ITUPublications
Recommendations

International Telecommunication Union
Standardization Sector

Recommendation

# ITU-T F.748.19 (12/2022)

SERIES F: Non-telephone telecommunication services

Multimedia services

# Framework for audio structuralizing based on deep neural networks

ITU-T F-SERIES RECOMMENDATIONS

**NON-TELEPHONE TELECOMMUNICATION SERVICES**

| | |
|---|---|
| TELEGRAPH SERVICE | |
| Operating methods for the international public telegram service | F.1–F.19 |
| The gentex network | F.20–F.29 |
| Message switching | F.30–F.39 |
| The international telemessage service | F.40–F.58 |
| The international telex service | F.59–F.89 |
| Statistics and publications on international telegraph services | F.90–F.99 |
| Scheduled and leased communication services | F.100–F.104 |
| Phototelegraph service | F.105–F.109 |
| MOBILE SERVICE | |
| Mobile services and multidestination satellite services | F.110–F.159 |
| TELEMATIC SERVICES | |
| Public facsimile service | F.160–F.199 |
| Teletex service | F.200–F.299 |
| Videotex service | F.300–F.349 |
| General provisions for telematic services | F.350–F.399 |
| MESSAGE HANDLING SERVICES | F.400–F.499 |
| DIRECTORY SERVICES | F.500–F.549 |
| DOCUMENT COMMUNICATION | |
| Document communication | F.550–F.579 |
| Programming communication interfaces | F.580–F.599 |
| DATA TRANSMISSION SERVICES | F.600–F.699 |
| **MULTIMEDIA SERVICES** | **F.700–F.799** |
| ISDN SERVICES | F.800–F.849 |
| UNIVERSAL PERSONAL TELECOMMUNICATION | F.850–F.899 |
| ACCESSIBILITY AND HUMAN FACTORS | F.900–F.999 |

*For further details, please refer to the list of ITU-T Recommendations.*

# Recommendation ITU-T F.748.19

## Framework for audio structuralizing based on deep neural networks


**Summary**

Recommendation ITU-T F.748.19 presents an overview of the framework for audio structuralizing based on deep neural network. It provides a high-level description of architecture, processing flows, data categories, audio processing tasks and requirements for data management.



**History**

| Edition | Recommendation | Approval | Study Group | Unique ID* |
|---|---|---|---|---|
| 1.0 | ITU-T F.748.19 | 2022-12-14 | 16 | 11.1002/1000/15196 |

**Keywords**

Audio structuralizing, deep neural network.


* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

## FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

## NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

## INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents/software copyrights, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the appropriate ITU-T databases available via the ITU-T website at http://www.itu.int/ITU-T/ipr/.

## Table of Contents


| | | Page |
|---|---|---|
| 1 | Scope | 1 |
| 2 | References | 1 |
| 3 | Definitions | 1 |
| | 3.1 Terms defined elsewhere | 1 |
| | 3.2 Terms defined in this Recommendation | 1 |
| 4 | Abbreviations and acronyms | 1 |
| 5 | Conventions | 2 |
| 6 | Background of audio structuralizing | 2 |
| 7 | Technical framework of the audio structuralizing system | 2 |
| | 7.1 Metadata extraction subsystem | 3 |
| | 7.2 Deep neural network-based audio processing subsystem | 3 |
| | 7.3 External system | 3 |
| | 7.4 Data integration and storage subsystem | 3 |
| | 7.5 Data capability opening subsystem | 4 |
| 8 | Categories of data in the audio structuralizing system | 4 |
| | 8.1 Metadata | 4 |
| | 8.2 Intelligent data | 4 |
| | 8.3 Transaction data | 4 |
| | 8.4 Integration data | 4 |
| 9 | Deep neural network-based audio processing tasks | 5 |
| | 9.1 Standalone task | 5 |
| | 9.2 Multitask | 8 |
| 10 | Requirements for data management in the audio structuralizing system | 8 |
| | 10.1 Data collection | 8 |
| | 10.2 Data storage | 8 |
| | 10.3 Data integration | 9 |
| | 10.4 Data open service | 9 |
| Bibliography | | 10 |

# Recommendation ITU-T F.748.19

## Framework for audio structuralizing based on deep neural networks

### 1 Scope

This Recommendation presents an overview of the framework for audio structuralizing based on deep neural network. It provides a high-level description of architecture, processing flows, data categories, audio processing tasks and data management.

