INTERNATIONAL STANDARD

ISO/IEC 19794-13

First edition
2018-03

Information technology — Biometric data interchange formats —

Part 13: Voice data

Technologies de l'information — Formats d'échanges de données biométriques —

Partie 13: Données relatives à la voix

Reference number
ISO/IEC 19794-13:2018(E)

COPYRIGHT PROTECTED DOCUMENT

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Fax: +41 22 749 09 47
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see the following URL: www.iso.org/iso/foreword.html.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 37, *Biometrics*.

A list of all the parts in the ISO/IEC 19794 series, can be found on the ISO website.

Introduction

This document assumes that the voice data interchange record is to be attributed to a single individual and recorded in a single session. Voice data is a time record of audible, acoustic vibrations produced by a human in the course of a verbal interaction and will generally contain both speech and non-speech vocal sounds, as well as non-vocal sounds to be considered "noise" in this context. In addition to serving the linguistic function of semantic information transfer, voice data contains both acoustic and semantic information that can be used to recognize speakers. It is the collection, storage and transmission of voice data containing speech for the purpose of recognizing individuals that is the focus of this document.

This format is designed specifically to support a wide variety of automatic speaker recognition applications, including both text-dependent and text-independent Speaker Identification and Verification (SIV) and enrolment, with minimal assumptions made regarding the voice data capture conditions or the collection environment. This document is intended to be sufficiently general that speaker recognition applications beyond traditional SIV could also be supported, such as linking utterances to the same unknown speaker, and determining that a known speaker is not the source of an utterance. The differentiation between speech used to create the reference for future comparisons (which in some applications is called "enrolment"), and that used to create voice representations (VRs) queried against the references, might occur only at the point of application, thus requiring each stored speech record to potentially support either reference or query creation. Further, automated speaker recognition might incorporate related technologies, such as speech and language recognition, not only in current algorithms and applications, but in future ways that cannot be anticipated. Therefore, this document is written from a very broad perspective with the intent of supporting the broadest possible range of speaker recognition applications and technical approaches.

Information technology — Biometric data interchange formats —

Part 13:
Voice data

1 Scope

This document specifies a data interchange format that can be used for storing, recording, and transmitting digitized acoustic human voice data (speech) assumed to be from a single speaker recorded in a single session. This format is designed specifically to support a wide variety of Speaker Identification and Verification (SIV) applications, both text-dependent and text-independent, with minimal assumptions made regarding the voice data capture conditions or the collection environment. Other uses for the data encapsulated in this format, such as automated speech recognition (ASR), may be possible, but are not addressed in this documnet. This document also does not address handling of data that has been processed to the feature or voice model levels. No application-specific requirements, equipment, or features are addressed in this document. This document supports the optional inclusion of non-standardized extended data. This document allows both the original data captured and digitally-processed (enhanced) voice data to be exchanged. A description of any processing of the original source input is intended to be included in the metadata associated with the voice representations (VRs). This document does not address data streaming.

Provisions that stored and transmitted biometric data be time-stamped and that cryptographic techniques be used to protect their authenticity, integrity and confidentiality are out of the scope of this document.

Information formatted in accordance with this document can be recorded on machine-readable media or can be transmitted by data communication between systems.

A general content-oriented subclause describing the voice data interchange format is followed by a subclause addressing an XML schema definition.

This document includes vocabulary in common use by the speech and speaker recognition community, as well as terminology from other ISO standards.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 19794-1, *Information technology — Biometric data interchange formats — Part 1: Framework*

ISO/IEC 19785-1, *Information technology — Common Biometric Exchange Formats Framework — Part 1: Data element specification*

ISO/IEC 2382-37, *Information technology — Vocabulary — Part 37: Biometrics*

3 Terms and definitions

For the purposes of this document, the terms and definitions in ISO/IEC 19794-1 and the following apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— IEC Electropedia: available at http://www.electropedia.org/

— ISO Online browsing platform: available at http://www.iso.org/obp

3.1
analog-to-digital converter (ADC) resolution
exponent of the base 2 representation (the number of bits) of the number of discrete amplitudes that the analog-to-digital converter is capable of producing

Note 1 to entry: Common values for ADC resolution for sound-cards are: 8, 16, 20 and 24.

3.2
audio duration
duration of the complete audio containing all voice representation utterances, e.g. whole call recordings

3.3
audio encoding
encoding used by the data capture subsystem, e.g. cellphone

Note 1 to entry: The voice signal is encoded before being transmitted over a channel. There are many formats in use today and the number is likely to continue to change as telephones and transmission channels evolve. Formats include PCM(ITU-T G.711) and ADPCM(ITU-T G.726) for wave encoding and ACELP(ITU-T G.723.1) and CS-ACELP(ITU-T G.729 Annex A) for AbS encoding. A-law PCM and mu-law PCM are included in ITU-T G.711.

Note 2 to entry: A comprehensive overview list is provided in 7.4.3.2.

3.4
compression
process that reduces the size of a digital file and, accordingly, the data rate required for transmission

Note 1 to entry: Some audio encodings include compression and some do not. Compression is almost always "lossy" and, therefore, has an impact on the speech signal.

3.5
cut-off frequency (lower/upper)
frequency (below/above) which the acoustic energy drops 3dB below the average energy in the pass band

3.6
far-field
region far enough from the source where the angular field distribution is independent of the distance from the source

3.7
interactive voice response
IVR
predicate title for a telephony based computer that is used to control the flow of telephone calls and to provide voice based self-service

Note 1 to entry: Technology that allows a computer to detect voice and keypad inputs.

Note 2 to entry: IVR systems deal with several real-world and constrained-content effects, such as emotional voices, varying environmental noises, recording of free speech, but also hotwords (e.g., yes, no, digits, keywords).

Note 3 to entry: IVRs apply ASR for user navigation, where on secure applications SIV becomes relevant e.g., financial transactions via telefone. IVR systems may combine ASR and SIV to detect audio sample replays and detect user liveness by introducing on-time generated knowledge to the user that should be spoken.

3.8
microphone
data capture subsystem that converts the acoustic pressure wave emanating from the voice into an electrical signal

3.9
mid-field
region between the near-field and the far-field which has a combination of the characteristics found in both the near-field and the far-field

3.10
near-field
region in an enclosure in which the direct energy at the microphone from the primary source is greater than the reflected energy from that source

Note 1 to entry: In a free field, the near-field is the region close enough to the source that the angular energy distribution is dependent upon the distance from the source.

3.11
public switched telephone network
channel based technology used to switch analogue signal, typically telephone calls, through a network from a source such as a telephone to a destination such as another telephone

Note 1 to entry: Knowledge about the channel where a telephone call originates is useful because, historically, noise and other channel characteristics vary from country to country. The advent and growth of VoIP and other digital telephone networks has attenuated the impact of national telecommunications networks becausethey are not constrained by national boundaries. For example, a call originating in the United States might traverse Canada before arriving at its destination, which could be within the United States (also see Voice over IP).

3.12
representation duration
duration of a single voice representation utterance

3.13
sampling rate
number of samples per second (or per other unit) taken from a continuous signal to make a discrete signal

Note 1 to entry: When the rate is per second, the unit is Hertz (Hz).

Note 2 to entry: Equal to the sampling frequency.

Note 3 to entry: The rate of sampling needs to satisfy the Nyquist criterion.

