Investigatory Voice Biometrics Committee Report

Development of a Draft Type-11 Voice Signal Record

09 March, 2012

Version 1.8

Contents

Summary

The idea of automated and semi-automated (human-assisted) speaker recognition for forensic and investigatory applications goes back to World War II. Considerable government monies have been spent over the intervening 70 years in developing technical approaches, speech databases and testing programs. Missing from these efforts, however, has been the development of a forensic voice recording interchange format comparable to the interchange formats that currently exist for fingerprint, palmprint, face, iris, scar/mark/tattoo, and DNA data used for the purpose of human recognition. The Investigatory Voice Biometrics Committee (IVBC) was created by the FBI in early 2011 to take on the task of initiating development of a voice recording format to allow the interchange of voice data within the ANSI/NIST ITL forensic biometric data interchange standard [1], the current *de facto* international standard for exchange of biometric data for law enforcement and national security applications. This document is a report on those efforts and contains the first draft of a voice recording interchange format, which will be known as a "Type-11 Voice Signal Record" within the parlance of the ANSI/NIST ITL community. The Type-11 record is modeled roughly after existing record types in the 2011 version of the ANSI/NIST standard, but will allow exchange of both digital and analog data using both electronic and physical media. This Type-11 record is designed to be used within ANSI/NIST formatted transactions for law enforcement and homeland security-type speaker recognition applications. It is not designed for speaker recognition within logical or physical access control, "time-and-attendance", point-of-sale, or other consumer applications. It is the intent of the IVBC to submit this draft into the ANSI/NIST ITL process in early 2012 as a proposed starting point for the development of a speaker recognition supplement to the existing 2011 version of the standard.

Introduction

Speaker recognition presents some unique challenges not found in other forms of human recognition, such as fingerprint, iris or face. The human voice, generally carrying both speech and non-speech sounds, propagates varying distances through air to reach acoustic transducers (usually microphones) of varying amplitude and phase response. We will consider a "speaker" to be any person producing voice sounds ("vocalizations"), although with the current state of technology, speaker recognition usually requires vocalizations containing speech (linguistic content). We will also consider an automated interlocutor to be a "speaker" in this context, although such a speaker will not be the primary subject of a speaker recognition transaction.

When voice sounds carry speech, that speech usually occurs within a social context involving more than one speaker. Consequently, a speech signal collected *in situ* will generally contain the voices of multiple speakers, each voice signal with its own transfer function between the speaker and the transducer. Segmenting and de-conflicting overlapped voice signals ("speaker separation") through automation is currently an unsolved problem in the general case, thus

Comment [1]: General Comment (GC) 1: The number of "mandatory" items needs to be limited if it is to be used in our environment, since we often have no idea of most of the information in the record. Example - speech from a video that is downloaded from the Internet. We would normally have no info on what recorder was used to record it, which is currently noted as a mandatory item

Comment [2]: GC 2: I can see the value of this tour de force record in an LE environment, but here it can be hard to get analysts (myself included) to fill out the 20 or so pieces of metadata present in the DB for our SID system. If this record is associated with an automated system, that problem would be largely mitigated by having the machine fill out the majority of the record, with humans adding information iteratively (linguists checking/adding language and dialect, others adding mental state, etc.).

MARK: This appears as an editorial comment. Shall we record it for history in a diary and move on?

Comment [3]: General Comment 3: I presume that a similar record will be developed for speaker models. It is far more likely in the IC that models ...

Comment [4]: General Comment 4: I was wondering whether metric units will be used where applicable.

Comment [5]: GC 5: My only question is whether or not it makes sense to include anything about the certainty associated with the speaker label. Suppose that we assigned a ...

Comment [6]: GC 6: I am not completely happy with presenting this document only as a format for speaker identification (the very beginning, Summary, line 76). I think it could equally serve other purpo ...

Comment [7]: GC 7: The document needs a list of abbreviations, I did not have problem with the technical ones, but there is another set that is related to the organizational schemes, data-flows, etc., an ...

Comment [8]: GC 8: It is not clear, where the information for automatic speaker recognition software (features, statistics, low-dimensional ...

Comment [9]: Is there a reason defense applications are not mentioned. ...

Comment [10]: (lacking context) .. similar to those used in IVR systems?

Comment [USS11]: I was hoping this wasn't the only description of how speech is "unique" but it appears that it is. And this suggests the writers of the document want to portray spee ...

Comment [12]: The document counts on speaker physically producing the vocalizations (105). This might not be always true – think of a tapped conference with remote participants, when ...

Comment [13]: Some non-speech features have been used for speaker reco .. pause distributions in turn taking, rhythm/cadence, laughter and throat clearing. ...

implying that operational applications of speaker recognition technology will involve audio recordings containing multiple speakers and multiple acoustic transmission paths.

The ANSI/NIST ITL standard was originally developed for the interchange of fingerprint data, whether collected from latent prints lifted from crime scenes, scanned off of ink-based fingerprint cards or taken directly from electronic "live" scanners. The standard, therefore, is explicitly restricted to cases where, "All records in a transaction shall pertain to a single subject". This restriction presents special challenges for use of the standard for interchange of natural voice signals, containing both speech and non-speech sounds, collected in a social, multi-speaker context and stored either digitally or in analog form and either electronically or on physical media. Therefore, a voice signal record type will have to accommodate: 1) bespoke recordings of single speaker voice signals for the specific purpose of speaker recognition; 2) conversational and interview scenario voice signals, digitized and segmented into clips, or "snips", restricted to speech from the single speaker of interest (the voice data subject); 3) unsegmented natural voice signals on digital or analog media, with or without an accompanying timing diary of the segments attributable to speech from the single speaker of interest; 4) unannotated speech segment(s) for input to annotation work-flow tools. In all cases, the voice samples referred to in the Type-11 record must accommodate signals collected non-continuously and stored in multiple segments, a requirement that has been encountered before in other ANSI/NIST record types. For example, the Type-14 (variable-resolution fingerprint images) record has the capacity to carry multiple fingerprint samples in one image with segment boundary information for each finger in the image, albeit from a single individual, and serves as a model in this regard.

There are other challenges facing a speaker recognition standard. The most significant ones include:

- Voice signals generally contain both speech and non-speech elements, either of which might be useful in speaker recognition applications.
- Unlike other modalities, voice signals are collected in time, not spatial, dimensions and will not have a single "time of collection".
- In mobile applications, even a single segment of a voice signal may not be linkable to a single geographic location.
- Voice signals containing speech have direct informational content. Unlike other forms of biometric recognition, the speech itself means something and, even if stripped of all personally identifiable information including the acoustic content itself, may require protection for privacy or security reasons.

Consequently, creating a Type-11record for voice signal transmission with the ANSI/NIST ITL context is more complicated than simply copying an existing ANSI/NIST record type and changing terminology (for example, substituting "voice" for "fingerprint" and "signal" for "image") as has been often suggested by the standards community. In the case of DNA Type-18 records, ANSI/NIST has previously shown significant flexibility in dealing with record types which carry non-spatial data with significant content beyond that required for the recognition of individuals. Consequently, we fully expect that the current ANSI/NIST structure can flex to accommodate the "nuisances" of voice.

Comment [14]: That is not the case for all operational applications, i.e., in certain applications you often find speakers (and speech) in isolation. For example, four-wire phone calls, interviews, propaganda videos, and speeches, to mention a few. (relative to two sections of this page)

Comment [USS15]: See above, but yes, overlapping speakers are an issue. However, this accounts for very little of the challenges within speaker recognition in the law enforcement domain.

Comment [16]: There are Auditory Scene Analysis systems (and other mic arrays) that do record (or estimate) speaker location or direction of speech sources.

And

Not always true for certain applications, it all depends on the communication platform or geocoding capabilities.

Comment [USS17]: No mention whatsoever of the more obvious and challenging speech variability concerns related to social and behavioral mismatch.

Investigatory Voice Committee Membership

Joseph Campbell, MIT
Carson Dayley, FBI
Craig Greenberg, NIST
Peter Higgins, Consultant
Alysha Jeans, FBI
Ryan Lewis, FBI
Jim Loudermilk, FBI
Kenneth Marr, FBI
Alvin Martin, Consultant
Hirotaka Nakasone, FBI
Mark Przybocki, NIST (IVBC Chair)
Vince Stanford, NIST
Pedro Torres-Carrasquillo, MIT
James Wayman, Consultant
Bradford Wing, NIST

Definitions of Specialized Terms Used in this Document

Acoustic signal

Pressure waves in a media with information content.

Note: In this document, acoustic signals will be restricted to pressure waves in air.

Comment [18]: "pressure waves in air" does not take into account the usage of laryngophones (that are later mentioned as one possible type of mikes).

Audio signal

Information in analog or digital form that contains acoustic content (voice or otherwise), capable of being transduced into an audible acoustic signal.

Note: By "audible" we mean "capable of being heard by humans".

Contemporaneous

Existing at or occurring at the same period of time.

Note: In this record type, the phrase "contemporaneous capture of a voice signal" indicates recording of the voice signal at the time of the speaker vocalization.

Diary

List giving the start and stop times of speech segments of interest pertaining to the primary voice signal subject within the voice signal.

Note: Diarization of segments from multiple speakers requires multiple Type-11 records, one for each speaker.

Known Voice Signal

A voice signal from an individual who has been "identified", or individuated in a way that allows linking to additional, available information about that individual.

Metadata

Documentation about the voice signal necessary or helpful in supporting the types of speaker recognition transactions likely to be encountered in law enforcement and homeland security applications.

Physical medium

Any external storage material of the voice signal in either analog or digital form. Examples include reel-to-reel recording tape, cassette tape, Compact Disc, phonograph record.

Quality

An estimate of the usefulness of a voice signal for the purpose of speaker recognition.

Questioned Voice Signal

A voice signal from an individual who is unknown and cannot currently be linked to any previously encountered individual.

Note: The task of speaker identification is to link a questioned voice sample to a known voice sample ~~so that the speaker can be linked to additional, available information~~.

Record (n)

An ANSI/NIST biometric data format type, in its entirety, within an ANSI/NIST transaction.

Note 1:In this document, this will be the Type-11 record unless otherwise stated.

Note 2: An ANSI/NIST transaction might contain multiple Type-11 records, as well as other record types, including the mandatory Type-1 record. In the current FBI Electronic Biometric Transmission Standard (EBTS), a transaction will also contain the mandatory Type-2 record.

Record (v)

The act of converting an acoustic voice signal directly from an individual into a storage media, perhaps through contemporaneous, intermediate (transient) signal types.

Note: We maintain this definition because of its entrenchment in natural language use. Consequently, a record (n) is not recorded, it is created.

Note: Transcoding is the term used for further processing of the voice signal and any digital or analog representation of that signal.

Record creation

The act of creating a Type-11 record pertaining to a voice signal(s).

Recording (n)

A stored acoustic signal in either analog or digital form.

Redaction

Over-writing of segments of a voice signal for the purpose of masking speech content in a way that does not disrupt the time record of the original recording.

Snip (n)

A segment of a voice signal extracted from a larger voice signal recording.

Note: Also called a “clip” in some communities.

Snip (v)

Extraction of segments of a voice signal in a way that disrupts the continuity and time record of the original recording.

Comment [19]: (lacking context) For both Signal Related Info and Content related info?

Comment [USS20]: I do not agree with this definition. A questioned voice signal is from a speaker that is simply unknown (or their identity can not be confirmed)

Comment [USS21]: I do not agree with this definition. The task of speaker ID is to determine if the questioned voice sample(s) shares the same source as a known voice sample. The last half of that definition must be agency specific.

MARK: I agree and believe we can delete as suggested.

Speaker
A vocalizing human, whether or not the vocalizations contain speech.
Note: An interlocutor might be a synthesized voice, which can be considered a "speaker" within the context of this report.

Speech
Audible vocalizations made with the intent of communicating information through linguistic content.
Note 1: Nonsensical vocalizations with linguistic content will be considered as speech.
Note 2: Speech can be made by humans or by machine synthesizers.

Subject of the Type-11 record
The person to whom the voice data in the Type-11 record applies.
Note: Because a transaction can include Type-11 records for interlocutors and others not named as the subject of the transaction, the subject of the Type-11 record need not be the subject of the transaction.

Subject of the transaction
The person to whom the transaction applies.
Note: The primary or only speaker in a Type-11 record need not be the subject of the transaction.

Transaction
A transmission between sites or agencies comprised of records, types of which are defined in the ANSI/NIST ITL 1-2011 [1].
Note: An ANSI/NIST-ITL transaction is called a file in Traditional encoding and an Exchange Package in XML encoding.

Transcoding
Any transfer, compression, manipulation, re-formatting or re-storage of the original recorded material.
Note 1: Transcoding is not the first recording of the acoustic signal.
Note 2: Transcoding can be lossless or lossy.

Voice data file
The digital, encoded file primarily containing the sounds of vocalizations of both speech and non-speech content, convertible to an acoustic signal replicating the original acoustic signal.
Note 1: A voice data file is extracted from an audio signal, but not all audio signals contains voice data and not all voice data is speech.
Note 2: A physical medium, such as a cassette tape, contains a voice signal but is not a voice data file.

Voice recording
A signal, stored on a digital or analog medium, of vocalizations containing both speech and non-speech content.

Comment [22]: (lacking context of where on page) Are you prescribing particular methods?

MARK: the answer is surely no. We should check if this page is giving that perception.

Comment [23]: -speech can also be produced by replay devices, not just by humans or synthesizers (see my comment #3 above)

MARK: Should we replace "or by machine synthesizers" with "or by other means"

Voice signal subject
The single speaker of interest in the Type-11 record.
Note 1: This may not be the subject of the transaction.
Note 2: The voice signal subject may be known or unknown.

Relationship Between the Type-11 Record and Other Record Types and Documents

A Type-11 record would never be used in isolation, but would be placed in the context of an ANSI/NIST ITL transaction, which would by necessity contain at least the mandatory Type-1 record. A single ANSI/NIST transaction might contain several Type-11 records: for example, one or more Type-11 records of "Known" voice signals perhaps from different individuals and one or more Type-11 records with "Questioned" signals. In an FBI EBTS transaction, there will be a separate Type-2 record for the "Questioned" and each "Known" speaker. Further, the specifics of the implementation of the standard to support various types of transactions between agencies, including the mapping of subjects identified in Type-2 records to their roles in voice recordings, would be specified in "exchange agreements", such as the Electronic Biometric Transmission Specification (EBTS) [2] used by the FBI and DOD. The EBTS specifies by Type of Transaction (TOT) which fields and record types will be mandatory and which optional.

