Data Exchange File Format for SIV

Internal Working Draft ver. 1.2— March 10, 2006

VoiceXML Forum
Speaker Biometrics Committee

Authors:
Judith Markowitz, PhD, J. Markowitz, Consultants

Table of Contents

Table of Figures

About the VoiceXML Forum

Voice Extensible Markup Language (VoiceXML) is a markup language for creating voice user interfaces that use automatic speech recognition (ASR) and text-to-speech synthesis (TTS). Since its founding in March 1999, the VoiceXML Forum has continued to develop, promote and to accelerate the adoption of VoiceXML-based technologies via more than 150 member organizations worldwide.

Tens of thousands of commercial VoiceXML-based speech applications have been deployed across a diverse set of industries, including financial services, government, insurance, retail, telecommunications, transportation, travel and hospitality. Millions of calls are answered by VoiceXML applications every day.

The Forum's primary focus areas include:

- Promoting the adoption of VoiceXML-based technologies
- Cultivating a global VoiceXML ecosystem
- Actively supporting standards bodies and industry consortia, such as the W3C and IETF, as they work on VoiceXML and related standards, such as CCXML, X+V, MRCP, and speech biometrics.

For more information on the VoiceXML Forum visit the website at **http://www.voicexml.org**.

Disclaimers

This document is subject to change without notice and may be updated, replaced or made obsolete by other documents at any time.

The VoiceXML Forum disclaims any and all warranties, whether express or implied, including (without limitation) any implied warranties of merchantability or fitness for a particular purpose.

The descriptions contained herein do not imply the granting of licenses to make, use, sell, license or otherwise transfer any technology required to implement systems or components conforming to this specification. The
VoiceXML Forum, and its member companies, makes no representation on technology described in this specification regarding existing or future patent rights, copyrights, trademarks, trade secrets or other proprietary rights.

By submitting information to the VoiceXML Forum, and its member companies, including but not limited to technical information, you agree that the submitted information does not contain any confidential or proprietary information, and that the VoiceXML Forum may use the submitted information without any restrictions or limitations.

Revision History

Date	Description
February 17, 2006	Internal Working draft version 1.0
March 6, 2006	Internal Working draft version 1.1
March X, 2006	Internal Working draft version 1.1

1. Goals of This Document

This document has three primary objectives:

1. To present the rationale for developing a data-exchange file format (DEFF) for speaker verification and identification (SIV).

2. To present the rationale for developing a DEFF for speaker verification and identification (SIV) that complies with ANSI and ISO's Common Biometric Exchange Formats Framework (CBEFF) standard.

3. To provide a draft DEFF for comment.

2. Rationale for Developing a DEFF

As its name indicates, a data-exchange file format is a communication tool. It is intended to provide the following primary benefits:

- complement, and supplement the VoiceXML SIV standard specification;
- support interoperability among SIV vendors;
- facilitates communication between components of a single application.

It does not attempt to translate among different SIV algorithms.

In order to achieve these benefits it is critical that SIV technology and application developers participate in the creation a DEFF for SIV.

2.1 Complement, and Supplement the VoiceXML SIV Standard Specification

TO BE DEVELOPED

2.2 Support Interoperability among SIV Vendors

The unpredictability of the marketplace means that SIV deployments may outlive the SIV they use. This normal marketplace situation represents a tremendous source of risk for customers who are concerned about being faced with customer re-enrollment following the demise of their SIV technology supplier. This situation has promoted a "wait and see" attitude among potential customers and headaches for early adopters whose deployments are no longer supported.

One solution to this problem is to establish a meta-level "translation" voice model capable of converting between the models of different vendors. There are numerous challenges to this approach, including abiding vendor opposition. The marked differences between commercial algorithms make creating a translation problematic. The translation would need to support a variety of text-dependent, text-prompted (challenge-response), and text-independent technologies, It would also have to handle both language dependent and. language independent technologies plus representations for algorithms as different as Gaussian classifiers, neural networks, and dynamic time warping as well as blended approaches, such as neural-tree networks,

Creation of a DEFF minimizes represents a simpler, easier, and more straightforward solution to the risk associated with losing a vendor. The DEFF enables an SIV application to process data from enrollment and verification transactions performed by other products. It can accomplish this because a DEFF contains enrollment and/or verification data accompanied by information about those data that makes it easier to interpret and use the data in an existing deployment. This can be extended to seamless forward compatibility for a single vendor who may have made major changes or improvements to its algorithm.

