CJIS Division

US Government Interagency Symposium for Investigatory Voice Biometrics

Use Case Committee Report

Version 1.0

10/5/2009

Prepared by:

Use Case Committee

Produced for:
Federal Bureau of Investigation

Criminal Justice Information Services Division
1000 Custer Hollow Road, Clarksburg, WV 26306

TABLE OF CONTENTS

LIST OF TABLES

1 Executive Summary

US Government interest in the development of automated techniques to recognize people by their voices has a history of nearly 70 years. Although significant challenges remain, the consensus is that sufficient progress has been made to enable US Government agencies in general, and the FBI specifically, to further consider fielding speaker recognition technology in support of their missions. Therefore, the FBI Science and Technology Branch Biometric Center of Excellence (BCOE) asked the National Institute of Standards and Technology (NIST) to launch a program directed toward the development of voice biometric collection and interoperability standards capable of supporting the common investigatory needs of all interested US Government agencies.

To begin this process, NIST organized a 2-day Interagency Symposium for Investigatory Voice Biometrics on 24-25 March 2009 that was attended by about 80 international stakeholders from government, academia and industry. The symposium marked the beginning of a multi-year program to develop investigatory voice biometric collection and interoperability standards by establishing 4 committees (Use Case, Interoperability, Collection Standards, and Science & Technology), each assigned to create a "challenge" document to be delivered to a Government Steering Committee.

This document is the first of those 4 challenge documents and outlines several government use cases or scenarios. A taxonomy of voice biometric applications and collection conditions was created by the Use Case Committee and seven current and potential application scenarios were identified and addressed in this report. One of the scenarios, Scenario 6 – Intelligence Case 1, has been used an exemplar for applying the taxonomy to use cases. The Use Case Committee members will apply the taxonomy to the other six scenarios after this initial report is published.

A brief discussion is provided regarding the potential for developing Criminal Justice Information System (CJIS) voice biometric services comparable to current CJIS fingerprint identification services. Some science and technology gaps directly related to current voice applications are also identified in this report.

The report concludes with 5 recommendations:

1. This document and its Appendices should be developed further and converted into a survey instrument that can be circulated to agencies and departments that already perform voice-biometrics activities and those that would like to migrate to these technologies. This will help to

prioritize requirements for improving ASR (automatic speaker recognition) technology.

2. The FBI should establish a Special Interest Group (SIG) on their law enforcement on-line (LEO) system for distribution of the survey as well as for sharing it with major agencies and departments in other countries.

3. This Committee should work with CJIS and the Steering Committee to further develop use case information to include review of the responses to the survey.

4. Prototyping of voice sample collection as part of the arrest booking cycle should be carried out to help determine the "best practices" guidance for the collection of voice biometrics by law enforcement agencies.

5. Develop a long-range scientifically based program in accord with the National Academy of Sciences (NAS) report on "Strengthening Forensic Science in the United States" [7] for independent and rigorous scientific evaluation, including appropriate corpora, of the enabling technologies to support the FBI's voice biometrics mission. This should include creation of a scientific working group – a voice to resolve the issues of procedures for voice comparison, training and certification of examiners.

2 Introduction

The US Government Interagency Symposium for Investigatory Voice Biometrics started with a goal of identifying use cases across and between application domains that could lead to common collection standards, recommended best collection practices, and data exchange formats.

This Use Case Committee report provides a look at the discussions of use cases at the symposium, those in the weeks immediately after the symposium, and a way forward as part of a larger FBI CJIS study of the possible provision of voice biometrics services to law enforcement and intelligence communities. The CJIS Division has a major upgrade of their identification systems known as Next Generation Identification (NGI) with specific biometric modalities (e.g., automated facial image matching) scheduled for introduction over roughly the next 5 years. As a practical matter it is unlikely that CJIS would develop and offer an investigatory voice biometric service operationally before 2015. The Symposium is part of a forward looking effort by CJIS to determine what such a service might look like and what steps they would need to take over the next few years to

position themselves to offer that service in the next decade. The place to start that determination is with understanding today's uses of the technology.

The term "use cases", while selected by the steering committee has turned out to be better thought of as use scenarios. "Use cases", as a term of art in systems engineering, implies a description of a systems behavior in response to user scenarios[1]7. While this report and the committee will be referred to as the Use Case Report and Use Case Committee, the substantive contents of the report will focus on scenarios. Scenarios can best be thought of as accounts or synopses of possible investigatory courses of action, written in plain language and with minimal technical details, so that stakeholders can have common examples upon which to focus their discussions.

2.1 Use Case Panel

The use case panel at the symposium consisted of the following presenters:

- Peter T. Higgins, Chair, Higgins & Associates, International, USA
- David van Leeuween, TNO, Netherlands
- Fred Goodman, MITRE Corporation, USA
- Joaquin Gonzalez-Rodriquez, ATVS, Spain

2.2 Use Case Committee

As part of the symposium, participants were invited to volunteer for committees that would follow up on the symposium, one of which is the Use Case Committee. The Steering Committee tasked the Use Case Committee to use the panel presentations and resultant discussions to determine and document the goals and requirements for moving toward a large-scale voice biometrics capability within the FBI and other agencies. The Use Case Committee consists of the following people:

- Peter T. Higgins, Chair, USA
- Joseph P. Campbell, MIT Lincoln Laboratory, USA
- Carson R. Dayley, FBI, USA

[1] "A use case in software engineering and systems engineering is a description of a system's behavior as it responds to a request that originates from outside of that system. In other words, a use case describes "who" can do "what" with the system in question. The use case technique is used to capture a system's behavioral requirements by detailing scenario-driven threads…" [1]

- David van Leeuween, TNO, Netherlands
- James R. Luther, SAIC, USA
- Judith Markowitz, J. Markowitz Consultants, USA
- John Mears, Lockheed-Martin, USA
- Antonio Moreno, Agnitio, Spain
- James L. Wayman, USA

