



IEEE Standard Adoption of Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) Technical Specification Context-based Audio Enhanced (CAE) Version 1.4

Developed by the
Collaborative Activities Governance Board

IEEE SA Industry Affiliated Network (IAN)
Base specification contributed by MPAI community



IEEE Std 3302™-2022

IEEE Standard Adoption of Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) Technical Specification Context-based Audio Enhanced (CAE) Version 1.4

Developed by the

Collaborative Activities Governance Board
of the
IEEE SA Board of Governors

Approved 3 December 2022

IEEE SA Standards Board

Abstract: This standard adopts MPAA Technical Specification Version 1.4 as an IEEE Standard. The Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAA) Technical Specification Context-based Audio Enhancement (CAE) Version 1.4 is a collection of four use cases specifying AI-based technologies for audio-related applications including entertainment, communication, post-production, teleconferencing, and restoration.

Keywords: AI Module, AI Workflow, ARP, Audio Recording Preservation, EAE, EES, Emotion-Enhanced Speech, Enhanced Audio-Conference Experience, IEEE 3302™, MPAA-CAE

The Institute of Electrical and Electronics Engineers, Inc.
3 Park Avenue, New York, NY 10016-5997, USA

Copyright © 2023 by The Institute of Electrical and Electronics Engineers, Inc.
All rights reserved. Published 28 April 2023. Printed in the United States of America.

IEEE is a registered trademark in the U.S. Patent & Trademark Office, owned by The Institute of Electrical and Electronics Engineers, Incorporated.

The MPAA specification, Copyright ©MPAA 2021-2022. All rights reserved.

PDF: ISBN 978-1-5044-9342-0 STD25887
Print: ISBN 978-1-5044-9343-7 STDPD25887

IEEE prohibits discrimination, harassment, and bullying.

For more information, visit <https://www.ieee.org/about/corporate/governance/p9-26.html>.

No part of this publication may be reproduced in any form, in an electronic retrieval system or otherwise, without the prior written permission of the publisher.

Important Notices and Disclaimers Concerning IEEE Standards Documents

IEEE Standards documents are made available for use subject to important notices and legal disclaimers. These notices and disclaimers, or a reference to this page (<https://standards.ieee.org/ipr/disclaimers.html>), appear in all standards and may be found under the heading “Important Notices and Disclaimers Concerning IEEE Standards Documents.”

Notice and Disclaimer of Liability Concerning the Use of IEEE Standards Documents

IEEE Standards documents are developed within IEEE Societies and subcommittees of IEEE Standards Association (IEEE SA) Board of Governors. IEEE develops its standards through an accredited consensus development process, which brings together volunteers representing varied viewpoints and interests to achieve the final product. IEEE Standards are documents developed by volunteers with scientific, academic, and industry-based expertise in technical working groups. Volunteers are not necessarily members of IEEE or IEEE SA and participate without compensation from IEEE. While IEEE administers the process and establishes rules to promote fairness in the consensus development process, IEEE does not independently evaluate, test, or verify the accuracy of any of the information or the soundness of any judgments contained in its standards.

IEEE makes no warranties or representations concerning its standards, and expressly disclaims all warranties, express or implied, concerning this standard, including but not limited to the warranties of merchantability, fitness for a particular purpose and non-infringement. In addition, IEEE does not warrant or represent that the use of the material contained in its standards is free from patent infringement. IEEE standards documents are supplied “AS IS” and “WITH ALL FAULTS.”

Use of an IEEE standard is wholly voluntary. The existence of an IEEE Standard does not imply that there are no other ways to produce, test, measure, purchase, market, or provide other goods and services related to the scope of the IEEE standard. Furthermore, the viewpoint expressed at the time a standard is approved and issued is subject to change brought about through developments in the state of the art and comments received from users of the standard.

In publishing and making its standards available, IEEE is not suggesting or rendering professional or other services for, or on behalf of, any person or entity, nor is IEEE undertaking to perform any duty owed by any other person or entity to another. Any person utilizing any IEEE Standards document, should rely upon his or her own independent judgment in the exercise of reasonable care in any given circumstances or, as appropriate, seek the advice of a competent professional in determining the appropriateness of a given IEEE standard.

IN NO EVENT SHALL IEEE BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO: THE NEED TO PROCURE SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE PUBLICATION, USE OF, OR RELIANCE UPON ANY STANDARD, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE AND REGARDLESS OF WHETHER SUCH DAMAGE WAS FORESEEABLE.

Translations

The IEEE consensus development process involves the review of documents in English only. In the event that an IEEE standard is translated, only the English version published by IEEE is the approved IEEE standard.

Official statements

A statement, written or oral, that is not processed in accordance with the IEEE SA Standards Board Operations Manual shall not be considered or inferred to be the official position of IEEE or any of its committees and shall not be considered to be, nor be relied upon as, a formal position of IEEE. At lectures, symposia, seminars, or educational courses, an individual presenting information on IEEE standards shall make it clear that the presenter's views should be considered the personal views of that individual rather than the formal position of IEEE, IEEE SA, the Standards Committee, or the Working Group. Statements made by volunteers may not represent the formal position of their employer(s) or affiliation(s).

Comments on standards

Comments for revision of IEEE Standards documents are welcome from any interested party, regardless of membership affiliation with IEEE or IEEE SA. However, **IEEE does not provide interpretations, consulting information, or advice pertaining to IEEE Standards documents.**

Suggestions for changes in documents should be in the form of a proposed change of text, together with appropriate supporting comments. Since IEEE standards represent a consensus of concerned interests, it is important that any responses to comments and questions also receive the concurrence of a balance of interests. For this reason, IEEE and the members of its Societies and subcommittees of the IEEE SA Board of Governors are not able to provide an instant response to comments, or questions except in those cases where the matter has previously been addressed. For the same reason, IEEE does not respond to interpretation requests. Any person who would like to participate in evaluating comments or in revisions to an IEEE standard is welcome to join the relevant IEEE working group. You can indicate interest in a working group using the Interests tab in the Manage Profile & Interests area of the [IEEE SA myProject system](#).¹ An IEEE Account is needed to access the application.

Comments on standards should be submitted using the [Contact Us](#) form.²

Laws and regulations

Users of IEEE Standards documents should consult all applicable laws and regulations. Compliance with the provisions of any IEEE Standards document does not constitute compliance to any applicable regulatory requirements. Implementers of the standard are responsible for observing or referring to the applicable regulatory requirements. IEEE does not, by the publication of its standards, intend to urge action that is not in compliance with applicable laws, and these documents may not be construed as doing so.

¹ Available at: <https://development.standards.ieee.org/myproject-web/public/view.html#landing>.

² Available at: <https://standards.ieee.org/content/ieee-standards/en/about/contact/index.html>.

Data privacy

Users of IEEE Standards documents should evaluate the standards for considerations of data privacy and data ownership in the context of assessing and using the standards in compliance with applicable laws and regulations.

Copyrights

IEEE draft and approved standards are copyrighted by IEEE under US and international copyright laws. They are made available by IEEE and are adopted for a wide variety of both public and private uses. These include both use, by reference, in laws and regulations, and use in private self-regulation, standardization, and the promotion of engineering practices and methods. By making these documents available for use and adoption by public authorities and private users, neither IEEE nor its licensors waive any rights in copyright to the documents.

Photocopies

Subject to payment of the appropriate licensing fees, IEEE will grant users a limited, non-exclusive license to photocopy portions of any individual standard for company or organizational internal use or individual, non-commercial use only. To arrange for payment of licensing fees, please contact Copyright Clearance Center, Customer Service, 222 Rosewood Drive, Danvers, MA 01923 USA; +1 978 750 8400; <https://www.copyright.com/>. Permission to photocopy portions of any individual standard for educational classroom use can also be obtained through the Copyright Clearance Center.

Updating of IEEE Standards documents

Users of IEEE Standards documents should be aware that these documents may be superseded at any time by the issuance of new editions or may be amended from time to time through the issuance of amendments, corrigenda, or errata. An official IEEE document at any point in time consists of the current edition of the document together with any amendments, corrigenda, or errata then in effect.

Every IEEE standard is subjected to review at least every 10 years. When a document is more than 10 years old and has not undergone a revision process, it is reasonable to conclude that its contents, although still of some value, do not wholly reflect the present state of the art. Users are cautioned to check to determine that they have the latest edition of any IEEE standard.