2.3 Facilitates Communication between components of a single Application

The SIV process requires the VoiceXML browser to communicate with backend applications that may contain sensitive data. The formalization of a DEFF can standardize, support, and secure such communications within an SIV application or audits of that application. Support of intra-application communication also includes shipping a voice model or secured voice model from one application component to another.

The DEFF approach would be especially useful for VoiceXML version 3.0 which introduces the data-flow-presentation model (DFP). In the DFP model individual modules of an application (called “presentations”) exchange data with each other in a fashion that is mediated by SCXML. The use of a DEFF would support the shipping a voice model, a secured voice model, and supporting information from one application component to another or one presentation to the next.

It has also been suggested that a DEFF could be used to enhance support for assistive use of SIV within VoiceXML applications. TO BE FURTHER DEVELOPED

3. Rationale for Developing a CBEFF-Compliant DEFF

The Common Biometric Exchange Formats Framework (CBEFF) is an open, non-proprietary standard of the American National Standards Institute (ANSI) and the International Standards Organization (ISO). It is a “technology blind” interoperability tool. It defines a set of data elements that can be placed in a single file and used to exchange biometric information between applications, systems components, and organizations.

- facilitates multi-biometric applications;
- brings SIV into alignment with ANSI and ISO standards (e.g., BioAPI);
- opens new market opportunities for SIV.

3.1 Facilitates Multi-biometric Applications

The technology-blind CBEFF approach promotes interoperability of biometric-based application programs and systems. This goal of CBEFF is consistent with the rationale for developing a DEFF (see section 1. above). Because there are CBEFF DEFFs for all other biometric technologies, development of a CBEFF-compliant DEFF would extends the goal of interoperability to include multi-biometric applications

CBEFF is a template that describes the general structure and content of biometric-specific DEFFs. The actual content is determined by those who develop each CBEFF-compliant format. This freedom allows the speech-processing industry to craft a CBEFF-compliant DEFF that is tailored to the needs of our community while, at the same time, supporting the more general CBEFF goals.

The growing acceptance of biometrics has brought with regulations, such as HIPAA and FIL 103-2005, that require multi-factor authentication. The first wave of multi-factor authentication generally combines a biometric factor, such as SIV, with non-biometric factors, such as a password. The coming wave includes multi-biometric deployments. A CBEFF-compliant format facilitates such combinations by standardizing the way in which the various authentication factors communicate.

3.2 Brings SIV into Alignment with ANSI and ISO standards

The use of a CBEFF-compliant format

- supports BioAPI (INCITS 358-2002) compliance and other biometric standards,
- Simplifies software and hardware integration,
- Supports the creation of SIV databases for use in evaluating multiple algorithms,
- supports the development of multi-vendor SIV databases of voice models

The VoiceXML Speaker Biometrics Committee (SBC) identified BioAPI as one of the standards with which the VoiceXML SIV extension should comply. Since the SBC was formed, the chair of the BioAPI Consortium has been an active participant in SBC's meetings. Compliance with BioAPI requires the development of a CBEFF-compliant format for SIV because a CBEFF format is one of the outputs of a BioAPI-compliant system.

As figure 1 reveals, CBEFF lies at the heart of a broad landscape of biometric standards activities being performed by INCITS/M1 and ISO.