2.3 Scope and Approach

After careful consideration of the traditional methods of classifying applications of biometrics, the committee determined that "verification" and "identification" did not present a complete taxonomy for classifying Investigatory Voice Biometrics scenarios. These scenarios, examples of which are provided in this report, tend to have overlapping aspects, purposes, and techniques and are not just focused on "verification" and "identification" as commonly understood by many in biometrics community. In fact, the committee realized that in traditional biometrics a subject normally initiates verification to an access control system (e.g., access to a computer system) or to get a benefit (e.g., pass through a border checkpoint). Whereas here the verification is initiated, typically without the subject even being present or aware, by an investigator to determine if a sample matches previously collected voice samples of a subject under investigation. Additionally we see a potential future where voice samples are collected and enrolled but not searched or processed until the subject becomes a person of interest. The committee thus proposes a somewhat different taxonomy of voice recognition applications - identification, verification, and enrollment, given in this report at Appendix B.

The first agreement was on the meaning of the phrase - "investigatory voice biometrics". The first word of this phrase, "investigatory", was taken to encompass the use of speaker recognition technology in criminal and intelligence investigations and analysis. In US Federal Courts, the admissibility of scientific evidence is determined by the presiding judge, who is guided by Federal Rules of Evidence (FRE) 104 [2] and 702 [3], among others. FRE 702 notes the Daubert Criteria, which states the following factors must be met: the technique has been tested and subjected to peer review and publication; has a known error rate and standards controlling its use; is generally accepted in the scientific community [4]. The Symposium committee members believe that automatic speaker recognition technology has not yet reached the maturity to satisfy the Daubert Criteria. We noted that fingerprints have recently been the subject of Daubert challenges in the federal courts and related challenges (Frye hearings) in

state courts. Future research and evaluation are needed to advance speaker recognition technology to satisfy scientific-evidence admissibility requirements for the US Federal Courts [5].

In this critical phrase, “investigatory voice biometrics”, the committee also addressed the meaning of the words “voice biometrics”. This is not a term commonly used by either the forensic laboratory community or by intelligence investigators who use voice samples as part of their analysis. The term “biometrics” tends to be used more in the border control, physical and logical access control, and national identity program communities. The term “voice biometrics” implies a move toward fully automated speaker recognition, rather than the current, very hands-on process of skilled technicians arriving at conclusions through semi-automated tools and analysis. The very definition of biometrics shows us this: *“automated recognition of individuals based on their behavioral and biological characteristics*” [6].

Thus the Symposium’s focus was on the possibility of a future where automated techniques and tools could support more automated investigatory efforts leading to broader prosecutorial use. Of course, we realize that trained examiners testify in courts, not analytic tools or their reports – the recent Supreme Court case, Melendez-Diaz v. Massachusetts (June 2009), has made this even more important. So even achieving fully automated speaker recognition does not obviate the need for procedures and reviews to ensure the results and decisions are explainable and meet the FREs. The consensus was that the scenarios to be described would be predominantly based on current uses with one scenario describing a possible future CJIS voice biometric service.

The committee focused on the two most basic attributes of voice biometric applications that both address the collection of voice samples:

1. Cases employing controlled collection, which could be overt or covert but where the environment, transducer, and recording equipment are mostly known and controlled.

2. Cases employing uncontrolled collection, to include court-authorized surveillance, intercepts, and wiretaps, and various opportunistic sample collection.

The committee recognized that audio samples typically contain additional information beyond just voice data (e.g., gunshots) that can be of use to other investigatory disciplines. Consequently, to the extent practical, all information in audio samples should be preserved in the collection phase. The committee also recognized that current speaker recognition technologies have advanced beyond simple acoustic voice models –

now combining speaker, speech, language, dialect, and other higher-level information obtained from the samples, either automatically or more typically, semi-automatically.

In both classes (controlled and uncontrolled collection), the scenarios are aligned with the purposes of the investigative techniques applied. These techniques are often determined by the quality and amount of data collected. Ideally voice biometric-based speaker recognition efforts would lead to the identification of all known speakers (i.e., those represented in the corpus to be searched). We acknowledged that signal quality, language, sex of speaker, sample duration, and other factors lead to missed identifications, therefore limiting the possible application of voice biometrics in many cases. This is not dissimilar to noise, dirt, and other externalities limiting the identification of latent fingerprint sources in many criminal investigations.

The committee agreed that a good approach would be to identify the major motivations for forensic voice investigations and use these as the purpose-based scenario sets:

1. Identify one or more speakers
2. Determine the language(s) and dialects being spoken
3. Determine how many different speakers are in a sample
4. Link speakers across samples
5. Verify that a voice sample is from a certain individual under investigation – speaker detection
6. Determine that a voice sample is not from an individual under investigation
7. Disprove a claim by an individual that the voice on an intercept belongs to another specific individual and therefore could not be his/hers

The committee realized full well that a typical investigation might involve more than one scenario set. In fact it was felt that a single investigation might cycle through several investigative scenarios sequentially. An example would be an application with purpose #6 in the first analysis that proves the hypothesis (that the voice of the subject under investigation is not represented in the sample), thus necessitating an application of purpose #4 to see if the represented speaker has previously been encountered.

One of the high-priority tasks recommended later in this report is to survey domain experts to prioritize requirements for improving automatic speaker recognition (ASR) technology in support of an evolutionary shift to investigatory voice biometrics.

2.4 References

This report makes extensive use of standards documents, some of which are new working drafts, and authoritative references. There are citations throughout this report to the references section. One additional document, the ANSI/NIST ITL-1 standard used by the FBI for transmission of fingerprint, face and other biometric data, although not discussed in this text, is included as reference [10] for completeness.