Comment [24]: (lacking context of where on page) What are these? Because of storage constraints?

The Type-11 record will also utilize the new Type-20 record included in the 2011 version of ANSI/NIST ITL. Type-20 is a broadly defined record type for unedited and unmodified source data from which data subject-specific, segmented biometric data can be derived. The Type-11 record can take advantage of the availability of the Type-20 records as a method for transmitting unedited or unmodified voice signals when in digital form.

For EBTS users, development of the Type-11 record must take place within the context of the existing Type-1, Type-2 and Type-20 records and the current EBTS, all of which are living documents subject to modification. Domains of interest, such as the Criminal Justice Information Services Division of the FBI (CJIS), may in the future update their EBTS specification to reflect use of Type-11 records within their domains.

This report of the IVBC contains no information about recording or transmission "best practices", although we acknowledge the extreme importance of these issues. It is the intention of the committee to develop such documents in the future.

Some Types of Transactions Supported by a Type-11 Record

The IVBC considered various types of voice signal transactions currently supported and anticipated by the FBI. The committee recommends that the current FBI EBTS document be updated to support these voice signal transactions. This committee also recommends that a best practices document for use by the FBI EBTS community of interest be developed describing how

to use Type-11 and other record types for commonly expected scenarios of operation. This work can be accomplished at a speed determined appropriate by EBTS community. A partial list of potential "types of transactions" (TOTs) would be:

1. Voice model creation and storage for a known speaker
2. Voice model creation and storage for an unknown speaker.
3. Comparison of the speakers in two audio recordings.
4. Comparison of the voice in an audio recording to the voice models from a list of known speakers.
5. Converting an analog audio recording into digitized voice data file(s).
6. Duplicating or transcoding an audio recording.
7. Finding and isolating voice signals in an audio recording.
8. Finding and isolating speech signals within an audio recording.
9. Determination of the distinct speakers in an audio recording.
10. Indexing an audio recording into voice segments attributable to distinct speakers.
11. Creation of a diary, attributing speech segments to a speaker of interest.
12. Creation of word or phone level transcriptions, in the language spoken, of segments of speech attributable to a single speaker.
13. Redaction of an audio recording to remove sensitive speech segments.
14. Snipping of an audio recording to remove segments of non-speech, speech not attributable to the subject of interest, or speech not of interest to the transaction.
15. Enhancing the speech segments in an audio recording for return to the submitting agency for use in human-assisted or automated speaker recognition applications.
16. Authentication of an audio recording as containing the continuous speech of a single speaker without deletions or insertions.
17. Transfer voice recording to an archive for permanent storage.

The above list of potential "TOTs" will have to be further refined to determine which transaction types require responses from the receiving agency and which do not.

The Type-1 record within an ANSI/NIST transaction contains Field 1.004 for specifying the TOT – the purpose for which the transaction was generated. This field may be left blank in an ANSI/NIST transaction to indicate that the submitting agency is asking for a transaction type not formally listed.

This voice signal record type must support all of these transactions originating from submitting agencies with little or no capability in digitizing audio signals or in speech analysis, as well as inter- and intra-laboratory transmissions on fully or partially processed voice recordings. Further, the type of transactions ultimately to be performed on the voice recording might not be fully known at the time the Type-11 record is created. Therefore, the voice recordings referred to in the record must be accompanied by documentation, when available, to support a very wide variety of potential transactions. In this report, this documentation will be referred to as "metadata" and will be of four basic types:

- Administrative metadata: who initiated the transaction, why and with what authority?

Comment [25]: (lacking context) This issue should be addressed given collections in place and planned.

Comment [26]: -the storage does not need to be permanent – it can be well mandated to store it for a limited time.

MARK: This is a partial list so I think we can delete the comment (record in a diary) and move on.

Comment [USS27]: Characterization of the social/behavioral dynamic in the recording, i.e: accommodation, interlocutor issues (authoritative dynamic between speaker and interlocutor), etc.

MARK: This comment is asking for us to add to our partial list.

- Speaker metadata: what is known about the speaker of interest and their physical and psychological condition at the time of the speech?
- Content metadata: what language is being spoken, when was the original content spoken under what conditions, and what content information is available that might help in the speaker recognition process?
- Audio technology metadata: how was the voice signal collected, stored and processed and what technical parameters will help in the faithful reproduction and analysis of the signal within the storage medium?

Some of this metadata, such as the time and date of the original recording, might only be known from external sources. Some of the metadata, such as the language being spoken, might be discernible from the voice recording itself. Much of the metadata might not be known or available to the various agencies creating the audio recording, the Type-11 record and the ANSI/NIST transaction. All of the metadata, however, could be useful in the processing of the audio recording given the potential for widely varying transactions and should be made readily available to the receiving agency without requiring the reprocessing of the audio recording. Consequently, our goal is to create as many non-redundant metadata fields as possible to permit transmission of documentation of potential future interest, even if the metadata could potentially be recovered from the audio recording itself. Most of these fields will be optional because much of the potentially relevant metadata may be unknown to the various agencies involved in the transaction.

Comment [28]: (Lacking Context of where in page – now just (LC!)) Are you prescribing particular (approved/suggested methods?

Mark: The answer is surely no.

Scope of the Type-11 Record

This record type is intended to support the transmission of audio recordings containing speech by one or more speakers and noise (data of no interest to the transaction, whether speech, non-speech voice data, or non-voice data) for forensic and investigatory purposes in the context of an ANSI/NIST ITL transaction pertaining to a single, perhaps unknown, individual. These transmissions are intended to support transactions related to detecting and recognizing speakers, extracting from an audio recording speech segments attributable to a single speaker, and linking speech segments by speaker, whether these functions are to be accomplished through automated means (computers), human experts, or hybrid human-assisted systems. Related functions, such as redaction, authentication, phonetic transcription and enhancement, while also supported, are not the primary concern of this record type, although audio recordings supporting these related functions may be transmitted via Type-11 records. This record type is not designed for use in logical or physical access control, “time-and-attendance”, “point-of-sale”, or other consumer or commercial applications, and does not support streaming transactions. This record does not define the transmission of features or models extracted from voice data. This record type does not restrict the media by which the audio recording will be transmitted, but will support digital transmission of transaction information regardless of the audio recording media. A best practices document for voice recording will be created in a future effort.

Source Documents

1. ANSI/NIST ITL 1-2011, "Data Format for the Interchange of Fingerprint, Facial & Other Biometric Information", NIST Special Publication 500-290, November, 2011
2. "Electronic Biometric Transmission Specification", Criminal Justice Information Services Division, IAFIS-DOC 01078-9.2, 9 December, 2011
3. Collaborative Digitization Program, Digital Audio Working Group, "Digital Audio Best Practices", version 2.1, October, 2006, http://ucblibraries.colorado.edu/systems/digitalinitiatives/docs/digital-audio-bp.pdf
4. Audio Engineering Society, "AES standard for audio metadata - Audio object structures for preservation and restoration", AES57-2011, Sept. 21, 2011
5. Audio Engineering Society, "AES standard for audio metadata -Core audio metadata", AES60-2011, Sept. 22, 2011

Administrative Metadata Requirements

The IVBC identified the following requirements for administrative metadata for transactions containing audio data:

Comment [29]: Would this include speaker-specific traits, e.g. speech impediments, accent, dialect or any other salient feature?

Requirement 1: Point-of-Contact (POC) Name
Requirement 2: Agency
Requirement 3: Phone number
Requirement 4: Originating agency case ID
Requirement 5: Transaction ID
Requirement 6: Embed Case ID
Requirement 7: Email address of submitter
Requirement 8: Alternative POC
Requirement 9: User defined fields, such as "FBI/non-FBI case"

The above information is required to promote traceability of the audio data. There are at least three levels of traceability – to the submitting, compiling/post-processing and collecting agencies. It is possible for all three agencies to be the same in some transactions, but they will often be different.

The submitting agency is one approved to submit a transaction to a receiving agency for the processing requested in the "Type of Transaction" (TOT) field within the Type-1 record. Within the ANSI/NIST structure, the submitting agency is denoted in a Type-1 field, Field 1.008 (Originating Agency/ORI), which contains the identifying number of the agency that served as a channel for the request to the receiving agency for processing. The name is contained in Field 1.017 Agency Names/ANM, information item originating agency name/OAN. Other record Types also include Fields XX.004 (Submitting Agency/SRC) and XX.993 (Source agency name / SAN). Within the U.S. law enforcement domain of interest, the relevant EBTSs establish that ORI must be from an National Crime Information Center (NCIC)-authorized agency.

Type-18 (DNA) contains a more comprehensive Field 18.003 (DNA Laboratory Setting/DLS) that will serve as a model for metadata on the lab or agency that created the voice recording.

As defined in the FBI EBTS (not in the ANSI/NIST ITL document itself), the agency that compiled the raw data into a search transaction is identified in the Type-2 field, 2.073 (Controlling Agency Identifier/CRI). By FBI EBTS, the CRI itself has three levels and the ORI and top-level CRI are usually the same[1].

The collecting agency is not specified in either Type-1 or Type-2 (as defined by EBTS) records. The original source of the voice recording, which might be a local law enforcement office, 911 call center, or other non-law enforcement group, will be included specifically within the Type-11 record as Field 11.004 (Voice Laboratory Setting/VLS), modeled after Field 18.003 (DNA Laboratory Setting/DLS) and after Field 19.004 (Plantar images: source agency/SRC)

We can map these three levels as follows.
Submitting agency – ORI Field 1.008; ANM_OAN Field 1.017
Compiling agency – CRI Field 2.073
Collecting agency – VLS Field 11.004; SAN Field 11.993

This complex system is used where a county / city has an AFIS that submits to the state and in turn submitted by the state to the FBI. All of this structure will have to be re-thought for voice and policies established.

So, with regard to the specific metadata requirements identified above by the IVBC, in FBI-centric applications:

Response to requirement 1: Although neither the current Type-1 nor the FBI EBTS-defined Type-2 records allow for inclusion of the agency responsible for the collection of an audio or voice recording, this information is included in the Type-19 record field 19.004. Information about the specific individual or individuals responsible for the collection and serving as a Point of Contact will be a subfield of Field 11.004. Type-19 Field 19.004 contains additional information about the original source (agency, non-agency) of the data, including multiple subfields about the source. So for voice, Field 11.004 would contain data about the original source and well as the information identified in Administrative Metadata Requirement1 above.

Response to requirement 2: By FBI EBTS definition, the originating agency case ID is Field 2.009 and thus does not have to be defined as part of the Type-11 record.

Response to requirement 3: Phone number of the original source would be an additional subfield of 11.004

Comment [30]: (LC) and (deleted) an (inserted)

Comment [31]: AFIS not explained. Expand and add to list of abbreviations.

MARK: We can do this.

[1] The only time these would be different in an FBI EBTS-based transaction is for a user that is authorized by the State Identification Bureau (SIB) to submit friction ridge searches directly to CJIS without passing through the state system. In this case the ORI would be for the state, while the top-level CRI would be for the agency directly submitting the transaction

Response to requirements 4 and 5: Transaction ID is Field 1.009 Transaction control number/TCN. This can be the case ID of the originating agency. Additionally, Type-1 Field 1.010 Transaction control reference / TCR may be used to reference the TCN of a previous transaction involving an inquiry or other action that required a response.

Response to requirement 6: Provision for embedding an FBI file number is made by the FBI EBTS in Field 2.003. Field 2.012 is for the "FBI Latent Case Number". These numbers are in a format controlled by the FBI Latent fingerprint Section, a corollary set of numbers will have to be established for voice records. Outside of the Type-11 record, the various EBTS controlling authorities will have to add a fieled for voice cases.

Comment [32]: Field.

MARK: Corrected

Response to requirement 7: As there is no provision currently within Type-1 for an email address or alternate POC. However, Field 18.003 contains parallel information about the DNA laboratory processing the data. For voice, this information would be placed in subfields of 11.004.

Response to requirement 8: An alternate point of contact can be specified in a subfield of 11.004.

Response to requirement 9: The record type will accommodate user defined fields as defined in exchange agreements.

Also included in Type-1 is Field 1.006, which gives a priority (an integer 1…9) by the originating agency for the processing of the data. The lower the number, the higher the priority, as per the ANSI/NIST Standard. This field will be of interest in voice data transactions. Receiving laboratories will have to provide guidance on their priority schema, perhaps in the appropriate EBTS.

The security classification of both data and metadata is a concern for all ANSI/NIST record types and is being addressed by US government committees through the use of a uniform "wrapper" to the records to identify classification level. Consequently, security classification issues will not be addressed in this document.

Speaker and Content Metadata Requirements

The IVBC has identified some of the requirements for metadata about the data subject and the subject's speech, which are listed below. We recognize that the distinction between "permanent" and "temporary" attributes might be elusive in many cases.

Comment [33]: body height could be listed as another permanent speaker attribute. As is known from the literature, correlations between body height and acoustic properties such as formant frequencies are not impressive, but not fully absent either. If body height information can be obtained I think it should be stored because it can prove useful (probably more so in voice profiling than in voice comparison).

MARK: Body Height is not permanent.

1. Identifier
2. Permanent Attributes
    a. gender

b. accent[2]
c. date of birth
d. native language
e. educational level (may not be permanent)
f. primary location where they grew up
g. speech impediment (may be intermittent)

3. Temporary attributes
   a. Impairment/intoxication
   b. Language being spoken
   c. Health status
   d. Intelligibility
   e. Style (public speech, conversation, read, prompted, interview, other)
   f. Emotional state/vocal effort
   g. Citizenship

In the case of voice data records belonging to the subject of the transaction, three of these items are already included in EBTS-defined Type-2 fields: identifier (name) –2.018; gender(sex) – 2.024; date of birth – 2.022; citizenship -- 2.021.

In the case of voice data records of persons not the subject of the transaction, additional Type-2 records will be added to the transaction, but without altering the requirement that the transaction as a whole must pertain to a single subject.