In order to determine whether a given document is the current edition and whether it has been amended through the issuance of amendments, corrigenda, or errata, visit [IEEE Xplore](#) or [contact IEEE](#).³ For more information about the IEEE SA or IEEE's standards development process, visit the IEEE SA Website.

Errata

Errata, if any, for all IEEE standards can be accessed on the [IEEE SA Website](#).⁴ Search for standard number and year of approval to access the web page of the published standard. Errata links are located under the Additional Resources Details section. Errata are also available in [IEEE Xplore](#). Users are encouraged to periodically check for errata.

³ Available at: <https://ieeexplore.ieee.org/browse/standards/collection/ieee>.

⁴ Available at: <https://standards.ieee.org/standard/index.html>.

Patents

IEEE Standards are developed in compliance with the [IEEE SA Patent Policy](#).⁵

Attention is called to the possibility that implementation of this standard may require use of subject matter covered by patent rights. By publication of this standard, no position is taken by the IEEE with respect to the existence or validity of any patent rights in connection therewith. If a patent holder or patent applicant has filed a statement of assurance via an Accepted Letter of Assurance, then the statement is listed on the IEEE SA Website at <https://standards.ieee.org/about/sasb/patcom/patents.html>. Letters of Assurance may indicate whether the Submitter is willing or unwilling to grant licenses under patent rights without compensation or under reasonable rates, with reasonable terms and conditions that are demonstrably free of any unfair discrimination to applicants desiring to obtain such licenses.

Essential Patent Claims may exist for which a Letter of Assurance has not been received. The IEEE is not responsible for identifying Essential Patent Claims for which a license may be required, for conducting inquiries into the legal validity or scope of Patents Claims, or determining whether any licensing terms or conditions provided in connection with submission of a Letter of Assurance, if any, or in any licensing agreements are reasonable or non-discriminatory. Users of this standard are expressly advised that determination of the validity of any patent rights, and the risk of infringement of such rights, is entirely their own responsibility. Further information may be obtained from the IEEE Standards Association.

This IEEE standard is based on a technical specification developed by a consortium. Patent declarations may have been provided to the consortium as a membership requirement, or in compliance with the consortium's intellectual property rights (IPR) policies and procedures. Normative references in this standard may contain material subject to patent claims that are also subject to the consortium's or other organizations' IPR policies and procedures. Additional information about declarations may be available from the consortium's website:

MOVING PICTURE, AUDIO AND DATA CODING BY ARTIFICIAL INTELLIGENCE (MPAI)—
<https://mpai.community/standards/mpai-cae/>

IMPORTANT NOTICE

IEEE Standards do not guarantee or ensure safety, security, health, or environmental protection, or ensure against interference with or from other devices or networks. IEEE Standards development activities consider research and information presented to the standards development group in developing any safety recommendations. Other information about safety practices, changes in technology or technology implementation, or impact by peripheral systems also may be pertinent to safety considerations during implementation of the standard. Implementers and users of IEEE Standards documents are responsible for determining and complying with all appropriate safety, security, environmental, health, and interference protection practices and all applicable laws and regulations.

⁵ Available at: <https://standards.ieee.org/about/sasb/patcom/materials.html>.

Participants

At the time this standard was completed, the Context-based Audio Enhancement Working Group had the following membership:

Marina Bosi, *Chair*
Stephen Dukes, *Vice Chair*

<i>Organization Represented</i>	<i>Name of Representative</i>
0xSenses Corporation.....	Yu Yuan
1stCycle Corporation.....	Johnny Lin
Beijing Tsinghua Changgung Hospital.....	Bin Yang
Dreamerse.....	Stephen Dukes
MPAI/Stanford University	Marina Bosi
PetroChina Planning and Engineering Institute.....	Kai Pan

The following members of the entity Standards Association balloting group voted on this standard. Balloters may have voted for approval, disapproval, or abstention.

0xSenses Corporation	Dreamerse	PetroChina Planning and
1stCycle Corporation	MPAI/Stanford University	Engineering Institute

When the IEEE SA Standards Board approved this standard on 3 December 2022, it had the following membership:

David J. Law, *Chair*
Ted Burse, *Vice Chair*
Gary Hoffman, *Past Chair*
Konstantinos Karachalios, *Secretary*

Edward A. Addy	Johnny Daozhuang Lin	Mark Siira
Ramy Ahmed Fathy	Kevin Lu	Dorothy V. Stanley
J. Travis Griffith	Daleep C. Mohla	Lei Wang
Guido R. Hiertz	Andrew Myles	F. Keith Waters
Yousef Kimiagar	Damir Novosel	Karl Weber
Joseph L. Koepfinger*	Annette D. Reilly	Sha Wei
Thomas Koshy	Robby Robson	Philip B. Winston
John D. Kulick	Jon Walter Rosdahl	Daidi Zhong

*Member Emeritus

Introduction

This introduction is not part of IEEE Std 3302-2022, IEEE Standard Adoption of Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) Technical Specification Context-based Audio Enhanced (CAE) Version 1.4.

This IEEE standard is based on a technical specification developed by a consortium. Patent declarations may be provided to the consortium as a membership requirement, or in compliance with the consortium's intellectual property rights (IPR) policies and procedures. Normative references in this standard may contain material subject to patent claims that are also subject to the consortium's other organizations' IPR policies and procedures. Additional information about declaration may be available from consortium's website.

MOVING PICTURE, AUDIO AND DATA CODING BY ARTIFICIAL INTELLIGENCE (MPAI)—
<http://mpai.community/standards/mpai-cae/>.



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

MPAI Technical Specification

Context-based Audio Enhancement MPAI-CAE

V1.4

WARNING

Use of the technologies described in this Technical Specification may infringe patents, copyrights or intellectual property rights of MPAI Members or non-members.

MPAI and its Members accept no responsibility whatsoever for damages or liability, direct or consequential, which may result from use of this Technical Specification.

Readers are invited to review Annex 1 – Notices and Disclaimers.

Context-based Audio Enhancement

Version V1.4

1	Introduction (Informative).....	4
2	Scope of standard	5
2.1	Emotion-Enhanced Speech (EES).....	6
2.2	Audio Recording Preservation (ARP).....	6
2.3	Speech Restoration System (SRS)	6
2.4	Enhanced Audioconference Experience (EAE)	6
2.5	Normative content of the Use Cases	7
3	Terms and Definitions	7
4	References	9
4.1	Normative References	9
4.2	Informative References	10
5	Use Case Architectures	10
5.1	Emotion-Enhanced Speech (EES).....	10
5.1.1	Scope of Use Case.....	10
5.1.2	I/O data	10
5.1.3	Implementation Architecture.....	11
5.1.4	AI Modules.....	11
5.2	Audio Recording Preservation (ARP).....	12
5.2.1	Scope of Use Case.....	12
5.2.2	I/O data	12
5.2.3	Implementation Architecture.....	12
5.2.4	AI Modules.....	14
5.3	Speech Restoration System (SRS)	14
5.3.1	Scope of Use Case.....	14
5.3.2	I/O Data	15
5.3.3	Implementation Architecture.....	15
5.3.4	AI Modules.....	16
5.4	Enhanced Audioconference Experience (EAE)	16
5.4.1	Scope of Use Case.....	16
5.4.2	I/O data	17
5.4.3	Implementation Architecture.....	17
5.4.4	AI Modules.....	18
6	AIMs.....	19
6.1	AIM Interoperability	19
6.2	AIMs and their data.....	19
6.2.1	Emotion Enhanced Speech.....	19
6.2.2	Audio Recording Preservation (ARP).....	19
6.2.3	Speech Restoration System (SRS)	20
6.2.4	Enhanced Audioconference Experience (EAE)	20
6.3	Data Formats	20
6.3.1	Access Copy Files	21
6.3.2	Audio Scene Geometry	21
6.3.3	Damaged List	23
6.3.4	Denoised Speech	23
6.3.5	Editing List.....	24

