|  |  |
| --- | --- |
| ­ | Moving Picture, Audio and Data Coding by Artificial Intelligencewww.mpai.community |

|  |
| --- |
| **Public document** |
| **N151** | 2021/02/17 |
| **Source** | MPAI Audio-Events-Data |
| **Title** | MPAI-CAE Use Cases and Functional Requirements |
| **Target** | MPAI Members |

# Introduction

Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) is an [international association](http://mpai.community/) with the mission to develop *AI-enabled data coding standards*. Research has shown that data coding with AI-based technologies is *more efficient* than with existing technologies.

The MPAI approach to developing AI data coding standards is based on the definition of *standard interfaces* of *AI Modules (AIM).* AIMs operate on input and output data both having a standard format. AIMs can be *combined* and *executed* in an MPAI-specified *AI-Framework* according to the emerging standard MPAI-AIF. A [Call for MPAI-AIF Technologies](https://mpai.community/standards/mpai-aif/) [2] with associated [Use Cases and Functional Requirements](https://mpai.community/standards/mpai-aif/#Requirements) [1] was issued on 2020/12/16 and is now closed.

While AIMs must expose standard interfaces to be able to operate in an MPAI AI Framework, their performance may differ depending on the technologies used to implement them. MPAI believes that *competing* developers striving to provide more performing *proprietary* and *inter­operable* AIMs will promote *horizontal markets* of *AI solutions* that build on and further promote AI *innovation*.

This document is a collection of Use Cases and corresponding Functional Requirements in the MPAI Context-based Audio Enhancement (MPAI-CAE) application area. The Use Cases in this area help improve the audio user experience for several application spaces that include enter­tain­ment, communication, teleconferencing, gaming, post-production, restoration etc. in a variety of contexts such as in the home, in the car, on-the-go, in the studio etc.

Currently MPAI has identified four Use Cases falling in the Context-based Audio Enhancement application area:

1. Emotion-Enhanced Speech (EES)
2. Audio Recording Preservation (ARP)
3. Enhanced Audioconference Experience (EAC)
4. Audio-on-the-go (AOG)

This document is to be read in conjunction with the MPAI-CAE Call for Technologies (CfT) [3] as it provides the functional requirements of all technologies identified as required to implement the current MPAI-CAE Use Cases. Respondents to the MPAI-CAE CfT are requested to make sure that their responses are aligned with the functional requirements expressed in this document.

In the future MPAI may issue other Calls for Technologies falling in the scope of MPAI-CAE to support identified Use Cases. Currently these are:

1. Efficient 3D sound
2. (Serious) gaming
3. Normalization of TV volume
4. Automotive
5. Audio mastering
6. Speech communication
7. Audio (post-)production

It should also be noted that some technologies identified in this document are the same, similar, or related to technologies required to implement some of the Use Cases of the companion document [MPAI-MMC Use Cases and Functional Requirements](https://mpai.community/standards/mpai-mmc/#UCFR) [5]. Readers of this document are advised that being familiar with the content of the said companion document is a prerequisite for a proper understanding of this document.

This document is structured in 7 chapters, including this Introduction.

|  |  |
| --- | --- |
| Chapter 2 | briefly introduces the AI Framework Reference Model and its six Components |
| Chapter 3 | briefly introduces the 4 Use Cases. |
| Chapter 4 | presents the 4 MPAI-CAE Use Cases with the following structure1. Reference architecture
2. AI Modules
3. I/O data of AI Modules
4. Technologies and Functional Requirements
 |
| Chapter 5 | identifies the technologies likely to be common across MPAI-CAE and MPAI-MMC, a companion standard project whose Call for Technologies is issued simul­taneously with MPAI-CAE’s. |
| Chapter 6 | gives relevant references |
| Chapter 7 | gives a basic list of relevant terms and their definition |

For the reader’s convenience, *Table 1* introduces the meaning of the acronyms used in this doc­ument.

*Table 1 – MPAI-CAE acronyms*

|  |  |
| --- | --- |
| **Acronym** | **Meaning** |
| AI | Artificial Intelligence |
| AIF | AI Framework |
| AIM | AI Module |
| AOG | Audio-on-the-go |
| ARP | Audio Recording Preservation |
| CfT | Call for Technologies |
| DP | Data Processing |
| EAE | Enhanced Audioconference Experience |
| EES | Emotion-Enhanced Speech |
| KB | Knowledge Base |
| ML | Machine Learning |

# The MPAI AI Framework (MPAI-AIF)

Most MPAI applications considered so far can be implemented as a set of AIMs – AI, ML and even traditional DP-based units with standard interfaces assembled in suitable topol­ogies to achieve the specific goal of an application and executed in an MPAI-defined AI Frame­work. MPAI is making all efforts to identify processing modules that are re-usable and upgradable without necessarily changing the inside logic. MPAI plans on completing the development of a 1st generation MPAI-AIF AI Framework in July 2021.

The MPAI-AIF Architecture is given by *Figure 1*.



*Figure 1 – The MPAI-AIF Architecture*

MPAI-AIF is made up of 6 Components:

1. *Management and Control* manages and controls the AIMs, so that they execute in the correct order and at the time when they are needed.
2. *Execution* is the environment in which combinations of AIMs operate. It receives external inputs and produces the requested outputs, both of which are Use Case specific, activates the AIMs, exposes interfaces with Management and Control and interfaces with Communic­ation, Storage and Access.
3. *AI Modules* (AIM) are the basic processing elements receiving processing specific inputs and producing processing specific outputs.
4. *Communication* is the basic infrastructure used to connect possibly remote Components and AIMs. It can be implemented, e.g., by means of a service bus.
5. *Storage* encompasses traditional storage and is used to e.g., store the inputs and outputs of the individual AIMs, intermediary results data from the AIM states and data shared by AIMs.
6. *Access* represents the access to static or slowly changing data that are required by the application such as domain knowledge data, data models, etc.

# Use Cases

## 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** [3]. 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.

The AI Modules identified in the Emotion-Enhanced Speech (EES) Use Case considered in this document will make it possible to create virtual agents communicating in a more natural way, and thus improve the quality of human-machine interaction, by making it closer to a human-human interaction [8].

EES’s ultimate goal is to help realise a user-friendly system control interface that lets users gen­erate speech with various – continuous and real time – expressiveness control levels.