### 2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

None.

### 3 Definitions

#### 3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

**3.1.1 speech** [b-ITU-T H.703]: Speech is the vocalized form of human communication.

**3.1.2 structured data** [b-ITU-T Y.4500.1]: Data that either has a structure according to a specified information model or is otherwise organized in a defined manner.

#### 3.2 Terms defined in this Recommendation

This Recommendation defines the following term:

**3.2.1 audio structuralizing**: A method that makes unstructured audio data structured and provides forms of representation that allows interpretation of the meaning of information.

### 4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

API Application Programming Interface

ASR Automatic Speech Recognition

CNN Convolutional Neural Network

GFCC Gammatone Frequency Cepstral Coefficients

LSTM Long Short-Term Memory

MFCC Mel Frequency Cepstral Coefficients

PLP Perceptual Linear Prediction

RNN Recurrent Neural Network

TDNN Time Delay Neural Network

## 5 Conventions

The following conventions are used in this Recommendation:

– The keywords "**is recommended**" indicate a requirement that is recommended but which is not absolutely required. Thus, this requirement need not be present to claim conformance.

## 6 Background of audio structuralizing

In recent years, with the development of computer science, Internet and telecommunication technology, the amount of multimedia data such as video, audio and image has exploded. As one of the main forms of multimedia, audio plays an increasingly pervasive role in our lives.

Faced with massive amounts of audio content, how to make full and effective use of existing content and explore the value of content has become an urgent problem to be solved. Manually labelling and classifying audio content are very time-consuming. Also, as human hearing fatigues, the accuracy of manual classification will decrease.

As a non-semantic symbolic representation and unstructured binary stream, audio lacks a clear description of structured information and content semantics. This Recommendation provides a framework for audio structuralizing and aims to make unstructured audio data structured and provide forms of representation that allow interpretation of the meaning of information. The value of audio structuralizing is reflected in the following aspects:

a) Audio structuralizing provides different processing methods for different types of audio data. For example, automatic speech recognition (ASR) can be used to automatically create written records of broadcast news; targeted noise reduction processing can be adopted for different environmental sounds.

b) Audio structuralizing plays an important auxiliary role in grading and indexing video. For example, detecting, locating and identifying violent segments such as gunshots, explosions and screams in multimedia.

c) Audio structuralizing establishes the association between the underlying structural units of audio and high-level semantic content, which facilitates in-depth analysis and processing of audio information. For example, quick retrieval of the audio segment of a specified speaker by applying voiceprint recognition technology.

## 7 Technical framework of the audio structuralizing system

The audio structuralizing system is a set of audio data analysis and management subsystems that combines audio processing, deep learning and big data technology. The system can effectively analyse, organize and manage audio data, and can be easily accessed by downstream applications. An overview of the recommended technical framework of the audio structuralizing system is shown in Figure 1.

