AI-based media coding standards

Andrea Basso, Paolo Ribeca, Marina Bosi, Niccolo Pretto, Gérard Chollet, Michelangelo Guarise, Miran Choi, Leonardo Chiariglione, Roberto Iacoviello, Franesco Banterle, et al.

AI-Based Media Coding Standards

By Andrea Basso, Paolo Ribeca, Marina Bosi, Niccolò Pretto, Gérard Chollet, Michelangelo Guarise, Miran Choi, Leonardo Chiariglione, Roberto Iacoviello, Franesco Banterle, Alessandro Artusi, Francesco Gissi, Attilio Fiandrotti, Giovanni Ballocca, Marco Mazzaglia, and Scott Moskowitz

Abstract

Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) is the first standard organization to develop data coding standards that have artificial intelligence (AI) as their core technology. MPAI believes that universally accessible standards for AI-based data coding can have the same positive effects on AI as standards had on digital media. Elementary components of MPAI standards–AI modules (AIMs)–expose standard interfaces for operation in a standard AI framework (AIF). As their performance may depend on the technologies used, MPAI expects that competing developers providing AIMs will promote horizontal markets of AI solutions that build on and further promote AI innovation. Finally, the MPAI framework licences (FWLs) provide guidelines to intellectual property right (IPR) holders facilitating the availability of compatible licenses to standard users.

Keywords

Artificial Intelligence (AI), audio enhancement, IP license, standards, video compression.

MPAI has adopted an innovative standard development approach relying on components called *AI modules (AIMs)*; their aggregation is executed in AI frameworks (AIFs). An MPAI standard normatively defines the functionality and input/output data formats of an AIM, the topology of the AIMs in the AIF, and the functionality and input/output data of the AIF implementing each use case.

Introduction

Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) is an international, unaffiliated standard organization. Its mission is to promote the efficient use of any kind of data by standardizing coding technologies, especially those based on artificial intelligence (AI), and to facilitate the integration of such data coding components into complete systems.[1]

Two-fold motivations drove the establishment of MPAI. The first is rooted in the belief that, while data processing technologies have enabled massive use of digital technologies benefitting industry players and consumers alike, the propulsive force of those technologies is reaching its limits. On the contrary, the scope, performance, and applicability domain of technologies such as AI is growing. The second motivation is the recognition of serious signs of wear and tear in the system that has governed so far the transfer of innovation to standards and products/services. A revision of past practices is needed to guarantee the irreplaceable role of intellectual property (IP) and to iron out obstacles currently blocking the path to its dissemination.

To perform its mission, MPAI has worked out a process that allows it to reach out to those in need of a solution and those possessing technologies that can satisfy the former. The process is designed to be community-centric, especially during the phase where needs and functional requirements are identified; to allow all MPAI members to participate in the development of standards fulfilling such requirements; and to delegate the development of IP guidelines to members who have elected to perform the task (principal members).

MPAI has adopted an innovative standard development approach relying on components called *AI modules (AIMs)*; their aggregation is executed in AI frameworks (AIFs). An MPAI standard normatively defines the functionality and input/output data formats of an AIM, the topology of the AIMs in the AIF, and the functionality and input/output data of the AIF implementing each use case.

MPAI is well aware of the impact that AI technologies embedded in some of its standards will have on everyone's day-to-day life. Therefore, it has developed guidelines to test the conformance of AIMs and AIFs implementing use cases and to assess the performance of AIMs and AIFs. In MPAI parlance, performance is a synonym for reliability, robustness, and fairness.

Digital Object Identifier 10.5594/JMI.2022.3160793
Date of publication: XX XXX XXXX

This article introduces standards working at different levels of maturity: the first standard called *AIF (MPAI-AIF)*, the second and third standards called *context-based audio enhancement (MPAI-CAE)* and *multimodal conversation (MPAI-MMC)*, respectively, AI-enhanced video coding (MPAI-EVC), and server-based predictive multiplayer gaming (MPAI-SPG). Finally, MPAI's framework licence (FWL) approach to standardization is presented.

AI Framework

The reference model of MPAI AIF is depicted in **Fig. 1**. The scope of the MPAI-AIF standard[2] is to provide a framework that allows easy interconnection and interoperability of AI and data processing technologies implemented both in hardware (HW) and software (SW), when they are encapsulated in modules with standard interfaces called *AIMs*. MPAI-AIF adopts the philosophy of components-based development (CBD), enabling the reuse of independent components (AIMs) into systems.

In MPAI-AIF, the AIMs communicate with one another via standardized interfaces; they specify the services that other components can use and how such use should happen. As a result, there is no need to know the details of a specific implementation of an AIM to use it.

MPAI-AIF standard specifies architecture, interfaces, protocols, and APIs of an AIF capable of executing AI-based products, services, and applications. MPAI-AIF has the following main features: 1) it is component-based, 2) it defines the interfaces between its components, 3) it is secure as its components operate in a trusted zone, 4) it supports mixed hardware–software implementations, 5) it supports distributed and local execution environments, 6) it natively supports machine learning functionality, and 7) it supports the operation of proximity-based distributed AIFs. The extensive usage of expressive metadata minimizes the issues related to differences in implementations. The current version of MPAI-AIF has been developed by the MPAI AIF Development Committee (AIF-DC). MPAI-AIF supports event-based as well as signal-based signaling and provides mechanisms for resource management that are particularly useful in resource-constrained environments. Resource policies are enforced at both AIW and AIM levels. Finally, MPAI AIFs interface via a specific API to the MPAI store; the latter is a distribution system of AIW and AIM implementations from which a compliant AIF implementation can download the AIWs and AIMs it needs.