## Audio Recording Preservation (ARP)

Preservation of audio assets recorded on a variety of media (vinyl, tapes, cassettes etc.) is an important activity for a variety of application domains, in particular cultural heritage, because preservation requires more that “neutral” transfer of audio information from the analogue to the digital domain. For instance,

1. It is necessary to recover and preserve context information, obviously, but not exclusively, audio.
2. The recording of an acoustic event cannot be a neutral operation because the timbre quality and the plastic value of the recorded sound, which are of great importance in, for example, contemporary music, are influenced by the positioning of the microphones used during the recording.
3. The processing carried out by the Tonmeister, i.e., the person who has a detailed theoretical and practical knowledge of all aspects of sound recording. However, unlike a sound engineer, the Tonmeister must also be deeply trained in music: music­ological and historic-critical competence are essential for the identification and correct cataloguing of the information contained in audio documents [9].
4. As sound carriers are made of unstable base materials, they are more subject to damage caused by inadequate handling. The commingling of a technical and scientific formation with historic-philol­ogical knowledge (an important element for the identification and correct cataloguing of the infor­mation contained in audio documents) becomes essential for preservative re-recording oper­ations, going beyond mere A/D conversion. In the case of magnetic tapes, the carrier may hold important information: the tape can include multiples splices; it can be annotated (by the composer or by the technicians) and/or display several types of *irregularities* (e.g., corruptions of the carrier, tape of different colour or chemical composition).

In this Audio Recording Preservation Use Case, currently restricted to magnetic tapes, audio is digitised and fed into a preservation system. The audio information is supplemented by the information coming from a video camera pointed to the head that reads the magnetic tape. The output of the restoration process is composed by:

1. Preservation digital audio
2. Preservation master file that contains, next to the preservation audio file, several other inform­ation types created by the preservation process.

The introduction of this use case in the field of active preservation of audio documents opens the way to respond in an effective way to the methodological questions of reliability with respect to the recordings as documentary sources, while clarifying the concept of “historical faithfulness”.

The goal is to cover the whole “philologically informed” archival process of an audio document, from the active preservation of sound documents to the access to digitized files.

## Enhanced Audioconference Experience (EAE)

Often, the user experience of a video/audio conference is far from satisfactory. Too much background noise or undesired sounds can lead to participants not understanding or even misun­derstanding what participants are saying, in addition to creating distraction.

By using AI-based adaptive noise-cancellation and sound enhancement, those kinds of noise can be virtually eliminated without using complex microphone systems that capture environment char­acteristics.

In this use case, the system receives microphone sound and microphone geometry information which describes number, positioning and configuration of the microphone or the array of micro­phones. Using this information, the system is able to detect and separate audioconference speech information from spurious sounds. It is to be noted that Microphone Physical information (i.e., frequency response and deviation of the microphone) might be added, but that will likely be an overkill for this scenario. The resulting speech then undergoes Noise Cancellation. The resulting output is equalized based on the output device characteristics, fetched from an Output Device Acoustic Model Knowledge Base, which describes the frequency response of the selected output device. This way the speech can be equalized removing any coloration from the output device, resulting in an optimally delivered sound experience.

## Audio-on-the-go (AOG)

While biking in the middle of city traffic, the user should enjoy a satisfactory listening experience without losing contact with the acoustic surroundings.

The microphones available in earphones or earbuds capture the signals from the environment. The relevant environment sounds (i.e., the horn of a car) are selectively recognised and the sound rendition is adapted to the acoustic environment, providing an enhanced audio experience (e.g., performing dynamic signal equalization) and an improved battery life.

In this use case, Microphone sound captures the surrounding environment noise, together with geometry information (which describes number, positioning and configuration of the microphone or the array of microphones).

The sounds are then categorized. The result is an array of sounds with their categorization.

Sounds not relevant for the user in the specific moment are trimmed out and the rest of the sound information undergoes dynamic signal equalization using User Hearing Profile information.

Finally, the resulting sound is delivered to the output via the most appropriate the delivery method.

# Functional Requirements

## Introduction

The Functional Requirements developed in this document refer to the individual technologies identified as necessary to implement Use Cases belonging to given MPAI-CAE application area using AIMs operating in an MPAI-AIF AI Framework. The Functional Requirements developed adhere to the following guidelines:

1. AIMs are defined to allow implementations by multiple technologies (AI, ML, DP)
2. DP-based AIMs need interfaces such as to a Know­ledge Base. AI-based AIM will typically require a learning process, however, support for this process is not included in the document. MPAI may develop further requirements covering that process in a future document.
3. AIMs can be aggregated in larger AIMs. Some data flows of aggregated AIMs may not neces­sarily be exposed any longer.
4. AIMs may be influenced by the companion MPAI-MMC Use Cases and Functional Requ­ir­ements [5] as some technologies needed by some MPAI-MMC AIMs share a significant number of functional requirements.
5. Current AIMs do not feed information back to AIMs upstream. Respondents to the MPAI-CAE Call for Technologies [3] are welcome to motivate the need for such feed-back data flows and propose assoc­iated requirements.

The Functional Requirements described in the following sections are the result of a dedicated effort by MPAI experts over many meetings where different partitioning in AIMs have been proposed, discussed and revised. MPAI is aware that alternative partitioning or alternative I/O data to/from AIMs are possible. Those reading this document for the purpose of submitting a response to the [MPAI-CAE Call for Technologies](https://mpai.community/standards/mpai-cae/#Technologies) (N152) [2] are welcome to propose alternative partitionings or alternative I/O data in their submissions. In this case, however, they are required to justify their alternatives and determine the functional requirements of the relevant technol­ogies. The evaluation team, to which proponents can, if they so wish, be members, will study the proposed alternative arrangement and may decide to accept all or part of the proposed new arrangement.

## Emotion-Enhanced Speech

### Reference architecture

This Use Case can be implemented as in *Figure 2* and *Figure 3*. The two figures differ in the use of legacy DP technology vs AI technology:

1. In *Figure 2* the Speech analysis AIM is implemented with legacy Data Processing technologies.
2. In *Figure 3* the Speech analysis AIM is implemented as a neural network which incorporates the Emotion KB information.



