SOUND AND MUSIC COMPUTING USING AI: DESIGNING A STANDARD

Marina BOSI[1], **Niccolò PRETTO**[2], **Michelangelo GUARISE**[3], and **Sergio CANAZZA**[2]

[1] *Center for Computer Research in Music and Acoustics (CCRMA)*, **Stanford University**, Stanford, USA
[2] *Centro di Sonologia Computazionale (CSC)*, **University of Padova**, Padua, IT
[3] **Volumio**, Florence, IT

ABSTRACT

While there are currently various approaches that define and adapt the conditions in which the user experiences content or service for several music and audio-related applications including entertainment, communication, audio documents preservation/restoration, we are missing worldwide accepted standards that enable data exchange and interoperability based on common interfaces for such applications. The Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) is an international non-profit organization whose mission is to develop such standards. Relying on Artificial Intelligence (AI), MPAI creates a workflow of AI Modules (AIM) that are interchangeable and upgradable without necessarily changing the logic of the application. A specific area of work, MPAI Context-based Audio Enhancement (MPAI-CAE), is showing tremendous possibilities for the Sound and Music Computing (SMC) community. MPAI-CAE applies context information to the input content to deliver the audio output via the most appropriate protocol. Three MPAI-CAE case studies particularly relevant for the SMC community will be presented in this paper: Audio recording preservation (ARP), a use case that covers the whole "philologically informed" archival process of an audio document, from the active sound documents preservation to the access to digitized files; Audio-on-the-go (AOG), which aims to improve safety and listening quality for situations in which the users are in motion in different environments; and Emotion-enhanced speech (EES), a use case that implements a user-friendly system control interface that generates speech with various levels of emotions.

1. INTRODUCTION

Global-scope standardization projects not only offer an indication of the development of an industry but also allow for the partitioning of complex systems into components that can be provided by different sources. This, for example, was the case of media standards in the 1990's. While different companies developed prototypes, the international MPEG standard [1–3] was the catalyst that started a revolution in audio and media consumption. Similarly, MPAI is an international body with the mission to develop standards for data coding that have AI as its core technology. By data coding we mean the transformation of data preserving the semantic aspects that are important to a specific application.

Officially constituted on Wednesday 30 September 2020, MPAI has already produced a considerable body of work. After a Call for Technologies (CfT) for the general AI Framework (MPAI-AIF), MPAI established a Development Committee (MPAI-AIF DC) that is selecting technologies and is currently in the standard development phase (see also Section 2). The goal of the MPAI-AIF standard (see also Section 2 and Figure 1) is to enable the creation and automation of mixed Machine Learning (ML), AI, and legacy data processing modules (collectively AIMs) and to define their use as part of inference workflows. Innovations in AI have led to implementations that can now be found in a wide range of application areas including Speech, Audio, and Image Processing and Recognition. MPAI aims to identify and define interfaces to such implementations to make them usable across as many domains as possible.

MPAI-CAE defines the use of AI to improve the user experience for several audio and music-related applications including entertainment, communication, teleconferencing, gaming, post-production, restoration, etc. The currently available solutions, which adapt to various conditions in order to improve the ultimate user's experience, tend to be vertically integrated. Therefore it is difficult to re-use possibly valuable AI-based components for different applications and different platforms. MPAI-CAE intends to define interfaces between distinct stages (AIMs) to promote the development of horizontal markets of competing solutions tapping into and further promoting AI innovation. Adopting AIMs that are reusable, updatable and extensible, MPAI-CAE intends to define standards for AIM interfaces (i.e., input and output format) but is silent as to the AIM internals. Therefore, the performance of AIMs can continuously improve by incorporating new technologies. The performance evaluation process is under development in a separate standardization thread involving all the MPAI standards and it will be discussed in further work.

The potential impact on the SMC community is huge. MPAI-CAE will allow researchers to carry out sophisticated optimizations that enable a superior user experience. Such optimizations can then be implemented in AIMs by third party providers. Manufacturers and service providers

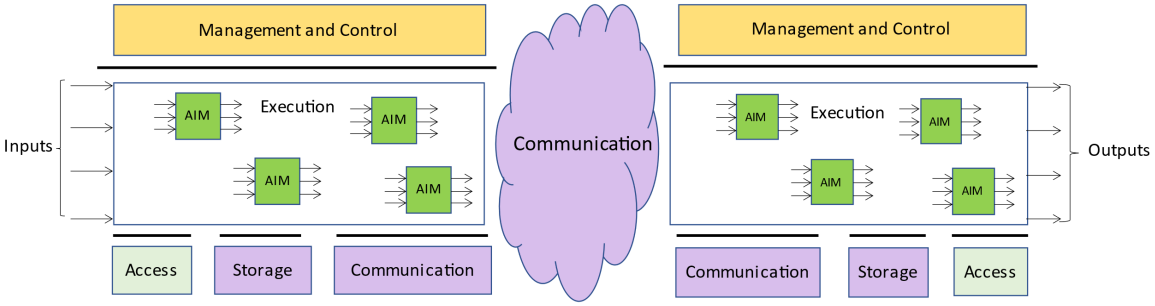

Figure 1. AI Framework schema.

can subsequently adopt these optimized AIMs in their products and services. In general, MPAI operates based on an open international collaboration of interested parties who support the MPAI mission and the means to accomplish it. The use cases described in this paper will help illustrate the core of the MPAI-CAE effort.