3.14
session
single capture process that takes place over a single, continuous time period

Note 1 to entry: In database collection, two sessions should have at least 3 weeks to 6 weeks in between, such that non-contemporary speech can be captured. However, in biometric systems a session can be interpreted as the time of recording one or more samples without the subject leaving the scene of the biometric capturing device, i.e. passing through a control stage/barrier infers the end of a session, while multiple rejects can occur during one session.

3.15
signal-to-encoding noise ratio
SNR
ratio of the pure signal of interest to the noise component that results from possible electronic noise sources

Note 1 to entry: SNR(dB) = 10 lg (*Ps*/*Pn*), where *Ps* is average signal power and *Pn* is average noise power, expressed as follows for digitized signals,

$$Ps = \frac{1}{N}\sum_{i=1}^{N} s(i)^2 \quad Pn = \frac{1}{N}\sum_{i=1}^{N} n(i)^2$$

Note 2 to entry: where *N* is the total number of digital samples.

Note 3 to entry: Usually measured in decibels (dB).

Note 4 to entry: For example, in PCM, the noise is caused by quantization and roughly calculated in Furui, Digital Speech Processing, Synthesis, and Recognition, (Dekker, 1989) as:

$$\mathrm{SNR(dB)} = 6\,\mathrm{B} - 7{,}2$$

Note 5 to entry: where B is quantization bits.

3.16
speaker identification
form of speaker recognition which compares a voice sample with a set of voice references corresponding to different persons to determine the one who has spoken

3.17
speaker recognition
process of determining whether two speech segments were produced by the vocal mechanism of the same data subject

3.18
speaker verification
speaker authentication
form of speaker recognition for deciding whether a speech sample was spoken by the person whose identity was claimed

Note 1 to entry: Speaker verification is used mainly to restrict access to information, facilities or premises.

3.19
speaker identification and verification
SIV
process of automatically recognizing individuals through voice characteristics

Note 1 to entry: The data format itself does not depend on the application purpose (active/passive SIV).

3.20
voice
speech
sound produced by the vocal apparatus whilst speaking

Note 1 to entry: Normally defined by phoneticians as the sound that emanates from the lips and nostrils, which comprises "voiced" and "unvoiced" sound produced by the vibration of the vocal folds and from constrictions within the vocal track and modified by the time varying acoustic transfer characteristic of the vocal tract.

Note 2 to entry: For the purposes of this document, speech and voice are used interchangeably.

3.21
speech signal bandwidth
range of speech frequencies between the upper and lower cutoff frequencies that are transmitted or recorded by a system

3.22
speech recognition
automatic speech recognition
conversion, by a functional unit, of a speech signal to a representation of the content of the speech

Note 1 to entry: The content to be recognized can be expressed as a proper sequence of words or phonemes.

3.23
streaming data
sequence of digitally encoded coherent signals (packets of data) used to transmit or receive information

3.24
text-independent recognizer
text-independent recognition system
speech recognizer that works reliably whether or not the received speech sample corresponds to a predefined message

3.25
text-dependent recognizer
text-dependent recognition system
speech recognizer that works reliably only when it receives a speech sample corresponding to a predefined message

3.26
text prompted
SIV technology that requires the data subject to repeat a sequence presented by the SIV system or to answer a question

Note 1 to entry: A synonym is “challenge-response”.

Note 2 to entry: “Text prompted” is often seen as a kind of text-independent interaction.

3.27
utterance
sequence of continuous speech units (e.g., phonemes, syllables, words) that is bounded by silence

3.28
voice over IP
digitized streaming speech carried over data channels as Internet Protocol packets

3.29
voice prompt
voice-response prompt
spoken message used to guide the user through a dialog with a voice response system

3.30
voice representation
VR
one or more voice utterances

3.31
volume
calculation of the “loudness” of the input signal (including speech)

Note 1 to entry: When it is known, volume is expressed in terms of the International Telecommunications Union’s P.56 algorithm[2].

Note 2 to entry: Volume level is a factor in the quality of the input utterances.

4 Abbreviated terms

ADC	Analog-to-Digital Converter
ADPCM	Adaptive Differential Pulse Code Modulation
ASR	Automatic Speech Recognition

bps	bits per second
BDIR	biometric data interchange record
CS-ACELP	Conjugate Structure Algebraic Code Excited Linear Prediction
dB	decibels, measured as a ratio between two energy levels (E1 and E2) as 10 lg(E1/E2)
Hz	Hertz (units of cycles per second)
ILBC	Internet Low Bitrate Codec
IP	Internet Protocol
IVR	Interactive Voice Response
PCM	Pulse Code Modulation
PSTN	Public Switched Telephone Network
SIV	Speaker Identification and Verification
SNR	Signal-to-encoding Noise Ratio (units of dB)
TTS	Text-To-Speech
URL	Uniform Resource Locator
VAD, SAD	Voice Activity Detection, Speech Activity Detection
VR	Voice representation
VoIP	Voice over IP
W3C	World Wide Web Consortium
XML	eXtensible Markup Language

5 Conformance

A biometric data record conforms to this document if it satisfies all of the normative requirements related to:

a) its data structure, data values and the relationships between its XML elements, as specified in ISO/IEC 19794-1 and throughout Clause 7 of this document; and

b) the relationship between its data values and the input biometric data from which the biometric data record is generated, as specified throughout Clause 6.

A system that produces biometric data records is conformant to this document if all biometric data records that it outputs conform to this document (as defined above) as claimed in the Implementation Conformance Statement associated with that system. A system does not need to be capable of producing biometric data records that cover all possible aspects of this document, but only those that are claimed to be supported by the system in the Implementation Conformance Statement.

A system that uses biometric data records is conformant to this document if it can read, and use for the purpose intended by that system, all biometric data records that conform to this document (as defined above) as claimed in the Implementation Conformance Statement associated with that system. A system does not need to be capable of using biometric data records that cover all possible aspects

of this document but only those that are claimed to be supported by the system in an Implementation Conformance Statement.

NOTE For details on the conformance testing methodology, see Annex A.

6 Processes and identifiers

6.1 Capture processes and utterances

6.1.1 Introduction

This clause defines the fundamental elements of SIV interactions called “capture process”, as defined in ISO/IEC 2382-37, and the VRs of data subject speech captured during those interactions or “sessions”. During a capture process voice sounds stemming not from the targeted speaker may be unintentionally recorded overlapping or not overlapping targeted speech sequences; this speech should be considered as noise. Compatible capture process structuring and acoustic signal descriptions are required for interoperability between and among SIV engines.

6.1.2 Voice utterance

A voice utterance is assumed to come from a single speaker for the purpose of recognizing individuals, (or to be used to create a reference for future comparisons). In the case that other voices from different individuals are included within the utterance, this information should be considered as noise, which might affect the SIV system. It is not the purpose of this document to specify how voice utterances will be demarcated, but they will generally be separated by: 1) a change in or repeat of a prompt; or 2) a pause of far longer duration than the inter-syllabic rate. There is no minimum or maximum length to a voice utterance.

6.1.3 Structure of a capture process

An SIV capture process is a verbal interaction which may be used for biometric enrolment, verification or identification that is conducted with a data subject by an automated system or another human. In general, a capture process may include background noise possibly from human sources.

