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

|  |  |
| --- | --- |
| N131 | 2021/01/19 |
| Source | MPAI Audio-Events-Data |
| Title | Draft 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 data having a standard format to provide output data having a standard format. AIMs can be *combined* and *executed* in an MPAI-specified *AI-Framework* called MPAI-AIF. A [Call for MPAI-AIF Technologies](https://mpai.community/standards/mpai-aif/) [1] is currently open.

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 Functional Requirements for the MPAI Context-based Audio Enhancement (MPAI-CAE) application area. The Use Cases in the MPAI-CAE standard help improve the audio user experience for several applications including 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 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) [2] as it provides the functional requirements of all the technologies that have been identified as required to implement the current MPAI-CAE Use Cases. Respondents to the MPAI-CAE CfT should 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 [3]. Readers of this document are advised that being familiar of 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 structure   1. 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 suggested references. Respondents are advised to become familiar with the references |
| Chapter 7 | gives a basic list of relevant terms and their definition |

# 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 Data Processing (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 AI Framework called MPAI-AIF in July 2021.

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



*Figure 1 – The MPAI-AIF Architecture*

Where

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 application specific interfacing with Management and Control and 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 required in several cases and can be implemented, e.g., by means of a service bus and may be used to connect with remote parts of the framework
5. *Storage* encompasses traditional storage and is 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.
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

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** [2]. 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 to improve the quality of human interaction with a machine, by making it closer to a human-human interaction [5].

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

## Audio Recording Preservation

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.

A totally neutral process in the analogue-to-digital (A/D) audio information transfer is not sufficient. It is necessary to recover and preserve context information, obviously, but not exclusively, audio. The recording of an acoustic event can never 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 already influenced by the positioning of the microphones used during the recording. In addition, 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 [6].

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 transfer. 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, audio is digitised and fed into a preservation system. The audio information is supplemented by the information coming from a video camera that is pointed to the head that reads the magnetic tape. The output of the restoration process is the preservation digital audio and a 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 effective answer to the methodological questions of reliability with respect to the recordings as documentary sources, also 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

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 to understand or even misun­derstand 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 goal is achieved by using a series of AIMs. The first AIM is fed with Microphone sound (which captures the conversation audio) and the according geometry information (which describes number, positioning and configuration of the microphone or the array of microphones). It is to be noted that also Microphone Physical information (frequency response and deviation of the microphone) might be added, but that will likely be an overkill for this scenario. The resulting output (Speech signal and Geometry information) is then fed to the Noise Cancellation AIM which performs de-noising of the conversation. The resulting output is then equalized based on the output device characteristics, fetched from the Output Device Acoustic Model KB, 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

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 and 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, the goal is achieved by using a series of AIMs. The first AIM (Environmental Sound Recognition) 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 KB, 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 KB and equalizes dynamically the sound taking into account the User’s specific hearing deviations.

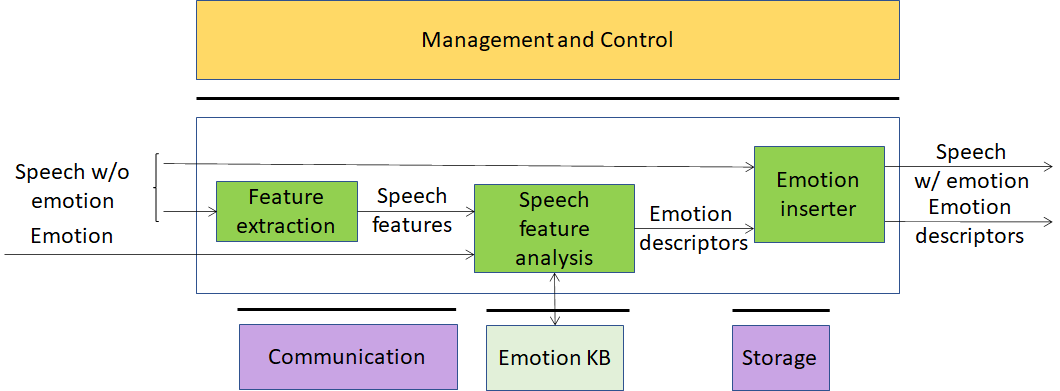
Finally, the resulting sound is delivered to the output via the most appropriate the delivery method.