*Figure 2 –* *Emotion-enhanced speech (using external Knowledge Base)*

**

*Figure 3 – Emotion-enhanced speech (fully AI-based)*

### AI Modules

The AI Modules perform the functions described in *Table 1*.

*Table 2 – AI Modules of Emotion-Enhanced Speech*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| **Speech feature analyser** | Computes Speech features, queries the Emotion KB and obtains Emotion des­criptors. Alternatively, Emotion descriptors are produced by an embedded neural network. |
| **Emotion KB** | Exposes an interface that allows Speech feature analyser to quey a KB of speech features extracted from recordings of dif­ferent speakers reading/reciting the same corpus of texts, with the standard set of emotions and without emotion, for different languages and genders.  |
| **Emotion inserter** | Inserts a particular emotional vocal timbre, e.g., anger, disgust, fear, happiness, sadness, and surprise into a neutral (emotion-less) synthesised voice. It also changes the strength of an emotion (from neutral speech) in a gradual fashion. |

### I/O interfaces of AI Modules

The I/O data of the Emotion Enhanced Speech AIMs are given in *Table 3*.

*Table 3 – I/O data of Emotion-Enhanced Speech AIMs*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| **Speech features analyser**  | Emotion-less speech EmotionEmotion descriptors | Emotion descriptorsSpeech features |
| **Emotion KB** | Speech features | Emotion descriptors |
| **Emotion inserter** | Emotion-less speechEmotion descriptors | Speech with EmotionEmotion descriptors |

### Technologies and Functional Requirements

#### Digital Speech

Speech should be sampled at a frequency between 8 kHz and 96 kHz and digitally represented between 16 bits/sample and 24 bits/sample (both linear). The frequency of 22.05 kHz should be used for the purpose of a response to the MPAI-CAE Call for Technologies. Demonstrations of a proposed technology for other sampling frequencies are welcome.

**To Respondents**

Respondents are invited to comment on these choices.

#### Emotion

By Emotion we mean a digital attribute that indicates an emotion out of a finite set of Emotions.

In EES the input speech – natural or synthesised – does not contain emotion while the output speech is expected to contain the emotion expressed by the input Emotion.

The most basic Emotions are described by the set: “anger, disgust, fear, happiness, sadness, and surprise” [10], or “joy versus sadness, anger versus fear, trust versus disgust, and surprise versus anticipation” [11]. One of these sets can be taken as “universal” in the sense that they are common across all cultures. An Emotion may have different Grades [12,13].

**To Respondents**

Respondents are requested to propose:

1. A minimal set of Emotions whose semantics are shared across cultures.
2. A set of Grades that can be associated to Emotions.
3. A digital representation of Emotions and their Grades (starting from [14]).

Currently, the MPAI-CAE Call for Technologies does not envisage to consider culture-specific Emotions. However, the proposed digital rep­resentation of Emotions and their Grades should either accommodate, or be extensible to accom­modate, culture-specific Emotions.

#### Emotion KB query format

To accom­plish their task, speech processing applications utilize certain features of speech signals. General speech features are described in [15,16]. The extraction of these features from a speech signal is known as speech analysis. Extraction can be done in the time domain as well as in the frequency domain.

*Time-domain features* are related to the waveform analysis in the time domain. Analysing speech in the time domain often requires simple calculation and interpretation. Time-domain features can be used to measure the arousal level of emotions.

Time-domain features carry information about sequences of short-time prosody acoustic features (features estimated on a frame basis). Example features modified by the emotional states are given by short-time zero crossing rate, short-term speech energy and duration [19].

*Frequency-domain features* can be computed using (short-time) Fourier transform, wavelet transform, and other mathematical tools [24]. Frequency domain operation provides mechan­isms to obtain some of the most useful parameters in speech analysis because the human cochlea performs a quasi-frequency analysis.

Initially, the time-domain signal is transformed into the frequency-domain, from which the features are extracted. Such features are highly associated with the human perception of speech. Hence, they have apparent acoustic characteristics. These features usually comprise formant frequency, linear prediction cepstral coefficient (LPCC), and Mel frequency cepstral coefficients (MFCC).

The frequency-domain features can carry information about:

1. The Pitch signal (i.e., the glottal waveform) that depends on the tension of the vocal folds and the subglottal air pressure. Two parameters related to the pitch signal can be considered: pitch frequency and glottal air velocity. E.g., high velocity indicates a speech emotion like hap­piness. Low velocity is in harsher styles such as anger [25].
2. The shape of the vocal tract that is modified by the emotional states. The formants (character­ized by a centre frequency and a bandwidth) could be a representation of the vocal tract reson­ances. Features related to the number of harmonics due to the non-linear airflow in the vocal tract. E.g., in the emotional state of anger, the fast air flow causes additional excitation signals other than the pitch. Teager Energy Operator-based (TEO) features measure the harmonics and cross-harmonics in the spectrum [26].

Example of features modified by the emotional states are given by the Mel-frequency cepstrum(MFC) [27].

Today, there is a variety of speech datasets available (online). Often, they consist of conversational setups and contain overlaps in speech as well as noise, or they are poor in expressiveness. Some datasets offer emotionally rich content with a high quality, but in a limited amount [e.g., 19,20,21,22]. To be effective, an Emotion KB should contain a large and expressive speech data­set.

Emotion KB contains speech features extracted from the speech recordings of speakers reading/ reciting the same corpus of texts with an agreed set of emotions and without emotion, for a set of languages and for different genders (voice performances by professional actors in comparison with the author’s spontaneous speech) [28, 29].

Emotion KB is queried by providing a vector of speech features. Emotion KB responds by prov­iding Emotion descriptors.

**To Respondents**

Respondents are requested to propose an Emotion KB query format satisfying the following requ­irements:

1. Accept as input:
	1. A vector of speech features capable of modelling:
		1. Non-extreme emotional states [17].
		2. Many emotional states with a natural-sounding voice [18].
	2. An Emotion.
2. Provide as output a set of Emotion descriptors.

When assessing proposed Speech features, MPAI may resort to objective testing.

Note: An AI-based implementation may not need Emotion KB.

#### Emotion descriptors

Emotion descriptors are features used to alter the prosodic characteristics, the pitch, and the for­mant frequencies and bandwidth of Digital speech.