Enhanced Audio Applications

A specific MPAI area of work, the MPAI-CAE, is showing tremendous potential for audio applications.[3] MPAI-CAE applies context-enhanced information to the input audio content to deliver the audio output via the most appropriate protocol. Four MPAI-CAE use cases have already been standardized.[4]

- *Audio Recording Preservation (ARP):* This use case provides a workflow for managing the overall process (analysis, restoration, etc.) to preserve and access open-reel audiotapes.
- *Emotion-Enhanced Speech (EES):* This use case implements a user-friendly system control interface that generates speech with various levels of emotions.
- *Speech Restoration System (SRS):* This use case aims to restore damaged parts of a speech by synthesizing it through neural networks speech models.
- *Enhanced Audio-Conference Experience (EAE):* This use case improves the auditory experience in an audio conference setting.

In this section, we will focus on the MPAI-ARP use case (**Fig. 2**), and we will describe a detailed approach to the preservation of analog audiotapes.[5]

A set of irregularities (e.g., damages in the carriers, splices, and marks) can be identified on the video of the tape flowing in the playback head of the tape recorder during the digitization process.[6] The video analyzer detects the significant frames from the preservation audio-visual file comparing consecutive frames. The detection is made in two areas: on the reading head and under the Pinch roller (**Fig. 3**). Some other valuable irregularities that cannot be detected by the video can be detected in the audio extracted from an open-reel

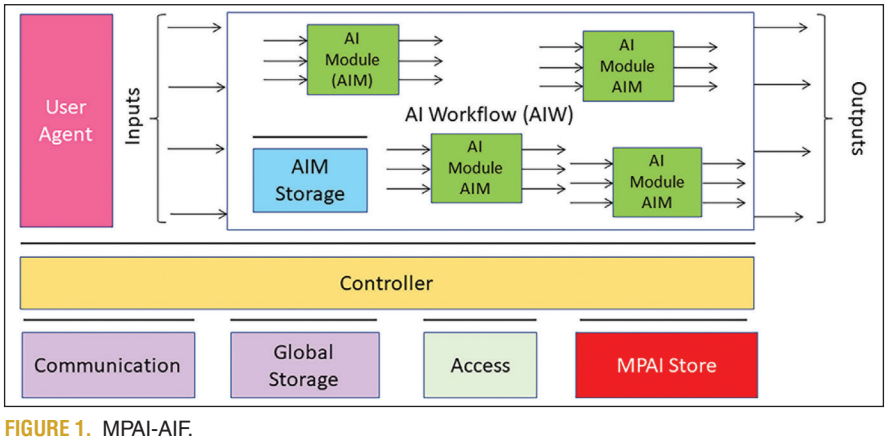


FIGURE 1. MPAI-AIF.

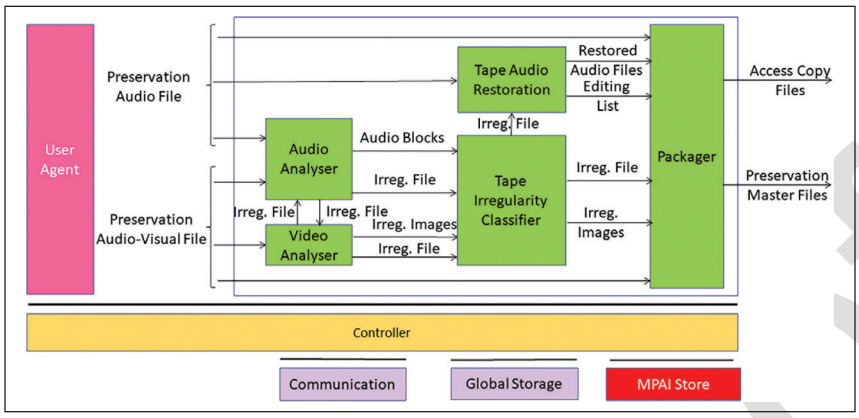


FIGURE 2. MPAI-CAE ARP workflow.

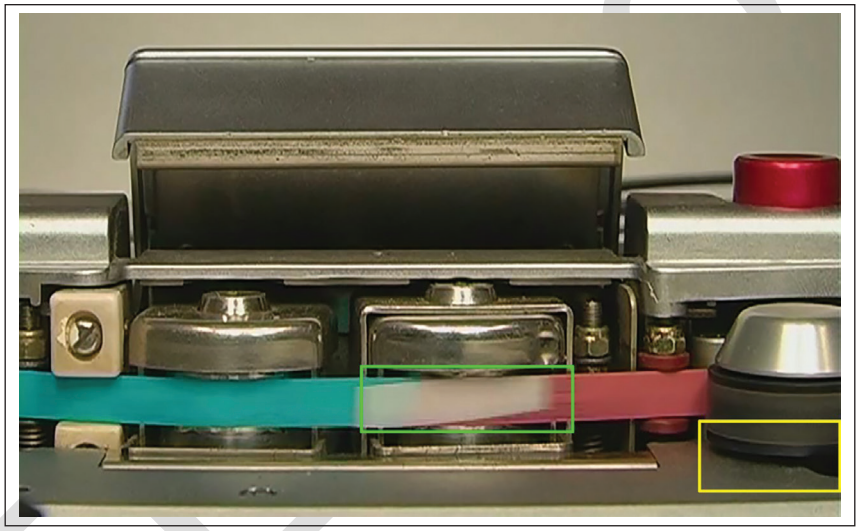


FIGURE 3. Area of the reading head of the tape recorder (green box), and area under the pinch-roller (yellow box).

audiotape (preservation audio file), for example, audio effects derived from noise generated by the play, pause, and stop buttons during the recordings or by changes in the configuration of the tape recorder. The extracted frames and audio segments from video analyzer and audio analyzer, respectively, are classified by the musicological classifier that selects only the most relevant irregularities. These irregularities are included in the preservation master file (set of files for long-term preservation) and used in the tape audio restoration. The latter corrects potential errors occurring during the transfer of the audio signal from the analog carrier to the digital file (e.g., speed and equalization configurations of the tape recorder; see Ref. 7), to make the audio content ready to be deployed. All the data output from the tape irregularity classifier, tape audio restoration, as well as the inputs (preservation audio and preservation audio-visual files) to the ARP are managed by the packager. The restored audio version is inserted in the access copy file, which is used for accessing the audio content. The packager also creates the preservation master files, with the original inputs and the irregularities analysis.