6.3.6	Emotion	26
6.3.7	Emotionless Speech.....	30
6.3.8	Interleaved Multichannel Audio.....	30
6.3.9	Irregularity File	30
6.3.10	Irregularity Image.....	33
6.3.11	Microphone Array Audio	33
6.3.12	Microphone Array Geometry	33
6.3.13	Mode Selection.....	35
6.3.14	Multichannel Audio Stream	36
6.3.15	Neural Network Speech Model.....	36
6.3.16	Preservation Audio File.....	37
6.3.17	Preservation Audio-Visual File.....	37
6.3.18	Preservation Master Files	37
6.3.19	Source Dictionary.....	37
6.3.20	Source Model KB Query Format	37
6.3.21	Speech Features.....	37
6.3.22	Spherical Harmonic Decomposition	39
6.3.23	Transform Denoised Speech	39
6.3.24	Transform Speech	39
6.3.25	Transform Multichannel Audio.....	40
6.3.26	Video	40
Annex 1	MPAI-wide terms and definitions	41
Annex 2	Notices and Disclaimers Concerning MPAI Standards (Informative).....	44
Annex 3	Patent Declarations.....	46
Annex 4	Examples (Informative).....	47
A4.1	Audio Scene Geometry	47
A4.2	Damaged List	47
A4.3	Editing List.....	47
A4.4	Irregularity File	48
A4.5	Microphone Array Geometry	49
A4.6	Speech Features 1	50
A4.7	Speech Features 2.....	50
Annex 5	AIW and AIM Metadata of CAE-EES.....	51
A5.1	AIW Metadata	51
A5.2	AIM Metadata	53
A5.2.1	Speech Feature Analyser1	53
A5.2.2	Speech Feature Analyser2.....	54
A5.2.3	Emotion Feature Producer.....	55
A5.2.4	Emotion Inserter1	56
A5.2.5	Emotion Inserter2.....	56
Annex 6	AIW and AIM Metadata of CAE-ARP	58
A6.1	AIW metadata	58
A6.2	AIM metadata.....	62
A6.2.1	Audio Analyser	62
A6.2.2	Video Analyser.....	63
A6.2.3	Tape Irregularity classifier	64
A6.2.4	Tape Audio Restoration	65
A6.2.5	Packager	66
Annex 7	AIW and AIM Metadata of CAE-SRS.....	68
A7.1	AIW metadata	68

A7.2	AIM metadata.....	70
A7.2.1	Speech Model Creation	70
A7.2.2	Speech Synthesiser	70
A7.2.3	Assembler.....	71
Annex 8	AIW and AIM Metadata of CAE-EAE	73
A8.1	AIW Metadata	73
A8.2	AIM Metadata	78
A8.2.1	Analysis Transform	78
A8.2.2	Sound Field Description	79
A8.2.3	Speech Detection and Separation	80
A8.2.4	Noise Cancellation	81
A8.2.5	Synthesis Transform.....	82
A8.2.6	Packager	83

1 Introduction (Informative)

In recent years, Artificial Intelligence (AI) and related technologies have been introduced in a broad range of applications, have started affecting the life of millions of people and are expected to do so even more in the future. As digital media standards have positively influenced industry and billions of people, so AI-based data coding standards are expected to have a similar positive impact. Indeed, research has shown that data coding with AI-based technologies is generally *more efficient* than with existing technologies for, e.g., compression and feature-based description.

However, some AI technologies may carry inherent risks, e.g., in terms of bias toward some classes of users. Therefore, the need for standardisation is more important and urgent than ever.

The international, unaffiliated, not-for-profit MPAI – Moving Picture, Audio and Data Coding by Artificial Intelligence Standards Developing Organisation has the mission to develop *AI-enabled data coding standards*. MPAI Application Standards enable the development of AI-based products, applications, and services.

As a rule, MPAI standards include four documents: Technical Specification, Reference Software Specifications, Conformance Testing Specifications, and Performance Assessment Specifications. The last type of Specification includes standard operating procedures to enable users of MPAI Implementations to make informed decision about their applicability based on the notion of Performance, defined as a set of attributes characterising a reliable and trustworthy implementation. In the following, Terms beginning with a capital letter are defined in *Table 1* if they are specific to this Standard and in *Table 21* if they are common to all MPAI Standards.

In general, MPAI Application Standards are defined as aggregations – called AI Workflows (AIW) – of processing elements – called AI Modules (AIM) – executed in an AI Framework (AIF). MPAI defines Interoperability as the ability to replace an AIW or an AIM Implementation with a functionally equivalent Implementation.

MPAI also defines 3 Interoperability Levels of an AIF that executes an AIW. The AIW and its AIMs may have 3 Levels:

Level 1 – Implementer-specific and satisfying the MPAI-AIF Standard.

Level 2 – Specified by an MPAI Application Standard.

Level 3 – Specified by an MPAI Application Standard and certified by a Performance Assessor.

MPAI offers Users access to the promised benefits of AI with a guarantee of increased transparency, trust and reliability as the Interoperability Level of an Implementation moves from 1 to 3. Additional information on Interoperability Levels is provided in reference [18].

Figure 1 depicts the MPAI-AIF Reference Model under which Implementations of MPAI Application Standards and user-defined MPAI-AIF Conforming applications operate.

MPAI Application Standards normatively specify the Syntax and Semantics of the input and output data and the Function of the AIW and the AIMs, and the Connections between and among the AIMs of an AIW.

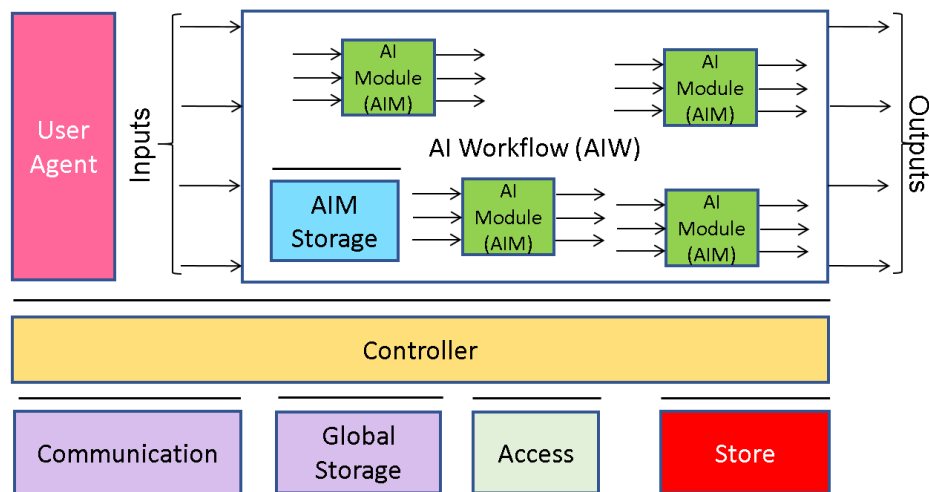


Figure 1 – The AI Framework (AIF) Reference Model and its Components

In particular, an AIM is defined by its Function and data, but not by its internal architecture, which may be based on AI or data processing, and implemented in software, hardware or hybrid software and hardware technologies.

MPAI Standards are designed to enable a User to obtain, via standard protocols, an Implementation of an AIW and of the set of corresponding AIMs, and execute it in an AIF Implementation. The Store in Figure 1 is an entity from which Implementations are downloaded. MPAI Standards assume that the AIF, AIW, and AIM Implementations may have been developed by independent implementers. A necessary condition for this to be possible, is that any AIF, AIW, and AIM implementations be uniquely identified. MPAI has appointed an ImplementerID Registration Authority (IIDRA) to assign unique ImplementerIDs (IID) to Implementers.¹

A necessary condition to make possible the operations described in the paragraph above is the existence of an ecosystem composed of Conformance Testers, Performance Assessors, and an instance of the IIDRA and of the Store. Reference [18] provides an informative example of such ecosystem.

The chapters and the annexes of this Technical Specification are Normative, unless they are labelled as Informative.

2 Scope of standard

The common characteristic shared by the MPAI-CAE Use Cases is the improvement of the user experience for audio-related 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. using context information to act on the input audio content, and potentially deliver the processed output via an appropriate protocol. These use cases are *Emotion Enhanced Speech (EES)*, *Audio Recording Preservation (ARP)*, *Speech Restoration System (SSR)*, and *Enhanced Audioconference Experience (EAE)*.

This version of the MPAI-CAE Technical Specification has been developed by the CAE-DC Development Committee. Future Versions may revise and/or extend the Scope of the Standard.