Speech analysis can use different strategies to render the emotion depending on:

1. The type of sentence (numbers of words, type of phonemes, etc.) to which an emotion is added
2. The emotions added to the previous and next sentence.

Emotion descriptors can be obtained by querying an Emotion KB (in the case of *Figure 2*) or from the output of a neural network (in the case of *Figure 3*).

**To Respondents**

Respondents should propose Emotion descriptors suitable to introduce Emotion into the specific emotion-less speech resulting in a speech that appears as “natural” to the listener.

When assessing proposed Speech features, MPAI may resort to subjective testing.

## Audio Recording Preservation

### Reference architecture

This Use Case is implemented as in *Figure 4* and *Figure 5*. The two figures differ in the use of legacy DP technology vs AI technology:

1. In *Figure 4* the Audio-video Analysis AIM is implemented with Data Processing Technol­ogies.
2. In *Figure 5* the Audio-video Analysis AIM is implemented as a neural network which incor­porates the Emotion KB information.



*Figure 4 – Tape Audio preservation (using external Knowledge Base)*

**

*Figure 5 – Tape Audio preservation (fully AI-based)*

### AI Modules

The AIMs required by this Use Case are described in *Table 4*.

*Table 4 – AI Modules of Audio Recording Preservation*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| **Audio enhancer**  | Produces Preservation audio using internal denoiser, finalized only to compensate (a) non-linear frequency response, caused by imperfect histor­ical recording equipment; (b) rumble, needle noise, or tape hiss caused by the imperfections introduced by aging. (see 4.3.5). |
| **Audio analyser**  | Produces audio excerpts based on signals from Video analysis. |
| **Video analyser**  | Extracts images from Video,queries the Tape irregularity KB and provides Images and Irregularities IDs. Alternatively, an embedded neural network produces images. |
| **Musicological classifier** | Produces relevant images from Digital video and Text describing images |
| **Packager** | Produces file containing:1. Digital audio
2. Input video
3. Audio sync’d images and text
 |
| **Tape irregularity KB** | Knowledge Base of visual (tape) and audio irregularities |

### I/O interfaces of AI Modules

The AIMs of Audio Recording Preservation are given in *Table 5*

*Table 5 – I/O data of Audio Recording Preservation AIMs*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| **Audio enhancer**  | Digital Audio | Preservation Audio  |
| **Audio analysis** | Preservation Audio Irregularity | Audio Excerpts |
| **Video analysis** | Digital Video Tape irregularity KB response | ImagesTape irregularity KB queryIrregularity IDs |
| **Musicological classifier** | Audio ExcerptsImagesIrregularity IDs | TextImages |
| **Packager** | Preservation Audio Digital VideoTextImages | Preservation Master |
| **Tape irregularity KB** | Query | Response |

### Technologies and Functional Requirements

#### Digital Audio

Digital Audio sampled from an analogue source (e.g., magnetic tapes, 78rpm phonographic discs) at a frequency in the 44.1-96 kHz range with at least 16 and at most 24 bits/sample [30].

**To Respondents**

Respondents are invited to comment on this choice.

#### Digital Video

Digital video has the following features.

1. Pixel shape: square
2. Bit depth: 8-10 bits/pixel
3. Aspect ratio: 4/3 and 16/9
4. 640 < # of horizontal pixels < 1920
5. 480 < # of vertical pixels < 1080
6. Frame frequency 50-120 Hz
7. Scanning: progressive
8. Colorimetry: ITU-R BT709 and BT2020
9. Colour format: RGB and YUV
10. Compression: uncompressed; if compressed AVC, HEVC

**To Respondents**

Respondents are invited to comment on these choices.

#### Digital Image

A Digital Image is

1. An uncompressed video frame with time information or
2. A JPEG-compressed video frame [32] with time information.

**To Respondents**

Respondents are invited to comment on this choice.

#### Tape irregularity KB query format

Tape irregularity KB contains features extracted from images of different tape irregularities [38].

The Irregularity KB is queried by giving a vector of Image features that describe [37]:

1. Splices of
	1. Leader tape to magnetic tape
	2. Magnetic tape to magnetic tape
2. Other irregularities such as brands on tape, ends of tape, ripples, damaged tapes, markings, dirt, shadows etc.

The Irregularity KB responds by providing the type of irregularity detected in the input Image.

**To Respondents**

Respondents are requested to propose a Tape irregularity KB query format satisfying the follow­ing requirements:

1. A complete set of audio tape irregularities and Image features that characterise them.
2. A response to a query shall indicate:
	1. Presence of irregularities or otherwise.
	2. Type of irregularity as output (if there are irregul­arities).

When assessing proposed Image features MPAI may resort to objective testing.

This CfT is specifically for of *audio tape* preservation. However, its scope may be extended if sufficient technologies covering other audio preservation instances are received. Any proposal for other audio preservation instances should be described with a level of detail comparable to this Use Case.

#### Text

Text should be encoded according to ISO/IEC 10646, Information technology – Universal Coded Character Set (UCS) to support most languages in use [39].

**To Respondents**

Respondents are invited to comment on this choice.

#### Packager

Packager takes Preservation Audio, Digital Video, Text and Images and produces the Preservation Master file.

**To Respondents**

Respondents should propose a file format capable to:

1. Support queries for irregularities, showing all the images corresponding to that given irregularity (splices, carrier corruptions, etc.)
2. Allow listening to the audio corresponding to a particular image.
3. Allow to annotate (with text) the audio signal, to support the musicological analysis
4. Support query on the annotation, returning the corresponding time (sec:ms:sample), the text, the audio signal excerpt and image (if any)
5. Support random access to a specified portion of video and/or audio providing.

Preference will be given to formats that have already been standardised or are in wide use.

### Information about Audio enhancement performance

A fifty-year-long debate around the restoration of audio documents has been ongoing inside the archivists’ and musicologists’ communities [33].

The Preservation audio produced by Audio enhancement must fulfil the requirements of accuracy, reliability, and philological authenticity.

In [34] Schuller makes an accurate investigation of signal alterations classified in two categories:

1. Intentional that includes recording, equalization, and noise reduction systems.
2. Unintentional further divided into those caused by:
	1. The imperfection of the recording technique of the time, resulting in various distortions.
	2. Misalignment of the recording equipment, e.g., wrong speed, deviation from the ver­tical cutting angle in cylinders, or misalignment of the recording in magnetic tape.