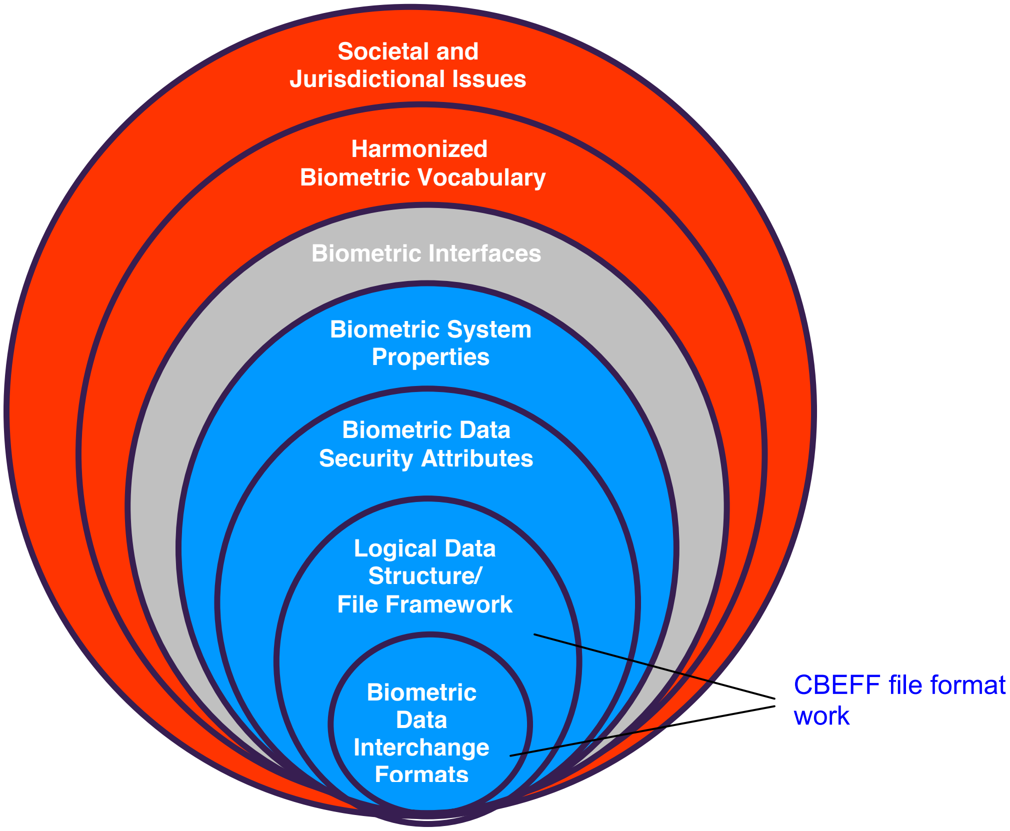


Figure 1. Importance of CBEFF DEFFs to the overall biometric standards landscape

3.3 Opens New Market Opportunities for SIV

Having a common method of representing the data from SIV transactions and the conditions under which the data were collected (e.g., the type of input device, audio quality) enhances the ability of other organizations and markets to use those data (e.g., law enforcement). In addition, some government agencies, such as the Department of Homeland Security, are required to use BioAPI and CBEFF and they include support of those standards in their requests for proposals (RFP).

4. Developing the DEFF

The CBEFF standard defines a publicly-available process by which any public or private organization can formally register specified DEFFs using some or all of these data elements. It has two forms:

- NISTIR 6529-A / ANSI INCITS 398-2005 / ISO 19785-1 2006 (CBEFF)
- OASIS XCBF

NISTIR 6529-A is the original CBEFF standard. It has three designations because it was first developed and approved by the US National Institute of Standards in Technology (NIST) and subsequently approved by the American National Standards Institute (ANSI) and the International Standards Organization (ISO). NISTIR 6529-A is a binary template for DEFFs. OASIS XCBF is an XML version of CBEFF. Additional support for translation from the binary ASN.1 format to XML is provided by a number of sources, including the IBM ASN.1/XML Translation program developed at IBM by Imamura and Maruyama (see Imamura & Maruyama, 2000)

4.1 Collaboration with M1

Development of a CBEFF-compliant format for SIV requires two kinds of expertise:

- Knowledge of the CBEFF template and
- knowledge about what needs to be included in a DEFF.

The M1 committee is responsible for the development of biometrics-related standards within ANSI. M1 is a division of ANSI's Information and Communication Technology (ICT) Standards (INCITS) committee. It has developed numerous CBEFF-compliant DEFFs (for fingerprint, finger, iris, face, signature, and hand geometry) it is the body that is the most knowledgeable about CBEFF formats and it is the organization that is best able to shepherd a CBEFF-compliant SIV format through INCITS and ISO.

The VoiceXML Forum, specifically the VoiceXML Speaker Biometrics Committee (SBC), and the speech-processing industry are the most knowledgeable sources about what needs to be included in an SIV format. The SBC created the SIV Requirements Document that was approved by the World Wide Web Consortium's (W3C) Voice Browser Working Group (VBWG) and precipitated the VBWG's work on an SIV specification that will be part of the next version of the VoiceXML markup language. The SBC is also preparing SIV best practices, application deployment, and security architecture documents for publication.