# Functional Requirements

## Emotion-Enhanced Speech

### Reference architecture

This Use Case is implemented as in *Figure 2*. The Speech analysis AIM can be implemented either as AI/ML or legacy DP modules. If this AIM is implemented as a neural network, access to Emotion KB may not be needed.



*Figure 2 –* *Emotion-enhanced speech*

### AI Modules

The AI Modules of *Figure 2* perform the functions described in *Table 1*.

*Table 1 – AI Modules of Emotion-Enhanced Speech*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| **Feature extraction** | Produces Speech features suitable for subsequent analysis |
| **Speech features analysis** | Produces Emotion descriptors by querying the Emotion KB. Alternatively, Emotion descriptors are produced by an embedded neural network. |
| **Emotion KB** | Allows Speech analysis to access features extracted from speech 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 2*.

*Table 2 – I/O data of Emotion-Enhanced Speech AIMs*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| **Feature extraction** | Emotion-less Digital Speech | Speech features |
| **Speech features analysis** | Speech features  Emotion  Emotion KB response | Emotion descriptors  Emotion KB query |
| **Emotion KB** | Query | Response |
| **Emotion inserter** | Emotion-less Digital Speech  Emotion descriptors | Speech with Emotion  Emotion descriptors |

### Technologies and Functional Requirements

#### Digital Speech

Emotion Enhanced Speech (EES) requires that speech be sampled at a frequency between 22.05 kHz and 96 kHz and digitally represented between 16 bits/sample and 24 bits/sample.

**To Respondents**

Respondents are invited to comment on these choices.

#### Emotion

By Emotion we mean an attribute that indicate 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” [7], or “joy versus sadness, anger versus fear, trust versus disgust, and surprise versus anticipation” [8]. 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 [9,10].

**To Respondents**

Respondents are invited 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 [11]).

Currently, culture-specific Emotions are not being considered. However, the proposed digital representation of Emotions and their Grades should either accommodate or be extensible to accommodate culture-specific Emotions.

#### Speech features

To accom­plish their task, speech processing applications utilize certain features of speech signals. General speech features are described in [12,13]. The extraction of these properties or features and how to obtain them from a speech signal is known as speech analysis. It can be done in the time domain as well as in the frequency domain. Analysing speech in the time domain often requires simple calculation and interpretation.

*Time-domain features* are related to the waveform analysis in the time domain. They 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 [16].

*Frequency-domain features* can be computed using (short-time) Fourier transform, wavelet transform, and other mathematical tools [21]. The frequency domain provides the 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 feature is 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 could 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 [22].
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 [23].

Example features modified by the emotional states are given by the Mel-frequency cepstrum(MFC) [24].

**To Respondents**

Respondents are expected to propose Speech features that are capable to model

1. non-extreme emotional states [14]
2. many emotional states with a natural-sounding voice [15].

#### Emotion descriptors

Emotion descriptors are a derivation of Speech features. They are used by the Emotion inserter to add the required emotion to the Digital speech.

By using frequency-domain and time-domain features a specific emotion can be added to a particular input 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 the output of a neural network or obtained by querying an Emotion KB.

**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.

#### Emotion KB query format

As of 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., 16,17,18,19]. 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 different female and male 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) [25, 26].