2.5 Vocabulary

Table 1 in Appendix A defines biometrics terms to permit a common understanding of their use in this report and in the speaker recognition community. As we know, the standards community has been working on a standard set of biometrics-related vocabulary terms for many years. Yet, today, within the forensic community there are numerous different vocabularies in use for similar functions and types of data samples.

2.6 Attributes of Voice Samples

Many factors influence the quality (as defined in Appendix A) of a voice recording, and like all biometrics, there is a solid relationship between input-signal quality and the samples to be matched from the same subject. In addition to quality factors there is the ability to train or "advise" an algorithm as to the conditions under which a collection was made. An analyst can better select appropriate algorithms and algorithm settings when aware of both the technical conditions of the collection and transmission, such as the frequency response of a microphone or a communications channel, sampling rate, compression, etc., and the nature of the speech itself, such as the language, speaking style, the speaker's stress, and the speaker's emotional state. The more we know about the collection, transmission and speaking conditions, the better we can tune the performance of the recognition algorithms.

At the symposium it was generally agreed that speaker samples typically exhibit variations in the speaker's cooperation, awareness, stress level, language, and style (e.g., read text or extemporaneous conversational speech) Variations in the recording equipment, environment, and channel all impact the samples in known and unknown ways. Session effects and what constitutes separate sessions were discussed. All these variations, and whether the samples are known or unknown and whether the samples are controlled or uncontrolled, were discussed. Whether a sample is to be used for enrollment (e.g., in training a model), to be used for testing, or to be used for both enrollment and testing, depends on the application and may not be known at the time of collection.

Modern ASR typically compares collected unknown voice samples, one at a time, to a model constructed from multiple known voice samples to estimate whether the speaker

in the unknown voice sample matches the speaker in the known voice samples. There are variations on this basic scenario, such as having multiple speakers in the voice samples, comparing unknown voice samples against multiple models (watch list), comparing unknown samples against each other, etc., depending on the Concept of Operations (CONOPS) to be discussed later.

2.7 Possible CJIS Voice Biometric Services

One potential goal in the law enforcement domain is to build large reference collections of known criminals and their speech data, from which enrollment records can be created. Spanish and other European police are already starting to do this. Based on their experiences and prior to an investment by the FBI CJIS division to start such a national service in the US, there is a need for:

- Operational concepts for the provision and use of such a service.
- Standards and policies for the collection and exchange of both enrollment and acquired speech data.
- Demonstration of the technical practicality of enrollments in typical law enforcement booking stations, etc.
- Demonstration of a more robust ability to search and match voice samples than has been demonstrated to-date.

With regard to the last point, the committee feels that demonstrations of large-scale search depth has been impeded by a lack of sufficiently large test-data sets to support research, algorithm training, testing, and demonstrations.

One operational concept that needs to be addressed is the relative mix of automation and expert human analysis, based on operational situations and data sample quality. This is to determine if the offered service will be primarily automated (similar to current tenprint fingerprint searches) or will require extensive expert intervention (similar to current latent search processes, where FBI forensic scientists do the analysis and make the decisions). The most likely scenario will be a combination of the two approaches, with the latter approach more prevalent with voice-based searches.

By way of explanation:

> Tenprint enrollments are submitted electronically via a secured network using appropriate standards. The CJIS fingerprint matchers then automatically characterize the fingerprints and search them without human intervention -- absent any data transaction processing errors. If

the matcher score is below a certain threshold then a response of "no match found" is automatically generated and returned to the submitting organization. If the score is above a second threshold then an automated response with the mated fingerprint related subject information is generated and transmitted. For those search results with a matcher score that falls in between the two thresholds, the candidates are sent to a trained fingerprint technician for determination of a match or no match decision.

Latent prints submitted to the FBI are sent to expert examiners who prepare the latents for searching by cleaning them up, isolating the region of interest from the rest of the information/noise in the image, orientating them correctly, and marking features either manually or using computer-based algorithms. The computer returns known or unknown finger (or palm) print candidate matches and the examiner reviews each and makes any identification decision.

As can be seen from the fingerprint model there are multiple paths to making identification decisions. One challenge for the FBI is how to offer voice biometric services: as a mostly automated process or as a skilled examiner service. It is anticipated that the recommendations from the Symposium will help in the formulation of the operational concepts and that those concepts will include both automated transactions and skilled specialists doing labor intensive pre-processing and analysis on the more challenging cases.

If the "tenprint" approach is applied to investigatory voice biometrics there must be sufficient experience and testing directed at developing thresholds dependent upon sample length and quality at which decisions of "no viable candidate" or "a match" can be automated. The less-decisive search results with scores between these two thresholds would then go to skilled technicians who would evaluate one or more candidates using appropriate technical tools.

The more challenging samples would go directly to technicians for pre-processing and possible end-to-end manual processing. If their results are to be admissible as evidence in the federal court system, protocols compliant with Federal Rules of Evidence will need to be developed.

2.8 High-Level Challenges

It is important to note that the long-term goal of offering voice biometric services presupposes that several challenges are successfully addressed:

1. Best practices for collection of suitable voice samples and related meta-data. The Collection Standards Committee will address these aspects.

2. Standards for transmission of known/enrolled samples and for transmission of unknown/intercept samples – in conformance with the best practices. The Interoperability Committee will focus on these issues.

3. Infrastructure for transmission of enrollment and probe transactions. The Interoperability Committee will focus on these issues.

4. Managing user expectations over the time it will take to develop such a service. The Steering Committee will coordinate with other committee chairs and members in address user expectations.

5. For processes assisted by human examiners, as in the fingerprint model, the development of examination protocols, training, and certification standards will be required. The FBI's OTD and CJIS are currently working on the initial drafting of such protocols, training, and certification standards.

6. For a large-scale forensic-style voice-search service (e.g., a national criminal voice biometrics service), there are multiple research, development, corpora, and evaluation challenges – a several year R&D effort. The Science & Technology Committee will address these challenges.

