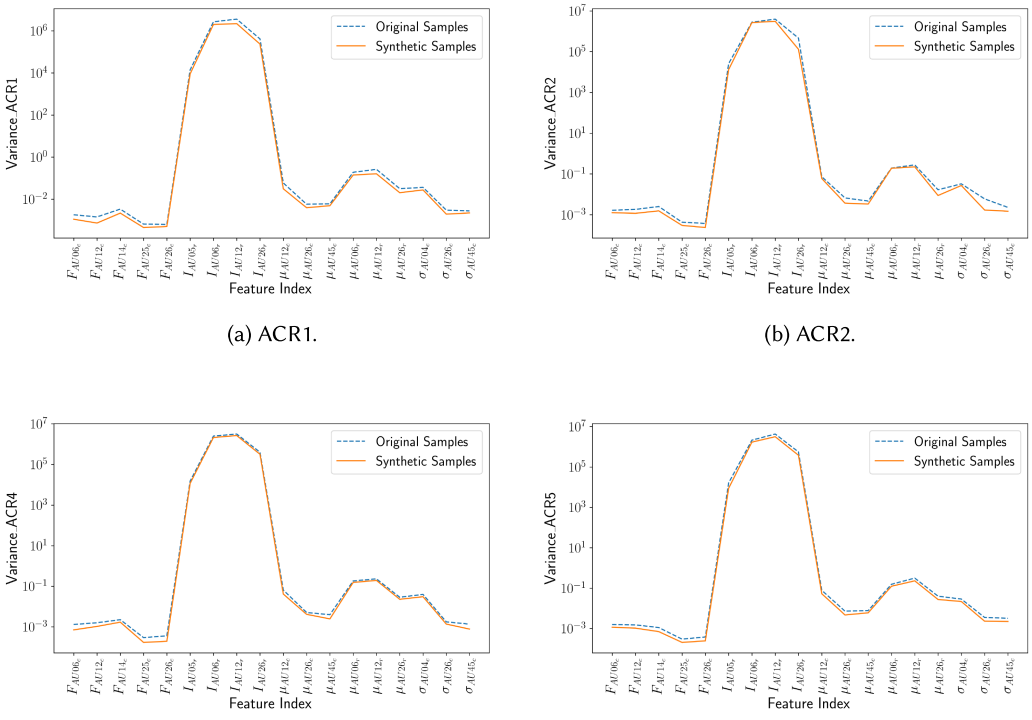
4.2 ML-based QoE Models Based on the Facial Features

Before training the ML-based models with the facial features' statistics identified in the previous section, we performed data augmentation to correct the dataset's class imbalance. Indeed, the extreme scores of the ACR scale were less used than the middle scores to rate the perceived QoE, as shown in Table 3. Therefore, we employed the **adaptive synthetic (ADASYN)** algorithm for constructing synthetic samples and achieving class over-sampling. In particular, ADASYN attempts to enhance class balance by adaptively producing new synthetic instances from the minority class, using linear weighted interpolation between existing minority class examples, to reduce the bias introduced by the imbalanced data distribution [23]. As shown in Table 3, the gap between the classes is decreased for the augmented dataset. Moreover, in Figure 3, we compare the variance of the original facial features' statistics with the variance of the synthetic facial features' statistics produced by ADASYN for the ACR scores 1, 2, 4, and 5. It can be seen that the synthetic samples follow the same data distribution as the original samples.

We then used the MATLAB software to train different ML algorithms on the augmented dataset using the significant features' statistics identified in Section 4.1 to estimate the QoE. In particular, we considered the following subsets of features (see Table 2) to train the classifiers:

— Subset1: The 9 S_1 feature's statistics.
— Subset2: The 6 S_2 feature's statistics.
— Subset3: The 3 S_3 feature's statistics.
— Subset4: The 18 $S_1 + S_2 + S_3$ feature's statistics.

The classifiers aim to find a pattern in the features dataset that describes a correlation between the facial feature's statistics and the provided ACR quality score. For all the ML algorithms, the 70%/30% training/validation rate was used with a 5-fold cross-validation. The **k-nearest**



(a) ACR1. (b) ACR2.

(c) ACR4. (d) ACR5.

Fig. 3. Variance of the original and synthetic facial expression features' statistics for the ACR scores 1, 2, 4, and 5. Note that a logarithmic scale was used for the y-axis.

neighbors (k-NN) and the **Support Vector Machine (SVM)** were the ML algorithms achieving the best QoE estimation performance for the four subsets among a set of diverse considered classifiers (e.g., decision trees, naive Bayes, ensembles, discriminant analysis).

In particular, the optimal classifier settings were as follows:

— Subset1 and Subset2: the k-NN with $n = 1$, utilizing the Euclidean distance metric and uniform weighting distance.
— Subset3: the k-NN with $n = 13$, utilizing the Euclidean distance metric and the inverse squared weighting distance.
— Subset4: the SVM with the linear kernel function and a kernel scale of 1.

Table 4 shows the QoE estimation performance achieved by these ML algorithms in terms of mean accuracy and precision, recall, and F1-score computed for the single ACR scores. It can be seen that the greatest mean accuracy is achieved when training with all the significant features' statistics included in Subset4. Also, this subset of features allows the SVM classifier to achieve the greatest values of precision, recall, and F1-score for the single ACR scores. This result is likely due because training with a larger number of significant features enables the ML models to achieve greater classification results. In particular, the recall achieved for the extreme classes (ACR 1 and 5) is close to 1 and decreases towards the middle class (ACR 3), achieving the lowest recall (0.35). However, intermediate classes (ACR 2 and 4) achieved good recall (around 0.8). This decreasing trend of the performance towards the middle class can also be observed when the ML algorithms were trained with the other three subsets of features, although with lower absolute performance

Table 4. QoE Estimation Performance Achieved by the ML-based QoE Models Trained with the Facial Features