Multimodal Conversation Applications

MPAI-MMC aims to enable human–machine conversation that emulates human–human conversation in completeness and intensity by using AI. Currently, the MPAI-MMC standard[8] includes five use cases: conversation with emotion (CWE), multimodal question answering (MQA), unidirectional speech translation (UST), bidirectional speech translation (BST), and one-to-many speech translation (MST).

In CWE (**Fig. 4**), the human side of the dialogue includes speech, video, and possibly text, while the machine responds with a synthesized voice, text, and video with an animated face. MPAI-MMC standard focuses on the description of output formats of each AIM. An important contribution concerns the list of main elements including the emotion element of the language understanding AIM, which formally describes the output of the module.

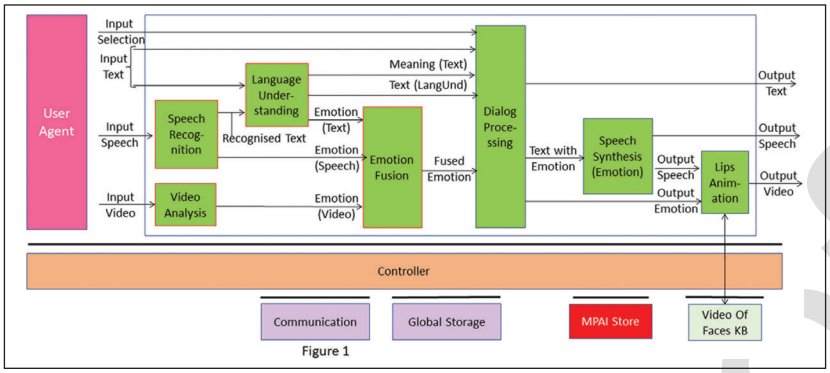


FIGURE 4. MPAI-MMC: CWEs

In MQA, a human requests information in natural language about a displayed object, and the machine responds with synthesized speech. MPAI worked on a formal description of output formats for each AIM in the workflow. Additionally, the question analysis AIM produces the intention of the question of the user by several elements including topic, focus, answer types, and domains of the question. The intention, together with the meaning of the question produced by the language understanding AIM, is used by the question answering AIM as the input to generate a reply to the question.

In UST, BST, and MST, a human-uttered sentence is translated by a machine using a synthesized voice that mimics the original speaker's speech features. The MPAI-MMC standard focuses on a formal description concerning the input and output of main AIMs. An important contribution of the standard concerns a list of speech features for the speech feature extraction AIM.

Utilizing an AI powerful speech analysis technology, in conjunction with multiple translation databases, personalized speech translation system aims to remove the language barrier found in traditional instant messaging programs, so you can have friends all over the world and communicate with ease. This system implementation leverages cloud hosted services and handles all translation/voice morphing on the server side to increase speed and minimize local app size. This system features an intuitive interface that allows users to easily select the appropriate language for each contact in their international contacts list and automatically applies that language translation function to each chat window. Users will have access to a face-to-face conversational capability for local translations while traveling, or favorite contact chat to talk to friends they have connected to around the world.

Unlike other traditional speech translation products, which utilize computer-generated voices, this next-generation proprietary technology allows users to talk to their contacts in their own voice. So now, you can actually hear what your friend would sound like if they were actually there in person and speaking your language. Also, unlike other standard translation products, this new technology implementation allows users to retain their original vocal inflection through the translation process, resulting in a more natural speech pattern.

Possible uses are: 1) voice can be morphed into many languages, 2) breaking down cross-language communication barriers, 3) online dating, same interest group conversations around the globe, 4) inviting friends from Facebook, phone or email address book, or Twitter, 5) chatting with foreign relatives, 6) keeping in touch with foreign business associates, and 7) online gamers who do not want to use limited in-game preset phrase translators.

Video Coding Applications

To face the challenges of offering more efficient video compression solutions, research effort is focused on radical changes to the classic block-based hybrid coding framework. AI approaches can play an important role in achieving this goal.

MPAI has recently carried out a literature survey on AI-based video coding. The result suggests that a performance enhancement of about 30% can be achieved. Therefore, MPAI is investigating whether it is possible to improve the performance of the essential video coding (EVC) modified by enhancing/replacing existing video coding tools, with AI tools keeping complexity increase to an acceptable level.[9] **Figure 5** describes the reference codec architecture. The red circles represent the data processing block candidates for enhancement/replacement with AI tools.

Currently, the group is working on two tools: intra prediction and super resolution.