3 Problem Statement

3.1 Identification of Capability Gaps

The symposium provided a good insight into where voice biometrics are today and what capabilities the various user domains desire.

3.1.1 Major applications today

Today there are successful applications of voice biometrics in identifying speakers by comparing a sample utterance from a tape recording, a 911 call, or a court-authorized wiretap to samples from a few suspects selected from criminal investigations or intelligence analysis. These are typically low volume, generally forensic or intelligence applications, where the analyst has hours or days to perform the identification, which is not always possible due to limited voice sample duration or quality.

3.1.2 Where we want to go

In the future we want to conduct more automated searches against larger reference files – similar to the way tenprint fingerprints are searched.

3.1.3 What holds us back

Some of the gaps that restrict the community from large scale enrollments and more automated voice biometrics are: lack of demonstrated maturity of the algorithms for searching large scale reference files, the lack of large collected repositories (corpora), and lack of interoperability based on common or compatible collection standards to include metadata about the collection as well as the sample utterance. In addition, there is a common consensus about the uncertainty of the error rates of the speaker recognition performance when operating under variations due to the speaker (e.g., stress) and variations not due to the speaker (e.g., channel distortion).

3.2 Additional Constraints

Voice has some challenges that other biometric technologies do not face or have already successfully addressed.

3.1.4 Enrollment Techniques

The presentation on the Spanish Guarda system provided excellent information on their approach to enrolling voice samples. The details are provided in Section 4.1, some highlights are:

- They use multiple enrollment systems – to include a GSM cell phone and a high quality microphone and have an acoustically prepared room. Their application provides some good ideas on enrollment facilities.

- Their enrollment time is about two minutes for an enrollment sample and one sample is sufficient but four are preferred. This is an acceptable even for high volume booking sites but is far shorter than the experience in many laboratory cases. At the symposium there was a brief discussion of enrollment protocols: whether the enrollment should be read speech or whether the subject should be engaged in a conversation.

- It was suggested that perhaps having a telephone connection to a voice enrollment service center where professionals do nothing else but enroll voice samples using brief conversations could accelerate the enrollment process.

The Committee feels that advancing speech data enrollment, as routine part of an arrest booking cycle will require additional analysis of police booking station workflows and physical layouts.

3.1.5 Handling Sensitive Information Content

Unlike fingerprint, face, or iris images, voice samples contain data that is entirely exogenous to the speaker's behavioral and physiological speech factors. The

information content can be used to help identify speakers using techniques associated with frequency of word usage. Speech can also contain sensitive or even classified information such as the names of other suspects that sometimes cannot legally be shared with other agencies or forensic examiners.

This implies a need to mask the content of speech data. This need can be partially addressed through the exchange of model data rather than the recorded sample. This only works when identical model capabilities are used in the two labs in question. Even in those cases, there is a risk that the speech content might be reconstructed from the model – as long as this possibility is not zero there will be legal considerations that must be taken into consideration.

3.1.6 Lack of Large Reference Files

Fingerprint and other governmental biometric systems can be high volume systems with very large reference files collected over many years. At the present time, however, there are no large reference files of voice data. If voice data were collected routinely, as fingerprints are today, repositories of voice data reference files could be built reasonably quickly. Collection of voice data in the arrest booking cycle could be prototyped with a booking station that supported high quality voice enrollment, but upgrading booking stations across the country to routinely collect such data would require appropriation of federal grant money.

3.1.7 Limitation on Search Depth

Voice systems currently do not have large reference libraries and typically can't search more than tens of thousands of voice models successfully. Searches of repositories exceeding 10,000 reference files have been successfully demonstrated under some conditions.

It is interesting to note that prior to the initial operations of the FBI's Integrated Automated Fingerprint Identification System (known as IAFIS) in the late 1990s, the FBI's latent fingerprint search capability was limited to a search depth of 100,000 known fingers. With investments by the FBI and NIST in technology, the automated fingerprint identification system (AFIS) industry has successfully scaled their technologies to now be able to search reference files with tens of millions of enrollments. Ideally a CJIS investigatory voice biometrics effort could lead to similar improvements for the speaker recognition community, augmenting the substantial DoD – NIST collaboration that has brought this technology such a long way over the past 20 years.

4 Sample Use Cases

The following examples of current use cases came from presentations at the symposium and supplemental material submitted after the symposium. As the Committee goes forward there will be a need to broaden this list to reflect other examples and to apply the taxonomy to each of them. Initially Scenario 6 was used to demonstrate the use of the taxonomy – see Appendix D.

4.1 Spanish Police Scenario

The Spanish Guardia Civil started to use Voice Biometric technology in 2000 for the production of evidence to be presented in court. By 2003 they were confident enough with the technology that they decided to start a project named SAIVOX (Automatic System for the Identification of Voices). The project included the creation of a database of well-known criminals to be available for searching, during the investigation of new cases, to identify newly collected but unknown voice samples.

Following the design and development of the system, they started the operational deployment by 2005. Currently, there are around 100 booking stations across Spain (between one and six stations per province). The booking station has an acoustically prepared room with a set of HW devices (microphone, phone card, and external HIFI sound card) to allow the collection of microphone and telephonic recordings. A recording protocol was prepared for use in the booking process in order to standardize the process and homogenize the recordings obtained. This protocol consists in the collection of a minimum of one recording per data subject with a recommended length of 2 minutes (microphone preferred) with an optimal of 3 recordings (microphone, telephonic, and cellular).

Currently the collection process is restricted to a specific group of crimes closely related with voice recordings and telephone transactions (terrorism, drug dealing, threats, black mail, kidnapping, etc.) to avoid filling the database with less useful recordings. During the past three years the database has grown to include reference recordings from over 1,500 individuals. In the period 2005-2008 the enrollments increased in significantly (21 in 2005, 314 in 2006, 407 in 2007, and 778 in 2008). The number of subjects enrolled is expected to double in 2009 relative to 2008.