SIV interactions as capture processes can be active or passive (user is aware of capture process or not), with or without behavioural adaptation of users (friendly/frequent users intend to adapt for performance purposes), and further with cooperative (friendly) and non-cooperative users.

An SIV capture process is known as a session. Example in Figure 3: the recording sample may cover the whole-call utterance of the enrolment call as well as single prompt utterances. An utterance is a continuous flow of vocalization stemming from one speaker; it may contain inter-syllabic or inter-word silence, and is bounded by pauses. Pauses are suspension of vocalization of perceptible duration, which are longer than inter-syllabic or inter-word silence, i.e. human-perceptible silence.

NOTE 1 Speech and non-speech sounds are uttered by biometric subjects and can be used for SIV purposes. Usually, an utterance is demarcated as an uninterrupted chain of speech, however applications can also intend the use of sub-utterances for VRs.

NOTE 2 Non-speech sounds do not indicate a suspension of vocalization.

NOTE 3 Utterances can cover temporary stops in action of speech, such as temporary interrupts, since the human perception may arguably still be “listening” rather than percepting a suspension of vocalization.

A single capture process generally takes place over a single, continuous time period (or “session”) and contains one or more utterances of voice data, known as voice representations (VR). A VR contains primarily the voice of one speaker and may be initiated by a prompt to the data subject requesting a response. Figure 1 illustrates a simple verification capture process with the voice utterance initiated by a prompt from an interactive voice response (IVR) system.

Prompt from IVR: Welcome to the ABC Bank home-banking security system.
Please say your account number.
Utterance1 of Speaker A: 357128999
Prompt from IVR: Thank you. Please say your passphrase
Utterance2 of Speaker A: lollapalooza
Prompt from IVR: Thank you

Figure 1 — Capture process 1: Basic speaker verification capture process in a text prompt technology

The capture process in Figure 1 represents a single session, which may contain one or two utterances of the speech of speaker A. These possibilities are shown in Figure 2. Figure 2 might show one representation or two representations.

VR1 — Utterance1: 35712899
Utterance2: lollapalooza

a) As one representation

VR1 — Utterance1: 35712899
VR2 — Utterance2: lollapalooza

b) As two representations

Figure 2 — Voice representations from voice utterances of capture process 1

This is an example from an access control application. In this example, the first voice utterance is the claimed reference pointer ("claim of identity") by the data subject "speaker A". A speaker independent automated speech recognition (ASR) system might be used to extract the content from the first utterance to determine the reference pointer. The second utterance is the "text-dependent" passphrase required to verify the claim using the stored voice model of the reference pointer. The capture process in Figure 1 would not need to change for data subjects interacting with humans (e.g., a call centre agent). Variants of capture process 1 include asking or allowing the data subject to input the reference pointer (account number) manually (e.g., using the touchtone keypad of the telephone). Prompts can be presented as audio by playing one or more sound files or by generating a TTS output for an internal string. Prompts may be presented as text displays (e.g. on PDAs, mobile, or smart devices).

From the data subjects' perspective, the simplest active SIV capture process would contain only one utterance. In capture process 1, this can be accomplished in two ways. Some applications use caller ID and/or other methods to implicitly establish the claim of identity. The result is a one-utterance capture process (utterance 2 only). The capture process may also be reduced to a single utterance (utterance 1 only) when ASR is used. In that utterance the IVR asks speaker A to say the account number. ASR decodes the digits and uses them to retrieve the biometric reference. Then it sends the same input to the SIV engine for biometric verification.

NOTE As Figure 3 reveals, the same capture process and utterance structure can also be used for enrolment.

Prompt from IVR:	Welcome to the ABC Bank voice enrollment system. Please say your account number.
Utterance1 of Speaker A:	357128999
Prompt from IVR:	Thank you. You will now be asked to repeat your passphrase four times. After the tone, please say your passphrase. [tone]
Utterance2 of Speaker A:	lollapalooza
Prompt from IVR:	After the tone, please say your passphrase. [tone]
Utterance3 of Speaker A:	lollapalooza
Prompt from IVR:	After the tone, please say your passphrase. [tone]
Utterance4 of Speaker A:	lollapalooza
Prompt from IVR:	After the tone, please say your passphrase. [tone]
Utterance5 of Speaker A:	lollapalooza
Prompt from IVR:	Thank you. You are now enrolled in the ABC Bank voice security system.

Figure 3 — Capture process 2: Enrolment

This capture process contains five utterances of speaker A. It first establishes the pointer to the claimed reference, which is followed by four repetitions of the passphrase prompted by a tone. The voice data acquired in these utterances compose the VRs, which are primary XML elements in the voice data BDB.

6.2 Registered format type identifiers

The registration listed in Table 1 has been made with the CBEFF registration process to identify the voice data record format. The CBEFF definition shall be in accordance with ISO/IEC 19785-1. The format owner is ISO/IEC JTC 1/SC 37 with the registered format owner identifier 31 ($001F_{Hex}$).

Table 1 — Format Type Identifiers

CBEFF BDB format Type identifier	Short name	Full object identifier
257 (0101_{Hex})	voice-data	{iso(1) registration-authority(1) cbeff(19785) biometricorganization(0) jtc1-sc37(257) bdbs(0) voice-data(31)}

7 General voice data interchange format (BDB)

7.1 Overview

This document will be implemented only in XML. In this clause, the voice-data specific header in the biometric data interchange record (BDIR), containing information about the VR collection conditions and any post-collection processing are discussed. It is not the purpose of this document to specify which of the data capture environments, the methods of data capture or any pre-processing (e.g. detection/segmentation, pre-emphasis filtering) are done on the utterances of voice data comprising the capture process.

The structure of XML elements is depicted in Figure 4. The record format is as follows:

— a Voice Record General Header containing information about the overall record (7.3),

— a representation element for each VR (7.4).

Each VR shall consist of:

— a VR header containing information about the data for a single representation,

— a VR data field,

where each header contains an element for extended vendor data (see Tables 2 and 6).