The ANSI/NIST ITL standard states that Type-2 Fields 2.003 and above are user defined fields. "Individual fields shall conform to the specifications set forth by the agency to which the transmission is being sent…." This implies that we might specify additional speaker metadata requirements as specific Type-2 fields to be added to the existing fields defined in the appropriate EBTS specifications. However, there is also precedence in Type-18 (DNA) for donor specific information the field 18.006. A logical approach would be to place the permanent attributes of the speaker into Type-2 fields and place the temporary attributes as subfields because those attributes are only pertinent to the transaction at hand. Consequently, we could put:

2. X00 NL native language
2.X01 EL educational level
2.X02 UP geographical location of first 12 years of life
2.X03 SI speech impairment

within the Type-2 record definition of the FBI EBTS document, while defining the remaining speaker and content metadata as fields within the Type-11 record.

The current EBTS has a field (2.037 RFP) assigned for "reason fingerprinted". Several implementations of the ANSI/NIST standard (e.g., the Afghan EBTS) have simply changed this

[2] Although this was an original requirement of the IVBC, we have been unable to find a metric or a means of codification of "accent".

Comment [USS34]: "Accent" or dialect is by no means a permanent attribute.

Comment [35]: There can be more complicated **language acquisition biographies** than just one first language (L1) and one second language (L2). An example from our casework is a speaker who grew up in Russia (L1 Russian), came to Germany as a young adult (L2 Germany, with fluent but slightly accented German), and spoke English in the speech samples (L3 "school"-English, nonfluent with a pretty bad accent). Other examples are Arabic (L1), Dutch (L2), German (L3) or Kurdish (L1), Turkish (L2), German (L3). Maybe there is a more flexible way to specify language biographies in the permanent speaker attribute list than is possible in the current scheme.

MARK: Is this an example of specific information that the record should allow to be captured in a freeform notes section?

Comment [USS36]: The correct term is "speech pathology" or "impairment" but this is not necessarily permanent either and brings up a whole host of issues about how a pathology is even defined.

Comment [37]: – Temporary attributes – we had a seminar of Dr. Svobodova, a Czech forensic expert and she told us a story about a heavily schizophrenic person that had also two speaker characteristics different to such extent that it fooled her as well as other experts. I'm asking myself if it would make sense to add a field "personality" to 11.037 or if it is so rare that it could be left to COM field …

MARK: Too rare. Leave for comment.

Comment [38]: **voice disguise** is a category that I see missing from the list of temporary attributes. I know, it's a difficult category because there can be many types of voice disguise (e.g. changing the voice source to falsetto; changing the filter with lip rounding or insertion of objects in the mouth; imitating different dialects or foreign accents; distorting speech rhythm; using technical voice-changing devices or software) and because it might not be easy to detect whether voice disguise took place at all. Yet, voice disguise is a typically forensic type of phenomenon and has regularly been discussed in the literature (though from our experience it is not all too common, occurring in perhaps five to ten percent of the cases). If you decide to include voice disguise in field 11.037 (pp. 50-52), I can give you some more input. A recent survey of voice disguise is presented by Eriksson (2010). …

Comment [39]: Temporary Attributes (Line 567) – What about adding dialect to this section (possibly as part of line 569)?

MARK: Second vote for dialect to be "temporary"

Comment [USS40]: Smoker or not, fatigue (possibly under health status), interlocutor effects, are very important under this heading

field to “reason enrolled”. The current FBI CJIS EBTS still uses fingerprints as identity foundations. The submittal of facial images from known collections, for instance, are always coupled with a set of fingerprints to ensure the facial image is linked to the correct FBI number. It is anticipated that many, if not all, voice submittals will not include fingerprints. Consequently, we could define within the FBI EBTS:

2.X04 RE reason enrolled

This field might additionally be used to contain the legal justification (i.e., Title III Omnibus Crime Control and Safe Street Act of 1968; Title 50 of the Foreign Intelligence Surveillance Act of 1978) for the existence of the voice data record. Alternatively, the Type-21 record could be used to contain the necessary legal documents, such as court orders, supporting the recording.

Audio Technology Metadata Requirements

The IVBC has identified the following requirements for metadata about the audio technology used to record the voice data:

Comment [41]: Are you considering naitiveness and proficiency in the spoken language as well?

1. Overall/Preliminary signal quality
2. Duration measured in seconds
3. Duration measured in samples
4. Encoding
5. Sampling rate
6. Bit depth (may be encoding dependent)
7. Recording (conversion of temporary to permanent storage)
8. Time/date of recording
9. Where recorded
10. Type of recorder
11. Make/model/serial number
12. Transducer characteristics
13. Transducer type: array, earbud, wire, microphone, handset, speaker phone,…
14. Channel information

All of these elements would most naturally become fields and subfields within the Type-11 record and have been included in the following draft.

Audit Logs

The Record Type-98, “Information assurance record”, allows special data protection procedures to ensure the integrity of the transmitted data and allows for the maintenance of an audit log. Field 98.900 (Audit log / ALF) may be used to indicate how and why a transaction was modified. The ALF is of particular use when a transaction is sent from one location to a second, where additional information is included, before sending the transaction to a final destination for processing. In the case of a voice recording, the ALF will be used to indicate how and why

redaction, snipping and diarization information was created or edited. See ANSI/ NIST ITL 2011, Section 8.22.

An example might be that a local police lab sends a transaction with multiple Type-11 records, containing voice signals of both known and unknown persons, to the appropriate FBI Field Office (FO) where additional information would be added, such as an FBI file number. The FO might also redact case-sensitive speech from the voice recordings referenced in the Type-11 records before sending the transaction to another FBI forensic unit for additional redaction. The forensic unit would then forward the updated transaction to the FBI Forensic Audio, Video and Image Analysis Unit (FAVIAU). FAVIAU might create diaries of the questioned voice samples in the Type-11 records, indicating which segments were from the speaker of interest, as recorded in a known voice sample in additional Type-11 records. The diaries might be revised after additional supervisory review. All of this would be documented in a Type-98 record included in the evolving transaction.

In contrast to the Type-98 record, which presents an audit log at the level of the entire transaction, Field 11.902, which is modeled after XX.902 fields in other record types, provides an audit log at the level of the Type-11 record. Field 11.902 lists the operations, such as redaction, snipping or diarization, performed on the original voice recording in order to prepare it for inclusion in the record type. See Section 7.4.1 of the ANSI/NIST ITL standard.

General Organization of the Type-11 Record

The Type-11 record is organized into 6 parts: I) mandatory fields; II) initial global fields, applying to the entire voice data record; III) indication of presence and definition of segments within the voice data record; IV) fields applying to the individual segments; V) additional global fields modeled on other Types in the ANSI/NIST standard; VI) fields containing or pointing to the voice recording.

I. Mandatory fields:
    01 Record header
    02 Information designation character
II. The initial global fields are:
    03 Audio object descriptor (internal or external digital file, external physical media containing digital/analog/unknown recording)
    04 Voice laboratory setting (source of the voice recording, phone numbers and POCs)
    05 Role of voice recording (known sample, unknown single speaker, unknown multiple speakers)
    06 Recorder (hardware/software)
    07 Type-11 record creation date
    08 Voice recording creation date
    09 Total signal duration
    10 Physical media object (tape, CD, phonograph record,...)
    11 Codec
    12 Preliminary signal quality (multiple quality metrics possible)

13-20 Fields reserved for future use

III. The presence and definition of segments within the audio file follow.

21 Redaction (yes/no, by whom?)

22 Redaction diary (where and why redaction occurred)

23 Snipping (yes/no, by whom?)

24 Snipping diary (separate snips/clips are numbered and identified by relative start/end times, comments)

25 Diarization (yes/no, by whom?)

26 Segment diary (segments are numbered with relative start/end times, labels of attributes attributed to the speech and speaker of each segment, and comments.)

27-30 Reserved for future use

IV. Repeating sets of sub-fields labeled by segment numbers as designated in the diarization. (If the segment number is "0", that becomes the default for all segments not otherwise listed.)

31 Date/time of recording of segment/snip and labeled date/time of recording

32 Geolocation of data subject of this Type-11 record at start of segment/snip

33 Segment/snip quality values (possible multiple values for each segment)

34 Vocal collision indicator (two or more persons speaking at once)

35 Processing priority of the segment/snip

36 Segment content (language, prompted/read/conversation, word transcript, phonetic transcript, translations)

37 Segment/snip speaker characteristics (impairment, intelligibility, health, emotion, vocal effort, vocal style)

38 Segment channel (transducer, capture environment, channel type)

39-50 Fields reserved for future use

V. More global fields modeled on other record types in ANSI/NIST ITL 2011:

51 Global comments

52 – 901 Fields reserved for future use

902 Annotation information

903 Device Unique Identifier

904 Make/Model/Serial

905-992 Fields reserved for future use

993 Source Agency Name

VI. The voice recording or pointers to that recording:

994 External file reference

995 Associated context (Type 21 record)

996 Voice data file hash Editor's note: Data integrity on the entire file is handled with a Type-98 record

997 Source representation (Type 20 record with original audio)

998 Field reserved for future use

999 Voice data file

Draft Record Type-11: Voice signal record

The Type-11 record shall be used to exchange a single voice data file or a physical medium containing a digital or analog voice recording, together with fixed and user-defined textual

information fields (referred to in this standard as “metadata”) pertinent for understanding and processing the voice signal.

A voice signal is defined in this standard as any audible vocalizations emanating from the human mouth with or without speech content. The Type-11 record will reference a recording of a voice signal stored as a digital voice data file within the record or external to the record. Information regarding the encoding type, the voice data file size, and other parameters or comments required to process the voice data file are given as fields within the Type-11 record. If the Type-11 record references a voice recording contained in a physical medium (i.e., an analog tape, a digital tape, a CD, a phonograph record), the label and location of that medium shall be indicated in this Type-11 record, along with the information necessary to render the stored recording as acoustic output.

Comment [42]: for physical media (such as phonograph record), it should be precisely defined what is “time zero” – needed for diaries. Or it should me mandatory to convert the record to digital form before specifying any timing …

A transmitted voice recording may be processed by the recipient agencies to isolate the voice signal of interest and to extract the desired feature or model information required for voice comparison, speaker detection, or speech attribution purposes.

A single ANSI/NIST transaction might contain multiple voice recordings, each as a separate Type-11 record within the transaction. Although the transaction pertains to a single person, the individual voice recordings in each of the Type-11 records required for the transaction may contain the speech of multiple speakers. For each known speaker in the recording, there should be a Type-2 record in the transaction.

If there are multiple speakers of interest in a voice recording supported by a Type-11 record, then a separate ANSI/NIST-ITL transaction may be created for each individual of interest, each transaction possibly containing the same Type-11 records. If the voice recording included in or pointed to by a Type-11 record has been extracted from a longer source recording, that source recording may be included in digital form within the transaction as a Type-20 record. Voice models or features extracted from voice data are not explicitly accommodated in this record, but may be transmitted in user-defined fields.

Table Sup:Voice 1 Type-11 record layout

Comment [43]: -many types of values for fields are not in the header of the table: U (user defined?), NS (number sequence?), H (hexa) etc.

Key for Character type: N=Numeric; A=Alphabetic; AN=Alphanumeric; B=Binary or Base64
Key for Cond. code: M=Mandatory; O=Optional; D = Dependent upon another value or condition described in the text;
M↑=Mandatory if the field/subfield is used; O↑=Optional if the field/subfield is used.

Field Number	Mnemonic	Content Description	Cond code	Character			Value Constraints	Occurrence	
				Type	Min #	Max #		Min #	Max #
11.001		**RECORD HEADER**	M	encoding specific: see Annex B: Traditional encoding or Annex C: NIEM-conformant encoding rules			encoding specific: see Annex B: Traditional encoding or Annex C: NIEM-	1	1

Field Number	Mnemonic	Content Description	Cond code	Character			Value Constraints	Occurrence	
				Type	Min #	Max #		Min #	Max #
							conformant encoding rules		
11.002	**IDC**	**INFORMATION DESIGNATION CHARACTER**	M	N	1	2	0 ≤ IDC ≤ 99 integer	1	1
11.003	**AOD**	**AUDIO OBJECT DESCRIPTOR**	M	N	1	1	See Table 11-1 1 ≤ AOD ≤ 4	1	1
11.004	**VLS**	**VOICE LABORATORY SETTING**	O					0	1
	LTY	lab type	O↑	A	1	1	LTY = G, I, P, O or U	0	1
	NOO	name of original source	O↑	U	1	400	none	0	1
	POC	point of contact	O↑	U	1	200	none	0	1
	CSC	code of sending country	O↑	A	1	3	value from *ISO-3166-1*	0	1
11.005	**ROL**	**ROLE OF VOICE RECORDING**	M	N	1	2	See Table 11.2 0 ≤ ROL ≤ 99	1	1
11.006	**REC**	**RECORDER**	M					1	1
	RTP	recorder type	O	U	1	4000	none	0	1
	MAK	recorder make	O	U	1	50	none	0	1
	MOD	recorder model	O	U	1	50	none	0	1
	SER	recorder serial number	O	U	1	50	none	0	1
	AQS	acquisition source	M	AN	1	2	value from **Table 83 except for 1 through 6 inclusive or 11; or AQS = MS**	1	1
	COM	comment	O	U	1	4000	none	0	1

ment [44]: (LC) Are you taking the nasal into account?

Field Number	Mnemonic	Content Description	Cond code	Character			Value Constraints	Occurrence	
				Type	Min #	Max #		Min #	Max #
11.007	**RCD**	**RECORD CREATION DATE**	M	See Section 7.7.2.4 Local date and time; encoding specific: see Annex B: Traditional encoding or Annex C: NIEM-conformant encoding rules			See Section 7.7.2.4 Local date and time; encoding specific: see Annex B: Traditional encoding or Annex C: NIEM-conformant encoding rules	1	1
11.008	**VRD**	**VOICE RECORDING CREATION DATE**	O	See Section 7.7.2.4 Local date and time; encoding specific: see Annex B: Traditional encoding or Annex C: NIEM-conformant encoding rules			See Section 7.7.2.4 Local date and time; encoding specific: see Annex B: Traditional encoding or Annex C: NIEM-conformant encoding rules	0	1
11.009	**TRD**	**TOTAL RECORDING DURATION**	O					0	1
	TIM	total time	O↑	N	1	14	1 ≤ TIM ≤ 99999999999999 (in microseconds) (no commas)	0	1
	CBY	compressed bytes	O↑	N	1	14	1 ≤ CBY ≤ 99999999999999 (no commas)	0	1
	SMP	total samples	O↑	N	1	14	1 ≤ SMP ≤ 99999999999999 (no commas)	0	1
11.010	**PMO**	**PHYSICAL MEDIA OBJECT**	D					0	1
	MTP	media type	O↑ – convener - M↑ ?	U	1	300	none	0 (0 is for O 1 would be for M)	1
	RSP	recording speed	O↑	NS	1	9	0.9999999 < RSP < 999999999 value may include a decimal point or be an integer (no commas)	0	1

mment [45]: -MTP – this filed should be itionally mandatory. I agree with the green ener's note in lines 945 and following ones.