With the introduction of “latent” voices (related with concrete cases but not identified) the first success histories appeared in 2007 with cases related to Islamic and National pro-independence (ETA) terrorism. Currently the system is successfully used in operational cases and is providing information that has led to the detention of persons identified through investigatory voice biometrics.

4.2 FBI Scenario 1: One-to-one Comparison

The Forensic Audio, Video, and Image Analysis Unit (FAVIAU), Operational Technology Division of the FBI, will receive a request from one of the field offices for a voice comparison examination to be conducted between an unknown voice on a 911 call and the known voice of a subject on a recording obtained by law enforcement officers.

The unknown recording was transferred from the 911 system, converted from the proprietary format used on the 911 system to a digital file that preserves the full fidelity of the evidence (e.g., for telephone speech, a .wav file using a sampling rate of 8,000 Hz, 16-bit samples, linear PCM encoding, single-channel mono, unless the evidence is in stereo) and then burned to a compact disc (along with the original file) to be submitted as evidence. FAVIAU has no expectations that the 911 system does not use some form of compression in their proprietary storage of the recorded calls, but understands that this is the best evidence available and that FAVIAU must work with the provided evidence. This 911 call will contain the voices of the 911 operator, the caller, and sometimes, additional voices and sounds from both the 911 operator's side of the call and the caller's side of the call.

The known exemplar provided is a jailhouse phone call made using the suspects assigned pin number from the county jail. For calls of this type, the suspect must identify himself verbally before the phone call is allowed to be connected and the far-party (person receiving the call) must accept the call. So, in this one call will be the voices of the subject, the operator (usually a recorded voice) informing the receiving party that the phone call is being made from a detention facility, and at least one party at the receiving end of the call. The audio data file resulting from this call may be provided to FAVIAU in many formats: it may be an audio cassette recording of the phone call; it may be a digital file that was created in the same manner as described for the 911 call; it may be a direct copy (no format conversions, no pre-processing, etc.) from the recording system used to record jailhouse calls; it may be an highly compressed MP3 file made from the jailhouse recording. Typically, if the examiner determines that the file is a digitally compressed version of the original recording, the examiner will request that the original, uncompressed file be provided for the voice comparison examination.

In this scenario, both the unknown and the known recording would be processed to separate the various speakers and the data for each speaker would be extracted into a separate file containing only that speaker. The examiner would then determine if any enhancement of the audio is required before further processing is done. Once the files have been "purified" (all unwanted loud bangs, clicks, handling noises, etc., attenuated),

and the “purified” voice samples are of sufficient quality, the examiner will conduct the aural and spectrographic comparison. The final examination would be automated voice identification.

Currently automated comparison is performed in criminal cases solely for research purposes and only if the evidence is deemed sufficient for such an examination. The extracted voices of speakers other than the “unknown speaker” and the suspect would be included in the models used in the automatic system to help determine if any channel effects are influencing the outcome of the automated process.

The examiner uses aural, spectrographic, and automatic comparisons. Some forensic laboratories include forensic-linguistic and forensic-phonetic analyses. The examiner reaches a decision based on an appropriate combination of these methods for a given case. The finding is typically on a multiple level scale and includes rejection of inappropriate samples.

4.3 FBI Scenario 2: Multiple-to-Multiple Comparisons

NOTE: This scenario is usually associated with intelligence cases.

In this scenario, there are multiple “unknown” people on multiple intercepted recordings. The field agent requests voice comparisons be performed to determine which speakers are present in each of the recordings and if any speaker is present in more than one recording. Voice exemplars of the suspected “known” speakers are provided.

These unknown recordings are usually telephone calls, but may be taken from videos where the speaker is off camera. The format of these recordings is usually received in the form of a .wav file, a .wma file, a .wmv file, a DVD movie, or on audiocassettes. The same procedures used in FBI Scenario 1 are performed on each of the recordings and each recording is processed as a separate comparison. In this simple case, each of the “unknown” and “known” voices is compared to all of the other “unknown” and “known” voices. FAVIAU will generate a matrix of N (unknown speakers) vs. M (known speakers) giving a total of ((N+M) * (N+M-1))/2 comparisons.

Therefore, one request for a voice comparison examination may entail many speaker-to-speaker comparisons being performed. The report of results would contain the findings of which speaker is present in which recording and which speaker in each recording is the same speaker in a different recording. At this number of comparisons, it becomes rapidly almost impossible to conduct the human-based examinations – definitely calling for an automated approach

4.4 FBI Scenario 3: Variation on Multiple to Multiple Analyses

FBI scenario 2 can be divided into three more specific applications that are quite common today but imply very different working points for the technology.

1. A 1:N search where we have only one unknown file and N possible targets represented in recordings of known persons. This environment is “AFIS like”. The objective is to identify potential suspects through voice matches to support criminal investigations.

2. The N:1 search, where we have the known voice of one target (a drug dealer, for instance) and have many (sometimes thousands) of phone recordings of unknown persons and want to know which could possibly include the voice of the known target. This is a “monitoring” environment. The idea here is to filter the large number of phone recordings to send for further analysis only the ones with a high probability of containing the voice of the target. At that point endogenous information such as time and phone number can help further direct the analysts to prioritize the intercepts for further analysis.

3. A derivative of the example above but using a short list of targets to find in the phone recordings of known or unknown persons. This is a N:M comparison where the M must be tens of subjects but not much more than that.

4.5 Scenario 5: Parolee Case

Enrolling prisoners’ voices prior to parole release permits the criminal justice system to perform two voice biometric functions:

1. Use of speaker recognition software to track compliance with any court ordered home incarceration or other constraints using telephone verification and speaker verification.

2. Use of speaker recognition to identify parolees engaging in telephone conversations with known criminals or other convicted felons contrary to the conditions of the parole.