¹ At the time of publication of this standard, the MPAI Store was assigned as the IIDRA. Information about the assignment of ImplementerIDs is available in Clause 1 of IEEE Std 3300-2022.

2.1 Emotion-Enhanced Speech (EES)

Speech carries information not only about its lexical content, but also about several other aspects including age, gender, identity, and **emotional state of the speaker**. Speech synthesis is evolving towards support of these aspects.

In many use cases, emotional force can usefully be added to speech which by default would be neutral or emotionless, possibly with grades of a particular emotion. For instance, in a human-machine dialogue, messages conveyed by the machine can be more effective if they carry emotions appropriately related to the emotions detected in the human speaker.

Emotion-Enhanced Speech (EES) enables a user to indicate a model utterance or an Emotion to obtain an emotionally charged version of a given utterance.

CAE-EES implementation can be used to create virtual agents communicating as naturally as possible, and thus improve the quality of human-machine interaction by bringing it closer to human-human interchange.

2.2 Audio Recording Preservation (ARP)

Preservation of audio assets recorded on analogue media is an important activity for a variety of application domains, in particular cultural heritage. Preservation goes beyond mere A/D conversion. For instance, the magnetic tape of an open reel may hold important information: it can carry annotations (by the composer or by the technicians) or it can include multiple splices and/or display several types of Irregularities (e.g., corruptions of the carrier, tape of different colour or chemical composition). This information shall be preserved for a correct playback. Nevertheless, some errors can occur during the digitisation as well as the digitisation could be partial because of the corruption of the carrier. These errors shall be restored to make the content listenable. The ARP Use Case (see 5.2) concerns the creation of a digital copy of the digitized audio of open reel magnetic tapes for long-term preservation and of an access copy (restored, if necessary) for correct playback of the digitized recording.

2.3 Speech Restoration System (SRS)

The goal of this use case is to restore a Damaged Segment of an Audio Segment containing only speech from a single speaker. The damage may affect the entire segment, or only part of it.

Restoration will not involve filtering or signal processing. Instead, *replacements* for the damaged vocal elements will be *synthesised* using a speech model. The latter is a component or set of components, normally including one or more neural networks, which accepts text and possibly other specifications, and delivers audible speech in a specified format – here, the speech of the required replacement or replacements. If the damage affects the entire segment, an entirely new segment is synthesized; if only parts are affected, corresponding segments will be synthesized individually to enable later integration into the undamaged parts of the Damaged Segment, with reference to appropriate Time Labels.

The speech segments necessary for the creation of the speech model can be flexibly resourced from undamaged parts of the input segment or from other recording sources that are consistent with the original segment's acoustic environment.

2.4 Enhanced Audioconference Experience (EAE)

The user experience of a video/audio conference is very often far from satisfactory due to multiple competing speakers, non-ideal acoustical properties of the physical spaces that the speakers occupy and/or background noise. These can lead to a reduction in intelligibility of speech resulting in participants not fully understanding what their interlocutors are saying, in addition to creating a distraction and eventually leading to what is known as *audioconference fatigue*. When microphone arrays are used to capture the speakers, most of the described problems can be resolved by

appropriate processing of the captured signals. The speech signals from multiple speakers can be separated from each other, the non-ideal acoustics of the space can be reduced and any background noise can be substantially suppressed.

Enhanced Audioconference Experience (EAE) aims to provide a complete solution to process speech signals recorded by microphone arrays to provide clear speech signals substantially free from background noise and acoustics-related artefacts to improve the auditory quality of audioconference experience.

The AIMs of the Enhanced Audioconference Experience (EAE) Use Case improve auditory experience in an audioconference, thereby reducing the effects of audioconference fatigue.

2.5 Normative content of the Use Cases

Each Use Case normatively defines:

1. The Functions of the AIW and of the AIMs.
2. The Connections between and among the AIMs.
3. The Semantics and the Formats of the input and output data of the AIW and the AIMs.

The word *normatively* implies that an Implementation claiming Conformance to:

1. An *AIW*, shall:
 - a. Have the AIW Function specified in the appropriate Section of Chapter 4.2.
 - b. Have all its AIMs and their Connections conforming with the AIW Reference Model specified in the appropriate Section of Chapter 4.2.
 - c. The AIW and AIM input and output data should have the Formats specified in the appropriate Subsection of Section 6.3.
2. An *AIM*, shall:
 - a. Have the AIM Function specified by the appropriate Section of Chapter 4.2.
 - b. Have input and output data Formats conforming with the appropriate Subsection of Section 6.3.
 - c. Receive as input and produce as output data having the Format specified in Section 6.3.
3. A data *Format*, the data shall have the Format specified in Section 6.3.

Users of this Technical Specification should note that:

1. This Technical Specification defines Interoperability Levels but does not mandate any.
2. Implementers are free to decide the Interoperability Level their Implementation should satisfy.
3. Implementers can use the Reference Software specification to develop their Implementations.
4. The Conformance Testing specification can be used to test the conformity of an Implementation to this Standard.
5. Performance Assessors can assess the level of Performance of an Implementation based on the Performance Assessment specification of this Standard.
6. Implementers and Users should consider the notices and disclaimers of Annex 2.

3 Terms and Definitions

The Terms used in this standard whose first letter is capital have the meaning defined in *Table 1*. The general MPAI Terms are defined in *Table 21*.

Table 1 – Table of terms and definitions

Term	Definition
Access Copy Files	Set of files providing the information stored in an audio tape recording, including Restored Audio Files, suitable for audio information access, but not for long-term preservation.

Audio	Digital representation of an analogue audio signal sampled at a frequency between 8-192 kHz with a number of bits/sample between 8 and 32.
Audio Block	A set of consecutive Audio samples.
Audio Channel	A sequence of Audio Blocks.
Audio File	A .wav file [5].
Audio Object	Direct audio source which is in the audible frequency band.
Audio Scene Geometry	Spatial information for the Audio Objects which are included in an audio scene.
Audio Segment	An Audio Block with Start Time and an End Time Labels corresponding to the time of the first and last sample of the Audio Segment, respectively.
Audio-Visual File	A file containing audio and video according to the MP4 File Format [9].
Capstan	The capstan is a rotating spindle used to move recording tape through the mechanism of a tape recorder.
Damaged List	A list of strings of Texts corresponding to the Damaged Segments (if any) requiring replacement with synthetic segments.
Damaged Section	An Audio Segment which is damaged in its entirety and is contained in a Damaged Segment.
Damaged Segment	An Audio Segment containing only speech (and not containing music or other sounds) which is either damaged in its entirety or contains one or more Damaged Sections specified in the Damaged List.
Degree	Strength of a feature, specifically, with respect to Emotion, “High,” “Medium,” or “Low.”
Editing List	The description of the speed, equalisation and reading backwards corrections occurred during the restoration process.
Emotion	One of the human emotions listed in <i>Table 15</i> , or in an augmented or alternate version of this <i>Table 15</i> .
Emotionless Speech	An Audio File containing speech without music and other sounds, and in which little or no identifiable emotion is perceptible by native listeners.
Interleaved Multichannel Audio	A data structure containing at least 2 time-aligned interleaved Audio Channels.
Irregularity	An event of interest to preservation from in <i>Table 17</i> and <i>Table 18</i>
Irregularity File	A JSON file containing information about Irregularities of the ARP inputs.
Irregularity Image	An Image corresponding to an Irregularity.
JSON	JavaScript object notation [12].
Microphone Array Geometry	Description of the position of each microphone comprising the microphone array and specific characteristics such as microphone type, look directions, and the array type.
Model Utterance	An Audio Segment used as a model or demonstration of the Emotion to be added to Emotionless Speech in order to produce Speech with Emotion.
Multichannel Audio	A set of multiple time-aligned Audio Channels
Multichannel Audio + Audio Scene Geometry	Multichannel Audio packaged with Audio Scene Geometry.