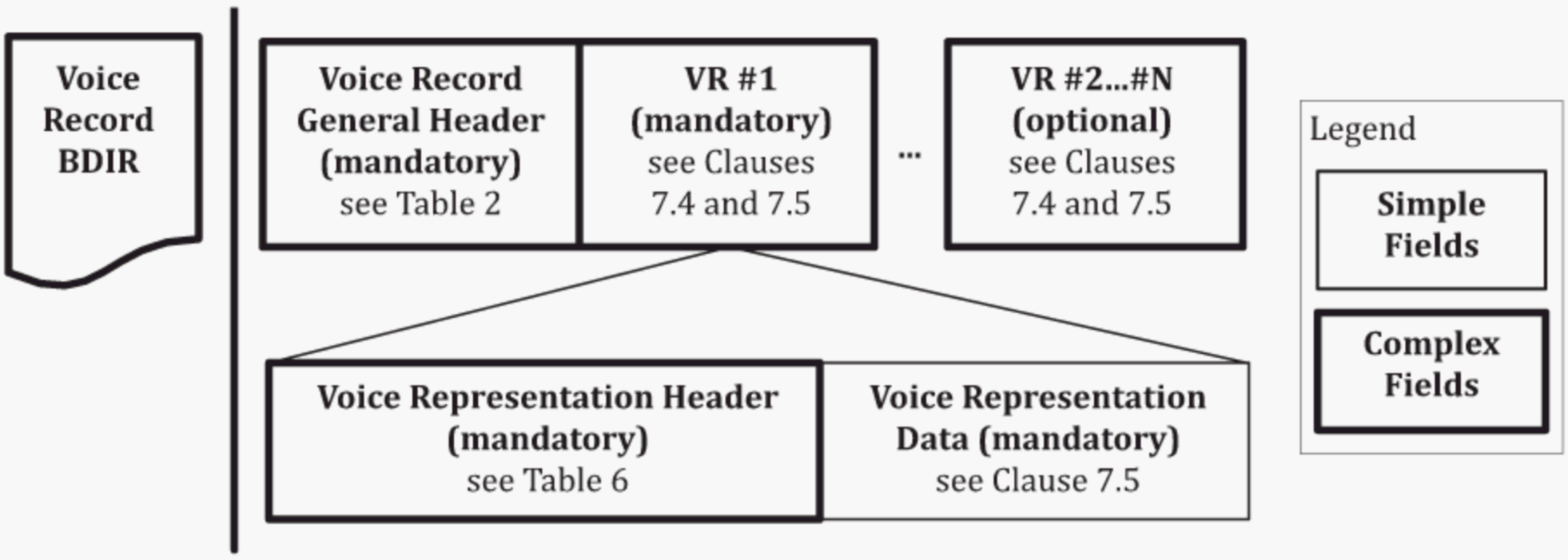

Figure 4 — Structure of XML elements

7.2 Conventions

Elements may be simple or complex. Complex elements contain other elements.

Elements may be mandatory or optional. Optional complex elements may contain both optional and mandatory elements and characteristics.

The naming convention for XML elements and characteristics used in this format shall consist of capital and small letters, such as NumberofVRs, with no hyphens or spaces. The printing convention for valid string values is to enclose each valid value in quotes.

7.3 Voice record general header

7.3.1 Overview

The header for the voice record elements is given below as Table 2. The first three fields of the voice records element schema are taken directly from ISO/IEC 19794-1. The remaining six fields are each complex elements, serving as the default for the capture process. Details within each representation, however, may vary. Therefore, the field values in the headers of the various VRs may be different from those in the general header.

Table 2 — Voice record general header

Field	Clause	Item type	Valid values	Optional/ Mandatory
Version	7.3.2	VersionType	see ISO/IEC 19794-1/Amd 2	M
Session ID	7.3.3	string	no limit string	O
Channel	7.3.4	ChannelType	see Table 3	M
Capture device	7.3.5	CaptureDeviceModelID	see ISO/IEC 19794-1/Amd 2	O
Transducer	7.3.6	TransducerType	see Table 4	O

Table 2 *(continued)*

Field	Clause	Item type	Valid values	Optional/ Mandatory
Audio meta information	7.3.7	AudioMetaInformationType	see Table 5	M
Capture process protocol	7.3.8	CaptureProcessProtocolType	no limit string	O
Extended vendor data	7.3.9	VendorSpecificDataType	see ISO/IEC 19794-1/Amd 2, max. 256	O

7.3.2 Version

Version of the associated entity (e.g., CBEFF version, patron/ data format specification).

[SOURCE: ISO/IEC 19794-1]

7.3.3 Session ID

Application-specific session identifier.

7.3.4 Channel

7.3.4.1 Overview

The Channel element shall describe the fields of the default channel from which the data were captured.

Table 3 — Description of the functionality of "ChannelType"

Field	Clause	Item type	Valid values	Optional/ Mandatory
Type	7.3.4.2	string	"Unknown" "Analog" "Digital" "NonVoIP" "DigitalVoIP" "Mixed"	M
Cutoff upper frequency	7.3.4.3	numeric	0 – 65535	O
Cutoff lower frequency		numeric	0 – 65535	
Country of origin	7.3.4.4	string	3 character string	O

7.3.4.2 Type

Type shall specify the kind of channel over which the data were captured. Types are Analog, Digital Non-VoIP, Digital VoIP, Mixed and Unknown. The default value is "Unknown".

7.3.4.3 Cutoff upper frequency and cutoff lower frequency

The voice elements record schema shall have an indicator of the upper and lower cutoff frequencies of the audio data. Both upper and lower cutoff frequencies shall be the integer that best represents the frequencies on the upper and lower ends of the audio band at which energy has fallen 3 dB below the average band energy. There is no default value. The value shall be 0 if unknown.

7.3.4.4 Country of origin

The Country element shall identify the country of origin of the channel, if known.

Country code of origin should be represented by an alpha code that complies with the two-letter country code of ISO 3166-1, which supports three kinds of country codes: two-letter, three-letter, and numeric.

7.3.5 Capture device

Registered identifier of the type of device used to capture the biometric data (BDIR):

— <Organization> identifies the device vendor;

— <Identifier> identifies the specific device type (e.g. maps to model).

[SOURCE: ISO/IEC 19794-1]

7.3.6 Transducer

The Capture Technology ID is a simple type for voice data that defines the characteristics of the signal collection transducer.

7.3.6.1 Overview

The transducer field shall specify the input device employed by the data subject. It is recognized that complex collection systems may consist of multiple transducers, to which the elements of this clause may not apply. In such cases, “unknown” is the default value.

NOTE This element is intended primarily to support R&D and engines that require device registration.

Table 4 — Description of the functionality of “TransducerType”

Field	Clause	Item type	Valid values	Optional/ Mandatory
Capture technology ID	7.3.6.2	string	“Telephone” “Microphone” “Handheld” “Mobile phone” “Stethoscope” “Other” “Unknown”	O
Microphone type	7.3.6.3	string	“Carbon” “Electret” “Other” “Unknown”	O
Manufacturer	7.3.6.4	string	no limit string	O
Model	7.3.6.5	string	no limit string	O
Mic cutoff upper	7.3.6.6	numeric	0 – 65535	O
Mic cutoff lower		numeric	0 – 65535	
Device info	7.3.6.7	string	no limit string	O

7.3.6.2 Capture technology ID

The voice record elements schema shall have a Capture Technology ID to specify the kind of input device used, if known. The default value is “telephone”.

7.3.6.3 Microphone type

The voice record elements schema shall indicate the type of microphone used in the input device, if known. Permitted values are carbon, electret, other, and unknown.

7.3.6.4 Manufacturer

The manufacturer field shall be a string identifying the manufacturer of the data subject’s input device.

7.3.6.5 Model

The model field shall be a string identifying the manufacturer of the data subject’s input device.

7.3.6.6 Mic cutoff upper and mic cutoff lower

The optional upper and lower cutoff frequencies shall both be an integer that best represents the frequencies on the upper and lower ends at which the capacity for energy conversion of the microphone has fallen 3 dB below the average band energy. There is no default value, but 0 shall indicate that the information is unknown.

7.3.6.7 Device info

Device info shall be reserved for additional information about the device, but not about the capture process or the data subject. It shall be limited to data that a recipient SIV engine or application is able to discern and use.

7.3.7 Audio meta information

7.3.7.1 Overview

This clause gives the technical specifications of the signal process used to capture all VRs in the record.

Table 5 — Description of the functionality of “AudioMetaInformationType”

Field	Clause	Item type	Valid values	Optional/ Mandatory
Channel count	7.3.7.2	numeric	1 – 15	M
Sampling rate	7.3.7.3	numeric	0 – 128000	M
Bits per sample	7.3.7.4	numeric	0 – 255	M
Audio duration	7.3.7.5	numeric	built-in type	M

7.3.7.2 Channel count

The voice elements record schema shall have a Channel Count field. This integer element gives the number of channels in the input stream. The default value shall be 1.