Field Number	Mnemonic	Content Description	Cond code	Character			Value Constraints	Occurrence	
				Type	Min #	Max #		Min #	Max #
	RSU	recording speed units	O↑ Convener - D ?	A	1	300	none	0	1
	EQ	equalization	O↑	AN	1	100	none	0	1
	TRK	tracks	O↑	N	1	2	1 ≤ TRK ≤ 99	0	1
	SPT	speaker track	O↑	NS	1	200	values between 1 and 99 inclusive that are separated by commas	0	99
	COM	comments	O↑	U	1	4000	none	0	1
	CDC	**CODEC**	D					0	1
	CDT	codec type	O↑ Convener M↑ ?	N	1	3	see external ~~**Tble of Codecs**~~	0 Convener: 1 ?	1
11.011	SRT	sampling rate	O↑	N	1	2	0 ≤ SRT < 100 expressed in MHz positive integer 0 = variable	0	1
	BIT	bit depth	O↑	N	1	2	0 ≤ BIT ≤ 60 positive integer 0 = variable	0	1
	NCH	number of channels	O↑	N	1	2	1 ≤ NCH ≤ 99	0	1
	COM	comment	D	U	1	4000	none	0	1
	PSQ	**PRELIMINARY SIGNAL QUALITY**	O					0	1
11.012		***Subfields: Repeating sets of information items***						1	9
	QVU	quality value	M↑	N	1	3	0 < QVU < 100 or QVU = 255 (quality not assessed) Integer	1	1

mment [46]: - TRK – correct typo in 'tracks'.

RK: corrected.

mment [48]: should be conditionally ndatory

mment [47]: -should contain a field for codec ate – for example, G726 allows for several bit-s.
011 should contain a field for **Endianedness** in of Linear PCM – I have seen so many errors mming from wrong order of bytes …

Field Number	Mnemonic	Content Description	Cond code	Character			Value Constraints	Occurrence	
				Type	Min #	Max #		Min #	Max #
	QAV	algorithm vendor identification	M↑	H	4	4	0000 ≤ QAV ≤ FFFF	1	1
	QAP	algoorithm product identification	M↑	N	1	5	0 < QAP < 65534 positive integer	1	1
	COM	comments	D	U	1	300	none	0	1
11.013--11.020		**RESERVED FOR FUTURE USE only by ANSI/NIST-ITL**							
11.021	**RED**	**REDACTION**	O					0	1
	RDI	redaction indicator	M↑	B	1	1	0=no 1=yes	1	1
	RDA	redaction authority	O↑	U	1	300	none	0	1
	COM	comment	O↑	U	1	4000	none	0	1
11.022	**RDD**	**REDACTION DIARY**	O					0	1
		Subfields: Repeating sets of information items	M↑					1	999
	RID	redaction identifier	M↑	N	1	3	1 ≤ RID ≤ 999	1	1
	RST	relative start time	M↑	N	1	14	1 < RST < 99999999999998	1	1
	RET	relative end time	M↑	N	1	14	99999999999999 ≥ RET > RST	1	1
	COM	Comment	O↑	U	1	4000	none	0	1
11.023	**SNP**	**SNIPPING SEGMENTATION**	O					0	1
	SGI	snipping indicator	M↑	N	1	1	0=no	1	1

Field Number	Mnemonic	Content Description	Cond code	Character			Value Constraints	Occurrence	
				Type	Min #	Max #		Min #	Max #
							1=yes		
	SPA	snipping authority	O↑	U	1	300	none	0	1
	COM	comment	O↑	U	1	4000	none	0	1
	SPD	**SNIPPING DIARY**	O					0	1
		Subfields: Repeating sets of information items						1	999
	SPI	snip identifier	M↑	N	1	3	1 ≤ SPI ≤ 999	1	1
11.024	RST	relative start time	M↑	N	1	14	99999999999998 ≥ RST ≥ 0	1	1
	RET	relative end time	M↑	N	1	14	99999999999999 ≥ RET > RST	1	1
	COM	comment	O↑	U	1	4000	none	1	1
	DIA	**DIARIZATION**	D					0	1
11.025	DII	diarization indicator	M↑	N	1	1	0=no 1=yes	1	1
	DAU	diarization authority	O↑	U	1	300	none	0	1
	COM	comment	O↑	U	1	4000	none	0	1
11.026	**SGD**	**SEGMENT DIARY**	D					0	1
	SID	segment identifier	M↑	N	1	3	1 ≤ SID ≤ 999	1	1
	TRK	track identifier	O↑	N	1	2	1 ≤ TRK ≤ 99	0	1
	RST	relative start time	M↑	N	1	14	99999999999998 ≥ RST ≥ 0	1	1

mment [49]: -11.024 and 11.026 (SPI, SID) – number of segments is currently limited to 999, h is **not enough**! In case there is a physical limit 27h of recording, there can be well more ments than 999 !

mment [50]: -11.024 and 11.026 (SPI, SID) – number of segments is currently limited to 999, h is **not enough**! In case there is a physical limit 27h of recording, there can be well more ments than 999 !

Field Number	Mnemonic	Content Description	Cond code	Character			Value Constraints	Occurrence	
				Type	MIn #	Max #		MIn #	Max #
	RET	relative end time	M↑	N	1	14	99999999999999 ≥ RET > RST	1	1
	COM	comment	O↑	U	1	10000	none	0	1
11.027 – 11.030		**RESERVED FOR FUTURE USE only by ANSI/NIST-ITL**							
11.031	**TME**	**TIME OF SEGMENT RECORDING**	D					0	1
		Subfield: repeating sets of information items	M↑					1	1998
	DIA	diary identifier	M↑	B	1	1	0=snip diary 1=segment diary	1	1
	SID	segment identifier	M↑	N	1	3	1 ≤ SID ≤ 999	1	1
	DOR	date of original recording	O↑	encoding specific: see **Annex B** or **Annex C**			encoding specific: see **Annex B** or **Annex C**	0	1
	TDT	tagged date	O↑	encoding specific: see **Annex B** or **Annex C**			encoding specific: see **Annex B** or **Annex C**	0	1
	SRT	start time of segment recording	O↑	encoding specific: see **Annex B** or **Annex C**			encoding specific: see **Annex B** or **Annex C**	0	1
	TST	tagged start time	O↑	encoding specific: see **Annex B** or **Annex C**			encoding specific: see **Annex B** or **Annex C**	0	1
	END	end time of segment recording	O↑	encoding specific: see **Annex B** or **Annex C**			encoding specific: see **Annex B** or **Annex C**	0	1
	TET	tagged end time	O↑	encoding specific: see **Annex B** or **Annex C**			encoding specific: see **Annex B** or **Annex C**	0	0

Field Number	Mnemonic	Content Description	Cond code	Character			Value Constraints	Occurrence	
				Type	MIn #	Max #		MIn #	Max #
	STM	source of time	O↑	U	1	300	none	0	1
	COM	comment	O↑	U	1	4000	none	0	1
	GEO	**SEGMENT GEOGRAPHICAL INFORMATION** (about person of interest at start of segment)	D					0	1
		Subfields: repeating sets of information items	M↑					1	1998
	DIA	diary identifier	M↑	N	1	1	0=snip diary 1=segment diary	1	1
	SID	segment identifiers	M↑	NS	1	3995	0 or a list of integers separated by commas	1	1
	SCT	segment cell phone tower code	O↑	U	1	100	none	0	1
11.032	LTD	latitude degree value	D	NS	1	9	-90 ≤ LTD ≤ 90	0	1
	LTM	latitude minute value	D	NS	1	8	0 ≤ LTM < 60	0	1
	LTS	latitude second value	D	NS	1	8	0 < LTS < 60	0	1
	LGD	longitude degree value	D	NS	1	10	-180 ≤ LGD ≤ 180	0	1
	LGM	longitude minute value	D	NS	1	8	0 ≤ LGM < 60	0	1
	LGS	longitude second value	D	N	1	2	0 ≤ LGS < 60 positive integer	0	1
	ELE	Elevation	O↑	NS	1	8	-422.000 < ELE < 8848.000 real number	0	1
	GDC	geodetic datum code	O↑	AN	3	6	value from Table 6	0	1

ment [51]: -032 SID – mismatch in max. – in case the max number is 999, the max should be coherent with this. T should be y higher, see above.

ment [52]: -what about is speech is ed in a coal mine that is deeper than 422m ? about recording in flight levels higher than Mt. t ?

K: Do we need a floor and ceiling?

Field Number	Mnemonic	Content Description	Cond code	Character			Value Constraints	Occurrence	
				Type	Min #	Max #		Min #	Max #
	GCM	geographic coordinate universal transverse mercator zone	D	AN	2	3	one or two integers followed by a single letter	0	1
	GCE	geographic coordinate universal transverse mercator easting	D	N	1	6	integer	0	1
	GCN	geographic coordinate universal transverse mercator northing	D	N	1	8	integer	0	1
	GRT	geographic reference text	O↑	U	1	150	none	0	1
	OSI	geographic coordinate other system identifier LANDMARK	O↑	U	1	10	none	0	1
	OCV	geographic coordinate other system value	D	U	1	126	none	0	1
	SQV	**SEGMENT QUALITY VALUES**	D					0	1
		Subfields: Repeating sets of information items	M↑					1	1998
	DIA	diary identifier	M↑	N	1	1	0=snip diary 1=segment diary	1	1
11.033	SID	segment identifiers	M↑	NS	1	3995	0 or a list of integers separated by commas	1	1
	QVU	quality value	M↑	N	1	3	0 ≤ QVU ≤ 100 or QVU = 255 (quality not assessed) Integer	1	1
	QAV	algorithm vendor identification	M ↑	H	4	4	0000 ≤ QAV ≤ FFFF	1	1
	QAP	algorithm product identification	M ↑	N	1	5	0 < QAP < 65534 positive integer	1	1

ment [53]: -11.033,034,035,036 SID – comment for max value as before.

Field Number	Mnemonic	Content Description	Cond code	Character			Value Constraints	Occurrence	
				Type	Min #	Max #		Min #	Max #
	COM	comment	D	U	1	300	none	1	1
11.034	**VCI**	**VOCAL COLLISION INDICATOR**	D					0	1
		Subfields: Repeating sets of information items						1	1998
	DIA	diary identifier	M↑	N	1	1	0=snip diary 1=segment diary	1	1
	SID	segment identifiers	M↑	NS	1	3995	0 or a list of integers separated by commas	1	1
11.035	**PPY**	**PROCESSING PRIORITY**	D					0	1
		Subfields: Repeating sets of information items						1	1998
	DIA	Diary identifier	M↑	N	1	1	0=snip diary 1=segment diary	1	1
	SID	segment identifiers	M↑	N	1	3995	0 or a list of integers separated by commas	1	1
	PTY	Priority	↑	N	1	1	1 ≤ PTY ≤ 9	1	1
11.036	**SCN**	**SEGMENT CONTENT**	D					0	1
		Subfields: Repeating sets of information items	M↑					0	1998
	DIA	diary identifier	M↑	N	1	1	0=snip diary 1= segment diary	1	1
	SID	segment identifiers	M↑	N	1	3995	0 or a list of integers separated by commas	0	1

ment [54]: -11.036 TRN – in case this filed ontain information for many (or even all) nts, the max value should be definitely higher 0000 characters.
6 should include more detailed information on nscription – automatics/manual, identification ware used, its configuration, name of person ribing, date, time, duration of transcription, (automatic, automatic manually corrected, manual), pointer to transcription guidelines, etc

Field Number	Mnemonic	Content Description	Cond code	Character			Value Constraints	Occurrence	
				Type	MIn #	Max #		MIn #	Max #
	TRN	transcript	O↑	U	1	10000	none	0	1
11.037	**SCT**	**SEGMENT SPEAKER CHARACTERISTIC**	D					0	1
		Subfields: Repeating sets of information items	M↑					1	1998
	DIA	diary identifier	M↑	N	1	1	0=snip diary 1=segment diary	1	1
	SID	segment identifiers	M↑	NS	1	3995	0 or a list of integers separated by commas	0	1
	IMP	impairment	O↑	N	1	1	$0 \leq IMP \leq 5$	0	1
	LBS	language being spoken	O↑	A	3	3	Value from *ISO 639-3*	0	1
	STY	style of speech	O↑	N	1	2	See **Table 11-3**	0	1
	INT	intelligibility	O↑	N	0	2	$0 \leq INT \leq 99$	0	1
	HST	health status	O↑	U	0	4000	none	0	1
	EM	emotional state	O↑	N	1	2	See **Table 11-4**	0	1
	VEF	vocal effort	O↑	N	1	1	$0 \leq VEF \leq 5$	0	1
	VSY	vocal style	O↑	N	1	2	See **Table 11-5**	0	1
	COM	comment	O↑	U	1	4000	none	0	1
11.038	**SCH**	**SEGMENT CHANNEL**	D					0	1
		Subfields: Repeating sets of information items	M↑					1	1000

Field Number	Mnemonic	Content Description	Cond code	Character			Value Constraints	Occurrence	
				Type	MIn#	Max#		MIn#	Max#
	DIA	diary identifier	M↑	N	1	1	0=snip diary 1=segment diary	1	1
	SID	segment identifiers	M↑	N	1	3995	0 or a list of integers separated by commas	0	1
	TYP	transducer type	O↑	N	1	2	See **Table 11-6**	0	1
	TRN	Transducer	O↑	N	1	1	unknown=0 carbon=1 electret=2 other=3	0	1
	ENV	capture environment	O↑	AN	1	4000	Text	0	1
	DST	distance to transducer	O↑	N	1	5	0 ≤ DST ≤ 99999	0	1
	ACS	acquisition source	O↑	N	1	2	See **Table 83**	0	1
	ALT	alteration	O↑	U	1	400	none	0	1
	COM	comment	O↑	U	1	4000	none	0	1
11.039-11.050		**RESERVED FOR FUTURE USE only by ANSI/NIST-ITL**							
11.051	COM	**COMMENT**	O↑	U	1	4000	none	0	1
11.052-11.099		**RESERVED FOR FUTURE USE only by ANSI/NIST-ITL**	Not to be used						
11.100-11.900	**UDF**	**USER-DEFINED FIELDS**	O	user-defined			user-defined		user-defined
11.901		**RESERVED FOR FUTURE USE only by ANSI/NIST-ITL**	Not to be used						
	ANN	**ANNOTATION INFORMATION**	O					0	1

mment [55]: -– I see this is reserved for e use, but still, if the fields are already there, is the difference between ANNOTATION and ANSCRIPTION ?