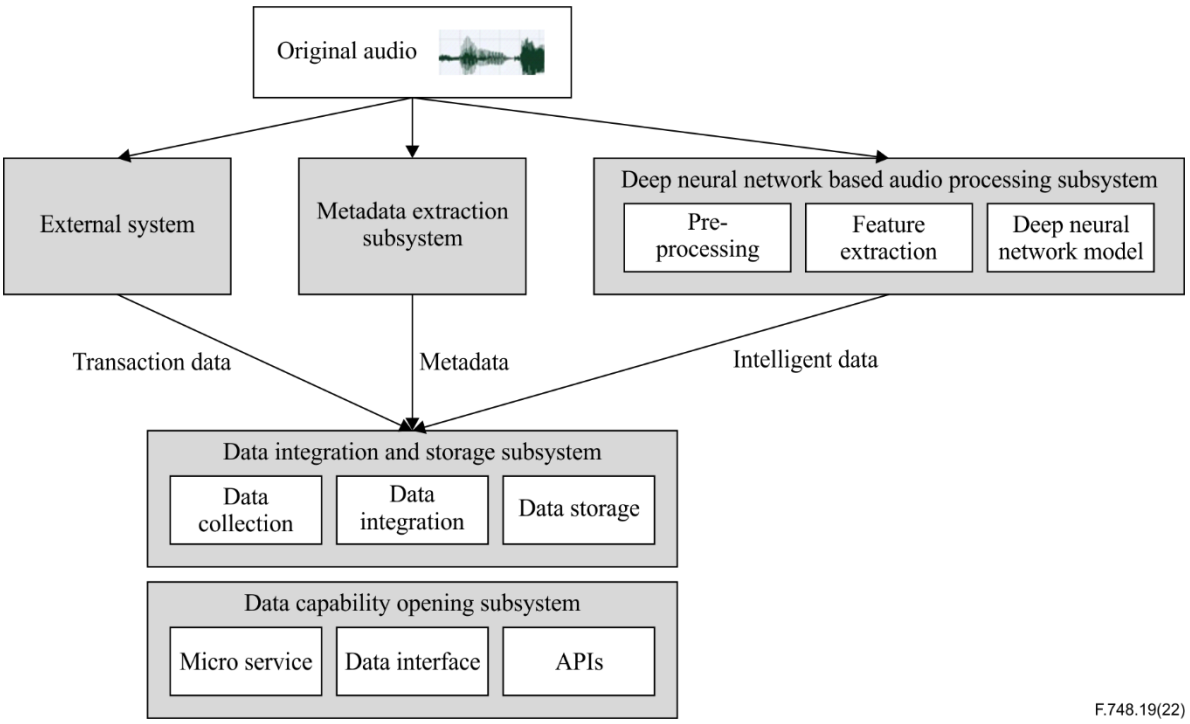


**Figure 1 – Technical framework of the audio structuralizing system**

### 7.1 Metadata extraction subsystem

Metadata extraction subsystem refers to the system that extracts the basic features of audio files.

### 7.2 Deep neural network-based audio processing subsystem

The function of this subsystem is to perform structural analysis of the original audio based on deep neural network. According to the different tasks, the corresponding structured data can be parsed and extracted. The general processing steps include, but are not limited to:

1) Audio pre-processing, such as voice enhancement, voice activity detection.

2) Feature extraction. The feature is generally a single-dimensional or multi-dimensional feature vector group arranged in time series. For example, mel frequency cepstral coefficients (MFCC) is a 39-dimensional feature vector, and perceptual linear prediction (PLP) is a 13-dimensional feature vector. The feature is generally extracted every 25 ms to 30 ms. Different tasks will select different relative features.

3) Modelling based on deep neural networks. Common deep neural networks include convolutional neural networks (CNNs), long short-term memory (LSTM), recurrent neural network (RNN), time delay neural network (TDNN), etc. Developers can select or design appropriate neural networks according to the data set and tasks to perform model training and extract structured data.

### 7.3 External system

External system refers to a system that contains additional valuable information. For example, a service platform that stores calling voice data includes calling number, called number, call duration, call time, call ID, etc.

### 7.4 Data integration and storage subsystem

After cleaning, associating and matching metadata, intelligent data and data from other systems, integration data is generated through technologies such as database table association and knowledge

graphs. Data is associated from multiple service dimensions to form rich and valuable integration data.

The integration data and original audio data are stored persistently. When the amount of data continues to accumulate, big data storage technology will be adopted to ensure the efficiency of external retrieval and query.

### 7.5 Data capability opening subsystem

The data capability opening subsystem provides data to the external system in the form of micro-services, data interfaces and application programming interfaces (APIs). At the same time, it can also provide structured audio big data capabilities for retrieval, query, statistics, analysis, reasoning and other services. Take the task of automatic speech recognition as an example. Automatic speech recognition can directly provide its speech recognition capabilities and/or combine the recognition results with metadata from other systems and/or structured data from other joint tasks to form new structured data ability.