Intra Prediction

We recast the task of generating an intra predictor for intra-coded CUs as a hole inpainting problem using the context encoder[10] depicted in **Fig. 6**. For the sake of simplicity, in the following, we exemplify for the case of a 32 × 32 predictor, yet the same method holds also for other CU sizes. The autoencoder receives in input a 64 × 64 patch representing the context available at the decoder (D0, D1, D2), whereas the 32 × 32 bottom-right corner (P3) is filled with black pixels, to represent the

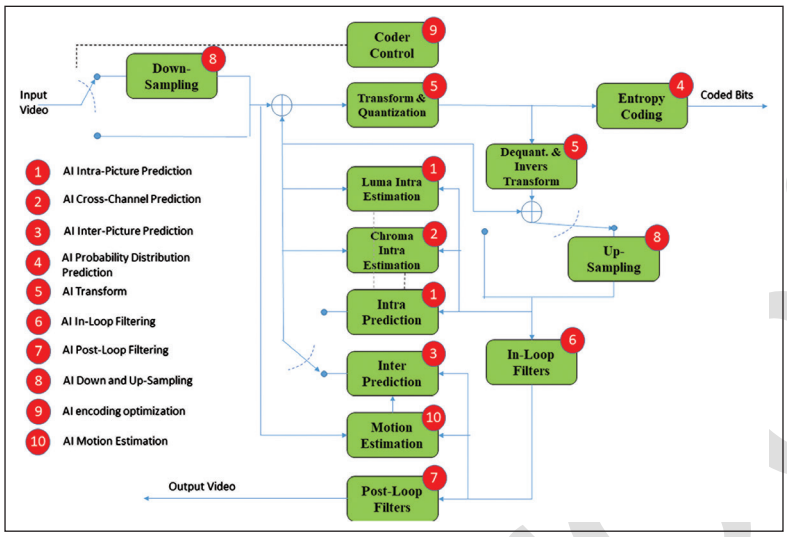


FIGURE 5. Schema of a hybrid video codec and encoding tools suitable for AI replacement.

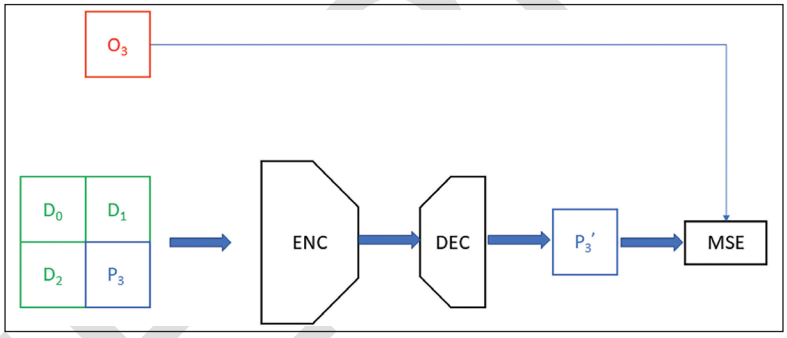


FIGURE 6. MPAI-EVC intra-predictor generation via context encoder.

area to inpaint (to predict, in our case). The autoencoder is trained to generate a 32 × 32 predictor (P3') minimizing the error between its output P3' and the original 32 × 32 block O3. With respect to similar works, for example, Ref. 11, in this case, we generate a new predictor rather than enhancing the encoder-generated predictor.

Toward training our network, we have set up a dataset of 1.5M patches of 32 × 32 predictors and 6M patches of 16 × 16 predictors extracted from the AROD dataset. The original-decoded-predicted (ODP) data structure has been used. These patches are arranged into 64 × 192 pixels images, constituted by three 64 × 64 pixels images, where the lower right-hand side patch is the codec-processed patch, while the surrounding patches are the neighborhood information (causal context).

The trained autoencoder is wrapped by a server process that interfaces with the EVC encoder via a UDP socket: this allows for easy experimentation with different frameworks (PyTorch, TensorFlow, Keras, etc.) with no need to modify the encoder. For each intra-coded CU, the EVC encoder sends to the server the 64 × 64 decoded context (D0, D1, D2, and P3). The server feeds the context to the autoencoder and returns the 32 × 32 autoencoder output P3' to the encoder. For these preliminary experiments, the EVC encoder simply replaces the mode 0 (DC) predictor with the autoencoder-generated predictor and the encoding proceeds as usual (no modification is required to the standard EVC mode signaling scheme). We experimented encoding the first frame of the standard JVET CTC sequences and the results are shown in **Table 1** (32 × 32 and 16 × 16 DC predictors only are autoencoder-generated, 8 × 8 and 4 × 4 predictors are the normal EVC DC predictors for the time being). This experiment report gains close to 2%, especially for sequences above 720p (top BD-Rate reduction is −6.20% for the 1080p BasketballDrive sequence) and at high QPs, while the bitstream is decodable at the receiver side.

Finally, we experiment with a perspective Oracle scheme where the DC intra predictor is replaced by the autoencoder-generated predictor only if that reduces the residual rate. While this method does not account for the signaling cost, it represents an upper boundary to the case where the autoencoder-generated predictor complements the five standard EVC intra modes. **Table 2** shows potential gains over 3%, prompting further experimentation with a proper sixth intra predictor and extending the scheme to 8 × 8 and 4 × 4 CUs.

Table 1. Autoencoder-generated intra predictor, decodable bitstream, QP 22–32 (left), 22–47 (center), 32–47 (right).

Sequence	BDRate	BDPSNR	Sequence	BDRate	BDPSNR	Sequence	BDRate	BDPSNR
Class A	–2.52	0.07	Class A	–3.08	0.10	Class A	–2.78	0.11
Class B	–2.39	0.10	Class B	–3.03	0.13	Class B	–3.29	0.14
Class C	–0.81	0.05	Class C	–1.05	0.06	Class C	–1.11	0.05
Class D	–0.18	0.01	Class D	–0.30	0.01	Class D	–0.52	0.02
Class E	–2.43	0.13	Class E	–3.20	0.18	Class E	–3.86	0.21
Class F	–0.46	0.03	Class F	–0.33	0.00	Class F	–0.35	0.03
AVG	**–1.47**	**0.07**	**AVG**	**–1.83**	**0.08**	**AVG**	**–1.99**	**0.09**

Table 2. Autoencoder-generated intra predictor, Oracle mode, QP 22–32 (left), 22–47 (center), and 32–47 (right).