Subset / ML model	Performance metric	ACR score 1	2	3	4	5
	Mean Acc.			0.70		
Subset1	**Precision**	0.71	0.72	0.49	0.65	0.86
k-NN	**Recall**	0.80	0.78	0.30	0.75	0.86
	F1-Score	0.75	0.75	0.37	0.70	0.86
	Mean Acc.			0.60		
Subset2	**Precision**	0.65	0.56	0.30	0.60	0.80
k-NN	**Recall**	0.72	0.57	0.20	0.65	0.88
	F1-Score	0.68	0.57	0.24	0.62	0.84
	Mean Acc.			0.42		
Subset3	**Precision**	0.45	0.36	0.17	0.38	0.67
k-NN	**Recall**	0.47	0.33	0.16	0.38	0.77
	F1-Score	0.46	0.34	0.17	0.38	0.72
	Mean Acc.			0.78		
Subset4	**Precision**	0.86	0.71	0.62	0.75	0.89
SVM	**Recall**	0.97	0.79	0.35	0.81	0.98
	F1-Score	0.91	0.75	0.45	0.78	0.93

results. In particular, Subset1 allows for achieving the best estimation performance among the first three subsets, which suggests the statistical method S_1 to be the most suitable when considering facial features. Subset2 achieved the second-best performance, followed by Subset3.

These results can be motivated by the fact that the lower number of significant features included in Subset2 and Subset3 compared to Subset1 is not compensated by a higher significance. Therefore, with a lower number of features for training the ML models, it is more difficult to achieve good estimation performance. This outcome is particularly highlighted when training with Subset3, which only includes three significant features. However, although the features included in Subset2 and Subset3 alone do not perform very well, they still provide important complementary information regarding facial features that can be exploited when grouped with the features included in Subset1 to reach enhanced estimation performance (Subset4, which includes all the features). The estimation performance of the ML model trained with all features is only limited by the low classification recall achieved for the middle class (ACR 3). This may be due to potential overfitting introduced by synthetic samples representing the minority classes (ACR 1 and 5, in particular). However, the capability to estimate with good recall if a user is satisfied (ACR 4 and 5) or annoyed (ACR 1 and 2) compared to the used audio-visual services is of extreme importance, especially if we consider that the estimation is solely based on the user's facial expressions and no feedback is requested. Finally, it is also interesting to note that for the first three subsets, the ML algorithm that achieved the best classification results was the k-NN, whereas when all the features were considered (Subset4), the SVM was the best estimator. This outcome would likely depend on the spatial distribution of the features, which in this case were better separated by the SVM functions, as they work better with large datasets.

5 QOE ESTIMATION BASED ON THE SPEECH FEATURES

In this section, we first describe how we applied statistical analysis to speech features (Section 5.1). Then, we present the performance of the ML-based QoE models trained on selected speech features (Section 5.2).

Table 5. ANOVA Results: LLDs that Contain Functional Statistical Features with a P-Value < 0.01 for the Three Speech Files: OS, NRS, and NSS

LLD	OS	NRS	NSS
audspec_lengthL1norm	7	4	1
audSpec_Rfilt	8	49	10
audspecRasta	-	7	4
F0final	-	2	3
jitterLocal	1	-	9
jitterDDP	-	-	3
logHNR	-	3	2
mfcc_sma	43	20	25
pcm_fftMag	44	16	14
pcm_RMSenergy	3	3	3
pcm_zcr	5	3	-
shimmerLocal	-	-	4
voicingFinalUnclipped	2	4	8
Total	**113**	**111**	**86**

5.1 Statistical Analysis of Speech Features

As we mentioned in Section 3, we recorded the participants' speeches while they were having the audio-visual conversations. Therefore, we considered three versions of the speech files:

— **Original Speech (OS)**: The original speech recorded from the participants during the experimental test. It contains the silent intervals (when the participant listens to the partner) and the background noise recorded during the conversation.
— **Noise Reduced Speech (NRS)**: We applied a non-stationary noise reduction method (called spectral gating) to the original speech file to reduce the background noise. We used the noisereduce[2] Python algorithm [42, 43], which reduces noise in time-domain signals by continuously adjusting the predicted noise threshold over time.
— **Non-Silent Speech (NSS)**: We removed the silent intervals from the OS using the librosa[3] Python library [36].

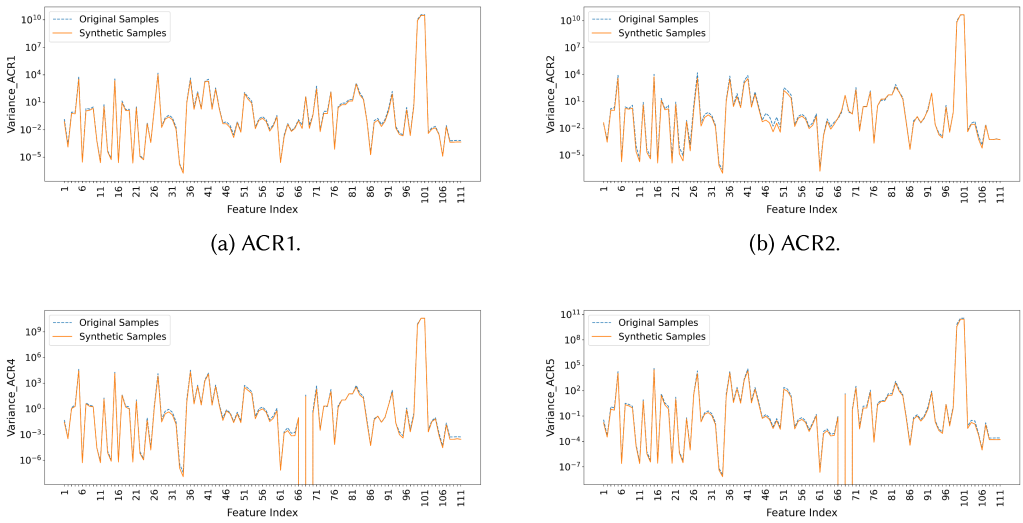
Utilizing the OpenSMILE feature extraction toolkit [19, 20], we extracted speech features from each of the three speech files. In particular, we analyzed speech features over **low-level descriptor (LLD)** contours with functionals relying on the ComParE acoustic feature set [46, 59]. For each speech file, OpenSMILE extracts 64 LLDs specifically related to the energy (4), spectral (55), and voicing (6) characteristics of the signal. Finally, a total of 6,373 functional statistical features are computed and provided for the considered LLDs. These features were computed for all three speech versions: OS, NRS, and NSS.