## 8 Categories of data in the audio structuralizing system

Clause 8 describes the categories of data in the audio structuralizing system, as shown in Figure 2.

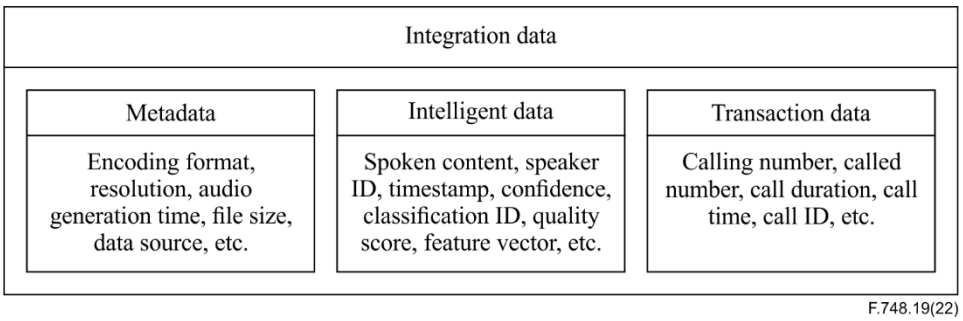


**Figure 2 – Categories of data in the audio structuralizing system**

### 8.1 Metadata

Metadata refers to the basic features in the audio files, such as encoding format, resolution, audio generation time, file size, data source, etc.

### 8.2 Intelligent data

Intelligent data refers to specific information extracted from original audio based on machine learning or deep learning algorithms according to different scenarios and tasks, such as semantic text extracted through ASR, character ID, timestamp and x-vector extracted through voiceprint recognition and voice separation.

### 8.3 Transaction data

Transaction data in the external system can be used as an input source of the data integration and storage subsystem. In addition to the original audio data and its derived data, data from other sources can bring additional information to the audio. For example, a customer service platform that stores call voice transaction data including calling number, called number, call duration, call time, call ID, etc.

### 8.4 Integration data

Integration data is formed by metadata, intelligent data and transaction data. For example, combining transaction data in the customer service system and intelligent data such as speech content and

voiceprint can determine whether the content of the conversation involves fraud and whether the voice is that of the alleged speaker.

## 9 Deep neural network-based audio processing tasks

According to the realization methods of tasks, audio processing tasks are divided into two categories: standalone task and multi-task. Standalone task is designed to tackle with a single service requirement.

Multi-task learning is different from single task learning in the training (induction) process in that inductions of multiple tasks are performed simultaneously to capture intrinsic relatedness. Multi-task learning can enable our model to generalize better on our original task.

The outputs of the tasks are divided into structured data and unstructured data. This clause analyses and explains various types of tasks in detail.

Organizations need to clearly inform and obtain consent of individuals when collecting face images and/or voiceprint, and implement solutions that ensure the protection of personally identifiable information.

The recommended outputs of standalone tasks are described in Figure 3.