Sequence	BDRate	BDPSNR	Sequence	BDRate	BDPSNR	Sequence	BDRate	BDPSNR
Class A	–3.33	0.09	Class A	–4.23	0.14	Class A	–4.42	0.17
Class B	–3.05	0.13	Class B	–4.10	0.17	Class B	–4.94	0.21
Class C	–1.05	0.07	Class C	–1.50	0.08	Class C	–2.02	0.09
Class D	–0.30	0.02	Class D	–0.50	0.03	Class D	–0.93	0.04
Class E	–3.08	0.17	Class E	–3.88	0.22	Class E	–4.84	0.26
Class F	–0.76	0.05	Class F	–0.87	0.05	Class F	–1.13	0.07
AVG	**–1.93**	**0.09**	**AVG**	**-2.51**	**0.12**	**AVG**	**–3.05**	**0.14**

Super Resolution

We added a super-resolution step after the postloop filters (**Fig. 5**, block 7) to improve the overall performance of the EVC decoding system. To achieve this, we have adopted a well-known deep learning algorithm, used in super-resolution studies such as the Densely Residual Laplacian Network.[12] This architecture is employed as an up-sampler whenever the input sequence, to the decoding system, has been down-sampled. Due to the complexity, in terms of memory and computational costs, of the deep-learning approach, we have designed an efficient training strategy. This has been solved by subdividing the input frame in crops and developing suitable training and validation sets based on the cropping strategy adopted. To achieve this, we have employed two strategies, both based on the entropy information of the input frame. This is calculated by estimating at each pixel position (i, j) the entropy of the pixel values within a 2-dim region centered at (i, j). The first strategy uses a random crop if, and only if, its average entropy exceeds a given threshold. The second strategy selects n crops, of the same size, from the total crops available in each frame. This is based on the sampling technique applied to the entropy values distribution of all crops in each frame. Particular attention needs to be given to the right combination of the crop and batch sizes in such a way that a trade-off with respect to GPU memory consumptions can be achieved. **Table 3** shows the tested combination.

To train the network, we have built three versions of the "Images 4K" Kaggle dataset with different resolutions, namely 4K-3840 × 2160, HD-1920 × 1080, and SD-960 × 540. Then, for each of them, the data have been encoded and decoded using four different values of quantization parameters (QPs), that is, 15, 30, 37, and 45, respectively, with deblocking option enabled. The training procedure has been built on a use case with upscaling factor of 2 from SD to HD resolution.

As described earlier, to efficiently perform the training tasks, two cropping strategies have been employed. The best hyperparameters and parameters identified and then used during the training phase were the following: learning rate (lr) $10e$-5, batch size 6, and two samples for image for the cropping strategy based on importance sampling, epochs 50, the resolution of the crop input was 128 × 128, the crop output was 256 × 256, while the dataset used was the one with deblocking option activated. The mean square error (MSE) metric was used as a loss function. The results using QP 15 from SD to HD for both cropping strategies are shown in **Fig. 7**, and similar results are obtained for the other QPs.

Based on the training and validation results, the important sampling-based cropping strategy has been showing better generalization capabilities, that is, validation and training curve have similar results. We experimented up-sampling, on all the sequence frames

for seven test sequences from the Kaggle dataset and the calculated BD-Rate, shown in **Table 4** for all the interpolated QPs.

This experiment report gains close to 4% (top BD-rate reduction is –22% for the Talk show sequence).

The BD-Rate calculation has been carried out using a "generalized Bjontegaard BD-PSNR approach." Bjontegaard's metric evaluation requires measurement of bitrate and PSNR at (at least) four QP values for both the reference curve and that under evaluation. The integral of the curves divided by the integration interval is calculated and the difference between the two values gives the average difference.

In our case, the reference curves are the BD-Rate curves made of the data points determined coding the HD sequences. The curves to be evaluated are made of SD bitrates and PSNR values for the upscaled sequences. To enable a fair comparison between the two curves (the data for upscaled SD content and reference HD content at the same QP will have dramatically different bitrate domain intervals), we fit the reference curves over 30 QP values, to be able to interpolate the "reference bitrate" at the PSNR value of the upscaled sequence.

Server-Based Predictive Multiplayer Gaming

Online gaming is a large and growing industry with billions of users. However, two problems beset this industry: network packet loss and cheating systems. MPAI-SPG is providing a solution for both problems: minimizing audio visual and gaming discontinuities caused by high latency or packet loss during realtime online gaming sessions and designing a system that intercepts anomalous (cheating) situations. Currently, prediction is done exclusively on the ongoing data of the game and is used by the clients to achieve a smooth gaming experience that does not depend only on the arrival of packets from the game server.

With MPAI-SPG (see reference model in **Fig. 8**), if the information of a client is missing, the data collected from the clients involved in a particular game is fed to an AI-based system that predicts the moves of the client whose data is missing. The neural networks of MPAI-SPG are trained using all the games that have been played to arrive at the most accurate prediction of the missing parts. In the case of anticheating, the neural network suggests to the online game server what the current state of the system might be according to the input data and the previous game state. The online game server compares the information coming from MPAI-SPG with its own. In the event of a significant deviation, the online game server takes action against a particular player who is sending altered information. The system will appear as a plug-in to be integrated into game engines according to the prevailing philosophy available in the Game Industry. Additional information and updates can be found in Ref. 13.