4.2. Design Methodology

The methodology for development of the SIV format will consist of an iterative application of the following steps:

1. SBC develops a draft SIV format
2. SBC actively solicits input from SIV vendors, integrators, and researchers to ensure that the draft format reflects their needs
3. SBC revises the draft, as needed
4. SBC presents the revised draft to M1.

This draft document is the product of Step 1 and is intended to be part of Step 2.

This work is scheduled to be completed by January, 2007 and will result in the creation of two variants of the SIV format:

- a binary version that will be published as an ANSI standard
- an XML standard that will be based on XCBF and published by the VoiceXML Forum.

This two-version approach reflects the divergent nature of M1 and VoiceXML Forum standards. The resulting standard will be proposed by M1 to ISO for acceptance as an international standard.

5. DEFF proposal for SIV

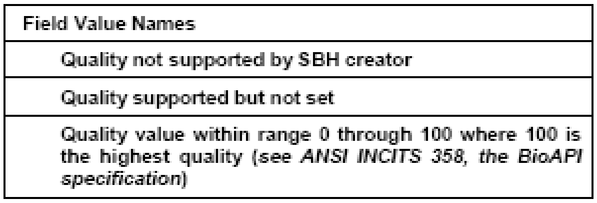
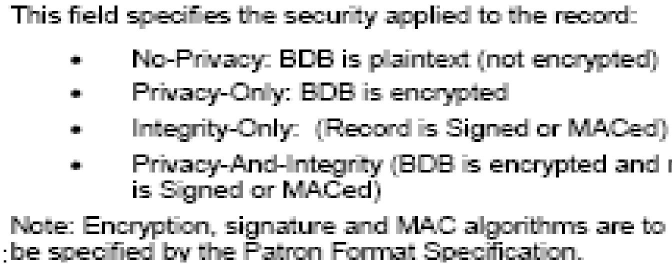
A CBEFF-compliant format consists of three data blocks:

- Standard Biometric Header (SBH)
- Signature Block (SB), and
- Biometric Data Block (BDB)

5.1 Standard Biometric Header (SBH)

The SBH is the header for the format. It is the portion of a CBEFF-compliant format that is the most generic and must adhere to an approved CBEFF SBH templates. (called *patron formats*). The one best suited to SIV is the ISO version of the BioAPI Patron Format shown in figure 2. This patron format is designed to support BioAPI-compliant applications. In its requirements the VoiceXML SBC listed BioAPI is one of the standards with which the VoiceXML SIV specification needs to comply.

Field name	**Description**
BioAPI_VERSION	The version of the BioAPI Patron Header used. The version represented here is BioAPI 2.0 (ISO/IEC 19784-1)
BioAPI_BIR*_DATA_TYPE	The kind of SIV data being transmitted: raw, intermediate (e.g., features), fully-processed model/template. The following types of data have been identified within the BioAPI patron format #define BioAPI_BIR_DATA_TYPE_INTERMEDIATE (0x02) #define BioAPI_BIR_DATA_TYPE_PROCESSED (0x04) #define BioAPI_BIR_DATA_TYPE_ENCRYPTED (0x10) #define BioAPI_BIR_DATA_TYPE_SIGNED (0x20) #define BioAPI_BIR_INDEX_PRESENT (0x80)