Neural Network Speech Model	A Neural Network Model trained on Speech Segments for Modelling and used to synthesize replacements for the entire Damaged Segment or Damaged Sections within it.
Passthrough AIM	An AIM with the same input and output data of an AIM without executing the Function of that AIM. E.g., a Noise Cancellation AIM that does not cancel the noise.
Preservation Audio File	The input Audio File resulting from the digitisation of an audio open-reel tape to be preserved and, in case, restored.
Preservation Audio-Visual File	The input Audio-Visual File produced by a camera pointed to the playback head of the magnetic tape recorder and the synchronised Audio resulting from the tape digitisation process.
Preservation Image	A Video frame extracted from Preservation Audio-Visual File.
Preservation Master Files	Set of files providing the information stored in an audio tape recording without any restoration. As soon as the original analogue recordings is no more accessible, it becomes the new item for long-term preservation.
Restored Audio Files	Set of Audio Files derived from the Preservation Audio File, where potential speed, equalisation or reading backwards errors that occurred in the digitisation process have been corrected.
Restored Audio Segment	An Audio Segment in which the entire segment has been replaced by a synthetic speech segment, or in which each Damaged Segment has been replaced by a synthetic speech segment.
Speech Segments for Modelling	A set of Audio Files containing speech segments used to train the Neural Network Speech Model.
Speech With Emotion File	An Audio File containing speech with emotional features.
Spherical Coordinate System	A coordinate system where the position of a point is specified by three numbers: the radial distance of that point from a fixed origin, its polar angle measured from a fixed zenith direction, and the azimuthal angle of its orthogonal projection on a reference plane.
Spherical Grid Resolution	The maximum spherical angle between any two neighbouring sampled points on a sphere.
Time Code	Number of ms from 1970-01-01T00:00:00.000 according to [3].
Time Label	A measure of time from a context-dependent zero time expressed as HH:mm:ss.SSS.
Transform Audio	A frequency representation of Audio
Transform Denoised Speech	Transform Audio whose samples are Denoised Speech samples.
Useful Signal	Digital signal resulting from the A/D conversion of the analogue signal recorded in an audio tape.

4 References

4.1 Normative References

This standard normatively references the following technical specifications, both from MPAI and other standard organisations:

1. MPAI; Technical Specification: Artificial Intelligence Framework (MPAI-AIF) V1; <https://bit.ly/3t3SNDT>.
2. A Universally Unique IDentifier (UUID) URN Namespace; IETF RFC 4122; July 2005.
3. Date and Time on the Internet: Time Stamps; IETF RFC 3339; July 2002.

4. Universal Coded Character Set (UCS): ISO/IEC 10646; December 2020.
5. WAVE PCM sound file format, <http://soundfile.sapp.org/doc/WaveFormat/>.
6. ISO/IEC 14496-10; Information technology – Coding of audio-visual objects – Part 10: Advanced Video Coding.
7. ISO/IEC 23008-2; Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 2: High Efficiency Video Coding.
8. ISO/IEC 23094-1; Information technology – General video coding – Part 1: Essential Video Coding.
9. ISO/IEC 14496-12; Information technology – Coding of audio-visual objects – Part 12: ISO base media file format.
10. ZIP format, <https://pkware.cachefly.net/webdocs/casestudies/APPNOTE.TXT>.
11. Neural Network Exchange Format; <https://www.khronos.org/registry/NNEF/specs/1.0/nnef-1.0.4.pdf>; Khronos.
12. The JavaScript Object Notation (JSON) Data Interchange Format; <https://datatracker.ietf.org/doc/html/rfc8259>; IETF rfc8259; December 2017.
13. BS EN 60094-1:1994, BS 6288-1: 1994, IEC 94-1:1981 - Magnetic tape sound recording and reproducing systems - Part 1: Specification for general conditions and requirements.
14. K. Bradley, IASA TC-04 Guidelines in the Production and Preservation of Digital Audio Objects: standards, recommended practices, and strategies, 2nd ed. International Association of Sound and Audiovisual Archives, (2009): 2014.
15. MPAA; The MPAA Statutes; <https://mpaa.community/statutes/>
16. MPAA; The MPAA Patent Policy; <https://mpaa.community/about/the-mpaa-patent-policy/>.
17. Framework Licence of the Context-based Audio Enhancement Technical Specification (MPAA-CAE); <https://mpaa.community/standards/mpaa-cae/framework-licence/>

4.2 Informative References

The references provided here are for information purpose.

18. MPAA; Technical Specification: The governance of the MPAA Ecosystem (MPAA-GME) V1; <https://bit.ly/3t3SNDT>.
19. Ekman, Paul (1999), "Basic Emotions", in Dalgleish, T; Power, M (eds.), Handbook of Cognition and Emotion (PDF), Sussex, UK: John Wiley & Sons.
20. B. Rafaely, Fundamentals of spherical array processing, Springer, 2018.

5 Use Case Architectures

5.1 Emotion-Enhanced Speech (EES)

5.1.1 Scope of Use Case

Emotion-Enhanced Speech (EES) converts an individual emotionless speech segment to a segment that has a specified emotion. Both input and output speech segments are contained in files. The desired emotion is expressed either as a tag belonging to a standard list of emotions or derived by extracting features from a model utterance. EES produces an output speech segment with emotion.

5.1.2 I/O data

Table 2 gives the input and output data of Emotion-Enhanced Speech.

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

Input data	Comments
Emotionless Speech	See definition in Table 1.

Emotion	See definition in <i>Table 1</i> .
Model Utterance	See definition in <i>Table 1</i> .
Output data	Comments
Speech with Emotion	See definition in <i>Table 1</i> .

5.1.3 Implementation Architecture

The Emotion-Enhanced Speech Reference Model depicted in *Figure 2* supports two Modes or pathways enabling addition of emotional charge to an emotionless or neutral utterance (Emotion-less Speech).

1. Along Pathway 1 (*Figure 2*), upper and middle left), a Model Utterance is input together with the neutral utterance Emotion-less Speech, so that features of the former can be captured and transferred to the latter.
2. Alternatively, along Pathway 2 (*Figure 2*), middle and lower left), neutral utterance Emotion-less Speech is input along with a specification of the desired Emotion. Speech Feature Analyser2 extracts Emotionless Speech Features from Emotionless Speech, which describe its initial state. These are sent to Emotion Feature Producer, which produces (emotional) Speech Features2 that can add the desired emotional charge to Emotionless Speech. These Speech Features2 are sent to Emotion Inserter2, which combines Emotionless Speech and the (emotional) Speech Features2 set. Speech with Emotion is then produced as output.

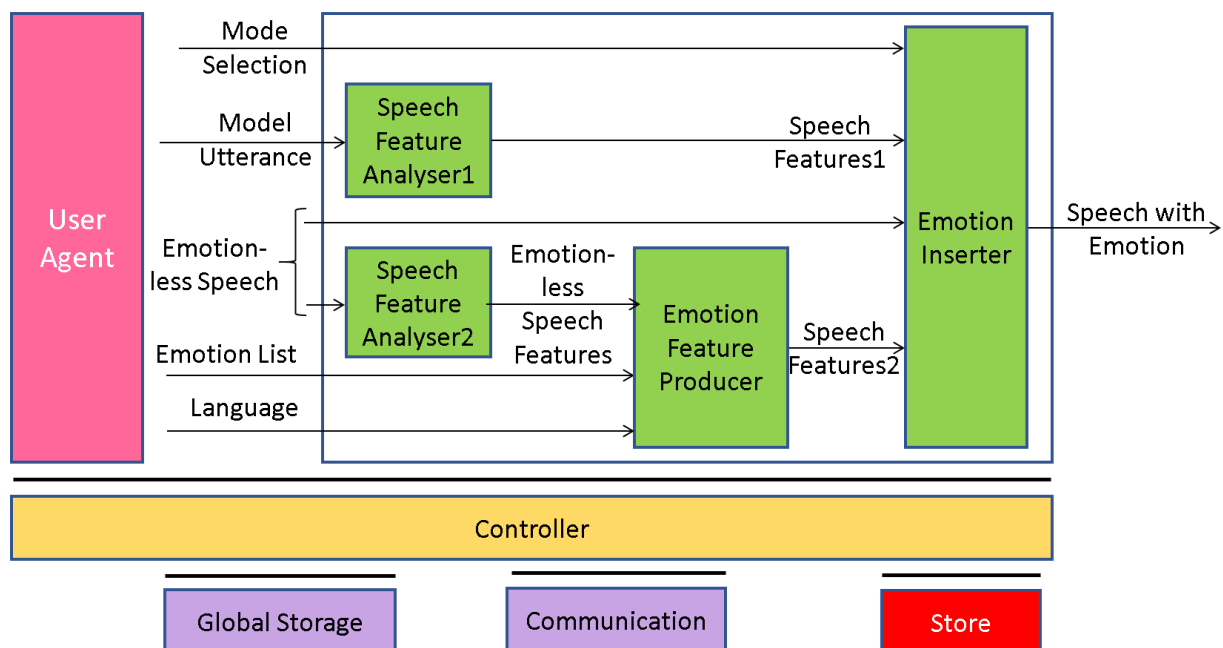


Figure 2 - Emotion-Enhanced Speech Reference Model

5.1.4 AI Modules

The AI Modules of *Figure 2* perform the functions described in *Table 3*.