The choice whether or not to compensate for these alterations reveals different restoration strat­egies: historical faithfulness can refer to the recording as it has been produced, precisely equalized for intentional recording equalizations, compensated for eventual errors caused by misaligned recording equipment (for example, wrong speed, deviation from the vertical cutting angle in cylinders, or misalignment of the recording in magnetic tape) and digitized using a modern equipment to minimize replay distortions.

There is a certain margin of interpretation because historical acquaintance with the document is called into question alongside with technical-scientific knowledge, for instance, to identify the equalization curves of magnetic tapes or to determine the rotation speed of a record. Most of the information provided is retrievable from the history of audio technology, while other information is experimentally inferable with a certain degree of accuracy.

The restoration must be focused to compensate non-linear frequency response, caused by imperfect historical recording equipment; rumble, needle noise, or tape hiss caused by the imperfections introduced by aging.

The restoration step can thus be carried out with a good degree of objectivity and represents an optimum level achievable by the original (analogue) recording equipment.

A legacy denoiser algorithm should [35,36]:

1. Use little a priori information.
2. Operate in real time.
3. Be based on frequency-domain methods, such as various forms of non-casual Wiener filtering or spectral subtraction schemes.
4. Include algorithms that incorporate knowledge of the human auditory system.

**To Proponents**

The CfT does not include technologies object of this AIM. However, respondents’ comments on the text above will be welcome.

## Enhanced Audioconference Experience

### Reference architecture

This Use Case is implemented as in *Figure 6*.



*Figure 6 – Enhanced Audioconference Experience*

### AI Modules

The AIMs required by the Enhanced Audioconference Experience are given in *Table 6*

*Table 6 – AIMs of Enhanced Audioconference Experience*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| **Speech detection and separation** | Separates relevant Speech vs non-speech signals |
| **Noise cancellation** | Removes noise in Speech signal |
| **Output dynamic noise cancellation** | Reduces noise level based on Output Device Acoustic Model |
| **Delivery** | Wraps De-noised Speech signal for Transport |
| **Output Device Acoustic Model KB** | Contains identifiers of all output devices of by manufacturer and their ID calibration test results  |

### I/O interfaces of AI Modules

The I/O data of Enhanced Audioconference Experience AIMs are given in *Table 7*.

*Table 7 – I/O data of Enhanced Audioconference Experience AIMs*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| **Speech detection and separation** | Microphone SoundGeometry Information | Digital SpeechGeometry Information |
| **Noise cancellation** | Digital SpeechGeometry Information | De-noised Speech |
| **Output dynamic noise cancellation** | De-noised Speech | Equalised Speech |
| **Delivery** | Equalised SpeechTransport info | Equalised Speech |
| **Output Device Acoustic Model KB** | Query | Response |

### Technologies and Functional Requirements

#### Digital Speech

Speech should be sampled at a frequency between 8 kHz and 96 kHz and the samples should be represented with a number of bits at least 16 bits/sample and at most 24 bit/sample (both linear).

**To Respondents**

Respondents are invited to comment on these two choices.

#### Microphone geometry information

Microphone geometry information is a descriptive representation of relative positioning of one or multiple microphones which describes physical characteristics of microphones such as type, pos­itioning, angle and their relative position and overall configuration such as Array Type. It allows to accurately reproduce a signal free of noise and distortion and to better separate noise from signal as required for proper working of EAE AIMs. Formats to represent microphone geom­etry infor­mation are: MPEG-H 3D Audio [40] and platform (Android, Windows, Linux) specific JSON Descriptors API [41].

**To Respondents**

Respondent are requested to:

1. Comment about MPAI’s choice of the two formats
2. Express their preference between the two formats.
3. Possibly suggest alternative solutions.

#### Output device acoustic model metadata KB query format

The Output device acoustic model KB contains a description of the output device acoustic model, such as frequency response and per-frequency attenuation.

The Output device acoustic model KB is queried by requesting the unique ID of a device, if available, or by providing a means to identify the model or unique reference to output device being considered. The Output device acoustic model KB responds with information about output device characteristics.

**To Respondents**

Respondents are requested to propose a query/response API satisfying the requirement that API shall provide:

1. Means to query the KB giving the device model as input to obtain the acoustic model.
2. Adequate schemas to represent the Output device acoustic model using, if necessary, current representation schemes.

#### Delivery

Equalised Speech needs to be transported using a transport protocol most appropriate for the environment.

**To Respondents**

Proponents are requested to identify the transport protocols suitable for the EAE Use Case and propose an extensible way to signal which transport mechanism is intended to be used.

## Audio-on-the-go

### Reference architecture

This Use Case is implemented as in *Figure 7* and in *Figure 8*. The two figures differ in the use of legacy DP technology vs AI technology:

1. In *Figure 7* Environment sound separation and Environment sound processing AIMs are implemented using legacy Data Proces­sing technology.
2. In *Figure 8* the Environment sound processing AIM is implemented as neural a network.



*Figure 7 – Audio-on-the-go (using external Knowledge Base)*

**

*Figure 8 – Audio-on-the-go (full AI-based solution)*

### AI Modules

The AIMs of Audio-on-the-go are given by *Table 8*.

*Table 8 – AIMs of Audio-on-the-go*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| **Environment sound separation** | Separates the individual sounds captured from the surrounding environment  |
| **Environment sound processing** | Determines which sounds are relevant to the user |
| **Sound categorisation KB** | Contains audio features of the sounds in the KB |
| **Dynamic signal equalization** | Dynamically equalises sound using information from User hearing profiles KB to produce the best possible quality output |
| **Delivery** | Wraps equalised sound for Transport |
| **User hearing profiles KB** | A dataset of hearing profiles of target users |

### I/O interfaces of AI Modules

The I/O data of Audio on the go AIMs are given by *Table 9*

*Table 9 – I/O data of Audio-on-the-go AIMs*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| **Environment sound separation** | Microphone Sound Geometry info | Sound array |
| **Environment sound processing** | Sound arraySound categorisation | Relevant soundsSound features |
| **Dynamic signal equalization** | Relevant soundsUser’s hearing profiles | Dynamically equalised soundUser ID |
| **Delivery** | Equalised SpeechTransport info | Equalised Speech |
| **Sound categorisation KB** | Sound features vector | Sound categorisation |
| **User hearing profiles KB** | Query | Response |

### Technologies and Functional Requirements

#### Digital Audio

Digital Audio is a stream of samples obtained by sampling audio at a frequency in the 44.1-96 kHz range with at least 16 and at most 24 bits/sample.