7.3.7.3 Sampling rate

The voice elements record schema shall have an integer characteristic giving the number of samples per second at which the original audio input stream was sampled.

7.3.7.4 Bits per sample

The voice elements record schema shall have an integer Bits per Sample characteristic. This integer gives the bit depth of a single sample of the audio signal. For formats that use variable bit depth, like Ogg Vorbis, this element is set to 0.

7.3.7.5 Audio duration

Audio Duration is an integer value that indicates duration of the utterance in milliseconds.

7.3.8 Capture process protocol

Capture Process Protocol shall be reserved for additional information about the capture process, but not about the data subject or data capture device. It shall be limited to data that a recipient SIV engine or application is able to discern and use.

7.3.9 Extended vendor data

This is used when non-standardized data, proprietary to a vendor/product, needs to be included.

[SOURCE: ISO/IEC 19794-1]

7.4 Voice representation header

7.4.1 Overview

The VR shall be the child of the capture process element that contains the elements and fields that may change in the course of a capture process. There shall be a minimum of one representation for each capture process. The VR elements are shown in Table 6.

NOTE Information regarding the spoken text, language, dialects, or a subject's gender are not considered for VR elements at all. If these or other information can aid the recognition process, analysts can use ASR, Automatic Language Recognition (ALR), or Automatig Gender Detection (AGD) software.

Table 6 — Voice representation header

Field	Clause	Item type	Valid values	Optional/ Mandatory
Date and time	7.4.2	DateAnd-TimeType	see Table 7	O
Audio content	7.4.3	AudioCon-tentType	see Table 8	M
Quality	7.4.4	VRQuality-Type	see Table 10	O
Signal enhancement	7.4.5	string	no limit string	O
Extended vendor data	7.4.6	VendorSpeci-ficDataType	see ISO/IEC 19794-1/ Amd 2, max. 256	O

7.4.2 Date and time

7.4.2.1 Overview

The date and time element shall indicate when the VR started and ended. The start time of the VR shall be considered its "capture time". The time specification should be in compliance with WC3 – XML 1.0.

Table 7 — Description of the functionality of "DateandTimeType"

Field	Clause	Item type	Valid values	Optional/ Mandatory
Start	7.4.2.2	dateTime	see ISO/IEC 19794-1/ Amd 2	O
End	7.4.2.3	dateTime	see ISO/IEC 19794-1/ Amd 2	O
Voice start time	7.4.2.4	dateTime	built-in type	O
Voice end time	7.4.2.5	dateTime	built-in type	O
Voice elapsed time	7.4.2.6	time	built-in type	O

7.4.2.2 Start

Each capture process record shall have a Start field in accordance with ISO 8601 that specifies the date and time that the representation began. The start time shall be considered as the "capture time" of the VR in 7.5.

7.4.2.3 End

This mandatory element shall specify the date and time that the representation ended. Because of the possible use of activity detection software, the length of the audio data in the representation may be shorter than the difference between start and end times.

7.4.2.4 Voice start time

This is the Start Time of the representation within the voice data of the capture process.

7.4.2.5 Voice end time

This is the End Time of the representation within the voice data of the capture process.

7.4.2.6 Voice elapsed time

In the case of spontaneous/free or conversational voice, characteristics are the start and end time of conversation.

7.4.3 Audio content

7.4.3.1 Overview

This clause gives the details of the audio content of the VR, including the mandatory information on audio encoding, the length in seconds, information on the speech content (if known), volume and SNR estimate.

NOTE A subject's verbalised text is not considered to be part of this data format, since it may reveal data about the subject possibly containing sensitive (privacy) data. Thus, neither ASR-based analyses nor the actual verbalised text is included. However, ASRs can aid SIVs as subsystems.

Table 8 — Description of the functionality of "AudioContentType"

Field	Clause	Item type	Valid values	Optional/ Mandatory
Audio encoding	7.4.3.2	string	"Linear PCM" "Mu-Law" "A-Law" "Non-streaming OGG Vorbis" "Speex" "ADPCM" "CS-ACELP" "PCM" "AMR" "ILBC" "MPEG" "AC3" "AAC" "AMR" "APE" "FLAC" "MMF" "M4A" "MP2" "MP3" "MP4" "RA" "Full-HD Voice" "other" "unknown"	M
Duration	7.4.3.3	numeric	built-in type	M
Conversation	7.4.3.4	complex	"Unknown" "Spontaneous/Free" "Reading" "Prompt" "Conversational" "Other" if "Prompt", see Table 9	O
Dominant language	7.4.3.5	string	3 character string	O

7.4.3.2 Audio encoding

Each VR shall name the audio encoding of the stored data from an enumerated list, limited to Linear PCM, Mu-Law, A-Law, non-streaming OGG Vorbis, Speex, ADPCM, CS-ACELP, PCM, AMR, ILBC, MPEG, AC3, AAC, AMR, APE, FLAC, MMF, M4A, MP2, MP3, MP4, RA, Full-HD Voice, "other" and "unknown".

7.4.3.3 Duration

Duration is an integer value that indicates the total time taken for the representation in milliseconds. The net result should allow back calculation of the sample rate.

7.4.3.4 Conversation

This field shall contain the type of speech recorded. Valid values are Unknown, Spontaneous/Free, Reading, Prompt, Conversational, and Other.

If the value is Prompt, it becomes a complex structure. The complex structure is a valid value of String Prompt Content and/or Audio Prompt Content. String Prompt Content shall contain a text of the prompt, if known. If an audio prompt was used and a URL containing that audio prompt is available, then the Audio Prompt Content element shall give the URL of the audio prompt or a depending identifier, if known. It is possible for both an audio prompt file and a transcription of the audio prompt to be available. In these cases, both of these fields may have content.

Table 9 — Description of the functionality of "ConversationRoot" if Conversation == "Prompt"

Field	Clause	Item type	Valid values	Optional/ Mandatory
String Prompt Content	7.4.3.4	string	no limit string	O
Audio Prompt Content		URL	no limit URL	O

7.4.3.5 Dominant language

This field shall contain the most dominant language in VR. The dominant language identifier should be a string in accordance with ISO 639-3 and IETF RFC 5646 codes.

NOTE Additional guidance for language codes can be found in ISO 639-3 and IETF RFC 5646[3].

7.4.4 Quality information

7.4.4.1 Overview

This element shall give the details of the audio quality of the VR, including the mandatory information on audio recording conditions and signal enhancements.

Table 10 — Description of the functionality of "QualityInformationType"

Field	Clause	Item type	Valid values	Optional/ Mandatory
Quality	7.4.4.2	QualityType	see ISO/IEC 19794-1/ Amd 2	M
Field	7.4.4.3	string	"Near-field" "Mid-field" "Far-field" "Other" "Unknown"	O
Microphone distance	7.4.4.4	string	"Close" "Mid-range" "Far"	O
Volume	7.4.4.5	float	built-in type	O
SNR	7.4.4.6	float	built-in type	O

7.4.4.2 Quality

This field is the quality of the biometric data and identification of the algorithm used to compute it.

If no quality calculation was attempted (and no value is available), then the quality element should not be present. When projects add this element, they should therefore set minOccurs=0.