The remainder of the manuscript is organized as follows. Section 2 reviews the basic structure of the MPAI process, whereas Section 3 offers a detailed description of three MPAI-CAE use cases particularly relevant for the SMC community (ARP, Section 3.1; AOG, Section 3.2; and EES Section 3.3). Section 4 concludes the paper.

2. STRUCTURE

The overall structure of MPAI relies on MPAI-AIF. The smallest units are the AIMs which are computational modules trained for specific tasks by exploiting AI, ML, and legacy data processing and that can be implemented in hardware, software and mixed hardware/software. MPAI does not define the internal behavior of the AIMs, nevertheless clearly specifies the syntax and semantics of the interfaces. AIMs operate in the standard AI framework and exchange data in specified formats. For this reason, AIMs are replaceable, re-usable and upgradable without changing the logic of the application, fostering the continuous improvement of the AI technology. Different AIMs can be seamlessly interconnected as in the examples provided in the next Sections.

The framework can create, compose, execute, and update multi-vendor AIMs. As can be seen in Figure 1, the Framework is composed by six main components: (a) *Management and Control* manages and controls the AIMs; (b) *Execution* is the environment in which combinations of AIMs operate; (c) *AIMs*, already described; (d) *Communication* is the basic infrastructure used to connect possibly remote Components and AIMs; (e) *Storage* encompasses traditional storage; (f) *Access* represents the access to static or slowly changing data that are required by the application.

The standardization process of MPAI follows an approach based on seven stages. The initial stage (stage 0) concerns the gathering of interest in developing use cases related to a specific topic. As the use cases are not part of the normative standard, they can be augmented later in the process. The information collected in the stage 0 is formalized in the 1^{st} stage. In this stage, the use cases are characterized and a detailed work plan is delineated. The 2^{nd} stage consists in the definition of the functional requirements for a specific work area. The 3^{rd} stage finalizes the commercial requirements, specifically the development of the framework license. The 4^{th} stage formalizes the CfT while the 5^{th} promotes the development of the standard. During the 6^{th} and last stage the standard is approved and published. As of today (March 14, 2021), MPAI-AIF is in the 5^{th} stage, while the MPAI-CAE project is in the 4^{th} stage.

From the bottom-up approach described above, a wide variety of high-tech schemes related to current hot topics in AI are developed. In this context, MPAI-CAE is one of the most promising areas of work. Three of the main MPAI-CAE use cases will be described in the next section.

3. USE CASES

3.1 Audio Recording Preservation (ARP)

MPAI-CAE covers several different SMC areas. The ARP use case represents an important example in the preservation of open-reel analog audio tapes. As shown in Figure 2, the input of this use case is the audio of a digitized tape and the video of that tape flowing on the magnetic head of the tape recording as described in [4].

The first module is the *Audio Enhancer*. This is an optional module consisting of a "denoiser" (in a broad sense). This stage compensates for eventual errors caused by misaligned recording equipment and/or for tape hiss caused by the imperfections introduced by aging (see Storm's Type B, that defines a historically faithful level of reproduc-

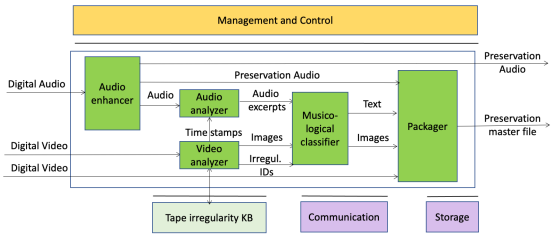

Figure 2. Audio Recording Preservation workflow.

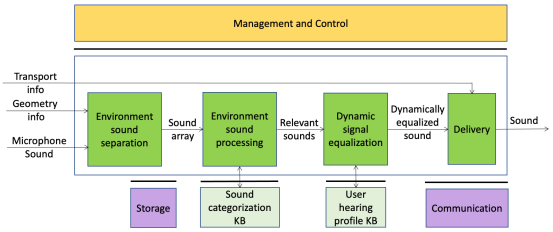

Figure 3. Audio-on-the-go workflow.