**To Respondents**

Proponents are invited to comment on this choice.

#### Microphone geometry information

Microphone geometry information is a descriptive representation of relative positioning of one or multiple microphones which describes physical characteristics of microphones such as type, pos­itioning, angle and their relative position and overall configuration such as Array Type. It allows to accurately reproduce a noise- and distortion-free signal and to better separate noise from signal as required for proper working of EAE AIMs. Formats to represent microphone geometry infor­mation are: MPEG-H 3D Audio [40] and platform (Android, Windows, Linux) specific JSON Descriptors API [41].

**To Respondents**

Respondents are requested to:

1. Express their preference between the two formats.
2. Comment about MPAI’s choice of the two formats.
3. Possibly suggest alternative solutions.

#### Sound array

The sounds identified in the Microphone sound are passed as an array of sounds represented as

1. Sound samples.
2. Encoding information (e.g., sampling frequency, bits/sample, compression method).
3. Relative metadata.
4. Duration.

**To Respondents**

Respondents are requested to propose:

1. A format to package a set of environment sounds with appropriate metadata.
2. An extensible identification of audio compression methods.

#### Sound categorisation KB query format

Sound categorisation KB contains audio features of the sounds in the KB. Sound categorisation KB is queried by providing a vector of Sound features. Sound categorisation KB responds by giving the category of the sound.

Sound features are extracted from samples of individual sounds in the Sound array for the purpose of querying the Sound categor­is­ation KB.

**To Respondents**

Respondents should propose a Sound categorisation query format satisfying the following requir­ements:

1. Use an extensible set of Sound features that satisfy the following requirements:
	1. Be suitable for identifying a sound.
	2. Be suitable as input to query the Sound categorisation.
2. Provide as output:
	1. The probability for the most relevant N categories.
	2. From which Sound categorisation KB this value has been derived.

When assessing proposed Sound features MPAI may resort to objective testing.

#### Sounds categorisation

Each vector in the sound array should be accompanied by an identifier of the category it belongs to.

**To Respondents**

Respondents should propose an extensible classification of all types of sound of interest [42]. Support of a set of sounds classified according to a proprietary scheme should also be provided.

#### User Hearing Profiles KB query format

User Hearing Profiles KB contains the user hearing profile for the properly identified (e.g. via a UUID or a third-party identity provider) specific user.

User Hearing Profiles KB is queried giving the User hearing profile ID as input. User hearing profile KB responds with the specific user hearing profile. The User hearing profile contains the hearing attenuation for a defined number of frequency spectrums or any representation able to determine the unique individual sound perception ability [43]. There are currently at least 2 SDKs on the matter: MIMI SDK, NURA SDK (both proprietary) [44].

**To Respondents**

Respondents should propose a query format which the following requirements:

1. Input: user identity, array of frequency values
2. Output: the values of the user’s sound perception ability at those frequency values

#### Delivery

Equalised Speech needs to be transported using a transport protocol most appropriate for the environment.

**To Respondents**

Proponents are requested to identify the transport protocol suitable for the AOG Use Case and propose an extensible way to signal which transport mechanism is intended to be used.

# Potential common technologies

*Table 10* introduces the acronyms representing the MPAI-CAE and MPAI-MMC Use Cases.

*Table 10 – Acronyms of MPAI-CAE and MPAI-MMC Use Cases*

|  |  |  |
| --- | --- | --- |
| **Acronym** | **App. Area** | **Use Case** |
| EES | MPAI-CAE | Emotion-Enhanced Speech |
| ARP | MPAI-CAE | Audio Recording Preservation |
| EAE | MPAI-CAE | Enhanced Audioconference Experience |
| AOG | MPAI-CAE | Audio-on-the-go |
| CWE | MPAI-MMC | Conversation with emotion |
| MQA | MPAI-MMC | Multimodal Question Answering |
| PST | MPAI-MMC | Personalized Automatic Speech Translation |

*Table 11* gives all MPAI-CAE and MPAI-MMC technologies in alphabetical order.

Please note the following acronyms:

|  |  |
| --- | --- |
| KB | Knowledge Base |
| QF  | Query Format |

*Table 11 – Alphabetically ordered MPAI-CAE and MPAI-MMC technologies*

|  |  |
| --- | --- |
| Notes | UC=Use case |
|  | UCFR=Use Cases and Functional Requirements document number |
|  | Section=Section of the above document |
|  | Technology=name of technology |

|  |  |  |  |
| --- | --- | --- | --- |
| **UC** | **UCFR** | **Section** | **Technology** |
| EAE | N151  | 4.4.4.4 | Delivery |
| AOG | N151  | 4.5.4.7 | Delivery |
| CWE | N153 | 4.2.4.9 | Dialog KB query format |
| ARP | N151  | 4.3.4.1 | Digital Audio |
| AOG | N151  | 4.5.4.1 | Digital Audio |
| ARP | N151  | 4.3.4.3 | Digital Image |
| MQA | N153 | 4.3.4.3 | Digital Image |
| EES | N151  | 4.2.4.1 | Digital Speech |
| EAE | N151  | 4.4.4.1 | Digital Speech |
| CWE | N153 | 4.2.4.2 | Digital Speech |
| MQA | N153 | 4.3.4.2 | Digital Speech |
| PST | N153 | 4.4.4.2 | Digital Speech |
| ARP | N151  | 4.3.4.2 | Digital Video |
| CWE | N153 | 4.2.4.3 | Digital Video |
| EES | N151  | 4.2.4.2 | Emotion |
| CWE | N153 | 4.2.4.4 | Emotion |
| EES | N151  | 4.2.4.4 | Emotion descriptors |
| CWE | N153 | 4.2.4.5 | Emotion KB (speech) query format |
| CWE | N153 | 4.2.4.6 | Emotion KB (text) query format |
| CWE | N153 | 4.2.4.7 | Emotion KB (video) query format |
| EES | N151  | 4.2.4.3 | Emotion KB query format |
| MQA | N153 | 4.3.4.4 | Image KB query format |
| CWE | N153 | 4.2.4.11 | Input to face animation |
| CWE | N153 | 4.2.4.10 | Input to speech synthesis |
| MQA | N153 | 4.3.4.7 | Intention KB query format |
| PST | N153 | 4.4.4.4 | Language identification |
| CWE | N153 | 4.2.4.8 | Meaning |
| MQA | N153 | 4.3.4.6 | Meaning |
| EAE | N151  | 4.4.4.2 | Microphone geometry information |
| AOG | N151  | 4.5.4.2 | Microphone geometry information |
| MQA | N153 | 4.3.4.5 | Object identifier |
| MQA | N153 | 4.3.4.8 | Online dictionary query format |
| EAE | N151  | 4.4.4.3 | Output device acoustic model metadata KB query format |
| ARP | N151  | 4.3.4.6 | Packager |
| AOG | N151  | 4.5.4.3 | Sound array |
| AOG | N151  | 4.5.4.4 | Sound categorisation KB query format |
| AOG | N151  | 4.5.4.5 | Sounds categorisation |
| PST | N153 | 4.4.4.3 | Speech features |
| ARP | N151  | 4.3.4.4 | Tape irregularity KB query format |
| ARP | N151  | 4.3.4.5 | Text |
| CWE | N153 | 4.2.4.1 | Text |
| MQA | N153 | 4.3.4.1 | Text |
| PST | N153 | 4.4.4.1 | Text |
| PST | N153 | 4.4.4.5 | Translation results |
| AOG | N151  | 4.5.4.6 | User Hearing Profiles KB query format |