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

**To Respondents**

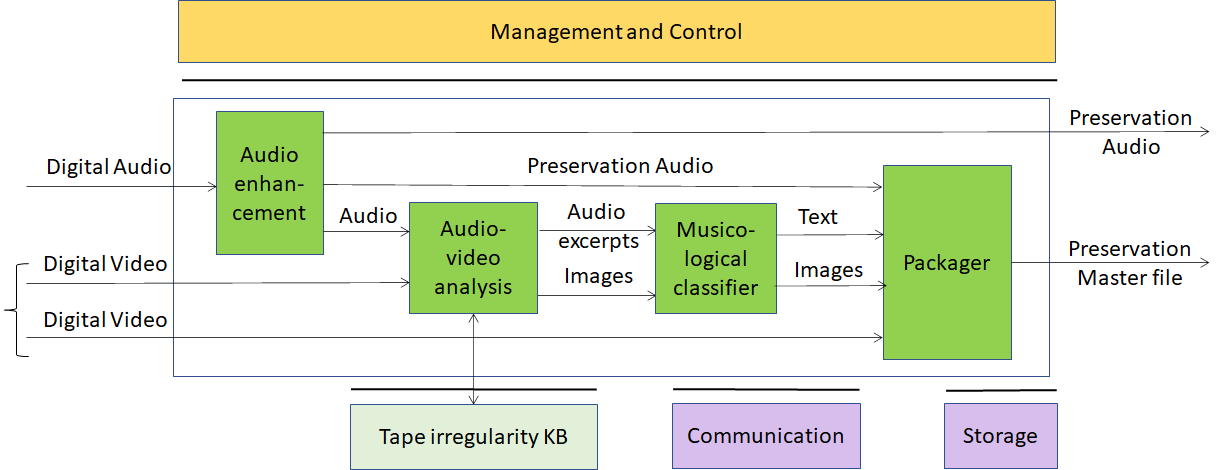
Respondents are requested to propose an Emotion KB query format satisfying the following requirements:

1. Accept a list of the speech features identified in 4.1.4.4
2. Provide as output a set of Emotion descriptors identified in 4.1.4.5.

## Audio Recording Preservation

### Reference architecture

This Use Case is implemented as in *Figure 3*. The Audio-video Analysis AIM can be implemented either using AI or legacy technologies. If this AIM is implemented as a neural network, access to the Tape irregularity KB may not be required.



*Figure 3 – Tape Audio preservation*

### AI Modules

The AIMs required by this Use Case are described in *Table 3*.

*Table 3 – AI Modules of Audio Recording Preservation*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| **Audio enhancement** | 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 4.2.4.4). |
| **Audio-video analysis** | Produces images and audio excerpts querying the Tape irregularity KB. Alternatively, an embedded neural network produces images and audio excerpts. |
| **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 and audio irregularities |

### I/O interfaces of AI Modules

The AIMs of Audio Recording Preservation are given in *Table 4*

*Table 4 – I/O data of Audio Recording Preservation AIMs*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| **Audio enhancement** | Digital Audio | Preservation Audio |
| **Audio-video Analysis** | Preservation Audio  Digital Video  Tape irregularity KB response | Audio Excerpts  Images  Tape irregularity KB query |
| **Musicological classifier** | Audio Excerpts  Images | Text  Images |
| **Packager** | Preservation Audio  Digital Video  Text  Images | 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 48-96 kHz range with at least 16 and at most 24 bits/sample [27].

**To Proponents**

Proponents 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 Proponents**

Proponents are invited to comment on these choices.

#### Digital Image

A Digital Image is

1. An uncompressed video frame with time information or
2. A video frame compressed with JPEG [29] with time information.

**To Proponents**

Respondents are invited to comment on this choice.

#### Image Features

Image Features are used to describe [34]

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.

**To Proponents**

Respondents are requested to propose

1. a complete set of irregularities from audio tapes
2. Image features that characterise them.

#### Tape irregularity KB query format

Tape irregularity KB contains features extracted from images of different tape irregularities [35].

The Irregularity KB is queried by giving the features of an Image. 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. Accept a list of the Image features identified in 4.2.4.4 as input
2. Responds with indication of presence of irregularities or otherwise and, if there are irregul­arities, with the type of irregularity identified in 4.2.4.4 as output

This CfT is specifically for preservation of audio tapes. 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 [36].