Table 3 – AI Modules of Emotion-Enhanced Speech

AIM	Function
-----	----------

Speech Feature Analyser 1	Extracts Speech Features ¹ of a model emotional utterance and transfers them to the Emotion 1Inserter ¹ .
Speech Feature Analyser2	Extracts Emotionless Speech Features of an emotionless input utterance, passing these to Emotion Feature Producer.
Emotion Feature Producer	Receives the Emotionless Speech Features produced by Speech Feature Analyser2 plus a list of Emotions to be added. (If the Degree of an Emotion is not specified, the Medium value is used.)
Emotion Inserter1	Integrates the (emotional) Speech Features ¹ with those of the Emotionless Speech input, yielding and delivering an emotionally modified utterance.
Emotion I2Inserter2	Integrates the (emotional) Speech Features ² with those of the Emotionless Speech input, yielding and delivering an emotionally modified utterance.

5.2 Audio Recording Preservation (ARP)

5.2.1 Scope of Use Case

In this Audio Recording Preservation Use Case, two files are fed into a preservation system:

1. A Preservation Audio File obtained by digitising the analogue tape audio recording composed of music, soundscape or speech read from a magnetic tape.
2. A Preservation Audio-Visual File produced by a camera pointed to the playback head of the magnetic tape recorder.

The following is not required:

1. Alignment of the start and end times of the two files. However, the maximum tolerated misalignment is 10s.
2. Presence of signal at the start and the end of the two files.
3. Alignment of the Useful Signal on both files.
4. The same time base for both files. However, the time difference between the same samples in two files shall not be more than 30ms for a 1-hour audio tape.

The output of the restoration process is composed by:

1. Preservation Master Files.
2. Access Copy Files.

5.2.2 I/O data

Table 4 gives the input and output data of Audio Recording Preservation.

Table 4 – I/O data of Audio Recording Preservation

Input	Comments
Preservation Audio File	See 6.3.16
Preservation Audio-Visual File	See 6.3.17
Output data	Comments
Preservation Master Files	See 6.3.18
Access Copy Files	See 6.3.1

5.2.3 Implementation Architecture

Figure 3 depicts the Audio Recording Preservation Reference Model.

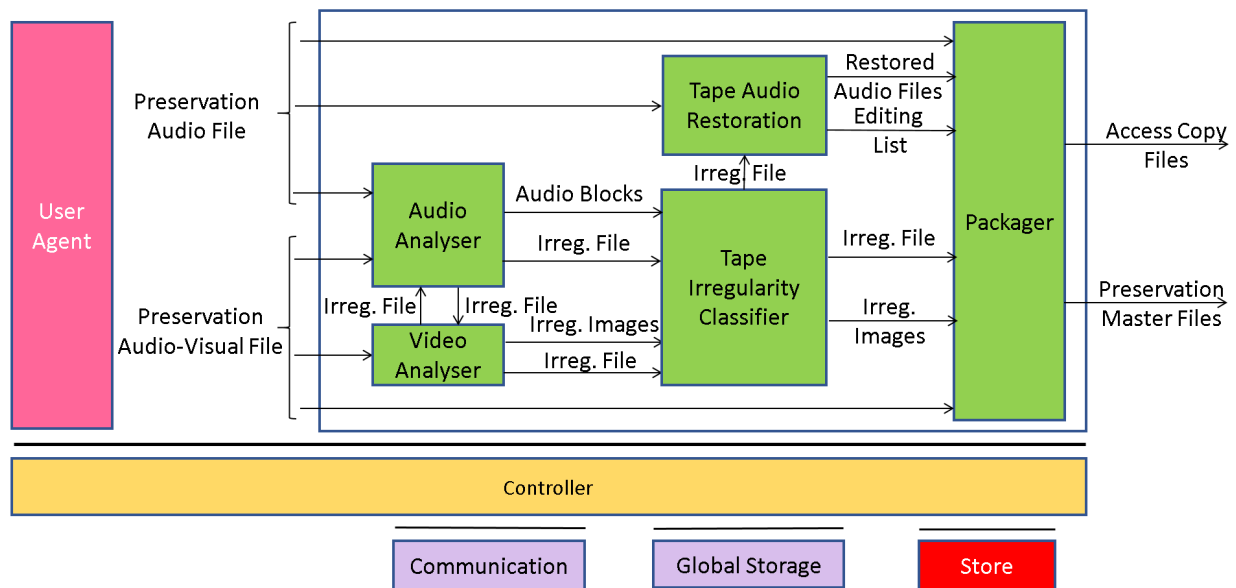


Figure 3 – Audio Recording Preservation Reference Model

The sequence of operations of the Audio Recording Preservation unfolds as follows:

1. The analogue audio signal from the open-reel tape recorder is digitised as Preservation Audio File.
2. The Preservation Audio-Visual File is the combination of:
 - a. The video camera pointed at the playback head of the open-reel tape recorder.
 - b. The analogue audio signal digitised with the same video clock.
3. The Video Analyser:
 - a. Detects Irregularities.
 - b. Assigns IDs to them that are unique to the analysed open-reel tape.
 - c. Receives an Irregularity File from the Audio Analyser and the offset between Preservation Audio File and the Preservation Audio-Visual File.
 - d. Extracts the Images corresponding to each Irregularity received or detected.
 - e. Sends the Irregularity Images and the Irregularity File related to all Irregularities to the Tape Irregularity Classifier.
4. The Audio Analyser:
 - a. Detects Irregularities.
 - b. Assigns IDs to them that are unique to the analysed open-reel tape.
 - c. Receives an Irregularity File from the Video Analyser, extracts the Audio Blocks corresponding to each Irregularity detected and each Irregularity File received or detected.
 - d. Sends the Audio Blocks and the Irregularity File related to all Irregularities to the Tape Irregularity Classifier.
5. The Tape Irregularity Classifier:
 - a. Receives an Irregularity File with the corresponding Images and Audio Blocks.
 - b. Classifies and selects the ones considered relevant.
 - c. If the Irregularity was detected by the Video Analyser, the selected Irregularity File and the corresponding Irregularity Images are sent to the Packager.
6. The Tape Audio Restoration uses the Irregularity File to identify and restore portions of the Preservation Audio File.
7. The Packager collects the Preservation Audio File, Restored Audio Files, the Editing List, the Irregularity File and corresponding Irregularity Images if detected by the Video Analyser, and

the Preservation Audio-Visual File and then it produces the Preservation Master Files and Access Copy Files.

5.2.4 AI Modules

The AIMs required by this Use Case are described in *Table 5*.

Table 5 – AI Modules of Audio Recording Preservation

AIM	Function
Audio Analyser	1. At the start, it calculates the offset between Preservation Audio and the Audio of the Preservation Audio-Visual File. 2. Detects Irregularities of the Preservation Audio File and produces the related Audio Blocks as well as the corresponding Irregularity File. 3. Sends the Irregularity File related to detected Irregularities to the Video Analyser. 4. Receives the Irregularity File detected by Video Analyser. 5. Extracts Audio Blocks corresponding to the Irregularities detected by Video Analyser. 6. At the end, it merges the produced Irregularity File with the one received from the Video Analyser, and it sends it with corresponding Audio Blocks to the Tape Irregularity Classifier.
Video Analyser	1. At the start, it detects Irregularities of the Preservation Audio-Visual File and produces the related Irregularity Images. 2. Sends the Irregularity File related to detected Irregularities to the Audio Analyser. 3. Receives the Irregularity File from the Audio Analyser. 4. Extracts the Irregularity Images corresponding to the Irregularities detected by Audio Analyser. 5. At the end, it merges the produced Irregularity File with the one received from the Audio Analyser, and it sends it with the corresponding Irregularity Images to the Tape Irregularity Classifier.
Tape Irregularities classifier	1. Receives all information from Audio Analyser and Video Analyser, classifies and selects the Irregularities of the Preservation Audio-Visual File and Preservation Audio File considered relevant. 2. Sends the Irregularity File related to the selected Irregularities and the corresponding Irregularity Images to the Packager. 3. Sends the Irregularity File related to the selected Irregularities to Tape Audio Restoration.
Audio Tape Restoration	1. Detects and corrects speed, equalisation and reading backwards errors in Preservation Audio File. 2. Sends Restored Audio Files and Editing List to Packager.
Packager	Produces Preservation Master Files and Access Copy Files.