BioAPI_BIR_BIOMETRIC_DATA_FORMAT	The DEFF used for the BDB
BioAPI_QUALITY	The quality of the data that are being transmitted. basic CBEFF quality fields are  Field Value Names Quality not supported by SBH creator Quality supported but not set Quality value within range 0 through 100 where 100 is the highest quality (*see ANSI INCITS 358, the BioAPI specification*) The BioAPI specification further defines relative quality ranges allowing speicification of quality for raw, intermediate, and processed data.
BioAPI_BIR_PURPOSE_	The intended use of the data. The following uses have been identified within the BioAPI patron format: #define BioAPI_PURPOSE_VERIFY (1) #define BioAPI_PURPOSE_IDENTIFY (2) #define BioAPI_PURPOSE_ENROLL (3) #define BioAPI_PURPOSE_ENROLL_FOR_VERIFICATION_ONLY (4) #define BioAPI_PURPOSE_ENROLL_FOR_IDENTIFICATION_ONLY (5) #define BioAPI_PURPOSE_AUDIT (6)
BioAPI_BIR_BIOMETRIC_TYPE	The kind of biometric used in the biometric record: Voice. The binary representation is "000004"
BioAPI_BIR_BIOMETRIC_PRODUCT_ID	. Indicates product owner and product type
BioAPI_DTG	Creation date and time. Date (year, month, day) and time (hour, minute, second)
BioAPI_BIR_SUBTYPE	This field is useful for indicating the kind of SIV data being transmitted. Its fillers are: • Text-dependent • Text-prompted/challenge-response • Text-independent
BioAPI_DATE	Expiration date (year, month, day)
BioAPI_BIR_SECURITY_BLOCK_FORMAT	This is specified in terms of "owner" (the organization that set up the security format being used) and type of format  This field specifies the security applied to the record: • No-Privacy: BDB is plaintext (not encrypted) • Privacy-Only: BDB is encrypted • Integrity-Only: (Record is Signed or MACed) • Privacy-And-Integrity (BDB is encrypted and record is Signed or MACed) Note: Encryption, signature and MAC algorithms are to :be specified by the Patron Format Specification. This information can be encrypted
BioAPI_UUID	SIV product registration (with IBIA) code - UUID

* BIR - **biometric identification record** refers to SIV utterance data included with the format.

Figure 2 BioAPI Patron Format ISO/IEC 19784-1

The only field that has been uniquely defined for SIV is BioAPI_BIR_SUBTYPE. This indicates the kind of utterance data that were elicited and accompany the DEFF.

5.2 Signature Block (SB)

The SB is optional and used only if security has been applied to the contents of the format. The BioAPI Patron format defines the following categories of security used

- No-privacy: the BDB is plain text (not encrypted)
- Privacy-only: the BDB is encrypted
- Integrity-only: the record is Signed or MACed
- Privacy-and-integrity: the BDB is encrypted and the record is Signed or MACed.

5.3 Biometric Data Block (BDB)

The BDB is the heart of the CBEFF format. As figure 3 reveals, the BDB contains the SIV/biometric data (raw, partially processed, or template). It has a header that contains all the information required to communicate effectively about the attached data. Some header fields are required; others are optional and are necessary for only a subset of the data communication operations (see Sections 2.1 and 2.2, above).

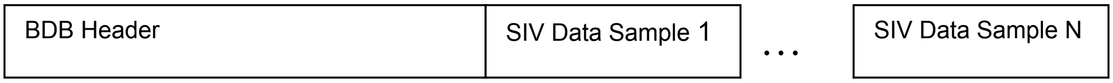


Figure 3 Sample BDB structure

The bulk of the development work that will be done by the SBC and the speech-processing community will be directed towards specifying and defining the fields to be included in the BDB. The BDB must include fields and field values needed by entities that will communicate about the SIV data. Figure 4 presents the current proposal.

One of the most challenging is determining the kind of data (BIR) to be transmitted. The growing diversity of voice models (called “templates” in CBEFF) does not support the transmission of the models themselves. This would violate the basic goal of interoperability.

Unlike the simpler data that would be transmitted for most other biometrics, SIV transactions generally involve dialogues that may be fairly lengthy and complex. Length and complexity intersect with size and transmission because raw spoken data quickly become very large. While this would argue for sending intermediate level data, such as features, the features that are extracted by different vendors are not entirely the same nor are they always processed in the same manner. The transmission of feature data may also represent a security vulnerability.

THESE POINTS NEED FURTHER ELABORATION

The following proposed fields

- Number of Utterances
- Utterance length
- Audio connection
- Audio format
- Audio channel
- Sampling rate

- Data quality
- Recording
- Type of data
- Country
- Language
- Language region
- Linguistic content
- Utterance Validation
- Simultaneous ASR and SIV
- Other factors

Each is defined and described in terms of its status (optional or required), the values needed, and any issues related to it.

The fields Utterance Length through Recording provide information about the underlying audio of the accompanying data (BIR)

The fields Type of Data through Linguistic Content provide information about the BIR data itself.