**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 [30].

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

In [31] 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 two groups:
   1. those caused by the imperfection of the recording technique of the time, resulting in various distortions
   2. those caused by misalignment of the recording equipment, for example, 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 [32,33]:

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.

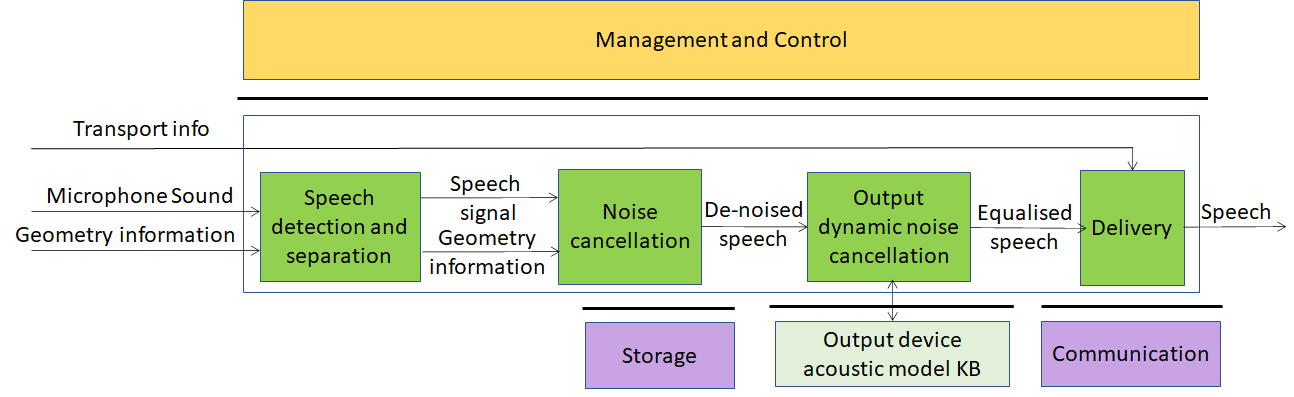
**To Proponents**

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

## Enhanced Audioconference Experience

### Reference architecture

This Use Case is implemented as in *Figure 4*.



*Figure 4 – Enhanced Audioconference Experience*

### AI Modules

The AIMs required by the Enhanced Audioconference Experience are given in

*Table 5 – 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 calibration test results for all output devices of a given manufacturer identified by their ID |

### I/O interfaces of AI Modules

The I/O data of Enhanced Audioconference Experience AIMs are given in *Table 6*.

*Table 6 – I/O data of Enhanced Audioconference Experience AIMs*

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

### Technologies and Functional Requirements

#### Digital Speech

Enhanced Audioconference Experience (EAE) requires that speech be sampled at a frequency between 22.05 kHz and 96 kHz and that the samples be represented with a number of bits at least 16 bits/sample and at most 24 bit/sample.

**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 [37] and platform (Android, Windows, Linux) specific JSON Descriptors API [38].

**To Respondents**

Respondent 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.

#### 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 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 following requirements: API shall provide

1. Means to enquiry for a specific device, model or family of models, if available.
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.