We obtained a total of 1,720,710 features' statistics values (15 TCs × 18 participants × 6,373 features) for OS, NRS, and NSS. Then, we performed the one-way ANOVA between the speech features (grouped for different TCs) and the corresponding ACR scores. Table 5 indicates the LLDs that contain functional statistical features with a p-value < 0.01 for each of the three speech files. Note that, compared to Section 4.1, we considered a lower significance level for the p-value in this case, because the number of speech features is much larger than that of the facial features. Indeed,

[2]https://pypi.org/project/noisereduce/
[3]https://github.com/librosa/librosa

Table 6. Number of ACR Scores (and Corresponding Speech Features) before and after Data Augmentation for the Three Speech Files: OS, NRS, and NSS

ACR score	Collected samples	Augmented Samples		
		OS	NRS	NSS
1	37	104	102	107
2	65	107	110	107
3	103	103	103	103
4	44	111	104	105
5	21	104	98	97
Total	**270**	**529**	**517**	**519**



(a) ACR1. (b) ACR2.

(c) ACR4. (d) ACR5.

Fig. 4. Variance of the original and synthetic speech features' statistics for ACR scores 1, 2, 4, and 5. Note that a logarithmic scale was used for the y-axis.

even with a lower threshold, we obtained a larger number of speech-significant features as a result of the ANOVA analysis. In particular, a total of 113, 111, and 86 functional statistical features were found to be statistically significant for the QoE when considering the OS, NRS, and NSS speech files, respectively.

5.2 ML-based QoE Models Based on the Speech Features

Similarly to Section 4.2, before training the ML-based models with the speech feature's statistics identified in the previous section and summarized in Table 5, we performed data augmentation using ADASYN to correct the dataset's class imbalance. The number of ACR scores before and after data augmentation for each of the three speech files is shown in Table 6. Note that the number of augmented samples differs for the three speech files, because it also depends on the different number of significant features found for each speech file. Moreover, in Figure 4, we compare the variance of the original speech features' statistics with the variance of the synthetic speech features' statistics produced by ADASYN for the ACR scores 1, 2, 4, and 5. It can be seen that the synthetic samples follow the same data distribution as the original samples.

Table 7. QoE Estimation Performance Achieved by the ML-based Models Trained with the Speech Features

Speech file / ML model	Performance metric	ACR score 1	2	3	4	5
	Mean Acc.			0.83		
OS	**Precision**	0.85	0.71	0.76	0.80	0.99
SVM	**Recall**	0.95	0.76	0.52	0.89	0.99
	F1-Score	0.90	0.73	0.62	0.85	0.99
	Mean Acc.			0.86		
NRS	**Precision**	0.98	0.89	0.61	0.99	0.99
SVM	**Recall**	0.93	0.70	0.88	0.83	0.99
	F1-Score	0.95	0.78	0.72	0.90	0.99
	Mean Acc.			0.85		
NSS	**Precision**	0.93	0.79	0.76	0.83	0.93
SVM	**Recall**	0.93	0.85	0.57	0.91	0.98
	F1-Score	0.93	0.82	0.65	0.87	0.95

We then used the MATLAB software to train different ML algorithms (e.g., SVM, K-NN, decision trees, naive Bayes, ensembles, and discriminant analysis) on the augmented datasets using the significant features' statistics identified in Section 5.1 to estimate the QoE. For all the ML algorithms, the 70%/30% training/validation rate was used with a 5-fold cross-validation.

In particular, we considered the following subsets of features to train the classifiers:

— OS: The 113 significant speech features' statistics.
— NRS: The 111 significant speech features' statistics.
— NSS: The 86 significant speech features' statistics.

The SVM with the linear kernel function and a box constraint level of 1 achieved the greatest estimation performance for all of the three subsets of features. The k-NN classifier with the Euclidean distance metric achieved the second-best results, followed by the SVM with a cubic kernel function. Table 7 provides the performance obtained with the best classifier (e.g., the SVM with the linear kernel) for the three speech files in terms of mean accuracy, and precision, recall, and F1 scores calculated for every class of ACR scores.

The SVM trained on the NRS's features' statistics achieved the greatest mean accuracy, although comparable to that achieved when training the SVM on the other two sets of features. Moreover, the single recall for each class of ACR scores is at least 0.70 for NRS, whereas, for OS and NSS, one of the single recall values (class 3) is lower than 0.60. Therefore, the speech feature's statistics extracted from the NRS file are the most suitable for building a QoE estimator. In general, all models achieved single recall higher than 0.80 for classes 1, 4, and 5, while classes 2 and 3 were the most difficult to predict for the QoE models. This can be due to the emotional speech features, which are more informative when the perceived QoE is clearly low or high, whereas it is more difficult to distinguish between poor and sufficient quality. The same trend was also obtained with the QoE estimators trained on facial features.

6 QOE ESTIMATION BASED ON THE FACIAL AND SPEECH FEATURES

Previous sections demonstrated that SVM-based models trained on individual datasets of facial- and speech-related features achieved a mean QoE estimation accuracy of up to 0.78 and 0.86, respectively. In this section, we investigate the utilization of data fusion techniques to implement

Table 8. Number of ACR Scores (and Corresponding Facial and Speech Features) before and after Data Augmentation

ACR	Collected samples	Augmented samples
1	37	108
2	65	108
3	103	103
4	44	106
5	21	96
Total	**270**	**521**

a QoE estimation model based on both the facial- and speech-related features with the aim of enhancing the QoE estimation performance.