The fields Utterance Validation and Other Factors deal with external resources that may be used in addition to the SIV biometric analysis. One point of discussion is whether these fields need to be included in a DEFF.

5.3.1 Number of Utterances

Field Name	Status	Definition	Values
Number of Utterances	Required	The number of utterances or turns included in the transmission	1

An utterance is a segment of spoken data from a user. An alternate field name could be “Number of Turns.”

The SBC recommends that only one turn be transmitted by each DEFF. The single-turn approach is the simplest and most straightforward. If a DEFF is restricted to a single utterance/turn then this field can be eliminated or retained with only the value “1” permitted (as shown above) as a reminder that only one turn is allowed.

Theoretically, the DEFF should not care whether the utterance that is being sent is a concatenation. It’s the work of the application and/or engine to determine how the data are to be treated. For example, a string of digits from a single turn can be sent as a unit and the engine/application may use speech recognition to recognize each digit separately.

Problems arise when the concatenation spans turns. If, therefore, the DEFF is allowed to support transmission of multiple turns the following issues will need to be addressed

1. It cannot be presumed that the receiving application or engine will be able to process multi-turn “utterance” data as a single unit. The DEFF would, therefore, require an additional field that would specify the turn separator. That separator would need to be a standard separator that could be easily identified and manipulated.

2. The dialogue between the user and the IVR may last longer than what is needed for verification and you don’t want to force the verifier to wait until the conversation is over. For text-independent

verification, for example, more data than a single turn may be needed in order to reach a decision but it is unlikely that it is unlikely that data from the entire dialogue will be required. Furthermore, the direction taken by the dialogue may depend upon a speedy verification decision.

3. Some of the properties used to describe the spoken data in one turn may differ from those that describe another turn of the same dialogue.

4. Text-independent engines may require 10 or more seconds of spoken data. Simply concatenating a series of "yes" and "no" responses or comparable "bursty" data is unlikely to provide the acoustic quality required to enable the engine to produce a reliable result.

5. In text-dependent or text-prompted enrollment an application would want to be able to analyze each "turn" separately – not concatenated into a single utterance unit.

5.3.2 Utterance Length

Field Name	Status	Definition	Values
Utterance Length	Required	The total length of the utterance data.	Numeric in Seconds

Utterance length is measured in seconds. One question that might be applied is whether the length refers to the recording length or only the segment of the recording that contains speech. Most likely it will be the latter.

5.3.3 Audio Connection, Audio Format

Field Name	Status	Definition	Values
Audio Connection	Optional	The audio connection used to capture the attached utterances.	Desktop, Telephone, Device
Audio Format	Optional	The input format used for the attached utterance data	ulaw, Alaw, 16bit linear, G.711, G.723.1, G.729A, DPCM, GSM
Audio channel	Optional	Microphone and transmission channel	Speaker phone, Cell phone, wireline phone, microphone into PC

Audio Connection specifies the general category of audio connection.

Audio Format identifies the format used to capture the utterances. Only one format should be specified. The list shown may need to be expanded or more generic depending upon what is most useful for the industry. This warrants more discussion.

Audio Channel indicates the type of microphone or telephone used to input the data. The microphone and channel have a strong impact on the attributes of the utterance data. Some engines and applications have different templates for different channels so it is useful to know the channel. Generally, this information is acquired through the use of handset detection technology or other methods external to the SIV.

It would be desirable to make these three fields required fields. Unfortunately, the IVR generally does not know this information so they must remain optional fields.

5.3.4 Sampling Rate

Field Name	Status	Definition	Values
Sampling rate	Required	The number of samples per second	Numeric in samples per second

This refers to the number of samples taken from the spoken data. Typically, the sampling rate for telephone channels is 8k per second. This means that, to some extent, the sampling rate is predictable from the audio channel.

5.3.5 Data Quality

Field Name	Status	Definition	Values
Data quality	Optional	Specifies the signal to noise ratio (SNR)	Integer 1-100

This field specifies the signal-to-noise ratio (SNR) of the utterance data. It is one of the properties that argue in favor of single-turn transmissions because SNR can vary from one turn to the next or even within a single turn (for example, when there is a bang in the background during a portion of the turn). This means it would have to be re-evaluated all the time.