4.6 Scenario 6: Intelligence Case

A high profile terrorist releases an audiotape (a broadcast) in which he makes a threat of violence or a statement signaling his followers. An intelligence agency could compare the sample on the tape to known reference samples of the terrorist to determine the authenticity of the speaker’s claimed identity. The primary purpose of such a comparison is not to provide information for the news services but rather to determine any speaker related location, health, thought pattern information that might be

gathered from the signals, the words used, etc. See Appendix D for a detailed description of this use case.

4.7 Scenario 7: European Combined Analysis

The European Network of Forensic Science Institutes (ENFSI), which includes most of the Police Forensic Labs in Europe, created the Forensic Speech and Audio Analysis Working Group (FSAAWG) over ten years ago. The objective of the group was the creation of common speech and audio analysis standards for the forensic laboratories across Europe. In the last few years, a consensus has developed that the most complete and useful analytic workflow is the “Combined Analysis” (sometimes called the “New Combined Analysis” as there already exists an older version not including automated methodologies).

This Combined Analysis includes the use of the three “classical” approaches to forensic speaker analysis (perceptive, phonetic/linguistic, and acoustic/spectrographic) and tools that use automated Voice Biometric technology. An increasing number of European labs (about 10) are applying this new paradigm in their protocols.

Currently, fusion of the four dimensional result (Perceptive, Morph-linguistic, Acoustic/Spectrographic and Automatic) into a single identification decision is not unified because of the very different nature of the metrics expected in the four analytic approaches. Fusion methodologies also depend on the experience level and history of each laboratory. The use of automatic tools to support expert reports on voice sample matches is accepted in the Courts of some European countries. Laboratories in Spain, France, Italy, Germany, Finland, and Romania are using these tools and laboratories in the UK, Portugal, and the Netherlands are evaluating these tools and are preparing to use them, according to Antonio Moreno, a member of this committee.

In US Federal Courts the admissibility of scientific evidence is determined by the Federal Rules of Evidence (FRE) 702 [3]. As previously mentioned, automatic speaker recognition technology is not yet able to satisfy the FRE 702 criteria. Future research, evaluation, and progress are needed to advance speaker recognition technology to this level of acceptance.

5 Conclusions

As a result of the symposium the committee feels that the next logical strip is to develop some profile information for the use cases discussed. We have developed a taxonomy of use case descriptor as provided in Appendix B.

6 Specific Recommendations

The following recommendations are proposed for review by the Steering Committee:

1. This committee recommends that this document and its Appendices be developed further and converted into a survey instrument that can be circulated to agencies and departments that already perform voice-biometrics activities and those that would like to migrate to these technologies. This will help to prioritize requirements for improving ASR (automatic speaker recognition) technology

2. The FBI establish a Special Interest Group (SIG) on their law enforcement on-line (LEO) system for distribution of the survey as well as for sharing it with major agencies and departments in other countries.

3. This Committee work with CJIS and the Steering Committee to further develop use case information to include review of the responses to the survey.

4. Prototyping of voice sample collection (enrollment) as part of the arrest booking cycle be carried out to help determine the "best practices" guidance for the collection of voice biometrics by law enforcement agencies.

5. Develop a long-range scientifically based program in accord with the NAS report on "Strengthening Forensic Science in the United States" [7] for independent and rigorous scientific evaluation, including appropriate corpora, of the enabling technologies to support the FBI's voice biometrics mission. This should include creation of a scientific working group – voice to resolve the issues of procedures for voice comparison, training and certification of examiners.

7 References

[1] "Use Case", Wikipedia, http://en.wikipedia.org/wiki/Use_case.

[2] FRE 104. Preliminary Questions, available: http://www.law.cornell.edu/rules/fre/rules.htm#Rule104.

[3] FRE 702, Testimony by Experts, available: http://www.law.cornell.edu/rules/fre/rules.htm#Rule702.

[4] Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993), see: http://www.law.cornell.edu/rules/fre/ACRule702.htm.

[5] Campbell, J.P.; Shen, W.; Campbell, W.M.; Schwartz, R.; Bonastre, J.-F.; Matrouf, D "Forensic Speaker Recognition," *IEEE Signal Processing Magazine, Special Issue on Digital Forensics,* vol 26, issue 2, March 2009, p. 95-103, available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?isnumber=4806187&arnumber=4806209&count=23&index=13.

[6] Harmonized biometric vocabulary; ISO/IEC JTC 1/SC 37 N 3068, working draft 2009-02-28, http://isotc.iso.org/livelink/livelink?func=ll&objId=2299739.

[7] Committee on Identifying the Needs of the Forensic Sciences Community, National Research Council, "Strengthening Forensic Science in the United States: A Path Forward", National Academies Press, 2009, available: http://www.nap.edu/catalog.php?record_id=12589.

[8] Tabassi, E., Wilson, C.L., and Watson, C.I., "Fingerprint Image Quality", NISTIR 7151, August 2004, available http://fingerprint.nist.gov/NFIS/ir_7151.pdf

[9] The NIST Year 2008 Speaker Recognition Evaluation Plan (SREP), dated 3 April 2008, which is available on the NIST Information Technology Laboratory (ITL) web site at: http://www.itl.nist.gov/iad/mig//tests/sre/2008/sre08_evalplan_release4.pdf.

[10] *Information Technology:* American National Standard for Information Systems—Data Format for the Interchange of Fingerprint Facial, & Other Biometric Information – Part 1 ANSI/NIST-ITL 1-2007 Revision of ANSI/NIST-ITL 1-2000.

8 Appendix A: Biometrics Vocabulary

The following terms are defined to permit a common understanding of their use in this report. Table 1 documents the speaker recognition/biometrics vocabulary used in the report. The use of specific terms and definitions in this report does not imply official FBI or NIST endorsement of those definitions or use.