6.1 Input Features and Data Augmentation

We define $\mathbf{FAC}_{(270,18)}$ the facial features dataset and $\mathbf{SP}_{(270,111)}$ the speech features dataset. 270 is the number of feature samples corresponding to the ACR scores collected during the subjective test, 18 are the facial features of Subset4 described in Section 4.2, and 111 are the most significant speech features extracted from the NRS file described in Section 5.2. We define $\mathbf{L}_{(270,1)}$ the matrix of 270 ACR scores (labels).

Before applying data fusion techniques on the **FAC** and **SP** datasets, we need to apply the ADASYN data augmentation technique to reduce the datasets' class imbalance of the ACR scores. We cannot rely on the individual augmented datasets discussed in Sections 4.2 and 5.2, because their number of class samples is different, which makes them not compatible for data fusion.

Therefore, we first horizontally concatenated the **FAC** and **SP** datasets as in Equation (1) to create the **FAC-SP** single dataset:

$$\mathbf{FAC\text{-}SP}_{(270,129)} = [\mathbf{FAC}_{(270,18)} | \mathbf{SP}_{(270,111)}]. \tag{1}$$

Then, we applied the ADASYN algorithm on the **FAC-SP** dataset and the corresponding **L** matrix, as in Equation (2), for creating synthetic features and label samples reducing class imbalance:

$$[\mathbf{FAC}_{aug}, \mathbf{SP}_{aug}, \mathbf{L}_{aug}] = \text{ADASYN}(\mathbf{FAC\text{-}SP}, \mathbf{L}), \tag{2}$$

where $\mathbf{FAC}_{aug(521,18)}$, $\mathbf{SP}_{aug(521,111)}$, and $\mathbf{L}_{aug(521,1)}$ are the augmented facial and speech datasets and the augmented matrix of labels, respectively. To obtain a normalized distribution of the data in each dataset, both $\mathbf{FAC}_{aug(521,18)}$ and $\mathbf{SP}_{aug(521,111)}$ have been normalized using the Z-score function. Table 8 summarizes the number of augmented samples for each class. This approach allowed us to create synthetic samples for both facial and speech features corresponding to the same label as well as augmented facial and speech features datasets including the same number of class samples.

6.2 Data Fusion

The SVM is the ML classifier that demonstrated to achieve the best QoE estimation performance on the individual facial and speech features datasets. Thus, we utilized an SVM as the ML classifier and we considered two data fusion approaches to fuse the $\mathbf{FAC}_{aug}$ and $\mathbf{SP}_{aug}$ datasets: the **Principal Component Analysis (PCA)** and the **Improved Centered Kernel Alignment (ICKA)**. The PCA is a popular and standard technique for analyzing, interpreting, reducing the dimensionality, and fusing large datasets containing a high number of features [31]. The ICKA is a kernel fusion technique developed for SVM classifiers [34].

Table 9. QoE Estimation Performance Achieved by the SVM Model Trained with Facial and Speech Features Fused with PCA and ICKA Techniques

Data fusion technique	Performance metric	ACR score 1	2	3	4	5
PCA	**Mean Acc.**			0.84		
	Precision	0.95	0.74	0.60	0.93	0.95
	Recall	0.90	0.85	0.71	0.86	0.90
	F1-score	0.93	0.78	0.68	0.87	0.93
ICKA	**Mean Acc.**			0.93		
	Precision	0.85	0.83	0.82	0.95	0.87
	Recall	0.92	0.92	0.91	0.99	0.93
	F1-score	0.87	0.85	0.80	0.95	0.88

6.2.1 Principal Component Analysis. The PCA is a statistics technique used to reduce the dimensionality of a dataset while retaining as much as possible of the original data variance. The PCA transforms the original dataset into a reduced dataset of new variables capturing the most important patterns and relationships in the data. This characteristic of the PCA can be exploited for data fusion, which is the process of combining data from multiple sources to create a unified view of the underlying phenomenon. Therefore, we first applied PCA separately on the $\mathbf{FAC}_{aug}$ and $\mathbf{SP}_{aug}$ datasets to extract the most important patterns and relationships. By applying the elbow method, we found the first four PCA components include the greatest percentage of data variance (95%) for both datasets. We define $\mathbf{FAC}_{aug}^{PCA}$ and $\mathbf{SP}_{aug}^{PCA}$ the augmented facial and speech features datasets transformed after the application of the PCA. Then, we horizontally concatenated the transformed datasets as in Equation (3):

$$\mathbf{FAC\text{-}SP}_{aug}^{PCA} = [\mathbf{FAC}_{aug}^{PCA} | \mathbf{SP}_{aug}^{PCA}]. \tag{3}$$

We trained the SVM on the $\mathbf{FAC\text{-}SP}_{aug}^{PCA}$ dataset with a 5-fold cross-validation approach by applying a 70%/30% training/validation split rate. To obtain the best-tuned SVM, we used the GreedSearchCV function.[4] We considered the following tuning parameters: the penalty parameter of the error term $C = \{0.1, 1, 10, 100, 1000\}$, the parameter $\gamma = \{1, 0.1, 0.01, 0.001, 0.0001\}$, and three kernels K: $gaussian$, $linear$, and $radial\ basis$. The results in Table 9 are obtained with $C = 10$, $\gamma = 1$, and $K = radial\ basis$.