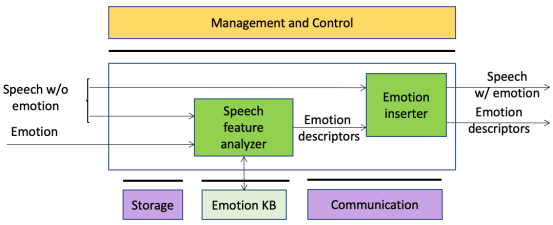

Figure 4. Emotion enhanced speech (EES) workflow.

tion [5]).

The video input is analyzed by an AI algorithm for detecting irregularities (splices, damages, marking, etc.) on the tape such as [4]. This module could interact with the external *Tape Irregularity Knowledge Base (KB)*. Its output consists of frames of the irregularities extracted from the video, their related ID, and the timestamp related to the irregularity. The video of the tape also includes a low quality audio that can be used to synchronize the high quality audio stream with the video itself. Relevant audio excerpts corresponding to the detected irregularities can be extracted and analyzed by the *Audio Analyzer* module.

Single frames concerning irregularities and the corresponding audio excerpts are then analyzed by the *Musicological classifier* [1] that aims to select and describe relevant irregularities. The resulting description and images will be part of the preservation master file created by the *Packager* module. Therefore the preservation master file, composed by audio, video, metadata as indicated in [6], will be provided as output.

In addition, an alternative output is provided. It consists of the digitized tape audio that could be used for accessing the audio content without using the preservation master.

3.2 Audio-on-the-go (AOG)

MPAI-CAE Audio-On-The-Go is a use case that aims to improve safety and listening quality in various situations in which users are on the move, like in a car, with a bike, running and so on. For example while biking in the middle of city traffic, the user should enjoy a satisfactory listening experience without losing contact with the acoustic surroundings. There will be sounds which are not relevant for safety (like wind noise) and sounds which are (like the horn of a car or incoming traffic), and therefore such sounds shall be selected and presented to the user only if relevant for safety [7]. This is achieved thanks to the microphones available in earphones and earbuds capturing the signals from the environment, the relevant environment sounds (i.e., the horn of a car) are then selectively recognized. In addition, the sound rendition is adapted to the acoustic environment, providing an enhanced audio experience (e.g., performing dynamic signal equalization) and allowing a more energy efficient operation resulting in an improved battery life. In this use case, the goal is achieved by using a series of AIMs. The first AIM (*Environmental Sound Separation*) is fed with Microphone sound which captures the surrounding environment noise, together with according geometry information (which describes number, positioning and configuration of the microphone or the array of microphones). The sounds are then categorized following prescriptions of a *Sound Categorization Knowledge Base* (queried by the corresponding AIM), resulting in a sounds array and their categorization. Sound samples might eventually be compressed to allow a cloud-processing procedure. The *Environmental Sound Processing* AIM, after fetching a list of relevant sounds from a KB, will trim sounds not relevant for the user in the specific moment and feed them to the next AIM, *Dynamic Signal Equalization*. This AIM fetches the *User Hearing Profile from a Knowledge Base* and equalizes dynamically the sound taking into account the user's specific hearing deviations. Finally, the resulting sound is delivered to the output via the most appropriate the *Delivery* method, such as Bluetooth 5.0 or any compatible protocol.

[1] The use of the term "musicological classifier" was selected because it specifically identifies a classification of interest for musicologists.

3.3 Emotion enhanced speech (EES)

Speech carries information not only about the lexical content, but also about a variety of other aspects such as age, gender, signature, and emotional state of the speaker, and this is an acknowledged issue. Speech synthesis is evolving towards supporting these aspects. There are many cases where a speech without emotion needs to be converted to a speech carrying an emotion, possibly with grades of a particular emotion. This is the case, for instance, of a human-machine dialogue where the message conveyed by the machine is more effective if it carries an emotion properly related to the emotion detected in the human speaker. MPAI-CAE EES use case aims to standardize a *natural* communication by virtual agents, and thus improve the quality of human-machine interaction, by making it closer to a human-human interaction (e.g., [8,9]). By means of EES anyone can realize a user-friendly system control interface that lets users generate speech with various — continuous and real-time — expressiveness control levels.

The MPAI-CAE EES can be implemented as in figure 4, using data processing technology or artificial intelligent technology, where a neural network incorporates the *Emotion Knowledge Base* information.

The inputs are: a *neutral* (without specific emotion) speech, synthesized or recorded; a text file with the annotation of which basic emotion [10,11] to insert (and where) into the speech signal.

The *Speech feature analyzer* extracts the speech features, queries the *Emotion KB* and obtains *Emotion descriptors* (a subset of speech features modified accordingly to the particular emotion). Alternatively, *Emotion descriptors* are produced by an embedded neural network.