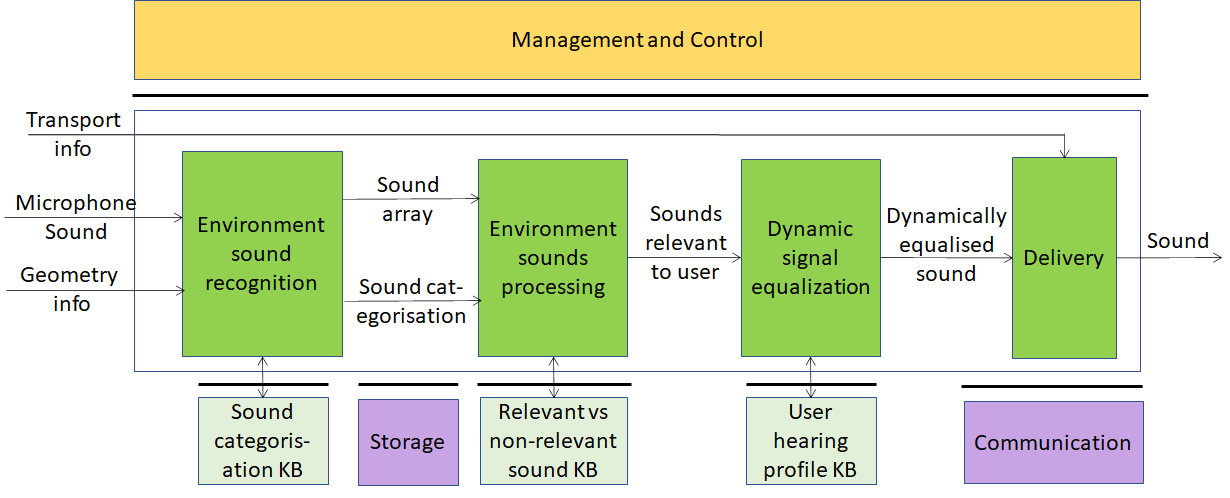
**To Respondents**

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 5*. Environment sound recognition and Environment sound processing AIMs can be implemented either using AI or legacy technology. If any of these AIMs are implemented as a neural network, access to the corresponding KB may not be needed.



*Figure 5 – Audio-on-the-go*

### AI Modules

The AIMs of Audio-on-the-go are given by *Table 7*

*Table 7 – AIMs of Audio-on-the-go*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| **Environment Sounds Recognition** | Recognises, separates and categorises sounds captured from the surrounding environment |
| **Environment Sound Processing** | Determines which sounds are relevant for the user vs sounds which are not |
| **Dynamic Signal Equalization** | Dynamically equalises the sound using information from the User hearing profiles KB to produce the best possible quality output |
| **Delivery** | Wraps equalised sound for Transport |
| **Sound categorisation KB** | Contains audio features of the sounds in the KB |
| **Relevant vs non-relevant sound KB** | Contains audio features of relevant sounds |
| **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 8*

*Table 8 – I/O data of Audio-on-the-go AIMs*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| **Environment Sounds Recognition** | Microphone Sound  Geometry info | Sound array  Sound categorisation |
| **Environment Sound Processing** | Sound array  Sound categorisation | Sound relevant to user |
| **Dynamic Signal Equalization** | Sound relevant to user | Dynamically equalised sound |
| **Delivery** | Equalised Speech  Transport info | Equalised Speech |
| **Sound categorisation KB** | Query | Response |
| **Relevant vs non-relevant sound KB** | Query | Response |
| **User hearing profiles KB** | Query | Response |

### Technologies and Functional Requirements

#### Digital Audio

Digital Audio sampled is a stream of samples obtained by sampling audio at a frequency in the 48-96 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 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 [1] and platform (Android, Windows, Linux) specific JSON Descriptors API [38].

**To Respondents**

Respondent 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

Respondents should propose a format to package a set of environment sounds with the requirements on being able to include the sound samples, encoding information (e.g., sampling frequency, bits per sample, compression method) and relative metadata, and duration.

**To Respondents**

Respondents are requested to propose an extensible identification of audio compression methods.

#### Sounds categorisation

Sounds captured by the microphone should be categorised.

**To Respondents**

Respondents should propose an extensible classification of all types of sound of interest [39]. Support of a set of sounds classified according to a proprietary scheme should also be provided.

#### Sound categorisation KB query format

Sound categorisation KB contains audio features of the sounds in the KB.

Sound categorisation KB is queried by giving features extracted from the input sound as input. Sound categorisation KB responds by giving the category of the sound.

**To Respondents**

Respondents should propose an extensible set of features to be used to query the Sound categorisation KB and obtain the categories of the sounds with following requirements

1. The confidence value for the most relevant N categories.
2. From which classification KB it has been extracted

#### Relevant vs non-relevant sound KB query format

Relevant vs non-relevant sound KB contains audio features of the relevant sounds.