6.2.2 Improved Centered Kernel Alignment. The ICKA [34] is a method used for ML feature fusion tasks. It is an extension of the kernel alignment method, which measures the similarity between two datasets by computing the inner product of their corresponding kernel matrices. The main idea behind ICKA is to compute the SVM kernel alignment between the ideal kernel blocks and the base kernel that are selected to be representative of the data. Then, we construct the fuse kernel by a weighted linear combination of multiple aligned kernels. Thus, we define the SVM optimization by approaching its dual problem formulation and applying the "kernel trick" to avoid mapping the features into the high dimensional space. Therefore, according to the SVM definition reported in Reference [58], we can define the SVM with the following optimization function:

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j k(x_i x_j), \tag{4}$$

[4] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

which is subject to

$$\alpha_i \geq 0, i = \{1, \ldots, m\}, \sum_{i=1}^{m} a_i y_i = 0, \tag{5}$$

where $\alpha_i \alpha_j$ are Lagrange multipliers, x is the set of samples, m is the total number of samples in the set of features x, y is the set of labels, and $k(x_i x_j)$ is the kernel trick non-linear mapping function. As introduced before, the fusion kernel can be obtained with a weighted sum of multiple base kernels to integrate the different types of features, as follows:

$$k\left(x_i, x_j\right) = \sum_{h=1}^{H} d_h k_h\left(x_i, x_j\right), s.t.\ d_h \geqslant 0, \text{and} \sum_{h=1}^{H} d_h = 1, \tag{6}$$

where d_h is the non-negative weight for the base kernel and H is the total number of base kernels. d_h is obtained as follows:

$$d_h = dist(K_h) / \sum_{h=1}^{H} dist(K_h). \tag{7}$$

The $dist$ function is the distance measured between the base kernel and the ideal kernel, which is defined as follows:

$$dist(K_h) = \sum_{c=1}^{C} S(|D_{hc}|)/S(|K_h|), \tag{8}$$

where $h = \{1, 2, \ldots, H\}$, $c = \{1, 2, \ldots, C\}$ identifies the number of classes, S is the sum function of all elements of the matrix D_{hc}, and D is the diagonal block element that identifies the class c and base kernel h. D_{hc} is obtained from the ideal kernel block K_{block} defined as:

$$K_{block}(X, X) = \begin{bmatrix} D_{h1} & & \\ & \ddots & \\ & & D_{hc} \end{bmatrix}, \tag{9}$$

where $h = \{1, 2, \ldots, H\}$ and $c = \{1, 2, \ldots, C\}$.

Thus, according to Equation (6), we can define the fuse obtained kernel as:

$$ICKA_{Kernel} = d_{face} k_{face}\left(x_i, x_j\right) + d_{speech} k_{speech}\left(x_i, x_j\right). \tag{10}$$

The $ICKA_{kernel}$ is used as the input kernel of the SVM. Similarly to the PCA approach, we tuned the SVM using the GreedSearchCV function, performing a 5-fold cross-validation, dividing the dataset with a 70%/30% split rate, and finding the best C and γ parameters, respectively, equal to 10 and 1.

6.3 Data Fusion Results

Table 9 shows the QoE estimation results in terms of mean accuracy, precision, recall, and F1-score obtained with the SVM classifier using the PCA and ICKA data fusion techniques. The PCA-based solution reached a mean accuracy of 0.84 that, compared to the ICKA-based method, is 9% lower. Comparing the recall results between the two approaches, the PCA-based one obtained results comparable to the ICKA-based method for classes 1 and 5. The remaining classes achieved worse and less stable results. ICKA recall results appear to be more stable and always greater than 0.90. The precision and the F1-score metrics reported that ICKA has fewer misclassification issues. Still, the results are always greater than 0.80, which makes the ICKA-based SVM suitable to predict the five rating scales of the perceived quality. Comparing the QoE estimation results achieved by the ICKA-based SVM with those achieved by the facial- and speech-based SVMs (reported in Tables 4 and 7, respectively), it can be noted that the data fusion technique provided a significant

Table 10. Comparison with the State-of-the-art in Terms of the Mean Accuracy

Method	MLQoE [11]	SpeechQoE [57]	Ours (speech only)	Ours (facial and speech)
Accuracy	0.76	0.81	0.86	0.93

performance improvement. Indeed, both the facial- and speech-based SVMs show weaknesses when predicting the middle class 3, highlighting a very unbalanced prediction in favor of the external classes. The best NRS SVM method obtained significantly lower scores of recall than the ICKA-based method, obtaining for the classes 2, 3, and 4 recall scores lower than 22%, 16%, and 22%, respectively.

Moreover, we have computed the **Pearson correlation coefficient (PCC)** and **Root Mean Square Error (RMSE)** between the actual ACR scores and the ACR scores estimated with the proposed PCA- and ICKA-based solutions. The PCC measures the linear correlation between two sets of data. It is essentially a normalized measurement of covariance, whose result always falls between −1 and 1. PCC values greater than 0.8 indicate a strong positive correlation between the two sets of data, whereas PCC values lower than −0.8 indicate a strong negative correlation. PCC values in the middle of these thresholds indicate low or no correlation between the datasets. The RMSE measures the quadratic mean of the differences between the true and estimated values. The lower the RMSE, the better the estimation quality. The ICKA-based solution achieved a PCC of 0.984 and an RMSE of 0.255, whereas the PCA-based one achieved a PCC of 0.923 and an RMSE of 0.561. These results highlight that both the solutions estimate ACR scores strongly correlated with the real ACR scores. However, the ICKA-based method achieves a lower RMSE (almost half) than that achieved by the PCA-based method. Thus, the ICKA-based solution is the most accurate solution to estimate single ACR scores based on speech- and facial-based features.