[SOURCE: ISO/IEC 19794-1]

7.4.4.3 Field

Field shall refer the conditions of the direct and reverberant sound field to the input microphone. It shall be a string value. Allowable values shall be:

— Near-field;

— Mid-field;

— Far-field;

— Other;

— Unknown.

The default value shall be “Near-field”.

NOTE Near-field and the other terms are general specifications of sound field (see definitions). “Mid-field” is included to cover earpiece and other microphones used with wireless telephones. This approach is preferred to using a numeric specification of actual distance because it is generally not possible to access that level of detail.

7.4.4.4 Microphone distance

“Close”: the case of using a hand-set or headset. This is the case of using equipment with an integrated speaker/microphone in which the voice is acquired while the headset is contacting the ear (about 0,05 m to 0,15 m).

“Mid-range”: the case of using a boundary microphone, hands-free phone (talk with looking at the display), or tablet. This is a case of inputting from the mike within the range where not "Near-field" but the speaker's hand reaches (about 0,1 m to 0,5 m).

“Far”: the case of inputting from the mike within the range where the speaker's hand doesn't reach (over 0,5 m).

7.4.4.5 Volume

When it is known, volume shall be expressed in terms of the International Telecommunications Union’s P.56 algorithm[3]. Otherwise, the value is set to Unknown.

7.4.4.6 SNR

In this optional field, only the coding noise is taken into account to SNR calculation for each VR.

7.4.5 Signal enhancement

Signal enhancement is pre-processing applied to the speech signal. This element is to inform post processing operations of the signal conditioning already applied. Signal enhancement may make the recognition rate better or worse. It should be made clear what type of enhancement, if any, has been applied to the voice signal.

Typical speech enhancement processing includes:

1) Noise suppression;
2) Echo suppression;
3) Echo cancellation;
4) Active Noise Control;
5) Speech emphasis (including the result of microphone arrays processing);
6) Automatic gain control (AGC);
7) Equalization, filter, pre-emphasis;
8) De-reverberation (remove reverberation);
9) End pointing;
10 Silence removal;
11) Other effects.

7.4.6 Extended vendor data

This is used when non-standardized data, proprietary to a vendor/product, needs to be included.

[SOURCE: ISO/IEC 19794-1]

7.5 Voice representation data

Provide either a URL pointing to the location of the VR data field or a quoted BLOB of the voice utterance in Base64 encoded format.

7.6 Schema

This schema shall be used to validate xml voice records encoded in an XML format. It is 19794-13_ed1.xsd.

An electronic version of the schema presented here is available at:

http://standards.iso.org/iso-iec/19794/-13/ed-1/en/

The user is permitted to use the schema in its original format without any modifications for the purposes specified in this document.

```
<?xml version="1.0" encoding="utf-8"?>
<!--
-->
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" xmlns="http://standards.iso.
org/iso-iec/19794/-13/ed-1" xmlns:vdi="http://standards.iso.org/iso-iec/19794/-13/ed-1"
xmlns:cmn="http://standards.iso.org/iso-iec/19794/-1/ed-2/amd/2" targetNamespace="http://
standards.iso.org/iso-iec/19794/-13/ed-1" elementFormDefault="qualified"
attributeFormDefault="unqualified">
    <xs:import namespace="http://standards.iso.org/iso-iec/19794/-1/ed-2/amd/2"
schemaLocation="19794-1_ed2_amd2.xsd"/>
    <xs:complexType name="VoiceRecordGeneralHeaderType">
        <xs:sequence>
            <xs:element name="Version" type="cmn:VersionType"/>
            <xs:element name="SessionId" type="xs:string" minOccurs="0"/>
            <xs:element name="Channel" type="ChannelType"/>
            <xs:element name="CaptureDevice" type="cmn:RegistryIDType" minOccurs="0"/>
            <xs:element name="Transducer" type="TransducerType" minOccurs="0"/>
            <xs:element name="AudioMetaInformation" type="AudioMetaInformationType"/>
```

```
            <xs:element name="CaptureProcessProtocol" type="xs:string" minOccurs="0"/>
            <xs:element name="ExtendedVendorData" type="cmn:VendorSpecificDataType"
minOccurs="0" maxOccurs="256"/>
        </xs:sequence>
    </xs:complexType>
    <xs:complexType name="ChannelType">
        <xs:sequence>
            <xs:element name="Type">
                <xs:simpleType>
                    <xs:restriction base="xs:string">
                        <xs:whiteSpace value="collapse"/>
                        <xs:enumeration value="Unknown"/>
                        <xs:enumeration value="Analog"/>
                        <xs:enumeration value="Digital"/>
                        <xs:enumeration value="NonVoIP"/>
                        <xs:enumeration value="DigitalVoIP"/>
                        <xs:enumeration value="Mixed"/>
                    </xs:restriction>
                </xs:simpleType>
            </xs:element>
            <xs:element name="CutoffUpperFrequency" type="CutOffBoundType" minOccurs="0"/>
            <xs:element name="CutoffLowerFrequency" type="CutOffBoundType" minOccurs="0"/>
            <xs:element name="CountryOfOrigin" type="CountryOfOriginType" minOccurs="0"/>
        </xs:sequence>
    </xs:complexType>
    <xs:simpleType name="CutOffBoundType">
        <xs:restriction base="xs:unsignedInt">
            <xs:maxInclusive value="65535"/>
        </xs:restriction>
    </xs:simpleType>
    <xs:simpleType name="CountryOfOriginType">
        <xs:restriction base="xs:string">
            <xs:maxLength value="3"/>
        </xs:restriction>
    </xs:simpleType>
    <xs:complexType name="TransducerType">
        <xs:sequence>
            <xs:element name="CaptureTechnologyID" type="CaptureTechnologyIdType"
minOccurs="0"/>
            <xs:element name="Microphone" type="MicrophoneType" minOccurs="0"/>
            <xs:element name="Manufacturer" type="xs:string" minOccurs="0"/>
            <xs:element name="Model" type="xs:string" minOccurs="0"/>
            <xs:element name="MicCutoffUpper" type="CutOffBoundType" minOccurs="0"/>
            <xs:element name="MicCutoffLower" type="CutOffBoundType" minOccurs="0"/>
            <xs:element name="DeviceInfo" type="xs:string" minOccurs="0"/>
        </xs:sequence>
    </xs:complexType>
    <xs:simpleType name="CaptureTechnologyIdType">
        <xs:restriction base="xs:string">
            <xs:whiteSpace value="collapse"/>
            <xs:enumeration value="Telephone"/>
            <xs:enumeration value="Microphone"/>
            <xs:enumeration value="Handheld"/>
            <xs:enumeration value="Mobile Phone"/>
            <xs:enumeration value="Stethoscope"/>
            <xs:enumeration value="Other"/>
            <xs:enumeration value="Unknown"/>
        </xs:restriction>
    </xs:simpleType>
    <xs:simpleType name="MicrophoneType">
        <xs:restriction base="xs:string">
            <xs:whiteSpace value="collapse"/>
            <xs:enumeration value="Carbon"/>
            <xs:enumeration value="Electret"/>
            <xs:enumeration value="Other"/>
            <xs:enumeration value="Unknown"/>
        </xs:restriction>
    </xs:simpleType>
    <xs:complexType name="AudioMetaInformationType">
        <xs:sequence>
            <xs:element name="ChannelCount" type="ChannelCountType"/>
```

```
            <xs:element name="SamplingRate" type="SamplingRateType"/>
            <xs:element name="BitsPerSample" type="BitsPerSampleType"/>
            <xs:element name="AudioDuration" type="xs:unsignedInt"/>
        </xs:sequence>
    </xs:complexType>
    <xs:simpleType name="ChannelCountType">
        <xs:restriction base="xs:unsignedInt">
            <xs:minInclusive value="1"/>
            <xs:maxInclusive value="15"/>
        </xs:restriction>
    </xs:simpleType>
    <xs:simpleType name="SamplingRateType">
        <xs:restriction base="xs:unsignedInt">
            <xs:maxInclusive value="128000"/>
        </xs:restriction>
    </xs:simpleType>
    <xs:simpleType name="BitsPerSampleType">
        <xs:restriction base="xs:unsignedByte">
            <xs:maxInclusive value="255"/>
        </xs:restriction>
    </xs:simpleType>
    <xs:complexType name="VoiceRepresentationHeaderType">
        <xs:sequence>
            <xs:element name="DateAndTime" type="DateAndTimeType" minOccurs="0"/>
            <xs:element name="AudioContent" type="AudioContentType"/>
            <xs:element name="Quality" type="QualityInformationType" minOccurs="0"/>
            <xs:element name="SignalEnhancement" type="xs:string" minOccurs="0"/>
            <xs:element name="ExtendedData" type="cmn:VendorSpecificDataType" minOccurs="0"
maxOccurs="256"/>
        </xs:sequence>
    </xs:complexType>
    <xs:complexType name="DateAndTimeType">
        <xs:sequence>
            <xs:element name="StartTime" type="xs:dateTime" minOccurs="0"/>
            <xs:element name="EndTime" type="xs:dateTime" minOccurs="0"/>
            <xs:element name="VoiceStartTime" type="xs:dateTime" minOccurs="0"/>
            <xs:element name="VoiceEndTime" type="xs:dateTime" minOccurs="0"/>
            <xs:element name="VoiceElapsedTime" type="xs:dateTime" minOccurs="0"/>
        </xs:sequence>
    </xs:complexType>
    <xs:complexType name="AudioContentType">
        <xs:sequence>
            <xs:element name="AudioEncoding" type="AudioEncodingType"/>
            <xs:element name="Duration" type="xs:unsignedInt"/>
            <xs:element name="Conversation" type="ConversationRoot" minOccurs="0"/>
            <xs:element name="DominantLanguage" type="DominantLanguageType"
minOccurs="0"/>
        </xs:sequence>
    </xs:complexType>
    <xs:simpleType name="AudioEncodingType">
        <xs:restriction base="xs:string">
            <xs:whiteSpace value="collapse"/>
            <xs:enumeration value="Linear PCM"/>
            <xs:enumeration value="Mu-Law"/>
            <xs:enumeration value="A-Law"/>
            <xs:enumeration value="Non-streaming OGG Vorbis"/>
            <xs:enumeration value="Speex"/>
            <xs:enumeration value="ADPCM"/>
            <xs:enumeration value="CS-ACELP"/>
            <xs:enumeration value="PCM"/>
            <xs:enumeration value="AMR"/>
            <xs:enumeration value="ILBC"/>
            <xs:enumeration value="MPEG"/>
            <xs:enumeration value="AC3"/>
            <xs:enumeration value="AAC"/>
            <xs:enumeration value="AMR"/>
            <xs:enumeration value="APE"/>
            <xs:enumeration value="FLAC"/>
            <xs:enumeration value="MMF"/>
            <xs:enumeration value="M4A"/>
            <xs:enumeration value="MP2"/>
```

```
            <xs:enumeration value="MP3"/>
            <xs:enumeration value="MP4"/>
            <xs:enumeration value="RA"/>
            <xs:enumeration value="Full-HD Voice"/>
            <xs:enumeration value="other"/>
            <xs:enumeration value="unknown"/>
        </xs:restriction>
    </xs:simpleType>
    <xs:complexType name="ConversationRoot">
        <xs:choice>
            <xs:group ref="case1"/>
            <xs:group ref="case2"/>
        </xs:choice>
    </xs:complexType>
    <xs:group name="case1">
        <xs:sequence>
            <xs:element name="SimpleCases" type="SimpleCasesType"/>
        </xs:sequence>
    </xs:group>
    <xs:group name="case2">
        <xs:sequence>
            <xs:element name="PromptCase" type="PromptCaseType"/>
        </xs:sequence>
    </xs:group>
    <xs:simpleType name="SimpleCasesType">
        <xs:restriction base="xs:string">
            <xs:whiteSpace value="collapse"/>
            <xs:enumeration value="Unknown"/>
            <xs:enumeration value="Spontaneous"/>
            <xs:enumeration value="Reading"/>
            <xs:enumeration value="Conversational"/>
            <xs:enumeration value="Other"/>
        </xs:restriction>
    </xs:simpleType>
    <xs:complexType name="PromptCaseType">
        <xs:sequence>
            <xs:element name="StringPromptContent" type="StringPromptContentType"
minOccurs="0"/>
            <xs:element name="AudioPromptContent" type="AudioPromptContentType"
minOccurs="0"/>
        </xs:sequence>
    </xs:complexType>
    <xs:simpleType name="StringPromptContentType">
        <xs:restriction base="xs:string"/>
    </xs:simpleType>
    <xs:simpleType name="AudioPromptContentType">
        <xs:restriction base="xs:string"/>
    </xs:simpleType>
    <xs:simpleType name="DominantLanguageType">
        <xs:restriction base="xs:string">
            <xs:maxLength value="3"/>
        </xs:restriction>
    </xs:simpleType>
    <xs:complexType name="QualityInformationType">
        <xs:sequence>
            <xs:element name="Quality" type="cmn:QualityType"/>
            <xs:element name="Field" type="FieldType" minOccurs="0"/>
            <xs:element name="MicrophoneDistance" type="MicrophoneDistanceType"
minOccurs="0"/>
            <xs:element name="Volume" type="xs:float" minOccurs="0"/>
            <xs:element name="SNR" type="xs:float" minOccurs="0"/>
        </xs:sequence>
    </xs:complexType>
    <xs:simpleType name="FieldType">
        <xs:restriction base="xs:string">
            <xs:whiteSpace value="collapse"/>
            <xs:enumeration value="Near-field"/>
            <xs:enumeration value="Mid-field"/>
            <xs:enumeration value="Far-field"/>
            <xs:enumeration value="Other"/>
            <xs:enumeration value="Unknown"/>
```

```
    </xs:restriction>
</xs:simpleType>
<xs:simpleType name="MicrophoneDistanceType">
    <xs:restriction base="xs:string">
        <xs:whiteSpace value="collapse"/>
        <xs:enumeration value="Close"/>
        <xs:enumeration value="Mid-range"/>
        <xs:enumeration value="Far"/>
    </xs:restriction>
</xs:simpleType>
<xs:complexType name="VoiceRepresentationType">
    <xs:sequence>
        <xs:element name="VoiceRepresentationHeader"
type="VoiceRepresentationHeaderType"/>
            <xs:element name="VoiceRepresentationData"
type="VoiceRepresentationDataTypeRoot"/>
        </xs:sequence>
    </xs:complexType>
    <xs:complexType name="VoiceRepresentationDataTypeRoot">
        <xs:choice>
            <xs:group ref="caseVRDataURL"/>
            <xs:group ref="caseVRDataBLOB"/>
        </xs:choice>
    </xs:complexType>
    <xs:group name="caseVRDataURL">
        <xs:sequence>
            <xs:element name="URL" type="xs:string"/>
        </xs:sequence>
    </xs:group>
    <xs:group name="caseVRDataBLOB">
        <xs:sequence>
            <xs:element name="BLOB" type="xs:base64Binary"/>
        </xs:sequence>
    </xs:group>
    <xs:element name="VoiceRecord">
        <xs:complexType>
            <xs:sequence>
                <xs:element name="VoiceRecordGeneralHeader"
type="VoiceRecordGeneralHeaderType"/>
                <xs:element name="VoiceRepresentation" type="VoiceRepresentationType"
maxOccurs="unbounded"/>
            </xs:sequence>
      <xs:attribute ref="cmn:SchemaVersion" use="required" />
    </xs:complexType>
    </xs:element>
</xs:schema>
```