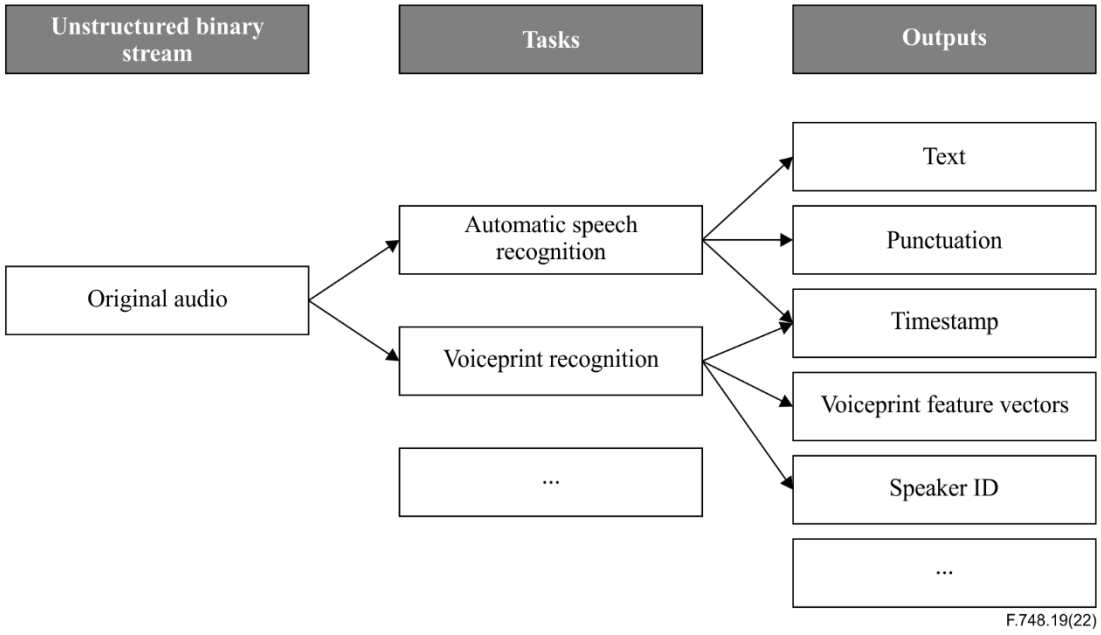


**Figure 3 – Overview processing flows of audio structuralizing based on deep neural network**

### 9.1 Standalone task

#### 9.1.1 Automatic speech recognition

Automatic speech recognition refers to the process of converting human voice signals into words or instructions. The output includes speech text, punctuation and the corresponding timestamp.

#### 9.1.2 Voiceprint recognition

The voiceprint feature is the spectrogram feature of the speaker's voice. Voiceprint recognition, also known as speaker recognition, refers to the process of identifying the speaker corresponding to the speech segment based on the voiceprint characteristics. The outputs are voiceprint feature vectors, speaker ID and timestamp.

### 9.1.3 Speaker diarization

Speaker diarization aims to solve the problem of "who spoke when", the outputs are speaker identities, their respective speech segments and timestamps.

### 9.1.4 Speech separation

Speech separation aims to extract the speaker's voice from the speech with background sound. The outputs can be speech files, speech streams or speech segments of speaker.

### 9.1.5 Audio scene classification

Audio scene classification aims to add semantic tags to the environment by analysing the environmental sounds, such as home scenes, vehicle scenes, natural environments and subway scenes. The output result is the category ID of the audio scene.

### 9.1.6 Sound event detection

Sound event detection aims to detect the presence of a target audio event and its occurrence time in a segment of audio, such as the sound of an explosion. The outputs are the event ID and its corresponding timestamp.

### 9.1.7 Specific speech detection

Specific speech detection aims to detect specific speech content. The outputs include the specific content classification ID and timestamp.

### 9.1.8 Spoofed and fake audio detection

Spoofed and fake audio detection aims to identify whether the voice is actually the alleged speaker. The common methods of speech spoofing are text-to-speech, voice conversion and replay.

### 9.1.9 Equipment fault diagnosis based on acoustics

Industrial inspection based on acoustics refers to detecting whether the sound collected in machine tools, machinery and equipment is abnormal. The output results include category ID of equipment fault.

### 9.1.10 Animal classification based on acoustics

Animal classification can determine the species and class of the animal. The output is animal category ID.

### 9.1.11 Keyword spotting

The purpose of keyword spotting is to detect all occurrences of specified words in the speech signal. The output is corresponding timestamp.

### 9.1.12 Speech enhancement

Speech enhancement refers to the process of beamforming for multi-channel speech, noise suppression, acoustic echo cancelling and dereverberation for single-channel speech. The signal-to-noise ratio of speech can be improved, and the output can be an enhanced speech time-domain or frequency-domain waveform.