Relevant vs non-relevant sound KB is queried by giving a sound as input. Relevant vs non-relevant sound KB responds by giving the relevant sound.

**To Respondents**

Respondents should propose a query format capable to provide a Boolean value (relevant/non-relevant) or a probability level (e.g., 70% relevant).

#### 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 [40]. There are currently at least 2 SDKs on the matter: MIMI SDK, NURA SDK (both proprietary) [41].

**To Respondents**

Respondents should propose a format which can convey the unique individual sound perception ability, in one of the following ways

1. The KB responds to a query with the values of the frequency perception of the user at a pre-defined set of frequency values
2. The KB responds to a query with the value of the frequency perception of the user at a specified frequency values with the query of a specific frequency value.

#### 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 9* introduces the acronyms representing the MPAI-CAE and MPAI-MMC Use Cases.

*Table 9 – 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 10* gives all MPAI-CAE and MPAI-MMC technologies in alphabetical order.

Please note the following acronyms

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

*Table 10 – Alphabetically ordered MPAI-CAE and MPAI-MMC technologies*

|  |  |  |
| --- | --- | --- |
| **UC** | **Technology** | **Description** |
| AOG | Delivery | Speech transport format |
| EAE | Delivery | Speech transport format |
| AOG | Digital Audio | PCM Audio 48-96 kHz/16-24 bit |
| ARP | Digital Audio | PCM Audio 48-96 kHz/16-24 bit |
| ARP | Digital Image | A (un)compressed digital video frame |
| MQA | Digital Image | (un)compressed image |
| CWE | Digital Speech | PCM speech 22.05-96kHz/16-24 bit |
| EAE | Digital Speech | PCM speech 22.05-96kHz/16-24 bit |
| EES | Digital Speech | PCM speech 22.05-96kHz/16-24 bit |
| MQA | Digital Speech | PCM speech 22.05-96kHz/16-24 bit |
| PST | Digital Speech | PCM speech 22.05-96kHz/16-24 bit |
| ARP | Digital Video | Digital Video |
| CWE | Digital Video | Digital Video |
| CWE | Emotion | Digital representation of emotion |
| EES | Emotion | Digital representation of emotion |
| EES | Emotion descriptors | Derivations of Speech features |
| CWE | Emotion KB (speech) QF | Provides emotion from speech features |
| CWE | Emotion KB (text) QF | Provides emotion from text features |
| CWE | Emotion KB (video) QF | Provides emotion from video features |
| EES | Emotion KB QF | Provides Emotion descriptors |
| ARP | Image Features | Features of tape irregularities Images |
| MQA | Image features | Features of object Images |
| MQA | Image KB QF | Provides object identifier |
| CWE | Input to speech synthesis | Plain text or concept |
| MQA | Intention | Information such as what, where, how |
| MQA | Intention KB QF | Provides Intention |
| PST | Language identification | Language identifier |
| CWE | Meaning | Information such as question, statement |
| MQA | Meaning | Information such as question, statement |
| AOG | Microphone geometry information | Description of microphone position |
| EAE | Microphone geometry information | Description of microphone position |
| MQA | Object identifier | Identifier of a physical object |
| MQA | Online dictionary QF | Provides paragraphs correlelated with questions |
| EAE | Output device acoustic model metadata KB QF | Provides output device metadata |
| ARP | Packager | Audio/Video/Images/Text Multiplexer |
| AOG | Relevant vs non-relevant sound KB QF | Provides relevant sound |
| AOG | Sound array | Vector of extracted sounds |
| AOG | Sound categorisation KB QF | Provides sound category |
| AOG | Sounds categorisation | Identifier of a type of sound |
| CWE | Speech features | Speech features containing emotion info |
| EES | Speech features | Features associated to speech analysis |
| PST | Speech features | Features of input speech |
| ARP | Tape irregularity KB QF | Provides image features |
| ARP | Text | Plain text |
| MQA | Text | Plain text |
| PST | Text | Plain text |
| CWE | Text features | Text features containing emotion info |
| AOG | User Hearing Profiles KB QF | Provides profile of identified user |
| CWE | Video features | Video features containing emotion info |

The following technologies are potentially applicable to different Use Cases.