Field Number	Mnemonic	Content Description	Cond code	Character			Value Constraints	Occurrence	
				Type	Min #	Max #		Min #	Max #
11.902		***Subfields: Repeating sets of information items***	M↑					1	*
	GMT	Greenwich mean time	M↑	encoding specific: see **Annex B** or **Annex C**			encoding specific: see **Annex B** or **Annex C**	1	1
	NAV	processing algorithm name version	M↑	U	1	64	none	1	1
	OWN	algorithm owner	M↑	U	1	64	none	1	1
	PRO	process description	M↑	U	1	255	none	1	1
11.903	**DUI**	**DEVICE UNIQUE IDENTIFIER**	O	ANS	13	16	first character = M or P	0	1
11.904	**MMS**	**MAKE/MODEL/SERIAL NUMBER**	O					0	1
	MAK	make	M↑	U	1	50	none	1	1
	MOD	model	M↑	U	1	50	none	1	1
	SER	serial ~~N~~number	M↑	U	1	50	none	1	1
	COM	comment	O↑	U	1	*	none		
11.905-11.992		**RESERVED FOR FUTURE USE only by ANSI/NIST-ITL**	Not to be used						
11.993	**SAN**	**SOURCE AGENCY NAME**	O	U	1	125	none	0	1
11.994	**EFR**	**EXTERNAL FILE REFERENCE**	D	U	1	200	none	0	1
11.995	**ASC**	**ASSOCIATED CONTEXT**	O					0	1
		Subfields: Repeating sets of information items	M ↑					1	255

Comment [56]: -11.902 – algorithms for what ? I did not have the Annex, so I don't know what is in Section 7.4.1.

omment [57]: -11.903 and 11.904 – not clear, explained in the text.

Field Number	Mnemonic	Content Description	Cond code	Character			Value Constraints	Occurrence	
				Type	MIn #	Max #		MIn #	Max #
	CAN	associated context number	M↑	N	1	3	1 ≤ ACN ≤ 255 positive integer	1	1
	ASP	associated segment position	O↑	N	1	2	1 ≤ ASP ≤ 99 positive integer	0	1
11.996	**HAS**	**HASH**	O	H	64	64	None	0	1
11.997	**SOR**	**SOURCE REPRESENTATION**	O					0	1
		Subfields: Repeating sets of information items	M↑					1	255
	SRN	source representation number	M↑	N	1	3	1 ≤ SRN ≤ 255 positive integer	1	1
	RSP	reference segment position	O↑	N	1	2	1 ≤ RSP ≤ 99 positive integer	0	1
11.998		RESERVED FOR FUTURE USE only by ANSI/NIST-ITL	Not to be used						
11.999	**DATA**	**VOICE DATA**	D	B	1	22	None	0	1

1. **Field 11.001: Record header**

The content of this mandatory field is dependent upon the encoding used. See the relevant annex of this standard for details. See **Section 7.1**.

2. **Field 11.002: Information Designation Character / IDC**

This mandatory field shall contain the **IDC** assigned to this Type-11 record as listed in the information item **IDC** for this record in **Field 1.003 Transaction content / CNT.** See **Section 7.3.1.** Editor's note: This field will allow linking of the Type-11 records to the appropriate Type-2 records in the transaction.

3. **Field 11.003: Audio Object Descriptor/AOD**

This mandatory field shall be a numeric entry selected from the attribute code column of Table 1-1. Only one value is allowed and indicates the type of audio object containing the voice recording which is the focus of this Type-11 record. Attribute code 0 indicates that the audio object of this record is a digital voice data file in the Field 11.999. Attribute code 1 indicates that the audio object is a digital voice data file at the URL given in Field 11.994. Attribute codes 2-4 indicate that the audio object is a physical media object at a location described in Field 11.994.

Table 11-1
Audio Object Descriptor

Audio Object	Attribute Code
Internal digital voice data file	0
External digital voice data file	1
Physical Media Object containing digital data	2
Physical Media Object containing analog signals	3
Physical Media Object containing unknown data or signals	4

4. Field 11.004: Voice Laboratory Setting/VLS

This is an optional field and shall contain information about the site or agency that created the voice recording pointed to or included in this record. In the case of files created from previous recordings, the VLS is not necessarily the source of the original transduction of the acoustic vocalizations from the person to whom the Type-11record pertains. The VLS need not be the same as the Submitting Agency of Field 1.008, any agency mentioned in the corresponding type-2 record or the Source Agency Name of Field 11.993. Editor's note: This field is modeled after 18.003

- The first information item is the **lab type / LTY** is optional. When present, this information item contains a single character describing the site or agency that created the voice recording:

  G = Government
  I = Commercial
  P = Private individual
  O = Other
  U = Unknown

- The second information item (**name of original source/ NOO**) is optional and shall be the name of the group, organization or agency that created the voice recording. There may be no more than one occurrence for this item. This is an optional information item is in Unicode characters and is limited to 400 characters in length.

- The third information item is the **point of contact / POC** who composed the voice recording. This is an optional information item that could include the name, telephone

number and e-mail address of the person or persons responsible for the creation of the voice recording. This information item may be up to 200 Unicode characters.

- The fourth information item is optional. It is the *ISO-3166-1* **code of the sending country / CSC**. This is the code of where the voice recording was created – not necessarily the nation of the agency entered in **Field 11.993: Source agency / SRC** . All three formats specified in *ISO-3166-1* are allowed (Alpha2, Alpha3 and Numeric). A country code is either 2 or 3 characters long.

5. Field 11.005: Role of Voice Recording/ROL

This is a mandatory field and shall be a numeric entry selected from the "attribute code" column of **Table 11-2**. Only one value is allowed and indicates the role of the voice recording (known or questioned) within the transaction.

Table 11-2
Role of the Voice Recording

Role	Attribute Code
No information	0
Known sample, single speaker, subject of transaction	1
Known sample, single speaker, interlocutor	2
Known sample, single speaker, other	3
Known sample, multiple speakers, including subject of transaction	4
Known sample, multiple speakers, excluding subject of transaction	5
Questioned sample, single speaker	6
Questioned sample, multiple speakers	7
Audio recording with unknown voice content	8
Other	9
User-defined	10-99

Editor's note: In the FBI EBTS environment, mapping of the various speakers in a multiple-speaker environment to their descriptions in Type-2 records will be handled by modification of the Type-2 record.

Editor's note: Do we need a comment field to handle the case where "other" is chosen? Convener recommends against user-defined fields, since the same number may be used by multiple organizations and cause confusion. However, there is nothing that is 'wrong' with user-defined fields, per se. If used, there should be a statement such that the user-defined codes are clearly explained in the application profile (such as EBTS). Convener recommends using code 9 and adding another information item to this field as Comment. Convener does not believe that another field is needed to add the comment.

Comment [58]: -- additional entries **will** be needed, there are lots of acquisition sources not mentioned in the table.

6. Field 11.006: Recorder / REC

This field is mandatory and shall indicate information about the recording equipment that created the voice recording contained in or pointed to by this record.

NOTE: As recordings or data files may be transcoded from previously recorded or broadcast content, this equipment may or may not be the equipment used to record the original acoustic vocalization of the person to whom Type-11 record pertains.

- The first information item (**recorder type/RTP)** is an optional text field of up to 4000 characters describing the recording equipment that created the voice recording. An example would be "Home telephone answering device".

- The second, third and fourth information items (**recorder make/MAK, recorder model/MOD, recorder serialnumber/SER)** are optional items of up to 50 characters each and shall contain the make, model and serial number, respectively, for the recording device. There may be no more than one entry for this item. See Section 7.7.1.2 for details. Editor's note: Field 11.904 is also make/model/serial number. We need to discuss what various hardware/software items these refer to.

- The fifth information item (**acquisition source/AQS**) is a mandatory and is an alpha-numeric item. If all of the audio signal in the voice recording comes from a single acquisition source, the item shall be a numeric entry selected from the "attribute code" column of **Table 83** of the Type-20 record. When multiple sources are used for various voice segments in the voice recording, the code "MS" shall be used and individual sources will be given in the following comment item. If "12" from **Table 83** is chosen indicating an analog recording, then **Field 11.003** will indicate "3", the recording will be described in **Field 11.010**, and the location of the physical medium will be recorded in **Field 11.994**. Note that codes 1 through 6 and 11 from Table 83 are inapplicable, and shall not be used as a value in this information item.

- The sixth information item (**comments/COM**) is an optional text string of a maximum length of 4000 characters may contain any additional information about the recorder used to create the voice recording, including information about the recording software. If AQS indicates multiple sources, "MS", this field should be used to summarize the known sources from which the voice recording was created.

Table 83

Editor's note: Copied from the Type-20 record, included here for convenience. Additional entries may be needed to support a voice recording record type

Acquisition Source

Acquisition source type	Attribute code
Unspecified or unknown	0
Static digital image from an unknown source	1
Static digital image from a digital still-image camera	2
Static digital image from a scanner	3

Single video frame from an unknown source	4
Single video frame from an analog video camera	5
Single video frame from a digital video camera	6
Video sequence from an unknown source	7
Video sequence from an analog video camera, stored in analog format	8
Video sequence from an analog video camera, stored in digital format	9
Video sequence frame from a digital video camera	10
Computer screen image capture	11
Analog audio recording device; stored in analog form (such as a phonograph record)	12
Analog audio recording device; converted to digital	13
Digital audio recording device	14
Landline telephone – both sender and receiver	15
Mobile telephone – both sender and receiver	16
Satellite telephone – both sender and receiver	17
Telephone – unknown or mixed sources	18
Television – NSTC	19
Television – PAL	20
Television – Other	21
Voice-over-internet protocol (VOIP)	22
Radio transmission: short-wave (specify single side band or continuous wave in FDN)	23
Radio transmission: amateur radio (specify lower side band or continuous wave in FDN)	24
Radio transmission: FM (87.5 MHz to 108 MHz)	25
Radio transmission: long-wave (150 kHz to 519 kHz)	26
Radio transmission: AM (570 kHz to 1720 kHz)	27
Radio transmission: Aircraft frequencies	28
Radio transmission: Ship and coastal station frequencies	29
Vendor specific capture format	30
Other	31

7. Field 11.007: Record Creation Date/RCD

This mandatory field shall contain the date and time of creation of this Type-11 record. This date will generally be different from the voice recording creation date and may be different from the date at which the acoustic vocalization originally occurred. See **Section 7.7.2.4 Local date and time** for details.

8. Field 11.008: Voice Recording Creation Date/VRD

This optional field shall contain the date and time of creation of the voice recording contained in the record. If pre-recorded or transcoded materials were used, this date may be different from the date at which the acoustic vocalization originally occurred. See **Section 7.7.2.4 Local date and time** for details.

9. Field 11.009: Total Recording Duration / TRD

This field is optional and gives the total length of the voice recording in time, compressed bytes and total samples. At least one of the three information items must be entered if this field is used.

- The first information item (**time/TIM**) is optional and gives the total time of the voice recording in microseconds. The size of this item is limited to 14 digits, limiting the total time duration of the signal to 99,999 seconds, which is approximately 28 hours.

- The second information item (**compressed bytes/CBY**) is optional and gives the total number of compressed bytes in the voice data file. Consequently, this information item applies only to digital voice recordings stored as voice data files. The size of this item is limited to 14 digits, limiting the total size of the voice data file to 99 terabytes.

- The third information item (**total samples/TSM**) is optional and gives the total number of samples in the voice data file after any decompression of the compressed signal. This information item applies only to digital voice recordings stored as voice data files. The size of this item is limited to 14 digits, limiting the total number of samples to 99 $x10^{12}$ samples

Comment [59]: I don't get how are you going from 14 digits to 28 hours?

10. Field 11.010: Physical Media Object/ PMO

This field is optional and identifies the characteristics of the physical media containing the voice recording. There can be only one physical media object per Type-11 record, but multiple Type-11 records can point to the same physical media object. This field only applies if Field 11.003 has an attribute code of 2,3 or 4. The location of the physical media object is given in Field 11.994.

- The first information item (**media type/MTP**) is optional text of up to 300 characters and describes the general type of media (i.e., analog cassette tape, reel-to-reel tape, CD, DVD, phonograph record) upon which the voice recording is stored. Editor's note: Should this be a table? Convener's note: This information item should be mandatory if this field is chosen. Table and text have not been changed to reflect this viewpoint. Also recommend sentences: If an analog media is used for storage, and AQS is 14, then a description of the digital to analog procedure should be noted in **Field 11.902** and the reasons for such a conversion noted in COM of this field.The second information item (**recording speed/RSP**) is optional and gives a numerical value to the speed at which the physical media object must be played to reproduce the voice signal content. This value may be integer or floating point. Convener's note:

- The third information item (**recording speed units/RSU**) is optional text of up to 300 characters that indicates the units of measure to which the second information item (RSP)

refers. Convener believes that this information item should be mandatory if RSP is entered, making it Condition code D.

- The fourth information item (e**qualization/EQ**) is optional and indicates the equalization that should be applied for faithful rendering of the voice recording on the physical media object.

Editor's note: How do we appropriately characterize EQ? AES 57 contains the following text:

4.4.17.4.12 equalization

The **`equalization`** element shall be used to indicate any inherent equalization curve that must be applied to the described audio object or region during playback to properly recover the recorded sound. Where possible, this information should be given by its internationally recognized standard name. If no equalization is required for proper playback of the described audio object, then the **`equalization`** element shall be omitted.

- The fifth information item (t**racks/TRK**) is an optional integer between 1 and 99, inclusive, that gives the number of tracks on the physical media object. For example, a stereo phonograph record will have 2 tracks.

- The sixth information item (s**peaker track/STK**) is an optional list of integers which indicate which tracks carry the voices of the speaker(s). Note that the speaker(s) may be identified by a Type-2 record linked to this record by having the same IDC.

- The seventh information item (c**omment/COM**) is optional and allows for additional comments of up to 4000 Unicode characters in length describing the physical media object.