6.4 Comparison with the State-of-the-art

Finally, in Table 10, we compare the performance of our method, in terms of the mean accuracy, with that achieved with state-of-the-art methods. Our method is the only one in the literature that estimates the QoE based on both facial and speech features, which achieved a mean accuracy of 0.93. By only considering the speech, our method achieved an accuracy of 0.86, as presented in Section 5. The MLQoE [11] takes as the input only the network conditions (Table 1 in this case) and employs multiple ML algorithms to estimate the QoE. It achieved an accuracy of 0.76. This lower performance may be due to speech-related features that are not being considered by this method. The SpeechQoE model [57] converts speech-based cues into a QoE score. It takes as the input the time-frequency domain speech spectrogram, from which it extracts implicit features by conducting multiple levels of non-linear operations. In this case, the inputs were the time-frequency domain speech spectrograms of the OS files described in Section 5.1. The SpeechQoE achieved an accuracy of 0.81. However, it must be said that the SpeechQoE model was trained by considering network conditions less disturbing than those considered in our study. In particular, the PLR ranged from 0 to 3% and the delay from 0 to 500 ms; also, the jitter was not considered. This may be the reason for the lower performance when compared to the performance achieved by our method.

7 CONCLUSION

In this article, we focused on the joint exploitation of voice and facial expression features to predict the quality of experience of the end-users during WebRTC video calls. This approach has the advantage of building personalized QoE models as the personal instinctive reaction of the user

that is seized by the facial expression and voice is extensively analyzed by the proposed method with a total of 129 features that have been found to be statistically significant (111 for the voice signal and 18 for the facial expression). This is not the case with most of the available models where the human influencing factors are often neglected due to the difficulty in collecting relevant parameters.

The proposed method has proved to be successful in predicting the QoE with an overall accuracy of 0.93. This is quite higher than the case where only the facial expression or the voice features are used, which achieved an accuracy of 0.78 and 0.86, respectively. Moreover, the proposed model outperformed the state-of-the-art models. Whereas the application of the approach requires the consent of the user (as is the case for most of the predictors), all the processing can be performed on the user side with the computational power of a normal smartphone or PC. Only the final predicted QoE level has to be shared with the service provider to take any action for service management. This data can be used in real-time to take any reactive action (e.g., allocate more resources for the service) in case of any quality issue or for offline analysis of the provided quality levels at different network and service conditions.

In the future, we plan to extend this approach to other application scenarios, as we believe that the correlation of the considered features with the QoE can be application-independent to a certain extent. Indeed, by feeding the system also with the parameters that characterize the application, it may be possible to achieve results comparable to the ones presented in this article with a generalized model. As the parameters considered may vary from one application configuration to another, we will be experimenting a **Multi-view (MV)** learning that allows for improving generalization efficiency by learning from multiple viewpoints.

REFERENCES

[1] Saeid Afshari and Naser Movahhedinia. 2016. QoE assessment of interactive applications in computer networks. *Multim. Tools Applic.* 75, 2 (2016), 903–918.

[2] L. Amour, M. I. Boulabiar, S. Souihi, and A. Mellouk. 2018. An improved QoE estimation method based on QoS and affective computing. In *International Symposium on Programming and Systems (ISPS'18)*. 1–6.

[3] T. Baltrušaitis, M. Mahmoud, and P. Robinson. 2015. Cross-dataset learning and person-specific normalisation for automatic Action Unit detection. In *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG'15)*, Vol. 06. 1–6.

[4] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L. Morency. 2018. OpenFace 2.0: Facial behavior analysis toolkit. In *13th IEEE International Conference on Automatic Face Gesture Recognition (FG'18)*. 59–66.

[5] Sabina Barakovic and Lea Skorin-Kapov. 2017. Survey of research on Quality of experience modelling for web browsing. *Qual. User Exper.* 6 (2017).

[6] Abhishek Bhattacharya, Wanmin Wu, and Zhenyu Yang. 2012. Quality of experience evaluation of voice communication: An affect-based approach. *Hum.-Cent. Comput. Inf. Sci.* 2, 1 (2012).

[7] Gülnaziye Bingöl, Simone Porcu, Alessandro Floris, and Luigi Atzori. 2022. QoE estimation of WebRTC-based audiovisual conversations from facial expressions. In *16th International Conference on Signal-Image Technology & Internet-based Systems (SITIS'22)*. 577–584.

[8] Gülnaziye Bingöl, Luigi Serreli, Simone Porcu, Alessandro Floris, and Luigi Atzori. 2022. The impact of network impairments on the QoE of WebRTC applications: A subjective study. In *14th International Conference on Quality of Multimedia Experience (QoMEX'22)*. 1–6.

[9] Gaetano Carlucci, Luca De Cicco, Stefan Holmer, and Saverio Mascolo. 2016. Analysis and design of the Google congestion control for web real-time communication (WebRTC). In *7th International Conference on Multimedia Systems (MMSys'16)*. DOI: https://doi.org/10.1145/2910017.2910605

[10] Pedro Casas, Michael Seufert, Florian Wamser, Bruno Gardlo, Andreas Sackl, and Raimund Schatz. 2016. Next to you: Monitoring quality of experience in cellular networks from the end-devices. *IEEE Trans. Netw. Serv. Manag.* 13, 2 (2016), 181–196. DOI: https://doi.org/10.1109/TNSM.2016.2537645

[11] Paulos Charonyktakis, Maria Plakia, Ioannis Tsamardinos, and Maria Papadopouli. 2016. On user-centric modular QoE prediction for VoIP based on machine-learning algorithms. *IEEE Trans. Mob. Comput.* 15, 6 (2016), 1443–1456. DOI: https://doi.org/10.1109/TMC.2015.2461216

[12] Yanjiao Chen, Kaishun Wu, and Qian Zhang. 2015. From QoS to QoE: A tutorial on video quality assessment. *IEEE Commun. Surv. Tutor.* 17, 2 (2015), 1126–1165. DOI: https://doi.org/10.1109/COMST.2014.2363139

[13] Federico Chiariotti. 2021. A survey on 360-degree video: Coding, quality of experience and streaming. *Comput. Commun.* 177 (2021), 133–155. DOI: https://doi.org/10.1016/j.comcom.2021.06.029

[14] Katrien De Moor, Sebastian Arndt, Doreid Ammar, Jan-Niklas Voigt-Antons, Andrew Perkis, and Poul E. Heegaard. 2017. Exploring diverse measures for evaluating QoE in the context of WebRTC. In *9th International Conference on Quality of Multimedia Experience (QoMEX'17)*. 1–3. DOI: https://doi.org/10.1109/QoMEX.2017.7965665

[15] Darragh Egan, Sean Brennan, John Barrett, Yuansong Qiao, Christian Timmerer, and Niall Murray. 2016. An evaluation of heart rate and electrodermal activity as an objective QoE evaluation method for immersive virtual reality environments. In *8th International Conference on Quality of Multimedia Experience (QoMEX'16)*. 1–6.