The following technologies are shared or shareable across Use Cases:

1. Delivery
2. Digital speech
3. Digital audio
4. Digital image
5. Digital video
6. Emotion
7. Meaning
8. Microphone geometry information
9. Text

Image features apply to different visual objects. The Speech features of all Use Cases are different.

However, respondents should consider the possibility of proposing a unified set of Speech features, e.g., as proposed in [45].

# Terminology

*Table 12* identifies and defines the terms used in the MPAI-CAE context.

*Table 12 – MPAI-CAE terms*

|  |  |
| --- | --- |
| **Term** | **Definition** |
| Access | Static or slowly changing data that are required by an application such as domain knowledge data, data models, etc. |
| AI Framework (AIF) | The environment where AIM-based workflows are executed |
| AI Module (AIM) | The basic processing elements receiving processing specific inputs and producing processing specific outputs |
| Audio enhancement  | An AIM that produces Preservation audio using internal denoiser |
| Communication | The infrastructure that connects the Components of an AIF |
| Data Processing (DP) | A legacy technology that may be used to implement AIMs |
|  |  |
| Delivery | An AIM that wraps data for transport |
| Digital Speech | Digitised speech as specified by MPAI |
| Dynamic Signal Equalization | An AIM that dynamically equalises the sound using information from the User hearing profiles KB |
| Emotion | A digital attribute that indicates an emotion out of a finite set of Emotions |
| Emotion Descriptor | A set of time-domain and frequency-domain features capable to render a particular emotion, starting from an emotion-less digital speech |
| Emotion inserter | A module to set time-domain and frequency-domain features of a neutral speech in order to insert a particular emotional intention. |
| Emotion KB | A speech dataset rich in expressiveness |
| Emotion KB query format | A dataset of time-domain and frequency-domain neutral speech features |
| Environment Sound Processing | An AIM that determines which sounds are relevant for the user vs sounds which are not |
| Environment Sounds Recognition | An AIM that recognises, separates and categorises sounds captured from the environment  |
| Execution | The environment in which AIM workflows are executed. It receives external inputs and produces the requested outputs both of which are application specific |
| Frequency-domain Features | Properties (descriptors) of the signal with respect to frequency |
| Emotion Grade | The intensity of an Emotion |
| Knowledge Base | Structured and unstructured information made accessible to AIM (especially DP-based) |
| Management and Control | Manages and controls the AIMs in the AIF, so that they execute in the correct order and at the time when they are needed |
| Musicological classifier | Algorithm that sorts unlabelled images from Digital Video into (relevant) labelled categories of information, linking them with text describing the images. |
| Noise cancellation | An AIM that removes noise in Speech signal |
| Output Device Acoustic Model KB | A dataset of calibration test results for all output devices of a given manufacturer identified by their ID |
| Output dynamic noise cancellation | An AIM that reduces noise level based on Output Device Acoustic Model |
| Packager | An AIM that packages audio, video, images and text in a file |
| Relevant vs non-relevant sound KB | A dataset of audio features of relevant sounds |
| Sound categorisation KB | Contains audio features of the sounds in the KB |
| Speech analysis | The AIM that extracts Emotion descriptors |
| Speech analysis | The AIM that understands the emotion embedded in speech |
| Speech analysis | The AIM that extracts the characteristics of the speaker (e.g., physiology and intention) |
| Speech and Emotion File Format | A file format that contains Digital speech and time-stamped Emotions related to speech |
| Speech detection and separation | AIM that separates relevant Speech vs non-speech signals |
| Speech Features | Speech features used to extract Emotion descriptors |
| Storage | Storage used to e.g., store the inputs and outputs of the individual AIMs, data from the AIM’s state and intermediary results, shared data among AIMs |
| Tape irregularity KB | Dataset that includes examples of the different irregularities that may be present in the carrier (analogue tape, phonographic discs) considered |
| Text | Characters drawn from a finite alphabet |
| Time-domain features | Properties (descriptors) of the signal with respect to frequency |
| User hearing profiles KB | A dataset of hearing profiles of target users |