Term	Definition
Biometrics	Automated recognition of individuals based on their behavioral and biological characteristics
Forensic	Forensic means pertaining to law. In our use it has to do with applying scientific knowledge in a legal setting. A setting established by or founded upon official or accepted man-made rules rather than scientific facts.
Forensic application	Applied sciences/techniques used in conjunction with courtroom proceedings as evidential material, or with criminal investigations as an investigative tool.
Forensic Voice Biometrics	Biometric technology based on voice used by the forensic scientists/analysts to present evidence in the court of law, or to provide guidance for investigative purposes.
Identify	Voice forensic analysis that leads to the provable conclusion that a previously unknown sample is from a known speaker.
Investigative process	Any process by which police gather evidence about crimes or suspected crime through continued observation of persons or places. Wiretapping, eavesdropping, electronic observation, tailing, and shadowing are all examples of this type of law enforcement procedure.
Investigative voice biometrics	Biometrics technology based on voice, used by trained analysts to provide guidance for criminal investigations and for intelligence purposes.
Known sample	A voice sample collected from a subject where at least a minimum amount of biographic and situational information is available to identify the person to some level, not necessarily sufficient to know their name. Compare to Questioned Sample.
Non-target (impostor) speaker	A hypothesized speaker of a test segment who is in fact not the actual speaker. [9]
Non-target (impostor) trial	A trial in which the actual speaker of the test segment is in fact not the target (hypothesized) speaker of the test segment. [9]
Probe	A sample that is being searched against known speaker models.
Quality	A predictor of matcher performance before a matching algorithm is applied [8]
Questioned sample	A voice sample collected from one or more subjects where typically only situational information is available, such as the sample was collected on a wiretap of a specific phone number at a specific time.
Scenario	An account or synopsis of a possible course of action or events. Scenarios are written in plain language, with minimal technical details, so that stakeholders can have common examples, which can be used to focus their discussions.
Segment speaker	The actual speaker in a test segment. [9]
Subject of Interest (SOI) Models	Models built from multiple samples from a known speaker. Analogous in some senses to a composite master fingerprint file entry.

Term	Definition
Target (model) speaker	The hypothesized speaker of a test segment, one for whom a model has been created from training data. [9]
Target (true speaker) trial	A trial in which the actual speaker of the test segment is in fact the target (hypothesized) speaker of the test segment.[9]
Test	A collection of trials constituting an evaluation component. [9]
Verify an identity	Voice forensic analysis that leads to the provable conclusion that a previously question sample from an unknown speaker is from a speaker specifically thought to be the unknown speaker prior to the comparison.

TABLE 1 VOCABULARY USED IN THIS REPORT

9 Appendix B: Taxonomy of Use Cases

Use cases, our scenarios, fall into three classes: forensic, investigatory and surveillance. In the biometrics world they map into identification searches, verification of identity, and enrollment into the reference file, or corpus, of known speakers. The taxonomy of use cases should contain the appropriate attributes from the biometric class it best fits in. The following is organized by biometrics functionality.

A. **Identification:** Search for the identity of an unknown voice in a database of voices (criminals, suspects, watch listed persons, etc.) (1:N). AFIS like (true speaker identification)

   i. **Intercept/surveillance** (e.g., capture of cell phone traffic)

   - Single channel telephony (wireline, wireless, VoIP)
     - Human-human dialogues
     - Human-machines dialogues (e.g., IVR)
   - Multi-channel telephony (wireline, wireless, VoIP)
     - Human-human dialogues
     - Human-machines dialogues (e.g., IVR)
   - Non-telephony
     - Wired surveillance placement
     - Wireless surveillance placement
     - Shotgun distance/far field surveillance
     - Phased array audio capture from a distance/far field
   - Transmitted voice files (analog or digital)
     - Telephony (e.g., voice mail, answering machines)
     - Non-telephony recordings (e.g., from tape recording)

   ii. **Data mining:** based in the comparison of a matrix of targets/unknown voices (as described in FBI Example 2) (N:M)

B. **Verification** Forensic 1 to 1 verification in order to present an evidence in court or support an investigation

   i. Intercepted (e.g., targeted surveillance) recorded (e.g., blackmail threat) telephone call.

   ii. Secondary verification for identification operations listed above

C. **Enrollment**

i. Cooperative user

- Human-human interaction
- Human-machine interaction (e.g., Interactive Voice response System - IVR)

ii. Uncooperative user

- User is aware of enrollment
  - Human-human telephone
  - Human-machine telephone (IVR)
- User is unaware of enrollment
  - Booking station
  - Interview room at a police station, prison, or secondary inspection at a border control point.

10 Appendix C: Use Cases Specification

The following table can be used to specify speaker recognition use cases. It is desirable to describe each use case, to the extent possible, using the characteristics given in table X. As discussed in this report, generally, the less controlled, known, and matched the voice sample collections are, the greater the challenge for the technology, but successful application also depends on the application. For instance, with an uncontrolled, opportunistic intercept of a voice sample, where few of the audio characteristics are known, the likelihood of successful identification of a speaker (given multiple reference samples/models to search against) could be fairly low. Yet, in this same situation, however, an analyst might be able to successfully exploit speaker recognition scores for sorting applications or to verify that a sample is from a certain individual at a level sufficient for investigatory applications. Table X shows the characteristics to describe each use case.

Table X. Use Case Table.