### 9.1.13 Audio feature extraction

The purpose of feature extraction is to extract new features closely related to a specific task from the original feature data. Commonly used methods include MFCC, FFT, PLP, gammatone frequency cepstral coefficients (GFCC). The output result can be a one-dimensional or multi-dimensional feature vector set arranged in time series.

### 9.1.14 Voice quality evaluation

Performs voice quality assessment on speech with background noise or reverberation, and the output can be a quality score.

### 9.1.15 Age classification based on voice

The age range of the speaker is predicted by analysing the speaker's voice sample, voice file or voice stream, and the output is the category ID of age range.

### 9.1.16 Gender classification based on voice

The gender of the speaker is predicted by analysing the speaker's voice sample, voice file or voice stream, and the output is the category ID of gender.

### 9.1.17 Musical instrument classification

Musical instrument classification aims to identify whether there are musical instruments in the audio clip and the type of musical instrument. The outputs include instrument category ID and the corresponding timestamp.

Table 1 lists the outputs of the above standalone audio task and whether it can be output as structured data.

**Table 1 – Standalone audio tasks and output**

| Standalone Task | Output | Types of data |
|---|---|---|
| Automatic speech recognition | Speech text, punctuation, timestamp | Structured data |
| Voiceprint recognition | Voiceprint feature vectors, speaker ID, timestamp | Structured data |
| Speaker diarization | Speaker ID, speech segment, timestamp | Structured data |
| Speech separation | Speech files, speech stream or speech segment of speaker | Unstructured data |
| Audio scene classification | Category ID of audio scene | Structured data |
| Sound event detection | Event ID, timestamp | Structured data |
| Specific speech detection | ID of specific speech, timestamp | Structured data |
| Spoofed and fake audio detection | Whether it is actually the alleged speaker | Structured data |
| Equipment fault diagnosis based on acoustics | Category ID of equipment fault | Structured data |
| Animal classification based on acoustics | Category ID of animals | Structured data |
| Keyword spotting | Timestamp | Structured data |
| Speech enhancement | Speech segments, speech files or speech streams after enhancement | Unstructured data |
| Audio feature extraction | One-dimensional or multi-dimensional feature vector | Unstructured data |
| Voice quality evaluation | Score of voice quality | Structured data |
| Age classification based on voice | Category ID of age range | Structured data |
| Gender classification based on voice | Female or male | Structured data |
| Musical instrument classification | Category ID of musical instrument | Structured data |

## 9.2 Multitask

In clause 9.2, several examples of multi-task learning are provided and explained.

### 9.2.1 Integration of ASR and voiceprint recognition

Refers to the simultaneous running of automatic speech recognition and voiceprint recognition in the same model. In the process of joint training, both automatic speech recognition data and voiceprint recognition data can provide more additional information for each other's tasks, and obtain better recognition results than separate training.

### 9.2.2 Joint recognition of face and voice

The joint recognition of face and voice means that the system recognizes the user's voice while recognizing the user's face, and obtains the face recognition result and the voice recognition result at the same time. One possible application scenario is to require the user to read a random verification code when face authentication is performed to improve the reliability of authentication.

### 9.2.3 Integration of voiceprint recognition and spoofed audio detection

Through multi-task learning, a model that can perform voiceprint recognition and spoofed audio detection at the same time is established, and output the authenticity of the voice and spoofed audio clips.

# 10 Requirements for data management in the audio structuralizing system

During the system operation, varying types of structured and unstructured data such as operational data, service reports, program logs and operational texts will be generated. These data are usually distributed in different subsystems, and most of them include hidden attributes and behaviour characteristics of users or services. Through the unified data collection, correlation, and analysis in the audio structuralizing system, enterprises will get useful intelligent data for better operation and provide powerful data support for third-party organizations in a capacity-opening architecture.

## 10.1 Data collection

The data collection process includes data review, extraction, cleaning and conversion. Data collection is recommended to support the operating data of different service sources in the system and follow the predetermined collection strategy. Converting data with different structures and formats into data with a unified format and clear association identification can reduce the complexity of subsequent data processing to a certain extent. The processed data is recommended to meet certain technical requirements, such as uniqueness and relevance. The data collection and pre-processing are recommended to complete the analysis and classification of the data model while retaining the safety and integrity of the original data.