Table 3. Tested combinations of crop and batch sizes.

Input size	Batch size	GPU memory usage (GiB)
48	16	9.7
72	16	18.0
96	10	19.7
128	6	20.0

Framework Licenses

Challenges in how to best assure competition for standards decisions were an underlying concern in developing MPAI framework license[s] (FWL). The nature of AI differs from traditional licensing agreements in software and technology generally. One concept, the performance warranty, seeks to provide guidance regarding how the software performs vs. its accompanying documentation and specification. As discussed herein, AI and MPAI's

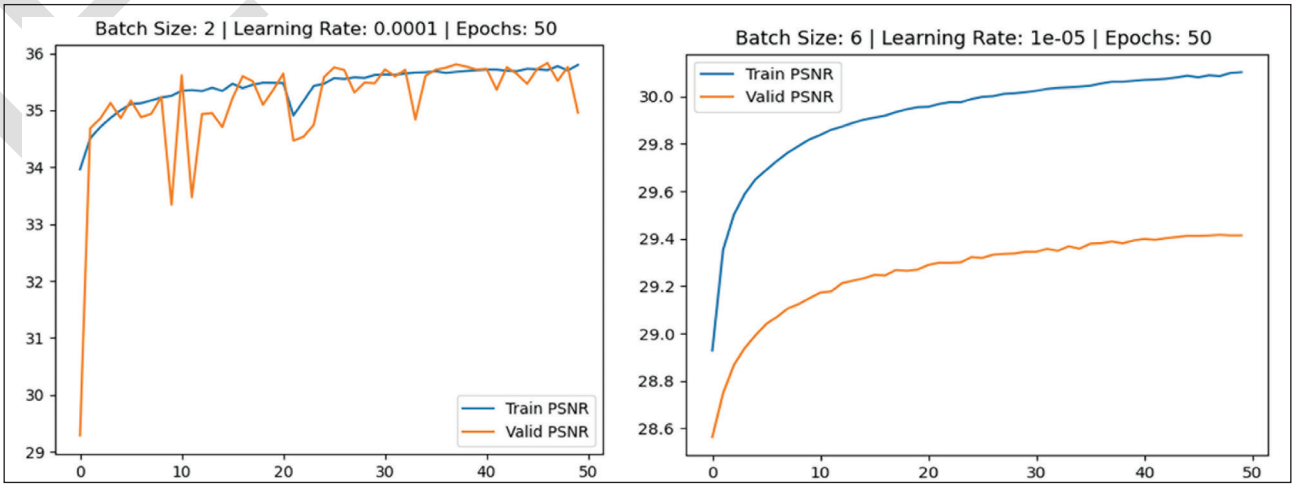


FIGURE 7. Training and validation results, as PSNR metric, using the two cropping strategies: (left) random cropping and (right) importance sampling-based cropping.

Table 4. BD-Rate computed for all the seven test sequences versus the ground truth, as average on all the sequence frames.

Sequence	Rome_1	Rome_2	Talk_	Rush_	Diego_	Crowd_	Parkjoy	Average
BDRate QP Averaged [%]	0.1902	–18.8094	–21.7534	4.9017	8.1107	–1.2430	–4.2977	–4.701

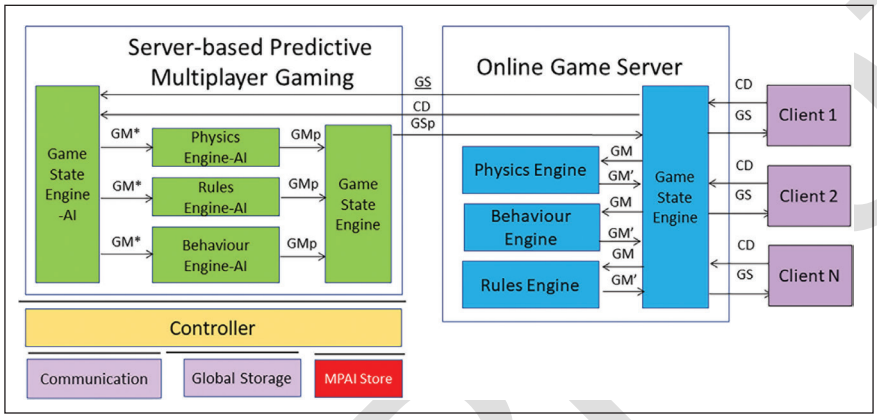


FIGURE 8. MPAI-SPG reference model.

standards efforts intend to provide greater flexibility given the changing nature of AI. Specifically, focusing on the desired outcomes by the parties making a particular claim or attesting to the same is of value to innovators and implementers alike.

In contrast with standards such as Moving Picture Experts Group (MPEG), MPAI places the business model part of a license into the process between defining requirements and calls for technologies. In other words, the MPAI approach is not a complete solution to providing timely licenses to data compression and representation standards, MPAI's FWLs facilitate at least one beneficial path forward. Significant differences in business models between adopters no longer dominate the standard discussions. Evaluation of functional and commercial requirements need not be undermined; FWLs instruct use cases and conditions, not specific cost.

FWLs thus serve to overcome the uncertainties associated with fair, reasonable, and nondiscriminatory (i.e., FRAND), and even provide transparency over standard essential patent (i.e., SEP) by establishing, to the extent possible, claims at the onset of the standardization process. In accordance with competition law and practice, MPAI replaced FRAND with FWLs developed by MPAI members with IP expertise as voluntary terms of use lacking any monetary consideration. In some cases, a range of values may provide a high and low estimate for understanding potential costs. FWL serves to provide guidance on business models for intellectual property rights (IPRs) holders by eliminating specific values such as percentages, royalty rates, cash values, dates, and simply proffering a cap for the cost of a given license based on comparable costs of similar standards and underlying technologies.