11. Field 11.011: Codec/CDC

This is an optional field that gives information about the Codec used to encode the voice data in the digital recording. This field is not used if the voice recording is stored on a physical media object as an analog signal. This field is only used if no header is read for the digital audio file when it is opened.

- The first information item (c**odec type/CDT)** is optional and indicates the single Codec type used for all audio segments in the record. This format does not accommodate multiple Codec types within a single record. It shall be a numeric entry selected from the "attribute code" column of the **Table of Codecs** that is available at http://www.nist.gov/itl/iad/ig/ansi_standard.cfm. These Codecs can be compressed (such as MP3) or uncompressed (such as WAV and AIF) file formats and are generally not open source, unlike OGG. For example, WAV, AIF and MP3 specifications are owned by MicroSoft, Apple, and the Fraunhofer Institute, respectively. OGG is a free, open container format maintained by the Xiph.Org Foundation. If the codec type is identified as "other" -- a value of 53, the fifth information item (**comment/COM**) shall be used to describe the codec.

  Editor's Note: This table is now an external reference to allow easy updating.
  Convener: This information item should be mandatory if this field is used.

1002
1003
1004

External Table of Codecs

Codec Type	Attribute Code
AAC	1
AC-3	2
AIF	3
AMR	4
AMR-WB	5
AMR-WB+	6
Apple Lossless	7
Asao	8
ATRAC	9
CELT	10
DRA	11
DTS	12
EVRC	13
EVRC-B	14
FLAC	15
GSM-EFR	16
GSM-FR	17
GSM-HR	18
HE-AAC	19
iLBC	20
iSAC	21
ITU-T G.711 (PCM)	22
ITU-T G.711.0 (PCM)	23
ITU-T G.711.1 (PCM)	24
ITU-T G.718	25
ITU-T G.719	26
ITU-T G.722	27
ITU-T G.722.1	28
ITU-T G.722.2	29
ITU-T G.723	30
ITU-T G.723.1 (ACELP)	31
ITU-T G.726 (ADPCM)	32
ITU-T G.728	33
ITU-T G.729	34
ITU-T G.729.1 (CS-ACELP)	35
Linear PCM	36
Monkey's Audio	37
MPEG (Surround)	38
MPEG-1 Layer I	39
MPEG-1 Layer II (multi-channel)	40
MPEG-1 Layer III (MP3)	41
MPEG-4 ALS	42

Comment [60]: AIFF?

MPEG-4 CELP	43
MPEG-4 DST	44
MPEG-4 HVXC	45
MPEG-4 SLS	46
MT9	47
Musepack	48
Non-streaming OGG Vorbis	49
OptimFROG	50
Opus	51
OSQ	52
Other	53
QCELP	54
RealAudio	55
RTAudio	56
SD2	57
SHN	58
SILK	59
Siren	60
SMV	61
Speex	62
SVOPC	63
TTA (True Audio)	64
TwinVQ	65
Unknown	66
USAC	67
VMR-WB	68
Vorbis	69
WavPack	70
WMA	71
WMA lossless	72
WMA Voice	73

Comment [61]: (LC) also known as Nellymoser audio codec

- The second item (**sampling rate/SRT**) indicates the number of digital samples that represent a second of analog voice data upon conversion to an acoustic signal. Acceptable values are between 0 and 100 MHz, but unknown sampling rates shall be given the value of 0. Editor's note: Re-sampling and changes to sampling rate should be logged in the Type-98 (which documents the entire ANSI/NIST transaction) or Field 11.902 (which documents only Type-11) audit logs. Each audio segment in the record is presumed to have the same sampling rate. Must we include ranges? There are oversampled 1-bit audio coders that, in theory, could be used for voice. For example, the SACD format uses 2.8224 MHz 1-bit sampling or 192 kHz 24-bit sampling. The DVD-Audio standard supports 192 kHz 24-bit coding Should there, if there isn't already, be a similar recommendation to indicate changes in coding? Of course, we want to encourage/require that the original format of the data be preserved, but in the process chain there could be a need for converting the coding – for example, a receiving agency's system might not be able to open files encoded with FLAC, which could be converted to LPCM that almost everything can open.

- The third item (**bit depth/BIT**) indicates the number of bits that are used to represent a single sample of voice data. Acceptable values are between 1 and 60, inclusive. Encoders of unknown or variable bit depth shall be given the value of 0. Nothing in this field is meant to be an indication of the dynamic range of the voice data. Changes to bit depth should be logged in Type-98 or **Field 11.902** audit logs.

- The fourth item (**number of channels/NCH**) is optional and gives the integer number of channels of data represented in the digital voice data file. The number of channels must be between 1 and 99, inclusive. If this item is not included, the voice data file will be assumed to have only one channel.

- The fifth item (**comment/COM**) is an optional unrestricted text string of up to 4000 characters in length that may contain additional information about the codec or reconstruction of audio output from the stored digital data. However, this information item shall be present if CDT = 53 (Other). This would include any noise reduction processing or equalization that must be applied to faithfully render the voice recording.

12. Field 11.012: Preliminary Signal Quality/PSQ

This field is optional and gives an assessment of the general "quality" of the voice recording. There may be as many as 9 PSQ fields for the audio file to indicate different types of quality assessments. Examples include signal-to-noise ratio and *Speech ~~Intelligibility~~Intelligibility Index, ANSI 3.5 1997*.

- The first information item (**quality value/QVU**) is mandatory and shall indicate the general quality value between ~~between~~ 0 (low quality) and 100 (high quality). A value of 255 indicates that quality was not assessed.

- A second information item is optional and shall specify the ID of the vendor of the quality algorithm used to calculate the quality score, which is an **algorithm vendor identification / QAV**. This 4-digit hex value (See **Section 5.5 Character types)** is assigned by IBIA and expressed as four characters. The IBIA maintains the Vendor Registry of CBEFF Biometric Organizations that map the value in this field to a registered organization. For algorithms not registered with the IBIA, the value of 0x00 shall be used.

- A third information item is optional and shall specify a numeric product code assigned by the vendor of the quality algorithm, which may be registered with the IBIA, but registration is not required. This is the **algorithm product identification / QAP** that indicates which of the vendor's algorithms was used in the calculation of the quality score. This information item contains the integer product code and should be within the range 1 to 65,534. For products not registered with the IBIA, the code 0 shall be used.

- The fourth information item (**comment/COM**) is an optional and should be used to give additional information about the quality assessment process. It shall be used to describe unregistered algorithms.

13. Fields 11.013-020: Reserved Fields

These fields are reserved for future use by ANSI/NIST-ITL.

14. Field 11.021: Redaction/ RED

This field is optional and indicates whether the voice recording has been redacted, meaning that some of the audio record has been overwritten ("Beeped") or erased to delete speech content without altering the relative timings within, or the length of, the segments. This field is not to be used to indicate that audio content has been snipped with the alteration of the relative timings in or length of the segment.

- The first information item (**redaction indicator/RDI)** is a binary variable and is mandatory if this field is used. It indicates whether the voice recording contains overwritten or erased sections intended to remove, without altering the length of the segment, semantic content deemed not suitable for transmission or storage. 0 indicates no redaction and 1 indicates that redaction has occurred.

- The second information item (**segmentation redaction authority/SGA)** is an optional text field of up to 300 characters in length containing information about the agency that directed, authorized or performed the redaction. Agencies undertaking redaction activities on the original speech should log their actions by appending to this item and noting the change of field contents in the Type-98 record and / or **Field 11.902** of this record

Comment [62]: –– I would use 'quality assessment algorithm' rather than 'quality algorithm'. The later could be interpreted in a way that it improves the quality which is certainly not true. Please check also other occurrences.

MARK: I do not have an issue with this.

Comment [63]: What is IBIA ? Expand and add to the list of abbreviations.

MARK: We should do this.

- The third information item (**comment/COM)** is an optional unrestricted text string of up to 4000 characters in length that may contain text information about the redactions affecting the stored voice data.

15. Field 11.022: Redaction Diary/RDD

This optional field (**redaction diary/RDD**) indicates the timings with the voice recording of redacted (overwritten) audio segments. The redactions need not be dominated by speech from the subjectof this transaction or record.. Three items (uniquely numbering the redactions and giving relative start and end times of each) are mandatory if this field is used and shall repeat for each redaction. A fourth item is optional and accommodates comments on the individual redactions. The record type accommodates up to 1000 redactions.

- The first item (**redaction identifier/RID**) is mandatory if this field is used and uniquely numbers the redactions to which the following items in the field apply. There is no requirement that the redactions be numbered sequentially. The **RID** may contain up to 3 digits, meaning that the number of redactions that may be identified is limited to 999.

- The second item (**relative start time/RST**) is a mandatory integer for every redaction identified by an **RID** and indicates in microseconds the time of the start of the redaction relative to the beginning of the voice recording. The item can contain up to 14 digits, meaning that the start of a redaction might occur anywhere within a voice recording limited to about 27 hours. It is not expected that redactions will overlap, meaning that the RST of a redaction is not expected to occur between the **RST** and **RET** of any other redaction, although this is not prohibited.

- The third item **(relative end time/RET**) is a mandatory integer for every redaction identified by an **RID** and indicates in microseconds the time of the end of the redaction relative to the beginning of the voice recording. The item can contain up to 14 digits, meaning that the end of a redaction might occur anywhere within a voice recording limited to about 27 hours As with the **RST**, it is not expected that redactions will overlap, although this is not prohibited.

- The fourth item (**comment/COM**) is an optional unrestricted text string of up to 4000 characters in length that allows for comments of any type to be made on a redaction.

16. Field 11.023: Snipping Segmentation/ SNP

This field is optional and indicates whether the voice recording referenced in this Type-11 record has had segments removed or contains segments that have been snipped from one or more longer voice recordings, in either case meaning that the voice signal is not a continuous recording in time. This field is used to indicate removal, for any reason, of audio signal from the original recording of the acoustic vocalizations in a way that disrupts time references.

- The first information item **(snip indicator/SGI)** is a binary variable and is mandatory if this field is used. It indicates whether the voice recording contains temporal discontinuities caused by snipping of segments from one or more longer recordings. 0 indicates no snipping and 1 indicates that snipping has occurred.

- The second information item (**snipping authority/SPA)** is an optional text field of up to 300 characters containing information about the agency that performed the snipping segmentation. Agencies undertaking snipping activities on the original speech should log their actions by appending to this item and noting the change of field contents in the Type-98 record and / or **Field 11.902** of this record.

- The third information item (**comment/COM)** is an optional unrestricted text string of up to 4000 characters that may contain text information about the snip activities affecting the voice recording.

17. Field 11.024: Snipping Diary/SPD

This optional field (**snip diary/SPD**) allows this type to document the snips obtained from larger voice recordings, which might themselves be included in the transaction as Type-20 records. Each snip diarized shall be dominated by speech from the subject of this Type-11 record, who may be described in a Type-2 record within the transaction with the same IDC as this record. Three items (uniquely numbering the snips and giving relative start and end times of each) are mandatory if this field is used and shall repeat for each snip identified. A fourth item is optional and allows for comments. The record type accommodates up to 999 snips. If there is no snipping (**Field 11.023**) indicated, then all of the data in the voice recording will be considered as in a single snip and the subfields will not repeat. There can be at most one snipping diary for each Type-11 record.

Comment [64]: 999 snips too few (see above)

- The first item (**snip identifier/SPI**) is mandatory if this field is used and uniquely numbers the snip to which the following items in the field apply. There is no requirement that the snips be numbered sequentially. The **SPI** may contain up to 3 digits, meaning that 999 snips may be identified. If **Field 11.023** indicates snipping, the voice recording must consist of at least one snip.

- The second item (**relative start time/RST**) is a mandatory integer for every snip identified by an **SPI** and indicates in microseconds the time of the start of the snip relative to the beginning of the voice recording. The item can contain up to 14 digits, meaning that the **RST** might occur anywhere within a voice recording limited to about 27 hours. Because each snip is obtained independently from a larger voice recording, snips shall not overlap, meaning that the **RST** of a snip shall not occur between the **RST** and **RET** of any other snip.

Comment [65]: As opposed to 28 hours quoted before?

- The third item **(relative end time/RET**) is a mandatory integer for every snip identified by an **SPI** and indicates in microseconds the time of the end of the snip relative to the

beginning of the voice recording. The item can contain up to 14 digits, meaning that the snip may end anywhere within the 27 hour voice recording. Because each snip is obtained independently from a larger voice recording, snips shall not overlap, meaning that the **RET** of a snip shall not occur between the **RST** and **RET** of any other snip.

- The fourth item **(comment/COM)** is an optional unrestricted text string of up to 4000 characters in length that allows for comments of any type to be made on a snip. This is allows for comments on a snip-by-snip basis, including comments on the source of each snip. This comment field could contain word- or phone-level transcriptions, language translations or security classification markings, as specified in exchange agreements.

Comment [66]: As opposed to 28 hours quoted before?

Comment [67]: I agree this should be included to benefit the receiving agency, especially if the segment is embedded in a large voice recording.

18. Field 11.025: Diarization/DIA

This field is optional and indicates whether the voice recording has been diarized, meaning that time markings are included in **Field 11.026** to indicate the speech segments of interest pertaining to the subject of this Type-11 record. Therefore, if this field is present, then **Field 11.026** shall also be present in the record.

- The first information item (**diarization indicator/DII)** is mandatory if this field is used. It is a binary variable that indicates whether the voice recording is accompanied a segment diaries in **Field 11.026** indicating speech segments from the voice signal subject of the Type-11 record. 0 indicates no accompanying diary and 1 indicates one or more accompanying diaries.

- The second information item (**diarization authority/DAU)** is an optional text field of up to 300 characters containing information about the agency that performed the diarization. Agencies undertaking diarization activities on the original speech should log their actions by appending to this item and noting the change of field contents in the Type-98 record and / or **Field 11.902** of this record
- The third information item (**comment/COM) is an optional** unrestricted text string of up to 4000 characters that may contain text information about the diarization activities undertaken on the voice data.

19. Field 11.026: Segment Diary/SGD

field only appears if **Field 11.025** is present and DII = 1. This field (**segment diary/SDI**) names and locates the segments within the voice recording of this Type-11 record associated with a single speaker. Within a Type-11 record, there may be only one segment diary describing a single speaker within the single voice recording. If additional diarizations of this voice recording are necessary -- for example, to locate segments of speech from a second speaker in the voice recording, additional Type-11 records must be created. Each segment diarized shall be dominated by speech from the subject of this record. The subject may be described in the Type-2 record with the same IDC value as this record. The first three items (uniquely numbering the segments and giving start and end times of each relative to the absolute beginning of the voice

recording) are mandatory if this field is used and shall repeat for each speech segment identified. A fourth item is optional and accommodates comments on the individual segments. This record type accommodates up to 999 speech segments. For voice recordings consisting of snips, the **SPD** may be included in the **SGD** as a subset and may be identical.