In CBEFF-compliant DEFFs for other biometrics quality is often specified using an integer range, usually from 1 (poorest) to 100 (best). Some indicate a breakpoint at which the quality is too poor to be processed. This presumes that the quality can be assessed at the point of reception.

SNR information is highly desirable which argues for making this a required field. Having information about the overall quality of the data provides important information for processing. For example, it helps set the “silence” level. It also provides valuable information for future analysis, for example auditing and re-analysis of the data.

That cannot be the case. Unlike most other biometrics the input device for voice can be virtually anything. They can be microphones into laptops, wireless phones with ear bots, landline phones, cordless phones, speaker phones, etc. Generally, the determination of quality is done after the processing begins rather than before it is initiated. That is, the system must process the data in order to find out about the quality. Depending upon the use to which the DEFF is put this may not be possible. For example, if the DEFF is used by a VoiceXML browser to communicate with an SIV application or engine.

The biggest problem is that currently there are many different approaches to interpreting signal-to-noise ratio (SNR). Some are applied to audio segments that contain voice data while others are done on audio segments that have no voice data present. The consistency of SNR calculations is likely to improve in the future. This should be encouraged and supported by having a field in the DEFF.

It was agreed to have a general measure such as the 1-100 ranking used in the BioAPI patron format accompanied (expanded) by “metadata” information or tags that will facilitate both immediate and future analyses of the data. The discussion about what would be useful metadata should continue.

5.3.6 Recording

Field Name	Status	Definition	Values
Recording	Optional	The recording technology used to capture the utterance data	To be determined

This field specifies the recording technology used to capture the utterance data. It will be done using the VoiceXML recording tag. The value of this field needs further examination and, if it is retained, the nature of the fillers that go into it need to be carefully and unambiguously defined. Additional information will be needed if it is determined that the recording technology adds noise or otherwise affects the audio.

5.3.7 Type of Data

Field Name	Status	Definition	Values
Type of data	Required	Specifies the kind of utterance included	Raw, Intermediate, Template

This field identifies the level of processing that has been applied to the utterance data. “Intermediate” refers to partially processed data. Most likely, intermediate data would be feature data. “Template” refers to complete processing into a voice model.

The default is “Raw” because all SIV vendors could process raw data. This means that a vendor could reanalyze raw data to rebuild templates, perform audits, and do a variety of other operations on the data. This is not the case for intermediate and template data.

The issues related to using raw data are

1. Large quantities of raw data require a great deal of storage and could be an issue for organizations with millions of users. At the same time, analysis of large quantities of raw data may be required for human-human interactions, such as speaker identification for the transcription of meetings.
2. There may be some privacy regulation issues related to storage and transmission of raw data.
3. The data will need to be encrypted
4. SIV engines differ in their data requirements and capabilities so the ability of different engines to handle raw data will vary.

The SBC supports the exclusive use of raw data for the primary DEFF. This conforms with standard practice for DEFFs for other biometrics. The SBC also recommends the development of another DEFF for transmission of intermediate or template data for communication between elements of a single system so that the data may be processed quickly. For example, if the transmission is between a distributed speech capture device (e.g., a cell phone) and its server.

5.3.8 Country

Field Name	Status	Definition	Values
Country	Optional	Indicates the country where the data were collected	ISO 3166 Country code

This field is useful for enhancing performance and accuracy for speech transmitted over telephone channels – primarily land-line channels.

5.3.9 Language and Language Region

Field Name	Status	Definition	Values
Language	Optional	The language spoken in the utterances	ISO 639 language code
Language region	Optional	Specific dialect	TBD

Some SIV engines are language dependent and, therefore, require knowledge of the language used in the utterance data in order to process them. Language Region may be needed for language-dependent SIV engines. This is a topic for discussion.

5.3.10 Linguistic Content

Field Name	Status	Definition	Values
Linguistic content	Required	The nature of the spoken material in the utterances	Digits, combination lock, other numbers, pass phrases, freeform speech

Linguistic content is critical for analysis by SIV engines that are text dependent and, therefore, restricted in their ability to analyze linguistic data. Note, that this field is tied to The "Biometric Data Type" field of the SBH which will likely indicate whether the utterances were collected as text-dependent, text-prompted, or text-independent data.