Emotion Knowledge Base exposes an interface that allows *Speech feature analyzer* to query a KB of speech features extracted from recordings of different speakers reading/reciting the same corpus of texts, with the standard set of basic emotions and without emotion, for different languages and genders. A set of acoustic cues are used to compare the voice quality characteristics of the speech signals on a voice corpus in which different emotions are reproduced. The psychoacoustic parameters of emotions in speech can be separated into two groups [12]: prosodic (rhythm, speed of speech, intonation and intensity) and vocal frequency-related parameters (timbre, fundamental tracking, position of the formants and distribution of the spectral energy).

Emotion inserter inserts a particular emotional vocal timbre, e.g., anger, disgust, fear, happiness, sadness, and surprise into a neutral (emotion-less) synthesized voice. It also changes the strength of an emotion (from neutral speech) in a gradual fashion.

4. CONCLUSIONS

A group of highly motivated experts in different fields has gathered within the MPAI community (`https://mpai.community/organisation/`) to develop use cases aggregated in areas where the MPAI standards can have a big impact. Thanks to the efforts of many, MPAI has reached several important milestones. For example, MPAI-AIF has already reached the standard development stage and multiple areas, including MPAI-CAE, have open call for technologies.

This paper presented three use cases of MPAI-CAE that are of particular interest to the SMC community. Other MPAI areas of work include Multi-modal conversation, AI-Enhanced traditional video coding, integrative AI-based analysis of multi-source genomic/sensor experiments, Compression and understanding of financial data, and Server-based predictive distributed multiplayer online gaming.

MPAI has introduced a number of innovative approaches in both the technologies that address specific industries and also in the development of licensing guidelines. MPAI is planning to develop for each standard a "framework licence" to overcome the ambiguities of the Fair, Reasonable and Non-Discriminatory (FRAND) model. Such framework licenses are already available for MPAI-AIF, MPAI-MMC and MPAI-CAE.

Moreover, MPAI pledges to address ethical questions raised by its technical work and is in the process of defining different procedures in this area.

Acknowledgments

The authors would like to thank the MPAI community for the invaluable discussions and, in particular, Leonardo Chiariglione, MPAI President, for his vision and his continued guidance.

5. REFERENCES

[1] ISO/IEC, "Coding of moving pictures and associated audio for digital storage media at up to about 1.5 mbit/s," *ISO/IEC 11172 - International Organization for Standardization/International Electrotechnical Commission and others*, 1993.

[2] ——, "Information technology — generic coding of moving pictures and associated audio information," *ISO/IEC 13818:1995 - International Organization for Standardization/International Electrotechnical Commission and others*, 1995.

[3] K. Brandenburg and M. Bosi, "Overview of MPEG audio: current and future standards for low bit-rate audio coding," *Journal of the Audio Engineering Society*, vol. 45, no. 1/2, pp. 4–21, 1997.

[4] N. Pretto, C. Fantozzi, E. Micheloni, V. Burini, and S. Canazza, "Computing methodologies supporting the preservation of electroacoustic music from analog magnetic tape," *Computer Music Journal*, vol. 42, no. 4, pp. 59–74, 2019, doi: 10.1162/comj_a_00487.

[5] F. Bressan and S. Canazza, "A systemic approach to the preservation of audio documents: Methodology and software tools," *Journal of Electrical and Computer Engineering*, 2013, doi: 10.1155/2013/489515.

[6] C. Fantozzi, F. Bressan, N. Pretto, and S. Canazza, "Tape music archives: from preservation to access," *International Journal on Digital Libraries*, vol. 18, no. 3, pp. 233–249, 2017, doi: 10.1007/s00799-017-0208-8.

[7] B. Schulte-Fortkamp, "Soundscape, standardization, and application," in *Proc. Int. Conf. Euronoise 2018*, Crete, 2018, pp. 2445–2450.

[8] J. E. Cahn, "The generation of affect in synthesized speech," *Journal of the American Voice I/O Society*, vol. 8, pp. 1–19, 1990.

[9] M. A. M. Shaikh, A. R. F. Rebordao, K. Hirose, and M. Ishizuka, "Emotional speech synthesis by sensing affective information from text," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–6.

[10] R. Plutchik, "A psychoevolutionary theory of emotions," *Social Science Information*, vol. 21, no. 4-5, pp. 529–553, 1982. [Online]. Available: https://doi.org/10.1177/053901882021004003

[11] T. Dalgleish and M. J. Powers, *Handbook of Cognition and Emotion*. Wiley, 1999.

[12] N. Tits, "A methodology for controlling the emotional expressiveness in synthetic speech-a deep learning approach," in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 2019, pp. 1–5.