- The first item (**segment identifier/SID**) is mandatory if this field is used and uniquely numbers the segment to which the following items in the field apply. There is no requirement that the segments be numbered sequentially. The **SID** may contain up to 3 digits, meaning that the number of segments identified is limited to 999.

- The second item (**track identifier/TRK)** is optional and indicates the track or channel of a multichannel voice recording upon which this segment is found. The number of tracks or channels on the recording is limited to 99, so this item may take any value between 1 and 99, inclusive.

- The third item (**relative start time/RST**) is a mandatory integer for every segment and indicates in microseconds the time of the start of the segment relative to the absolute beginning of the voice recording. The item can contain up to 14 digits, meaning that the segment can start at any time within the 27 hour voice recording. Because each segment is expected to be dominated by the primary subject of this Type-11 record, it is not expected that segments will overlap, meaning that the RST of a segment is not expected to occur earlier than the end of a previous segment, although this is not prohibited. In multiple transactions involving multiple speakers using the same voice data record, segments across the transactions may overlap during periods of voice collision. Editor's note: Neither "Start of the segment" nor "End of the segment" has been defined. Should they be?

- The fourth item **(relative end time/RET**) is mandatory for every segment and indicates in microseconds the time of the end of the segment relative to the absolute beginning of the voice recording. The item can contain up to 14 digits, meaning that the segment can end at any time within the 27 hour voice recording. As with the RST, it is expected that segments from the subject of this Type-11 record will not overlap, although this is not prohibited.

- The fifth item **(comment/COM**) is an optional unrestricted text string of a maximum of 10,000 characters in length that allows for comments of any type to be made on a segment. This comment field could contain word- or phone-level transcriptions, language translations or security classification markings, as specified in exchange agreements.

20. Field 11.027-030: Reserved Fields

These fields are reserved for future use by ANSI/NIST-ITL.

21. Field 11.031: Time of Segment Recording /TME

Comment [68]: You seem to categorically want to exclude **segments with overlapping speech**. For many methods and features in speaker identification, overlapping speech is indeed harmful (e.g. automatic speaker recognition) and should be avoided. However, there are other methods and features that are more tolerant in this regard. It might, for example, be possible to identify dialectal features from a portion of speech of the relevant speaker when another, irrelevant, speaker is talking at the same time. Perhaps if the speech material is long enough one can afford to ignore overlapping portions anyway, but in case the speech material is short or the amount of overlapping speech is massive, one should retain the option of including segments with overlap for analysis.

Comment [69]: 999 too few (see above)

Comment [70]: Be consistent, check throughout

This optional field (**Time of Segment Recording/TME**) refers to a segment identified in either the snip diary **SPD** of **Field 11.024** or the segment diary **SGD** of **Field 11.026** and gives the date, start, and end times of the original transduction of the contemporaneous vocalizations in the identified segment. This field is only present if **Field 11.024** or **Field 11.026** is present in this record. This field also accommodates circumstances in which the original voice signal recording was tagged with a time and date field. There is no requirement that the date and times for the original recording match the dates and times of the tags, if the tags have been determined to be inaccurate.

- The first item (**diary identifier/DIA**) is a mandatory and indicates the diary to which this field refers. If this item refers to a segment in the **SPD** of **Field 11.024**, the value is 0. If this item refers to a segment in the **SGD** of Field 11.026, the value is 1.

- The second item **(segment identifier/SID**) is mandatory and gives the segment identifier from the diary given in the first item above to which the values in this field pertain.

- The third item **(date of original recording/DOR**) is optional and gives the date of the original, contemporaneous capture of the voice data in the segment identified in the first item of this field. See **Section 7.7.2.3.**

- The fourth item **(tagged date/TDT**) is optional and gives the date tagged on the original, contemporaneous capture of the voice data in the segment identified in the first item of this field. This item may be different from the value of the **DAT** above, if the tag is determined to be inaccurate. See **Section 7.7.2.3.**

- The fifth item (**start time of recording/SRT**) is optional and gives the local start time of the original, contemporaneous capture of the voice data in the segment identified in the first item of this field. See **Section 7.7.2.4 Local date and time** for details.

- The sixth item (**tagged start time/TST**) is optional and gives the time tagged on original, contemporaneous capture of the voice data at the start of the segment identified in the first item of this field. This item may be different from the value of the **RST** above, if the tag is determined to be inaccurate See **Section 7.7.2.4 Local date and time** for details.

- The seventh item (**end time of recording/END**) is optional and gives the local end time of the original, contemporaneous capture of the voice data in the segment identified in the first item of this field. See **Section 7.7.2.4 Local date and time** for details.

- The eight item (**tagged end time/TET**) is optional and gives the time tagged on original, contemporaneous capture of the voice data at the end of the segment identified in the first item of this field. This item may be different from the value of the **END** above, if the tag is determined to be inaccurate. See **Section 7.7.2.4 Local date and time** for details.

- The ninth item (**Source of the time/STM**) is an optional character string that gives the reference for the values used for **TDT**, **SRT** and **END**.

- The tenth item (**comment/COM**) is an unrestricted text string of up to 4000 characters in length that allows for comments of any type to be made on the timings of the segment recording, including the perceived accuracy of the values of **TDT**, **SRT** and **END**.

22. Field 11.032: Segment Geographical Information/GEO

This field (**Segment Geographical Information/GEO**) refers to a segment identified in either the snip diary **SPD** of **Field 11.024** or the segment diary **SGD** of **Field 11.026** and gives geographical location of the primary subject of the Type-11record at the beginning of that segment.. This field is only present if **Field 11.024** or **Field 11.026** is present in this record.

- The first item (**diary identifier/DIA**) is a mandatory and indicates the diary to which this field refers. If this item refers to a segment in the **SPD** of **Field 11.024**, the value is 0. If this item refers to a segment in the **SGD** of Field **11.026**, the value is 1.

  The second item (**segment identifier/SID**) is mandatory and gives the segment identifiers from diary to which the values in this field pertain. The number of segment identifiers listed is limited to 999. A value of 0 in this field indicates the segment geographical information in this field shall be considered the default value for all segments not specifically identified in other occurrences of this field. If multiple segments are identified, they are designated as integers separated by commas.
- The third item (**segment cell phone tower code/SCT)** is optional and identifies the cell phone tower, if any, that relayed the audio data at the start of the segment. It is a text field.

- The next six items are latitude and longitude values. **See Section 7.7.3**

- The tenth information item (**elevation / ELE)** is optional. . It is expressed in meters. **See Section 7.7.3**

- The eleventh information item (**geodetic datum code / GDC)** is optional. **See Section 7.7.3.**

- The twelfth, thirteenth and fourteenth information items (**GCM**/**GCE**/**GCN**) are treated as a group and are optional. These three information items together are a coordinate which represents a location with a Universal Transverse Mercator (**UTM**) coordinate. If any of these three information items is present, all shall be present. **See Section 7.7.3**

- The fifteenth information item (**geographic reference text /GRT)** is optional. **See Section 7.7.3**

- A sixteenth information item (**geographic coordinate other system identifier / OSI)** is optional and allows for other coordinate systems and the inclusion of geographic landmarks. **See Section 7.7.3**

- A seventeenth information item (**geographic coordinate other system value / OCV)** is optional and shall only be present if **OSI** is present in the record**. See Section 7.7.3**

23. Field 11.033: Segment Quality Values/SQV

This field (**Segment Quality Values/SQV**) refers to a list of segments identified in either the snip diary **SPD** of **Field 11.024** or the segment diary **SGD** of **Field 11.026** and gives an assessment of the quality of the voice data within the segment. It only is present if Field 11.024 or Field 11.026 exists in the record. This contrasts with **Field 11.012** that gives the general quality across the entire audio recording. Values in this field dominate any values given in **Field 11.012**. It is possible for each segment given in the associated diary to have different quality. This field only accommodates a single quality value. If segments have multiple quality values based on different types of quality assessments, then multiple subfields ~~–~~are entered for those segments.

- The first item (**diary identifier/DIA**) is a mandatory and indicates the diary to which this field refers. If this item refers to a segment in the **SPD** of **Field 11.024**,~~-~~ the value is 0. If this item refers to a segment in the **SGD** of **Field 11.026**, the value is 1.

  The second item (**segment identifier/SID**) is a mandatory list of integers and gives the segment identifiers from diary to which the values in this field pertain. The number of segment identifiers listed is limited to 999. A value of 0 in this field indicates the segment quality information in this field shall be considered the default value for all segments not specifically identified in other occurrences of this field. If multiple segments are entered, they are listed as integers separated by commas.
- The third **information item (quality value/QVU**) is mandatory and shall indicate the segment quality value between 0 (low quality) and 100 (high quality). A value of 255 indicates that quality was not assessed. An example would be the *Speech Intelligibillity Index, ANSI 3.5 1997.*

- A fourth information item is mandatory and shall specify the ID of the vendor of the quality algorithm used to calculate the quality score, which is an **algorithm vendor identification / QAV**. This 4-digit hex value (See **Section 5.5 Character types)** is assigned by IBIA and expressed as four characters. The IBIA maintains the Vendor Registry of CBEFF Biometric Organizations that map the value in this field to a registered organization. A value of 0000 indicates a vendor without a designation by IBIA. In such case, ~~–~~an entry shall be made in COM of this field describing the ~~algorith~~algorithm and its owner / vendor.

- A fifth information item is mandatory and shall specify a numeric product code assigned by the vendor of the quality algorithm, which may be registered with the IBIA, but registration is not required. This is the **algorithm product identification / QAP** that indicates which of the vendor's algorithms was used in the calculation of the quality score. This information item contains the integer product code and should be within the range 0 to 65,534. A value of 0 indicates a vendor without a designation by IBIA. In

Comment [71]: quality algorithm -> quality assessment algorithm

MARK: Again, no issue for me to make this change.

such case, -an entry shall be made in COM of this field describing the ~~algorith~~algorithm and its owner / vendor.

- The sixth information item (**comment/COM**) is an optional but shall be used to provide information about the quality assessment process, including a description of any unregistered quality assessment algorithms used. (if QAV= 0000 or QAP = 0)

24. Field 11.034: Vocal Collision Indicator/VCI

This optional field (**Vocal Collision Indicator/VCI**) refers to a list of segments identified in either the snip diary **SPD** of **Field 11.024** or the segment diary **SGD** of **Field 11.026** and indicates that a vocal collision (two or more persons talking at once) occurs within the segment. This field shall only appear if **Field 11.024** or **Field 11.026** exists in this record.

- The first item (**diary identifier/DIA**) is a mandatory and indicates the diary to which this field refers. If this item refers to a segment in the **SPD** of **Field 11.024**, the value is 0. If this item refers to a segment in the **SGD** of **Field 11.0**26, the value is 1.

- The second item (**segment identifier/SID**) is a mandatory list of integers separated by commas and gives the segment identifiers from the diary named in the item above in which vocal collisions occur. There may be up to 999 segments identified in this field.

25. Field 11.035: Processing Priority /PPY

This optional field (**Processing Priority/PPY**) refers to a list of segments identified in either the snip diary **SPD** of **Field 11.024** or the segment diary **SGD** of **Field 11.026** and indicates the priority with which the segments named in those diaries should be processed. There may be two occurrences of this field: one for the segments identified in the snip diary, **Field 11.024**, and one for segments identified in the segment diary, **Field 11.026**. If this field exists, segments not identified should be given the lowest priority. This field is distinct from **Field 1.006**, which givesa priority for processing the entire transaction.

- The first item (**diary identifier/DIA**) is a mandatory and indicates the diary to which this field refers. If this item refers to a segment in the **SPD** of **Field 11.024**, the value is 0. If this item refers to a segment in the **SGD** of **Field 11.026**, the value is 1.

- The second item (**segment identifier/SID**) is a mandatory list of integers, separated by commas, and gives the segment identifiers from diary named in the first item above to which the values in this field pertain. There may be up to 999 values of this field, one for each segment identified in the diaries of **Field 11.024** or **Field 11.026**. A value of 0 in this field indicates the segment content information in this field shall be considered the default value for all segments not specifically identified in other occurrences of this field.

- The third information item (**processing priority/ PPY**) is optional and indicates the priority with which the segments identified in the second item should be processed. Priority values shall be between 1 and 9 inclusive. As with **Field 1.006**, 1 will indicate the highest priority and 9 the lowest.

26. Field 11.036: Segment Content/SCN

This optional field (**Segment Content/SCN**) refers to a segment identified in either the snip diary **SPD** of **Field 11.024** or the segment diary **SGD** of **Field 11.026** and gives an assessment of the content of the voice data within the segment and includes provision for semantic transcripts, phonetic transcriptions and translations of the segment. It may only appear if Field 11.024 or Field 11.026 is present in this record. There may be an up to 9 occurrences of the -subfield, one for each segment identified in related diary.

- The first item (**diary identifier/DIA**) is a mandatory and indicates the diary to which this field refers. If this item refers to a segment in the **SPD** of **Field 11.024**, the value is 0. If this item refers to a segment in the **SGD** of **Field 11.026**, the value is 1.

- The second item (**segment identifier/SID**) is a mandatory list of integers separated by commas and gives the segment identifiers from diary to which the values in this field pertain. There may be ~~an~~ 999 values of this field, one for each segment identified in related diary. A value of 0 in this field indicates the segment content information in this field shall be considered the default value for all segments not specifically identified in other occurrences of this field.

- The fourth information item (**transcript/TRN**) is an optional text field of up to 10,000 characters and may contain a semantic transcription, a phonetic transcription, translation, or comments on the segment. The text field should state the authority providing the transcription, translation or comments.

Editor's Note: We must fix the number of transcripts in each Field 11.036 Convener's note: If more than one transcript, a separate information item would have to be added for each. This is awkward, but doable.

27. Field 11.037: Segment Speaker Characteristics/SCC

This optional field (**Segment Speech Characteristics/SCT**) refers to a segment identified in either the snip diary **SPD** of **Field 11.024** or the segment diary **SGD** of **Field 11.026** and gives an assessment of the characteristics of the voice within the segment, including intelligibility, emotional state and impairment. This field shall only appear if Field 11.024 or Field 11.026 exists in the record.