5.3.11 Utterance Validation

Field Name	Status	Definition	Values
Utterance Validation	Required	Indicates whether ASR or other technology was used to process the utterance	Yes or No
Simultaneous ASR and SIV	Optional	Indicates whether ASR was used to decode the speaker's claim of identity	Yes or No

Speech recognition (ASR) and other techniques are often used to verify that the speaker said what the system expected him/her to say. For example, that the speaker correctly repeated a sequence of digits or words in a challenge-response sequence or that the speaker provided the correct password. This is a background (usually internal) validation function that reduces false rejections. The purpose of the "Utterance Validation" field is to indicate whether or not validation was done. This allows the system receiving the DEFF to know that it need not perform utterance validation. It is a required field with "yes" or "no" as fillers.

The argument against the inclusion of this field is that if the answer is "no" and the engine/application requires validation at least one other field would be needed to say what the system expected. Any engine or application that requires validation would still need that information whether or not there is a validation field. More discussion is needed to determine whether a second linguistic field is needed and, if so, how that information would be captured especially when the source is not configured to do validation or does validation using a variety of external techniques (e.g., calling number ID).

Utterance Validation is distinct from the common use of ASR in conjunction with SIV to conflate the claim of identity with a text-dependent password. That is the purpose of the "Simultaneous ASR and SIV" field. This field refers to the application of both ASR and SIV on the same utterance. ASR is used to decode the spoken claim of identity and SIV is applied to the same data to perform verification. This is an optional field even though it is included in the VoiceXML Requirements Document (see *Speaker Identification and Verification (SIV) Requirements for VoiceXML Applications*).

These two fields are distinct from the use of ASR and other technologies as additional factors. There are no fields in this DEFF for multi-factor SIV.

5.3.12 Other Factors

Field Name	Status	Definition	Values
Other Factors	Optional	Indicates whether other non-speech factors were used used as part of the SIV processing.	Yes or No

Other factors is similar to the Utterance Validation field in that it may not be necessary but could account for the inclusion of non-linguistic data if they are allowed.

6. References

CBEFF: Common Biometric Exchange Formats Framework – NISTIR 6529-A. Published April 5, 2004

Imamura, T. and H. Maruyama. Mapping Between ASN.1 and XML, IBM Research Report, RT0362, 2000.

Speaker Identification and Verification (SIV) Requirements for VoiceXML Applications. VoiceXML Forum Speech Biometrics Committee document. Published September 14, 2005

Use of BIP in remote authentication – INCITS document M1/05-0341. Published May 24, 2005

XML Common Biometric Format (XCBF) Committee Specification 1.1. OASIS Standard, ratified September, 2003 (http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xcbf)

7. Appendix – Table of Proposed BDB fields

Field Name	Status	Definition	Values
Number of Utterances	Required	The number of utterances or turns included in the transmission	1
Utterance Length	Required	The total length of the utterance data.	Numeric in Seconds
Audio Connection	Optional	The audio connection used to capture the attached utterances.	Desktop, Telephone, Device
Audio Format	Optional	This is the input format used for the attached utterance data	ulaw, Alaw, 16bit linear, G.711, G.723.1, G.729A, DPCM, GSM
Audio Channel	Optional	Microphone and transmission channel	Speaker phone, Cell phone, wireline phone, microphone into PC
Sampling Rate	Required	The number of samples per second	Numeric in samples per second
Data Quality	Optional	Specifies the signal to noise ratio (SNR)	Integer 1-100
Recording	Optional	The recording technology used to capture the utterance data	To be determined
Type of Data	Required	Specifies the kind of utterance included	Raw, Intermediate, Template
Country	Optional	Indicates the country where the data were collected	ISO 3166 Country code
Language	Optional	The language spoken in the utterances	ISO 639 language code
Language region	Optional	Specific dialect	To be determined
Linguistic content	Required	The nature of the spoken material in the utterances	Digits, combination lock, other numbers, pass phrases, freeform speech

Utterance Validation	Required	Indicates whether ASR or other technology was used to process the utterance	Yes or No
Simultaneous ASR and SIV	Optional	Indicates whether ASR was used to decode the speaker's claim of identity the speaker's claim of identity	Yes or No
Other Factors	Optional	Indicates whether other non-speech factors were used as part of the SIV processing.	Yes or No

Figure 4 Table of Proposed DEFF fields