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)

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

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:

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

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