- The first item (**diary identifier/DIA**) is mandatory and indicates the diary to which this field refers. If this item refers to a segment in the **SPD** of **Field 11.024**, the value is 0. If this item refers to a segment in the **SGD** of **Field 11.026**, the value is 1.

Comment [72]: Under the heading SCC only the temporary speaker attributes are included. Are the Permanent Attributes mentioned on p. 13f. coded elsewhere?

Comment [73]: Page 50-52: I have one general point about the Segment Speaker Characteristics as well as the permanent attributes mentioned on p. 13). Is perceptual rating by the person who makes the entry the only source of an entry or are other sources used as well? For example, are educational level or speech impediment inferred based on the perception of the speech sample or are they looked up from external sources, such as a CV or a medical file? Is there any other possible source for the entries? Perhaps there should be a place where it can be specified what the source of the entry is: is it due to the listening action of the person who makes the entry or did the person who makes the entry use a different source than his/her own auditory judgment? As another general point about the Segment Speaker Characteristics (but not including the permanent attributes), I am thinking about whether it would be useful to include a category **stress** as another information item (there is a recent survey by Kirchhübel et al. 2011). I am not sure, but excluding a category stress does make sense on a practical level because stress often goes hand in hand with other phenomena, in particular emotion (stress is often associated with fear and anger) and vocal effort (stress often leads to loud speech, especially in "real", threatening stress). It will therefore be difficult to perceive stress independently of fear/anger or increased vocal effort. What is left missing, though, if stress is not included, are certain physical stressors such as exposure to low temperature, high altitudes, acceleration, vibration etc. or task-related stress (physical or cognitive activity), e.g. talking while running, carrying something heavy or being involved with a cognitively demanding task. One way to capture those could be to subsume them under the IMP category. For example, if you accept my three-item list from above (effects of alcohol, effects of legal or illegal drugs/medication, fatigue/tiredness) you could add two further ones, such as "speaking under exposure to extreme environmental conditions" and "speaking during heavy physical or cognitive activity".

Just as an additional comment in this context: I am glad you included "**vocal effort**" as a separate item. In our experience this is one of the most relevant and frequently occurring factors. Most importantly, we include fundamental frequency in our set of measurements and even slight amounts of vocal effort increase lead to a rise in f0. But informal tests have shown that it also affects automatic speaker recognition. I also agree that it is important to scale the degree of vocal effort.

- The second item (**segment identifier/SID**) is a mandatory list of integers separated by commas and gives the segment identifiers from **Field 11.024** to which the values in this field pertain. There may be up to 999 values in this field, one for each segment identified in **Field 11.026**. A value of 0 in this field indicates the segment content information in this field shall be considered the default value for all segments not specifically identified in other occurrences of this field.
- The third information item (**impairment/IMP**) is optional and shall indicate an observed level of neurological diminishment, whether from disease, trauma, medication or substance abuse, across the speech segment. No attempt is made to differentiate the sources of impairment**.** The value shall be an integer between 0 (no noticed impairment) and 5 (assessed as life threatening), inclusive.

- The fourth item (**language being spoken/LBS**) is optional and gives the 3 character *ISO 639-3* code for the dominant language in the segments identified in the first item above.

- The fifth information item **(style of speech/STY**) is optional and shall be an integer as given in **Table 11-3**. There may only be one value for each identified segment and will indicate the dominant style of speech within the segment. If attribute code "8" is chosen to indicate "other", additional explanation should be included in the tenth item (**comment/COM**) below.

Table 11-3
Style of Speech

Style of Speech	Attribute Code
Unknown	0
Public speech (oratory)	1
Conversation	2
Familiar/intimate	3
Read	4
Prompted	5
Interview	6
Recited/memorized	7
Other	8
RESERVED FOR FUTURE USE only by ANSI/NIST-ITL	9-20

- The sixth information item (**intelligibility/INT**)

- The seventh information item **(health status/HST**) is optional text noting any observable health issues impacting the data subject during the speech segment.

- The eighth information item **(emotional state/EM**) is an optional integer giving an estimation of the emotional state of the data subject across the segment. Admissible attribute values are given in **Table 11-4**. Only one value for this item is allowed across

Comment [74]: The information item **IMP** (impairment) presents an interesting concept I haven't thought about this way yet. I see what you mean: there is an overall sense of physical impairment or weakness that can be inferred from the speech and yet it is difficult to determine its causes (alcohol, disease etc.). Because of this difficulty, you propose that no attempts shall be made to differentiate the qualitative sources of the impairment, and instead you offer a quantitative scale on which the perceivable impairment can be rated in terms of its severity. You are classifying those impairments as neurological diminishments, which again is an interesting idea. You are probably right there, however in lack of deeper knowledge of the underlying medical causes I would not exclude the possibility that other than neurological factors are involved as well (perhaps factors such as relaxation of muscle tonus, change in heart rate or blood pressure, or effects on the endocrinological system). When you recommend not to determine the sources of a perceivable impairment you do have a point. I am just thinking about well-known situations where for example, somebody who asked for help and sounded garbled was refused help because people thought he was drunk where in fact he had a severe low blood sugar episode or was in other ways sick. Still, your concept of avoiding qualitative classification differs from the one we use in our lab. We try to differentiate at least the following types:
- effects of alcohol
- effects of legal or illegal drugs/medication ...

Comment [75]: Under the influence of or abuse

Comment [USS76]: Exactly how does a life threatening impairment manifest itself in the speech signal? How about 0 = no noticed impairment to 5 = significant impairment.

Comment [77]: In the style of speech (**STY**) table, I recommend adding an entry "**Repeated**" in Table 11-3. Repeating along with or as an alternative to reading is a useful way to obtain text-identical material from the suspect. (We always also elicit spontaneous speech from the suspect where naturally there is no control over the text, but obtaining read/repeated speech with material that is text-identical to the speech of the questioned speaker ...

Comment [USS78]: There is also a significant overarching variable of "whether the person knows they are being recorded or not". This factor plays a huge role in speech style, etc.

Comment [USS79]: There are actually significant differences between conversational telephone speech and face-to-face conversation – might want to differentiate. Also familiar/intimate should probably be a layer separate from this heading.

Comment [80]: The category that belongs under the information item **HST** (health status) that occurs most frequently in our casework is **symptoms of the common cold** (e.g. hoarse voice, pitch lowering, increased nasality). Perhaps if you have the same experience, you could mention it in the text. There are of course many more health-related influences on speech (other than those addressed under the IMP ...

the segment. If attribute code "8" is chosen to indicate "other", additional explanation may be included in the tenth item (**comment/COM**) below.

Table 11-4
Emotional State

Emotional State	Attribute Code
Unknown	0
Calm	1
Hurried	2
Angry	3
Fearful	4
Agitated /Combative	5
Defensive	6
Crying	7
Other	8
RESERVED FOR FUTURE USE only by ANSI/NIST-ITL	9-20

- The ninth information item **(vocal effort/VEF**) is an optional integer between 0 (no effort or not talking) and 5(screaming/crying) which reports perceived vocal effort across the segment. Only one value is allowed for this item across the segment.

- The tenth information item (**vocal style/VSY**) is an optional integer assessing the predominant vocal style across the segment. The attribute value shall be chosen from **Table 11-5**. Only one value is allowed for this item across the segment.

Table 11-5
Vocal Style

Vocal Style	Attribute Code
Unknown	0
Spoken	1
Whispered	2
Sung	3
Chanted	4
Rapped	5
Mantra	6
Falsetto/Head voice	7
Megaphone/Public Address System	8
Other	9
RESERVED FOR FUTURE USE only by ANSI/NIST-ITL	10-20

- The eleventh information item (**comment/COM**) is optional and may be used to give additional information about the quality assessment process, including a description of any unregistered quality assessment algorithms used.

Comment [81]: -VEF specifies only values from normal to high vocal effort. Do you also want to provide values for low vocal effort ? The values could then run from -5 (LVE) thru 0 (NVE) to 5 (HVE)

Comment [82]: VSY (and EM): In the vocal style (**VSY**) table, I recommend including something involving **laughter**. There are many kinds of laughter, and laughter can be integrated with or isolated from speech to various degrees. There are also different levels of control over laughter; laughter can be a more or less direct emotional expression, but it can also be used as some form of conversational strategy. Perhaps one single entry such as "laughter (isolated or integrated in speech)" or "spoken with laughter" could do the job. I noticed that in the item **EM** (emotional state) you only have negative (or neutral) emotions. Perhaps another way to incorporate laughter is by including an entry such as "happy/joyful/laughing" in Table 11-4. Laughing, however is not always associated with joyfulness. Independently of the laughter question, I think one **positive emotion** should be included in Table 11-4.

Comment [USS83]: Shouting/yelling, while running or "in transit"

Comment [84]: Twelfth: The script for read or prompted speech should be included if at all possible.

28. Field 11.038: Segment Channel/SCH

This field (**Segment Channel/SCH**) refers to a segment identified in either the snip diary **SPD** of **Field 11.024** or the segment diary **SGD** of **Field 11.026** and describes the transducer and transmission channel within the segments. This field shall only be present if **Field 11.024** or **Field 11.026** appears in this record.

- The first item (**diary identifier/DIA**) is a mandatory and indicates the diary to which this field refers. If this item refers to a segment in the **SPD** of **Field 11.024**, the value is 0. If this item refers to a segment in the **SGD** of **Field 11.026**, the value is 1.

- The second item (**segment identifier/SID**) is a mandatory list of integers separated by commas, and gives the segment identifiers from diary to which the values in this field pertain. There may be an up to 999 values in this field. A value of 0 in this item indicates the segment content information in this field shall be considered the default value for all segments not specifically identified in other occurrences of this field.
- The third item **(transducer type/TYP) i**s an optional integer with attribute values given in **Table 11-6** –Editor's note: For most of the acquisition sources in Field 11.006 REC_AQS as specified by Table 83, we won't know the transducer type.

Table 11-6
Transducer Type

Transducer Type	Attribute Code
Unknown	0
Array	1
Multiple style microphones	2
Earbud	3
Body Wire	4
Microphone	5
Handset	6
Headset	7
Speaker phone	8
Lapel Microphone	9
Other	10
RESERVED FOR FUTURE USE only by ANSI/NIST-ITL	11-99

- The fourth item (**transducer/TRN)** is an optional integer that specifies the transducer type as unknown=0, carbon=1, electret=2, or other=3. Transducer arrays using mixed transducer types shall be designated "other".

- The fifth item (**capture environment/ENV**) is an optional text field of up to 4000 characters to describe the acoustic environment of the recording. Examples of text placed

in this field would be “reverberant busy restaurant”, “urban street”, “public park during day”.

- The sixth item (**distance to transducer/DST**) is an optional integer and specifies the approximate distance in centimeters, rounded to the nearest integer number of centimeters, between the speaker in the segment and the transducer. A value of 0 will be used if the distance is less than one-centimeter meter. Some example distances, handheld = 5cm, throat mic = 0cm, mobile telephone = 15cm, VOIP with a computer = 80cm, unless other information is available.

- The seventh item (**acquisition source/ACS**) is an optional integer that specifies the source from which the voice in the segment was received. Only one value is allowed. Permissible values are given in **Table 83** of the **Type-20** record. Any conflict between this value and **Field 11.006 REC_AQS** shall be resolved by taking this item to be correct for all segments identified in the second information item **SCH_SID** of this occurrence of **Field 11.038**.

- The eighth item (**alteration/ALT**) is an optional, unrestricted string for a description of any digital masking between transducer and recording, disguisers or other attempts to change the voice quality.

- The ninth information item (**comment/COM**) is an optional, unrestricted string for additional information to identify or describe the transduction and transmission channels of the segment.

29. Field 11.39-050: Reserved Fields

These fields are reserved for future use by ANSI/NIST-ITL.

30. Field 11.051: Comments/COM

This field (**Comments/COM**) is an optional unrestricted text string of up to 4000 characters in length that may contain comments of any type on the **Type 11** record as a whole. Comments on individual segments shall be given in **Field 11.024, SNP_COM**, or in **Field 11.026, SGD_COM**. This field should record any intellectual property rights associated with any of the segments in the voice recording, any court orders related to the voice recording and any administrative data not included in other fields.

31. Fields 11.052-099: Reserved Fields

These fields are reserved for future use by ANSI/NIST-ITL.

32. Fields 11.100-900: User-defined fields / UDF

These fields are user-defined fields. Their size and content shall be defined by the user and be in accordance with the receiving agency

33. Field 11.901: Reserved field

This field is reserved for future use by ANSI/NIST-ITL.

34. Field 11.902: Annotation information / ANN

This is an optional field, listing the operations performed on the original source in order to prepare it for inclusion in a biometric record type. This field logs information pertaining to this Type-11 record and the voice recording pointed to or included herein. See **Section 7.4.1.**

35. Field 11.993: Source agency name / SAN

This is an optional field. It may contain up to 125 Unicode characters. This is the name of the agency referred to in **Field 11.004** using the identifier given by domain administrator.

36. Field 11.994: External file reference / EFR

This conditional field shall be used to enter the URL / URI or other unique reference to a storage location for all source representations, if the data is not contained in **Field 11.999.** If this field is used, **Field 11.999** shall not be set. However, one of the two fields shall be present in all instances of this record type. A non-URL reference might be similar to: "Case 2009:1468 AV Tape 5". It is highly recommended that the user state the format of the external file in **Field 11.051: Comment / COM .**

37. Field 11.995: Associated context / ASC

This optional field refers to one or more **Record Type-21** with the same **ACN**. See **Section 7.3.3. Record Type-21** contains audio, video and images that are NOT used to derive the biometric data in **Field 11.999: Voice Record / DATA** but that may be relevant to the collection of that data.

38. Field 11.996: Hash/ HAS

This optional field shall contain the hash value of the data in **Field 11.999: Voice Data** of this record, calculated using SHA-256. See **Section 7.5.2**. Use of the hash enables the receiver of the data to check that the data has been transmitted correctly, and may also be used for quick searches of large databases to determine if the data already exist in the database. It is not intended as an information assurance check, which is handled by **Record Type-98**

39. Field 11.997: Source representation / SOR

This optional field refers to a representation in **Record Type-20** with the same **SRN**.

40. Field 11.999: Voice record / DATA

This field contains the voice data. See Section **7.2** for details.