5.3 Speech Restoration System (SRS)

5.3.1 Scope of Use Case

This Use Case addresses the need for restoration of a Damaged Segment, i.e., a segment containing speech which may be damaged in its entirety or only in part.

Restoration is carried out by synthesizing replacements for the damaged vocal elements as follows:

1. If the damage affects the entire segment, restoration will be carried out by synthesizing an entirely new segment version.
2. If the damage affects only parts of the segment, then those parts will be synthesized individually, and then integrated into the undamaged parts of the Damaged Segment in a final step, as indicated by appropriate Time Labels.

The Speech Segments for Modelling – Audio Segments necessary for the creation of the Neural Network Speech Model – may be obtained from any undamaged parts of the input speech segment; however, other Audio Segments consistent with the original segment’s sound environment can also be used.

5.3.2 I/O Data

Table 6 gives the input and output data of Speech Restoration System.

Table 6 – I/O data of Audio Recording Preservation

Input	Comments
Speech Segments for Modelling	See Table 1.
Text List	See Table 1.
Damaged List	See Table 1.
Damaged Segment	See Table 1.
Output	Comments
Restored Segment	See Table 1.

5.3.3 Implementation Architecture

The Reference Model of the Speech Restoration System is given by Figure 4.

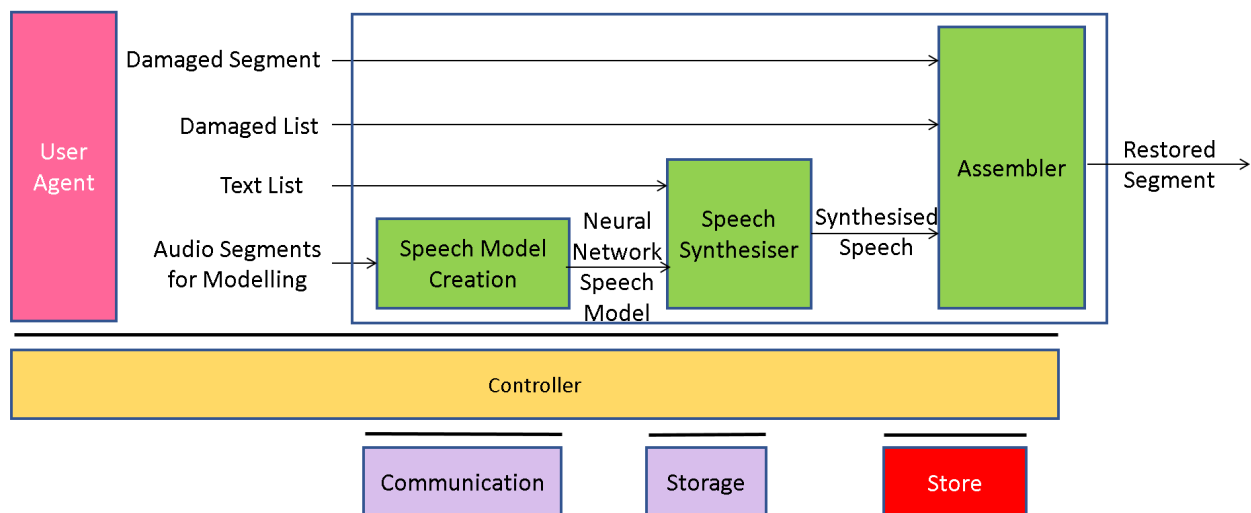


Figure 4 - Speech Restoration System (SRS) Reference Model

In the SRS use case, the entire Damaged Segment can be replaced by a synthesized segment, or parts within it can be synthesized to enable integration of the replaced segments. The sequence of events in this Use Case is as follows:

1. Speech Model Creation receives Audio Segments for Modelling, a set of recordings composing a corpus that will be used to train a Neural Network Speech Model in Speech Model Creation.
2. That Neural Network Speech Model is passed to the Speech Synthesiser AIM, which also receives a Text List as input. Each element of Text List is a string specifying the text of a damaged section of Damaged Segment (or of Damaged Segment as a whole). Speech Synthesiser produces synthetic replacements for each damaged section (or for Damaged Segment as a whole) and passes the replacement(s) to Assembler.
3. Assembler receives as input the entire Damaged Segment, plus Damaged List, a list indicating the locations of any damaged sections within Damaged Segment. The list will be null if Damaged Segment in its entirety was replaced.
4. Assembler produces as output Restored Segment, in which any repaired sections have been replaced by synthetic sections, or in which the entire Damaged Segment has been replaced.

5.3.4 AI Modules

The AIMs required by the Speech Restoration System Use Case are described in *Table 7*

Table 7 - AI Modules of Audio Recording Preservation

AIM	Function
Speech Model Creation	1. Receives in separate files the Audio Segments for Modelling, adequate for model creation. 2. Creates the current Neural Network Speech Model. 3. Sends that Neural Network Speech Model to the Speech Synthesiser.
Speech Synthesiser	1. Receives the current Neural Network Speech Model. 2. Receives Damaged List as a data structure: a. Containing one element if Damaged Segment is damaged throughout or b. Representing a list in which each element specifies via Time Labels the start and end of a damaged section within Damaged Segment. 3. Synthesizes each Damaged Section in Damaged List. 4. Sends the newly synthesised segments to the Assembler as an ordered list.
Assembler	1. Receives the Damaged Segment. 2. Receives the ordered list of synthetic segments. 3. Receives Damaged List Time Labels, indicating where the synthesized segments should be inserted in left-to-right order. In case Damaged Segment as a whole was damaged, the list contains one entry. 4. Assembles the final version of the Restored Segment.

5.4 Enhanced Audioconference Experience (EAE)

5.4.1 Scope of Use Case

The EAE use case addresses the situation where one or more speakers are active in a noisy meeting room and are trying to communicate with one or more interlocutors using speech over a network. The use case is concerned with extracting from microphone array recordings the speech signals from individual speakers as well as reducing the background noise and the reverberation that reduce speech quality. EAE also extracts the spatial attributes of the speakers with respect to the position of the microphone array to facilitate the spatial representation of the speech signals at the receiver side if necessary. These attributes are represented in a well-defined Audio Scene

Geometry metadata format and packaged in a format that is amenable to further processing for efficient delivery and further processing. The coding and compression of the extracted speech signals as well as their reconstruction/representation at the receiver side are outside the scope of this use case.

5.4.2 I/O data

Table 8 shows the input and output data for the Enhanced Audioconference Experience workflow.

Table 8 – I/O data of Enhanced Audioconference Experience

Inputs	Comments
Microphone Array Audio	See 6.3.11
Microphone Array Geometry	See 6.3.12
Outputs	Comments
Multi-channel Audio + Audio Scene Geometry	See Table 1.

5.4.3 Implementation Architecture

Figure 5 shows the Workflow for the EAE.