*Table 11 – Technologies potentially shared by MPAI-CAE and MPAI-MMC*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Function** | **EES** | **ARP** | **EAE** | **AOG** | **CWE** | **MQA** | **PST** |
| Delivery |  |  | X | X |  |  |  |
| Digital speech | X |  | X |  |  |  |  |
| Digital audio |  | X |  | X |  |  |  |
| Digital image |  | X |  |  |  | X |  |
| Digital video |  | X |  |  | X |  |  |
| Emotion | X |  |  |  | X |  |  |
| Image features |  | X |  |  |  | X |  |
| Meaning |  |  |  |  | X | X |  |
| Microphone geometry information |  |  | X | X |  |  |  |
| Speech features | X |  |  |  | X |  | X |
| Text |  | X |  |  | X | X | X |

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. Speech features are different for all Use Cases.

However, respondents should consider the possibility of proposing a unified set of Speech features as proposed in [42]

# Terminology

*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 |
| 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 | An 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 |
| 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 Call for Technologies; <https://mpai.community/standards/mpai-aif/#Technologies>
2. MPAI-CAE Call for Technologies; N131
3. MPAI-MMC Use Cases and Functional Requirements; N134
4. F. Burkhardt and N. Campbell, “Emotional speech synthesis,” in The Oxford Handbook of Affective Computing. Oxford University Press New York, 2014, p. 286
5. 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
6. T. W. Adorno, Philosophy of New Music, University of Minnesota Press, Minneapolis, Minn, USA, 2006
7. 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.
8. Plutchik R., Emotion: a psychoevolutionary synthesis, New York Harper and Row, 1980
9. Russell, James (1980). "A circumplex model of affect". Journal of Personality and Social Psychology. 39 (6): 1161–1178. doi:10.1037/h0077714
10. Cahn, J. E., The Generation of Affect in Synthesized Speech, Journal of the American Voice I/O Society, 8, July 1990, p. 1-19
11. https://www.w3.org/TR/2014/REC-emotionml-20140522/
12. Cahn, J. E., The Generation of Affect in Synthesized Speech, Journal of the American Voice I/O Society, 8, July 1990, p. 1-19
13. 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.
14. 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
15. 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
16. 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
17. 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
18. 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
19. 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
20. Mozziconacci, S. J. L., Speech Variability and Emotion: Production and Perception, PhD Thesis, Technical University Eindhoven, 1998
21. 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.
22. Cahn, J. E., The Generation of Affect in Synthesized Speech, Journal of the American Voice I/O Society, 8, July 1990, p. 1-19
23. 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]
24. 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
25. Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, Massimiliano Todisco. EMOVO Corpus: an Italian Emotional Speech Database.
26. 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
27. IASA-TC 05: Handling and Storage of Audio and Video Carriers. IASA Technical Committee (2014)
28. 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
29. ISO/IEC 10918-1:1994 Information Technology — Digital Compression And Coding Of Continuous-Tone Still Images: Requirements And Guidelines
30. 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
31. 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.
32. 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
33. S.J. Godsill and P.J.W. Rayner. Digital Audio Restoration - a statistical model-based approach (Berlin: Springer-Verlag 1998)
34. 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
35. 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
36. ISO/IEC 10646:2003 Information Technology — Universal Multiple-Octet Coded Character Set (UCS)
37. https://www.iis.fraunhofer.de/en/ff/amm/broadcast-streaming/mpegh.html
38. https://docs.microsoft.com/bs-cyrl-ba/azure/cognitive-services/speech-service/how-to-devices-microphone-array-configuration
39. https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01277/full
40. https://help.nuraphone.com/hc/en-us/articles/360000324676-Your-Profile
41. https://integrate.mimi.io/documentation/android/4.0.1/documentation
42. Problem Agnostic Speech Encoder; https://github.com/santi-pdp/pase