## 10.2 Data storage

The structured data after audio task processing is recommended to meet the persistence requirements. Due to the diversity of the output data from the different logical processing in each subsystem, the construction of multi-source and multi-structure data fusion requires coordination of multiple different storage categories. For example, the metadata, operation data and daily reports are mainly stored as structured data; the original audio and audio clips of the call system are stored in the form of files; the multi-dimensional matrix after the audio feature extraction is stored in a vector database.

The system is recommended to determine the storage types based on the characteristics and relationship types of each metadata. Generally speaking, the key-value database and document database in relational databases and non-relational databases can basically meet the storage requirements of audio structuralizing systems. Vector retrieval databases are recommended to be used for audio feature vector storage and retrieval.

### 10.3 Data integration

Data integration refers to the formation of a consensus model through certain strategies and methods of heterogeneous data collected and organized above. The goal of data integration is to provide a structured database that can be used directly and easily managed, which ultimately provides a basis for various models. This includes a series of data pre-processing and provides a foundation for subsequent data opening capabilities. In the pre-processing, second data cleaning is recommended, which includes filtering non-compliant data, deleting duplicate data, correcting data, and converting format. However, the data after processing may still be unrelated, redundant or missing some information. Therefore, the integration process is recommended to standardize and redefine the data in the repository, and perform subject division and data association if necessary. The processed data is gathered in the data sharing centre and used for subsequent modelling. The specification of data includes the designation of the correspondence between the basic dimensions of the service metadata and the service identifier of the fact table; it can also be the correspondence between the service attributes of the audio file and the structured attributes. In the end, the association and integration of audio attributes and service features will organize the diversified big data, and generate valuable data that satisfy the requirements of refined operation and production.

### 10.4 Data open service

The data open service will build a knowledge graph based on the data sharing centre. It not only provides users with model management, model exploration and data exploration, but also provides intelligent services such as mining analysis and professional modelling. The core knowledge graph is a giant knowledge network composed of nodes and edges. Nodes represent entities, and edges represent relationships between entities. Each entity describes the internal characteristics of the entity through key-value pairs. In addition, modelling can be performed based on core data such as entities and relationships in the knowledge graph, and high-level data mining analysis and processing can be performed. Therefore, unified data collection, data integration and data opening can construct an overall structured model.

# Bibliography


[b-ITU-T H.703] Recommendation ITU-T H.703 (2016), *Enhanced user interface framework for IPTV terminal devices*.

[b-ITU-T Y.4500.1] Recommendation ITU-T Y.4500.1 (2018), *OneM2M- Functional architecture*.

## SERIES OF ITU-T RECOMMENDATIONS

| | |
|---|---|
| Series A | Organization of the work of ITU-T |
| Series D | Tariff and accounting principles and international telecommunication/ICT economic and policy issues |
| Series E | Overall network operation, telephone service, service operation and human factors |
| **Series F** | **Non-telephone telecommunication services** |
| Series G | Transmission systems and media, digital systems and networks |
| Series H | Audiovisual and multimedia systems |
| Series I | Integrated services digital network |
| Series J | Cable networks and transmission of television, sound programme and other multimedia signals |
| Series K | Protection against interference |
| Series L | Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant |
| Series M | Telecommunication management, including TMN and network maintenance |
| Series N | Maintenance: international sound programme and television transmission circuits |
| Series O | Specifications of measuring equipment |
| Series P | Telephone transmission quality, telephone installations, local line networks |
| Series Q | Switching and signalling, and associated measurements and tests |
| Series R | Telegraph transmission |
| Series S | Telegraph services terminal equipment |
| Series T | Terminals for telematic services |
| Series U | Telegraph switching |
| Series V | Data communication over the telephone network |
| Series X | Data networks, open system communications and security |
| Series Y | Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities |
| Series Z | Languages and general software aspects for telecommunication systems |

Published in Switzerland
Geneva, 2023