As MPAI develops standards and market adoption begins, FWLs will guide licensing of MPAI standard-compliant technologies: FWLs will close the expectation gap between the innovators and licensors and the market adopters and implementers.

Conclusion

MPAI is a young organization that leverages a decade-long data processing-based experience in digital media compression to tackle the wider field of data coding. Albeit mainly focused on the use of AI, the approach is at the same time technology-agnostic. By adopting the philosophical definition of data coding as the transformation of data from a given representation to an equivalent one more suited to a specific application, MPAI can address disparate areas of data coding standardization with a unified approach. The definition of an AI system as a network of AIMs executed within an AIF, as opposed to a monolithic black box, allows MPAI to define explainability and reliability levels for the AI systems conforming to its standards. The article has given an overview of some of the MPAI media-coding-oriented standards. However, MPAI has already developed and approved several more standards in different domains (such as company performance prediction) and is addressing several additional use cases such as mixer-reality collaborative spaces, neural network watermarking, and others. Each MPAI activity is thoroughly described at https://mpai.community/.

Acknowledgments

The work of Marina Bosi was supported in part by the Center for Computer Research in Music and Acoustics (CCRMA), Stanford University. The work of Niccolò

Pretto was supported in part by the IT4aREC Project, granted by the Department of Information Engineering, University of Padova. The work of Miran Choi was supported by the Institute for Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean Government (MSIT) (Development of Knowledge Evolutionary WiseQA Platform Technology for Human Knowledge Augmented Services) under Grant 2013-2-00131. The work of Roberto Iacoviello was supported in part by the European Union's Horizon 2020 Research and Innovation programme, AI4MEDIA Project, under Grant 951911. The work of Alessandro Artusi was supported in part by the European Union's Horizon 2020 Research and Innovation Programme under Grant 73957 and in part by the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation, and Digital Policy.

References

1. A. Artusi et al., "Towards Pervasive and Trustworthy Artificial Intelligence: How Standards Can Put a Great Technology at the Service of Humankind," Amazon Publishing: Geneva, Switzerland, 2021.
2. MPAI Community, "Artificial Intelligence Framework (MPAI-AIF) V1. Accessed: Apr. 7, 2022. [Online]. Available: https://bit.ly/3t3SNDT
3. M. Bosi et al., "Sound and Music Computing Using AI: Designing a Standard," *Proc. 18th Sound Music Comput. Conf. (SMC'21)*, Jun. 2021.
4. MPAI Community, "Context-Based Audio Enhancement (MPAI-CAE) V1.1. Accessed: Apr. 7, 2022. [Online]. Available: https://bit.ly/3ii0h01
5. L. Chiariglione et al., "AI-Based Media Coding and Beyond," *Proc. Intl. Broadcast. Convention (IBC'21)*, Dec. 2021.
6. N. Pretto et al., "Computing Methodologies Supporting the Preservation of Electro-Acoustic Music From Analog Magnetic Tape," *Computer Music J.*, 42(4):59–74, Jun. 2019.
7. N. Pretto et al., "A Workflow and Novel Digital Filters for Compensating Speed and Equalization Errors on Digitized Audio Open-Reel Tapes," *Proc. Audio Mostly (AM'21)*, pp. 224–231, Sep. 2021.
8. MPAI Community, "Multi-Modal Conversation (MPAI-MMC) V1.1. Accessed: Apr. 7, 2022. [Online]. Available: https://bit.ly/3DH7Sit
9. MPAI Community, "AI-Enhanced Video Coding (MPAI-EVC). Accessed: Apr. 7, 2022. [Online]. Available: https://bit.ly/3tNjtId
10. T. Dumas et al., "Context-Adaptive Neural Network-Based Prediction for Image Compression," *IEEE Trans. Image Processing*, 29:679–693, 2019, doi: 10.1109/TIP.2019.2934565.
11. L. Wang et al., "Enhancing HEVC Spatial Prediction by Context-Based Learning," *Proc. IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP'19)*, pp. 4035–4039, May 2019, doi: 10.1109/ICASSP.2019.8683624.
12. Anwar et al., "Densely Residual Laplacian Super-Resolution," *IEEE Trans. Pattern Analysis Machine Intelligence*, 44(3):1192–1204, 2020, doi: 10.1109/TPAMI.2020.3021088.
13. MPAI Community, "Server-Based Predictive Multiplayer Gaming (MPAI-SPG). Accessed: Apr. 7, 2022. [Online]. Available: https://bit.ly/3I6yKcd

About the Authors

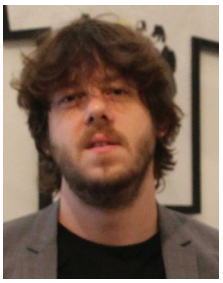
Andrea Basso received a PhD degree. He is an advisor at the Progress Tech Transfer fund, a chief technology officer at MITO Technology, Milan, Italy, a senior expert World Intellectual Property Organization/European Commission, and a director at Synesthesia, Turin, Italy. While at Bell and AT&T Labs, Murray Hill, NJ, USA, he contributed to Internet Engineering Task Force (IETF) and International Organization for Standardization/Motion Picture Experts Group (ISO/MPEG), and the International Multimedia Telecommunications Consortium (IMTC). He has authored or coauthored over 60 articles and holds 174 patents.

Paolo Ribeca is a physicist by education but for 15 years he has specialized in the analysis of high-throughput sequencing data, being particularly interested in the development of algorithms and workflows for alignment, de novo assembly, and compression. He is delighted whenever he can apply those techniques and answer complex biological questions.