[16] P. Ekman and W. Friesen. 1971. Constants across cultures in the face and emotion. *J. Person. Soc. Psychol.* 17, 2 (1971), 124–129.

[17] Paul Ekman and Wallace V. Friesen. 1978. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto.

[18] U. Engelke, D. P. Darcy, G. H. Mulliken, S. Bosse, M. G. Martini, S. Arndt, J. Antons, K. Y. Chan, N. Ramzan, and K. Brunnström. 2017. Psychophysiology-based QoE assessment: A survey. *IEEE J. Select. Topics Signal Process.* 11, 1 (2017), 6–21.

[19] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *21st ACM International Conference on Multimedia*. ACM, 835–838.

[20] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. OpenSMILE: The Munich versatile and fast open-source audio feature extractor. In *18th ACM International Conference on Multimedia*. ACM, 1459–1462.

[21] Boni García, Francisco Gortázar, Micael Gallego, and Andrew Hines. 2020. Assessment of QoE for video and audio in WebRTC applications using full-reference models. *Electronics* 9, 3 (2020). DOI: https://doi.org/10.3390/electronics9030462

[22] Vishal Gupta and Vinod Kumar Sharma. 2019. Skin typing: Fitzpatrick grading and others. *Clin. Dermatol.* 37, 5 (2019), 430–436. DOI: https://doi.org/10.1016/j.clindermatol.2019.07.010

[23] Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. 1322–1328.

[24] ITU. 1996. Methods for subjective determination of transmission quality. Recommendation ITU-T P.800. https://www.itu.int/rec/T-REC-P.800-199608-I

[25] ITU. 2015. The E-model: A computational model for use in transmission planning. Recommendation ITU-T G.107. https://www.itu.int/rec/T-REC-G.107-201506-I/en

[26] Sofiene Jelassi, Gerardo Rubino, Hugh Melvin, Habib Youssef, and Guy Pujolle. 2012. Quality of experience of VoIP service: A survey of assessment approaches and open issues. *IEEE Commun. Surv. Tutor.* 14, 2 (2012), 491–513. DOI: https://doi.org/10.1109/SURV.2011.120811.00063

[27] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. 2019. Speech emotion recognition using deep learning techniques: A review. *IEEE Access* 7 (2019), 117327–117345.

[28] N. Khan and M. G. Martini. 2016. QoE-driven multi-user scheduling and rate adaptation with reduced cross-layer signaling for scalable video streaming over LTE wireless systems. *EURASIP Journal on Wireless Communications and Networking*, 93 (2016). DOI: https://doi.org/10.1186/s13638-016-0584-6

[29] Shashidhar G. Koolagudi and K. Sreenivasa Rao. 2012. Emotion recognition from speech: A review. *Int. J. Speech Technol.* 15, 2 (2012), 99–117.

[30] Satish Kumar, Rajesh Devaraj, Arnab Sarkar, and Arijit Sur. 2019. Client-side QoE management for SVC video streaming: An FSM supported design approach. *IEEE Trans. Netw. Serv. Manag.* 16, 3 (2019), 1113–1126. DOI: https://doi.org/10.1109/TNSM.2019.2926720

[31] Takio Kurita. 2019. *Principal Component Analysis (PCA)*. Springer International Publishing, Cham, 1–4. DOI: https://doi.org/10.1007/978-3-030-03243-2_649-1

[32] Patrick Le Callet, Sebastian Möller, and Andrew Perkis. 2012. *Qualinet White Paper on Definitions of Quality of Experience*. European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland, Version 1.2, March 2013.

[33] Shan Li and Weihong Deng. 2022. Deep facial expression recognition: A survey. *IEEE Trans. Affect. Comput.* 13, 3 (2022), 1195–1215.

[34] Wenkai Liang, Yan Wu, Ming Li, and Yice Cao. 2022. Adaptive multiple kernel fusion model using spatial-statistical information for high resolution SAR image classification. *Neurocomputing* 492 (2022), 382–395.