# References

1. MPAI-AIF Use Cases and Functional Requirements, N74; <https://mpai.community/standards/mpai-aif/#Requirements>
2. MPAI-AIF Call for Technologies, N100; <https://mpai.community/standards/mpai-aif/#Technologies>
3. MPAI-CAE Use Cases and Functional Requirements, N151; <https://mpai.community/standards/mpai-cae/#UCFR>
4. MPAI-CAE Call for Technologies, N152; <https://mpai.community/standards/mpai-cae/#Technologies>
5. MPAI-MMC Use Cases and Functional Requirements, N153; <https://mpai.community/standards/mpai-mmc/#Requirements>
6. MPAI-MMC Call for Technologies, N154; <https://mpai.community/standards/mpai-mmc/#Technologies>
7. F. Burkhardt and N. Campbell, “Emotional speech synthesis,” in The Oxford Handbook of Affective Computing. Oxford University Press New York, 2014, p. 286
8. Noé Tits, A Methodology for Controlling the Emotional Expressiveness in Synthetic Speech - a Deep Learning approach, 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), September 2019, DOI: 10.1109/ACIIW.2019.8925241
9. T. W. Adorno, Philosophy of New Music, University of Minnesota Press, Minneapolis, Minn, USA, 2006
10. Ekman, P. (1999). Basic Emotions. In T. Dalgleish and T. Power (Eds.) The Handbook of Cognition and Emotion Pp. 45–60. Sussex, U.K.: John Wiley & Sons, Ltd.
11. Plutchik R., Emotion: a psychoevolutionary synthesis, New York Harper and Row, 1980
12. Russell, James (1980). "A circumplex model of affect". Journal of Personality and Social Psychology. 39 (6): 1161–1178. doi:10.1037/h0077714
13. Cahn, J. E., The Generation of Affect in Synthesized Speech, Journal of the American Voice I/O Society, 8, July 1990, p. 1-19
14. https://www.w3.org/TR/2014/REC-emotionml-20140522/
15. Cahn, J. E., The Generation of Affect in Synthesized Speech, Journal of the American Voice I/O Society, 8, July 1990, p. 1-19
16. Burkhardt, F., & Sendlmeier, W. F., Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis, ISCA Workshop on Speech & Emotion, Northern Ireland 2000, p. 151-156.
17. Scherer, K. R., Ladd, D. R., & Silverman, K., Vocal cues to speaker affect: Testing two models, Journal of the Acoustic Society of America, 76(5), 1984, p. 1346-1356
18. Kasuya, H., Maekawa, K., & Kiritani, S., Joint Estimation of Voice Source and Vocal Tract Parameters as Applied to the Study of Voice Source Dynamics, ICPhS 99, p. 2505-2512
19. S. R. Livingstone and F. A. Russo, “The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english,” PLOS ONE, vol. 13, no. 5, pp. 1–35, 05 2018
20. H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, “Crema-d: Crowd-sourced emotional multimodal actors dataset,” IEEE transactions on affective computing, vol. 5, no. 4, pp. 377–390, 2014
21. T. Banziger, M. Mortillaro, and K. R. Scherer, “Introducing the geneva multimodal expression corpus for experimental research on emotion perception.” Emotion, vol. 12, no. 5, p. 1161, 2012
22. F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of german emotional speech,” in Ninth European Conference on Speech Communication and Technology, 2005
23. Mozziconacci, S. J. L., Speech Variability and Emotion: Production and Perception, PhD Thesis, Technical University Eindhoven, 1998
24. Burkhardt, F., & Sendlmeier, W. F., Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis, ISCA Workshop on Speech & Emotion, Northern Ireland 2000, p. 151-156.
25. Cahn, J. E., The Generation of Affect in Synthesized Speech, Journal of the American Voice I/O Society, 8, July 1990, p. 1-19
26. Hamed Beyramienanlou, Nasser Lotfivand, "An Efficient Teager Energy Operator-Based Automated QRS Complex Detection", Journal of Healthcare Engineering, vol. 2018, Article ID 8360475, 11 pages, 2018. https://doi.org/10.1155/2018/8360475]
27. Davis S B. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 1980, 28(4):65-74
28. Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, Massimiliano Todisco. EMOVO Corpus: an Italian Emotional Speech Database.
29. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, pp. 3501–3504, May 2014. 2- Moataz El Ayadi, Mohamed S. Kamel, Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition Journal, Elsevier, 44 (2011) 572–587
30. IASA-TC 05: Handling and Storage of Audio and Video Carriers. IASA Technical Committee (2014)
31. Hamed Beyramienanlou, Nasser Lotfivand, "An Efficient Teager Energy Operator-Based Automated QRS Complex Detection", Journal of Healthcare Engineering, vol. 2018, Article ID 8360475, 11 pages, 2018. https://doi.org/10.1155/2018/8360475
32. ISO/IEC 10918-1:1994 Information Technology — Digital Compression And Coding Of Continuous-Tone Still Images: Requirements And Guidelines
33. Federica Bressan and Sergio Canazza, A Systemic Approach to the Preservation of Audio Documents: Methodology and Software Tools, Journal of Electrical and Computer Engineering, 2013. https://doi.org/10.1155/2013/489515
34. G. Boston, Safeguarding the Documentary Heritage. A Guide to Standards, Recommended Practices and Reference Literature Related to the Preservation of Documents of All Kinds, UNESCO, Paris, France, 1988.
35. S. Canazza. The digital curation of ethnic music audio archives: from preservation to restoration. International Journal of Digital Libraries, 12(2-3):121–135, 2012
36. S.J. Godsill and P.J.W. Rayner. Digital Audio Restoration - a statistical model-based approach (Berlin: Springer-Verlag 1998)
37. Pretto, Niccolò; Fantozzi, Carlo; Micheloni, Edoardo; Burini, Valentina; Canazza Targon, Sergio. Computing Methodologies Supporting the Preservation of Electroacoustic Music from Analog Magnetic Tape. In Computer Music Journal, 2018, vol. 42 (4), pp.59-74
38. Fantozzi, Carlo; Bressan, Federica; Pretto, Niccolò; Canazza, Sergio. Tape music archives: from preservation to access. pp.233-249. In International Journal On Digital Libraries, pp. 1432-5012 vol. 18 (3), 2017. DOI:10.1007/s00799-017-0208-8
39. ISO/IEC 10646:2003 Information Technology — Universal Multiple-Octet Coded Character Set (UCS)
40. https://www.iis.fraunhofer.de/en/ff/amm/broadcast-streaming/mpegh.html
41. https://docs.microsoft.com/bs-cyrl-ba/azure/cognitive-services/speech-service/how-to-devices-microphone-array-configuration
42. https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01277/full
43. https://help.nuraphone.com/hc/en-us/articles/360000324676-Your-Profile
44. https://integrate.mimi.io/documentation/android/4.0.1/documentation
45. Problem Agnostic Speech Encoder; https://github.com/santi-pdp/pase