7.7 Example

```
<?xml version="1.0" encoding="UTF-8"?>
<vdi:VoiceRecord xmlns:cmn="http://standards.iso.org/iso-iec/19794/-1/ed-2/amd/2"
xmlns:vdi="http://standards.iso.org/iso-iec/19794/-13/ed-1" xmlns:xsi="http://
www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://standards.
iso.org/iso-iec/19794/-13/ed-1 19794-13_ed1.xsd http://standards.iso.
org/iso-iec/19794/-1/ed-2/amd/2 19794-1_ed2_amd2.xsd " cmn:SchemaVersion="1.0">
<vdi:VoiceRecordGeneralHeader>
  <vdi:Version>
    <cmn:Major>1</cmn:Major>
    <cmn:Minor>0</cmn:Minor>
  </vdi:Version>
  <vdi:Channel>
    <vdi:Type>Unknown</vdi:Type>
  </vdi:Channel>
  <vdi:AudioMetaInformation>
    <vdi:ChannelCount>1</vdi:ChannelCount>
    <vdi:SamplingRate>8000</vdi:SamplingRate>
    <vdi:BitsPerSample>16</vdi:BitsPerSample>
    <vdi:AudioDuration>0</vdi:AudioDuration>
  </vdi:AudioMetaInformation>
</vdi:VoiceRecordGeneralHeader>
<vdi:VoiceRepresentation>
```

```
    <vdi:VoiceRepresentationHeader>
      <vdi:AudioContent>
        <vdi:AudioEncoding>Linear PCM</vdi:AudioEncoding>
        <vdi:Duration>0</vdi:Duration>
      </vdi:AudioContent>
    </vdi:VoiceRepresentationHeader>
    <vdi:VoiceRepresentationData>
<vdi:BLOB>UklGRiQAAABXQVZFZm10IBAAAAABAAEAQB8AAIA+AAACABAAZGF0YQAAAAA=</vdi:BLOB></
vdi:VoiceRepresentationData>
  </vdi:VoiceRepresentation>
</vdi:VoiceRecord>
```

Annex A
(normative)

Conformance testing methodology

A.1 Overview

This document specifies a biometric data interchange format for storing, recording, and transmitting one or more VR representations. Each representation is accompanied by modality-specific metadata contained in a header record. This annex establishes tests for checking the correctness of the record.

The objective of this document cannot be completely achieved until biometric products can be tested to determine whether they conform to those specifications. Conforming implementations are a necessary prerequisite for achieving interoperability among implementations; therefore there is a need for a standardised conformance testing methodology, test assertions, and test procedures as applicable to specific modalities addressed by each part of ISO/IEC 19794. The test assertions will cover as much as practical of the ISO/IEC 19794 requirements (covering the most critical features), so that the conformity results produced by the test suites will reflect the real degree of conformity of the implementations to ISO/IEC 19794 data interchange format records. This is the motivation for the development of this conformance testing methodology.

This normative annex is intended to specify elements of conformance testing methodology, test assertions, and test procedures as applicable to this document. For this edition of this document, the content of this Annex is not yet available, but when it is, it will be available as a separate document (Amendment) to supplement this document.

A.2 Conformance testing

Each part of ISO/IEC 19794 defining an XML schema shall require conformance testing in terms of strict validation of the XML schema definition. Additionally, it shall contain a table specifying test assertions for further requirements that are not explicitly covered by the schema validation process. Each part may contain a normative annex that defines an Extensible Stylesheet Language Transformations (XSLT) stylesheet for use in testing the level-2 conformance of valid XML documents claimed to conform to that XML schema.

Bibliography

[1] ISO/IEC SD 2:2007, *Text of Standing Document 2 (SD 2) Version 8, Harmonized Biometric Vocabulary. Geneva: International Standards Organization*

[2] ITU-T P.56:1993, *Objective measurement of active speech level. Geneva: International Telecommunication Union — Telecommunication Standardization Sector*

[3] IETF RFC 5646, *Tags for the Identification of Languages.* Edited by Phillips, A. 2009. Available at: http://www.ietf.org/rfc/rfc5646.txt

[4] WEB CONSORTIUM.EXTENSIBLE MARKUP LANGUAGE (XML). 1.0 (Fifth Edition) Available at: https://www.w3c.org/TR/2008/REC-xml-20081126/

[5] WEB CONSORTIUM.VOICE EXTENSIBLE MARKUP LANGUAGE (VOICEXML). Version 2.1, 2007. Available at: https://www.w3c.org/TR/2007/REC-voicexml21-20070619/

[6] WEB CONSORTIUM. *XML Key Management Specification (XKMS 2.0)* Edited by Hirsch, Frederick & Just, Mike. 2005. Available at :http://www.w3.org/TR/xkms2-req

[7] Directive 95/46/EC of the European Parliament and of the Council, 24 October 1995

[8] ISO 639-3, *Codes for the representation of names of languages — Part 3: Alpha-3 code for comprehensive coverage of languages*

[9] ISO 3166-1, *Codes for the representation of names of countries and their subdivisions — Part 1: Country codes*

[10] ISO 8601, *Data elements and interchange formats — Information interchange — Representation of dates and times*

[11] ISO 8879, *Information processing — Text and office systems — Standard Generalized Markup Language (SGML)*

[12] ISO/IEC 2382-29, *Information technology — Vocabulary — Part 29: Artificial intelligence — Speech recognition and synthesis*

[13] WEB CONSORTIUM: Extensible markup language (XML). 1.0 (Fifth Edition), 26 November 2008. Available at http://www.w3.org/TR/2008/REC-xml-20081126/

[14] Web Consortium: SCHEMA XML. 1.1 Part 2: Datatypes, W3C Recommendation, 5 April 2012, available at http://www.w3.org/TR/2012/REC-xmlschema11-2-20120405/

[15] Web Consortium: SCHEMA DEFINITION LANGUAGE XML (XSD) 1.1 Part 1: Structures, 5 April 2012, available at http://www.w3.org/TR/2012/REC-xmlschema11-1-20120405/

[16] WEB CONSORTIUM: XML Encryption Syntax and Processing Version 1.1, 2013, available at https://www.w3.org/TR/2013/REC-xmlenc-core1-20130411/

ICS 35.240.15

Price based on 26 pages