[35] Xiwen Liu, Xiaoming Tao, Mai Xu, Yafeng Zhan, and Jianhua Lu. 2020. An EEG-based study on perception of video distortion under various content motion conditions. *IEEE Trans. Multim.* 22, 4 (2020), 949–960.
[36] Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in Python. In *14th Python in Science Conference*. 18–25.
[37] Katrien De Moor, Filippo Mazza, Isabelle Hupont, Miguel Ríos Quintero, Toni Mäki, and Martín Varela. 2014. Chamber QoE: A multi-instrumental approach to explore affective aspects in relation to quality of experience. In *Human Vision and Electronic Imaging XIX*, Bernice E. Rogowitz, Thrasyvoulos N. Pappas, and Huib de Ridder (Eds.), Vol. 9014. SPIE, 204 – 217.
[38] Mumtaz Begum Mustafa, Mansoor A. M. Yusoof, Zuraidah M. Don, and Mehdi Malekzadeh. 2018. Speech emotion recognition research: An analysis of research focus. *Int. J. Speech Technol.* 21, 1 (2018), 137–156.
[39] Simone Porcu, Alessandro Floris, and Luigi Atzori. 2022. Analysis of the quality of remote working experience: a speech-based approach. *Qual. User Exper.* 7, 1 (2022).
[40] S. Porcu, A. Floris, J. N. Voigt-Antons, L. Atzori, and S. Möller. 2020. Estimation of the quality of experience during video streaming from facial expression and gaze direction. *IEEE Trans. Netw. Serv. Manag.* 17, 4 (2020), 2702–2716. DOI: https://doi.org/10.1109/TNSM.2020.3018303
[41] Jannik Rößler, Jiachen Sun, and Peter Gloor. 2021. Reducing videoconferencing fatigue through facial emotion recognition. *Fut. Internet* 13, 5 (2021).
[42] Tim Sainburg. 2019. *timsainb/noisereduce: v1.0*. DOI: https://doi.org/10.5281/zenodo.3243139
[43] Tim Sainburg, Marvin Thielk, and Timothy Q. Gentner. 2020. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS Computat. Biol.* 16, 10 (2020), e1008228.
[44] Sandvine. 2023. 2023 Global Internet Phenomena Report. Retrieved from https://www.sandvine.com/global-internet-phenomena-report-2023
[45] Marwin Schmitt, Judith Redi, Dick Bulterman, and Pablo S. Cesar. 2018. Towards individual QoE for multiparty videoconferencing. *IEEE Trans. Multim.* 20, 7 (2018), 1781–1795.
[46] Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K. Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini. 2016. The INTERSPEECH 2016 computational paralinguistics challenge: deception, sincerity & native language. In *Proc. Interspeech 2016*, 2001-2005. DOI: 10.21437/Interspeech.2016-129
[47] Michael James Scott, Sharath Chandra Guntuku, Yang Huan, Weisi Lin, and Gheorghita Ghinea. 2015. Modelling human factors in perceptual multimedia quality: On the role of personality and culture. In *23rd ACM International Conference on Multimedia*. ACM, 481–490.
[48] Michael Seufert, Sebastian Egger, Martin Slanina, Thomas Zinner, Tobias Hoßfeld, and Phuoc Tran-Gia. 2015. A survey on quality of experience of HTTP adaptive streaming. *IEEE Commun. Surv. Tutor.* 17, 1 (2015), 469–492.
[49] Lea Skorin-Kapov, Martín Varela, Tobias Hoßfeld, and Kuan-Ta Chen. 2018. A survey of emerging concepts and challenges for QoE management of multimedia services. *ACM Trans. Multim. Comput. Commun. Appl.* 14, 2s, Article 29 (May 2018), 29 pages. DOI: https://doi.org/10.1145/3176648
[50] Lohic Fotio Tiotsop, Antonio Servetti, Marcus Barkowsky, and Enrico Masala. 2022. Regularized maximum likelihood estimation of the subjective quality from noisy individual ratings. In *14th International Conference on Quality of Multimedia Experience (QoMEX'22)*. 1–4.
[51] Christos Tsiaras, Manuel Rösch, and Burkhard Stiller. 2015. VoIP-based calibration of the DQX model. In *IFIP Network Conference (IFIP Networking'15)*. 1–9. DOI: https://doi.org/10.1109/IFIPNetworking.2015.7145309
[52] Dimitris Tsolkas, Eirini Liotou, Nikos Passas, and Lazaros Merakos. 2017. A survey on parametric QoE estimation for popular services. *J. Netw. Comput. Applic.* 77, 1 (2017), 1–17.
[53] Sara Vlahovic, Mirko Suznjevic, and Lea Skorin-Kapov. 2022. A survey of challenges and methods for quality of experience assessment of interactive VR applications. *J. Multimod. User Interf.* (04 2022), 1–35.
[54] Dunja Vučić and Lea Skorin-Kapov. 2019. The impact of packet loss and Google congestion control on QoE for WebRTC-based mobile multiparty audiovisual telemeetings. In *MultiMedia Modeling*, Ioannis Kompatsiaris, Benoit Huet, Vasileios Mezaris, Cathal Gurrin, Wen-Huang Cheng, and Stefanos Vrochidis (Eds.). Springer International Publishing, Cham, 459–470.
[55] Dunja Vučić and Lea Skorin-Kapov. 2020. QoE assessment of mobile multiparty audiovisual telemeetings. *IEEE Access* 8 (2020), 107669–107684. DOI: https://doi.org/10.1109/ACCESS.2020.3000467
[56] Dunja Vučić, Lea Skorin-Kapov, and Mirko Sužnjević. 2016. The impact of bandwidth limitations and video resolution size on QoE for WebRTC-based mobile multi-party video conferencing. In *5th ISCA/DEGA Workshop on Perceptual Quality of Systems (PQS'16)*. 59–63.
[57] Chaowei Wang, Huadi Zhu, and Ming Li. 2023. SpeechQoE: A novel personalized QoE assessment model for voice services via speech sensing. In *20th ACM Conference on Embedded Networked Sensor Systems*. ACM, 305–319.
[58] Lipo Wang. 2005. *Support Vector Machines: Theory and Applications*. Vol. 177. Springer Science & Business Media.

[59] Felix Weninger, Florian Eyben, Björn Schuller, Marcello Mortillaro, and Klaus Scherer. 2013. On the acoustics of emotion in audio: What speech, music, and sound have in common. *Front. Psychol.* 4 (2013).
[60] E. Wood, T. Baltrušaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling. 2015. Rendering of eyes for eye-shape registration and gaze estimation. In *IEEE International Conference on Computer Vision (ICCV'15)*. 3756–3764.
[61] Huan Yan, Tzu-Heng Lin, Ming Zeng, Jing Wu, Yong Li, and Depeng Jin. 2021. Discovering usage patterns of mobile video service in the cellular networks. *IEEE Trans. Netw. Serv. Manag.* 18, 2 (2021), 1789–1802. DOI: https://doi.org/10.1109/TNSM.2020.3043482

Received 5 July 2023; revised 25 October 2023; accepted 17 December 2023