Marina Bosi is a pioneer in the development of digital audio coding, is a founding director of the Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) Organization, Geneva, Switzerland, and the chair of MPAI-Context-Based Audio Enhancement (MPAI-CAE). A fellow and past president of the Audio Engineering Society, Bosi was chief technology officer of MPEG LA, LLC, Denver, CO, USA, vice president-technology at DTS, Inc., Los Angeles, CA, USA, and was a member of the research team that created Dolby Digital at Dolby Laboratories, San Francisco, CA, USA, where she also led MPEG-2 AAC development.

Niccolò Pretto is a senior postdoctoral research fellow with the Department of Information Engineering, University of Padova, Padua, Italy. His research is primarily focused on sound and music computing, applying computer science to the preservation and access to historical audio documents and cultural heritage.

Gérard Chollet studied linguistics, electrical engineering, and computer science at the University of California at Santa Barbara, Santa Barbara, CA, USA, where he was granted a PhD in computer science and linguistics. In 1983, he joined a newly created Centre national de la recherche scientifique (CNRS) Research Unit at ENST (now Institut Polytechnique de Paris), Palaiseau, France. He supervised more than 40 doctoral theses. CNRS granted him an emeritus status in 2012. He is currently consulting for Intelligent Voice Ltd., London, UK, Speech Morphing Inc., San Jose, CA, USA, and Zaion.ai, Paris, France. His main publications are available from http://scholar.google.co.uk/citations?user=NakTCiYAAAAJ&hl=en



Michelangelo Guarise is a founder of Volumio, Florence, Italy, the market leader of hi-fi music playback ecosystem. He has been involved in physical computing and user experience research, now focusing on AI and machine learning (ML) on the field applications for high-performance and adaptive music playback systems.

Miran Choi is a principal researcher at the Electronics and Telecommunication Research Institute (ETRI), Daejeon, South Korea. Her research topics are natural language processing and human computer interaction (HCI). She is participating in standardization activities in ITU-T, International Organization for Standardization (ISO), and MPAI, where she develops standards on language-related technology and user interface.

Leonardo Chiariglione received a MS from the Polytechnic of Turin, Turin, Italy, and a PhD from the University of Tokyo, Tokyo, Japan. He founded and chaired the Moving Picture Experts Group (MPEG) for 32 years, and proposed and chairs the Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI), for AI-enabled data coding standards.

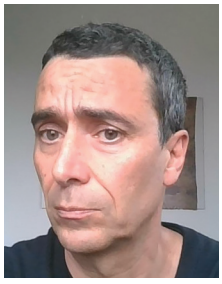
Roberto Iacoviello is a lead research engineer. He graduated from the Politecnico di Torino, Turin, Italy. Since 2007, he has been working at Rai Research and Development, Turin. His current research focuses on artificial intelligence applied to video compression, and augmented reality in broadcast television. He leads the Video Evaluation Group at the European Broadcasting Union (EBU) and actively participates in MPEG and MPAI meetings.

Franesco Banterle is a researcher with the Visual Computing Laboratory at Consiglio Nazionale delle Richerch- (CNR)-Institute of Information Science and Technologies (ISTI), Pisa, Italy. He received a PhD in engineering from the University of Warwick, Coventry, U.K., in 2009. He developed Inverse Tone Mapping, which bridges classic imaging and high-dynamic range (HDR) imaging. His main research fields are HDR imaging and deep learning.

Alessandro Artusi received a PhD in computer science from the Vienna University of Technology, Vienna, Austria, in 2004. He is the team leader of the DeepCamera Group at CYENS Nicosia, Cyprus. His research interests include image/video processing, computer graphics, computer vision, and color science, with a particular focus to deploy the next generation of imaging/video pipeline.

Francesco Gissi received an MSc in computer engineering. He is currently pursuing a PhD in information technology for engineering with the University of Sannio, Benevento, Italy. He is currently working as a data architectures, AI and advanced analyst engineer at Kebula, Naples, Italy. His principal research interests are application of AI models and approaches on video analytics tasks, audio-video compression pipelines via content-based analysis with deep learning techniques, and advanced learning algorithms.

Attilio Fiandrotti received a PhD in computer science from the Politecnico di Torino, Turin, Italy, in 2010. He is currently an assistant professor at the Dipartimento di Informatica, Università di Torino, Turin, and holds a position as maıtre de confèrènces with the Multimedia Group, LTCI, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France. His current research interests focus mainly on deep learning techniques for image and video analysis, compression, and synthesis and the distribution of multimedia contents over wireless packet networks.



Giovanni Ballocca received a master's in physics from the Università di Torino, Turin, Italy. He is a project manager with Sisvel Technology, None Torinese, IT, focusing on research and development projects on compression and distribution of multimedia content. He is also an expert in the field of intellectual property right (IPR) management and licensing in the consumer electronics field.

Marco Mazzaglia is a video game evangelist for Synesthesia, Italy. Since 2008, he has worked on more than 30 original titles and conversions on last-generation game consoles and for the B2B world. He is an adjunct professor of game design and gamification for the Polytechnic of Turin, Turin, Italy, at the master's degree courses for cinema and media engineering and computer engineering.

Scott Moskowitz graduated *cum laude* from the Wharton School of the University of Pennsylvania, Philadelphia, Pennsylvania. He holds more than 110 patents in digital watermarks, product license keys, address space layout randomization (ASLR), and so on. He presented his acclaimed "Cuba" photograph series at Leica Store Miami. He is a senior member of Institute of Electrical and Electronics Engineers (IEEE), Assoication for Computing Machinery (ACM), and Society of Photographic Instrumentation Engineers (SPIE). He is fluent in Japanese.