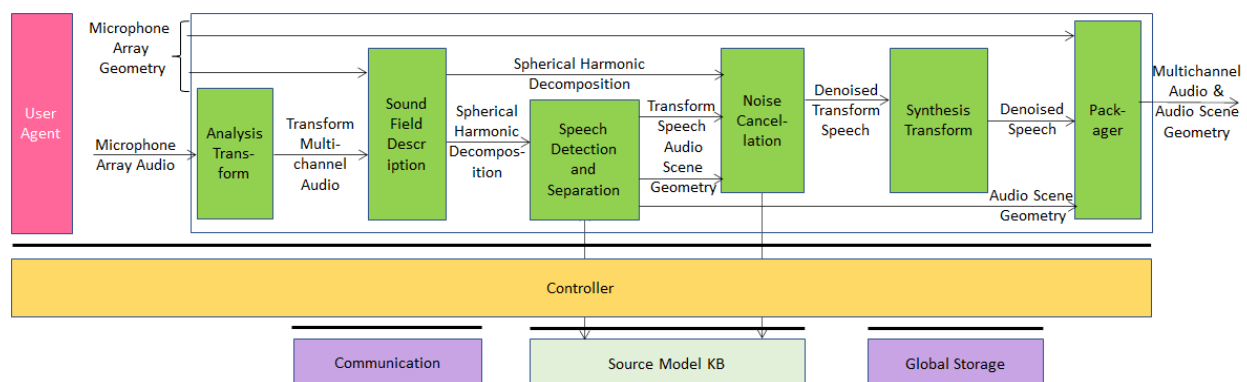


Figure 5 - Enhanced Audioconference Experience Reference Model

The EAE use case receives Microphone Array Audio and Microphone Array Geometry which describes the number, positioning, and configuration of the microphone(s). Using this information, the system can detect the relative directions of the active speakers according to the microphone array and separate relevant audioconference speech sources from each other and from other spurious sounds. Since audio conferencing is a real-time application scenario, the use case operates on Audio Blocks.

The Multichannel Audio is input to EAE as short Multichannel Audio Blocks comprising real valued time domain audio samples where the number of audio samples in each audio block is the same for all the microphones.

The sequence of operations of the EAE use case is the following:

1. **Analysis Transform** transforms the Multichannel Audio into frequency bands via a Fast Fourier Transform (FFT). The following operations are carried out in discrete frequency bands. When such a configuration is used a 50% overlap between subsequent audio blocks needs to be employed. The output is a data structure comprising complex valued audio samples in the frequency domain.

2. **Sound Field Description** converts the output from the Analysis Transform AIM into the spherical frequency domain [20]. If the microphone array used in capturing the scene is a spherical microphone array, Spherical Fourier Transform (SFT) can be used to obtain the Spherical Harmonic Decomposition (SHD) coefficients that represent the captured sound field in the spatial frequency domain. For other types of arrays, more elaborate processing might be necessary. The output of this AIM is $(M \times (N+1)^2)$ complex valued data frame comprising the SHD coefficients up to an order which depends on the number of individual microphones in the array.
3. **Speech Detection and Separation** receives the SHD coefficients of the sound field to detect directions of active sound sources and to separate them. Each separated source can either be a speech or a non-speech signal. Speech detection is carried out on an Audio Block basis by using on each separated source an appropriate voice activity detector (VAD) that is a part of this AIM. This AIM will output speech as an $(M \times S)$ Audio Block comprising transform domain speech signals and block-by-block the Audio Scene Geometry in JSON format comprising auxiliary information which contains a $(M \times 1)$ binary mask indicating the channels of the transform domain SHD coefficients that would be used by the Noise Cancellation AIM for denoising. **Speech Detection and Separation** AIM uses the **Source Model KB** which contains discrete-time and discrete-valued simple acoustic source models that are used in source separation.
4. **Noise Cancellation** eliminates background noise and reverberation which reduce the audio quality. If environmental conditions do not substantially add ambient noise to the desired speech, this AIM acts as a Passthrough AIM.
 - a. It receives Transform Speech from **Speech Detection and Separation** AIM and Acoustic Scene Metadata which includes attributes pertaining to the Audio Block being processed for denoising, and SHD coefficients.
 - b. It uses **Source Model KB**. The output of Noise Cancellation AIM is Denoised Transform Speech as an $(M \times S)$ complex-valued data structure which will in the next stage be processed through **Synthesis Transform** AIM to obtain Denoised Speech.
5. **Synthesis Transform** receives Denoised Transform Speech and outputs Denoised Transform Speech $(F \times S)$ by applying the inverse of the analysis transform.
6. **Packager**:
 - a. Receives Denoised Speech and Audio Scene Geometry.
 - b. Packages the Multichannel Audio stream and the Audio Scene Geometry.
 - c. Produces one interleaved stream which contains separated Multichannel Speech Streams and Audio Scene Geometry.

5.4.4 AI Modules

The AIMs required by the Enhanced Audioconference Experience are given in *Table 9*

Table 9 - AIMs of Enhanced Audioconference Experience

AIM	Function
Analysis Transform	Represents the input Multichannel Audio in a new form amenable to further processing by the subsequent AIMs in the architecture.
Sound Field Description	Produces Spherical Harmonic Decomposition of the Transformed Multichannel Audio.
Speech Detection and Separation	Separates speech and non-speech signals in the Spherical Harmonic Decomposition producing Transform Speech and Audio Scene Geometry.

Noise cancellation	Removes noise and/or suppresses reverberation in the Transform Speech producing Denoised Transform Speech.
Synthesis Transform	Effects inverse transform of Denoised Transform Speech producing Denoised Speech ready for packaging.
Packager	Packages Denoised Speech and the Audio Scene Geometry.

6 AIMs

6.1 AIM Interoperability

To the extent possible, AIM input and output data are specified in a way that is neutral to the technology used to implement the AIM internals. In some cases, however, AIM input and output data of strongly depend on whether the technology used is data processing or Artificial Intelligence. If an AIM is based on, e.g., a neural network, it will need either (1) a usable neural model whose training has included specifiable features, or (2) a precise specification of the features themselves plus an adequate training corpus, so that the AIM using that data can create its own usable model.

6.2 AIMs and their data

6.2.1 Emotion Enhanced Speech

Table 10 – CAE-EES AIMs and their data

AIM	Input Data	Output Data
Speech features Analyser1	Model Utterance	Speech Features1
Speech features Analyser2	Emotionless Speech	Emotionless Speech Features
Emotion Feature Producer	Emotionless Speech Features Emotion List Language	Speech Features2
Emotion Inserter1	Emotionless Speech Speech Features1	Speech with Emotion
Emotion Inserter2	Emotionless Speech Speech Features2	Speech with Emotion

6.2.2 Audio Recording Preservation (ARP)

Table 11 – CAE-ARP AIMs and their data

AIM	Input Data	Output Data
Audio Analyser	Preservation Audio File Preservation Audio-Visual File Irregularity File	Audio Blocks Irregularity File
Video analyser	Preservation Audio-Visual File Irregularity File	Irregularity File Irregularity Images
Tape Irregularity classifier	Audio Blocks Irregularity Images Irregularity File	Irregularity File Irregularity Images
Tape Audio Restoration	Irregularity File Preservation Audio File	Editing List Restored Audio Files
Packager	Preservation Audio File	Access Copy Files

	Restored Audio Files Editing List Irregularity File Irregularity Images Preservation Audio-Visual File	Preservation Master Files
--	--	---------------------------

6.2.3 Speech Restoration System (SRS)

Table 12 – CAE-SRS AIMs and their data

AIM	Input Data	Output Data
Speech Model Creation	Audio Segments for Modelling	Neural Network Speech Model
Speech Synthesiser	Text List Neural Network Speech Model	Synthesised Speech
Assembler	Damaged Segments Damaged List	Restored Segment

6.2.4 Enhanced Audioconference Experience (EAE)

Table 13 – CAE-EAE AIMs and their data

AIM	Input Data	Output Data
Analysis Transform	Multichannel Audio	Transform Multichannel Audio
Sound field Description	Transform Multichannel Audio Geometry Information	Spherical Harmonic Decomposition
Speech Detection and Separation	Spherical Harmonic Decomposition	Transform Speech Audio Scene Geometry
Noise Cancellation	Spherical Harmonic Decomposition Transform Speech Audio Scene Geometry	Denoised Transform Speech
Synthesis Transform	Denoised Transform Speech	Denoised Speech
Packager	Denoised Speech Audio Scene Geometry	Multichannel Audio Audio Scene Geometry

6.3 Data Formats

Table 14 lists all data formats specified in this Technical Specification.

Table 14 – Data formats

Data Format Name	Subsection	Use Case
Access Copy Files	6.3.1	ARP
Audio Scene Geometry	6.3.2	EAE
Damaged List	6.3.3	SRS
Denoised Speech	6.3.4	SRS
Editing List	6.3.5	ARP
Emotion	6.3.6	EES