Characteristic	Description
A. Challenge	
1. Problem statement	
2. Givens	a) Questioned: b) Known:
3. Objective (question(s) to be answered)	a) Determine
B. Audio Session Information (specify for the *known* and *questioned* samples)	
1. Sensor type (e.g., cell phone, wireline telephone, telephone intercept/tap, internal tape-recorder mic, internal digital-voice recorder mic, separate microphone, body/wire mic, covert room mic, laser vibrometer, accelerometer, fiber-optic stethoscope, unknown)	a) Questioned: b) Known:
2. Sensor placement (e.g., handset held close to mouth, desktop microphone 18” from lips, or unknown)	a) Questioned: b) Known:

3. Channel type and bandwidth (e.g., narrowband telephone, wideband broadcast TV, narrowband HF radio, cassette tape, digital audio tape, minidisc, microcassette, solid-state digital voice recorder)	a) Questioned: b) Known:
4. Channel conditions (e.g., clean, noisy, echo, dropouts, fading, etc.)	a) Questioned: b) Known:
5. Data a) File-based recordings (e.g., RIFF .wav or headerless) or streaming audio (e.g., Real…) b) Stream-based media - audio or audio/video (e.g., RealNetworks' RealAudio, streaming MP3, Macromedia's Flash and Director Shockwave, Macromedia/Adobe Flash Video H.263/H.264 VP6/HE-AAC, Microsoft's Windows Media Audio/Active Streaming Format, and Apple's QuickTime) c) Stream-based telephony VoIP (e.g., IP Phone, SIP Phone, Skype, AOL Voice Chat) d) Digital circuit switched (e.g., T1, T3, OC3, OC12…)	a) Questioned: b) Known:
6. Coding/compression (e.g., G.711 μ-law, G.711 A-law, GSM-EFR cellular voice coder, CELP voice coder, ACELP voice coder, G.726 ADPCM, G.722 split-band wideband ADPCM, MP2, MP3, AAC, MP4)	a) Questioned: b) Known:
7. Single channel (all talkers recorded on the same monaural channel) or multichannel (e.g., two talkers on separate stereo channels)	a) Questioned: b) Known:
8. Acoustic conditions (background noise and sounds - radio/TV/music, wind noise, background talkers, reverberation)	a) Questioned: b) Known:
9. Environment (e.g., home, office, car, outdoors, subway station, restaurant,	a) Questioned:

booking station, interrogation room)	b) Known:
10. Number of and durations of known samples and questioned samples	a) Questioned: b) Known:
11. Time span in between samples and range	a) Questioned: b) Known:
12. Additional information *(Note any mismatches between questioned and known samples' audio session information.)*	
C. Speaker Session Information (specify for the *known* and *questioned* samples)	
1. Style (e.g., spontaneous conversational telephone speech, face-to-face conversation, commands, read speech (what material?), question answering, broadcast speech, orated speech)	a) Questioned: b) Known:
2. Language(s)/dialects(s) spoken	a) Questioned: b) Known:
3. Speaker state (e.g., stress, emotion, mentally impaired, intoxicated, medicated)	a) Questioned: b) Known:
4. Cooperative or uncooperative	a) Questioned: b) Known:
5. Witting or unwitting	Questioned: Known:
6. Session data useful for processing this use case (e.g., date and time, telephone number, IP address, geographic location)	a) Questioned: b) Known:
7. Pointers to other sources that are typically linked to this kind of use case	a) Questioned: b) Known:

8. Additional information *(Note any mismatches between questioned and known samples' speaker session information.)*	
D. Speaker Information	
1. Speaker characteristics (e.g., name(s), sex, age/birth date, occupation, place of birth, place raised, race, ethnicity, years of education, native language/dialect, other language(s)/dialect(s), speech impairments/pathologies, social network)	
2. Additional information	

11 Appendix D: Use Case for Scenario 6: Intelligence Case

Characteristic	Description
A. Challenge	
4. Problem statement	A high profile terrorist releases an audiotape (a broadcast) in which he makes a threat or statement with the purpose of signaling his followers or raising an alarm in those who hear it through the distribution channels. An intelligence agency would compare the sample on the tape to known reference samples to determine the authenticity of the speaker's claimed identity. While these results are often discussed in public news broadcasts, the primary purpose is not to provide information for the news services, but rather to determine any speaker-related location, health, thought pattern information that might be gathered from the signals, the words used, and other factors.
5. Givens	a) Questioned: recording of a broadcasted audiotape from a claimed terrorist. b) Known: known reference samples of terrorist in question.
6. Objective (question(s) to be answered)	a) Determine the authenticity of the speaker's claimed identity. b) Determine threat or signal to followers (not biometric). c) Determine any speaker-related location, health, thought pattern information that might be

	gathered from the signals, the words used, and other factors (not biometric).
B. Audio Session Information (specify for the *known* and *questioned* samples)	
13. Sensor type (e.g., cell phone, wireline telephone, telephone intercept/tap, internal tape-recorder mic, internal digital-voice recorder mic, separate microphone, body/wire mic, covert room mic, laser vibrometer, accelerometer, fiber-optic stethoscope, unknown)	a) Questioned: unknown? b) Known:
14. Sensor placement (e.g., handset held close to mouth, desktop microphone 18" from lips, or unknown)	a) Questioned: unknown? b) Known:
15. Channel type and bandwidth (e.g., narrowband telephone, wideband broadcast TV, narrowband HF radio, cassette tape, digital audio tape, minidisc, microcassette, solid-state digital voice recorder)	a) Questioned: combined recorded broadcast audiotape channels. Bandwidth depends on combination of (unknown?) original audio tape, broadcast medium, and recording) b) Known:
16. Channel conditions (e.g., clean, noisy, echo, dropouts, fading, etc.)	a) Questioned: b) Known:
17. Data a) File-based recordings (e.g., RIFF .wav or headerless) or streaming audio (e.g., Real…) b) Stream-based media - audio or audio/video (e.g., RealNetworks' RealAudio, streaming MP3, Macromedia's Flash and Director Shockwave, Macromedia/Adobe Flash Video H.263/H.264 VP6/HE-AAC, Microsoft's Windows Media Audio/Active Streaming Format, and Apple's QuickTime) c) Stream-based telephony VoIP (e.g., IP Phone, SIP Phone, Skype, AOL Voice Chat) d) Digital circuit switched (e.g., T1, T3, OC3,	a) Questioned: b) Known:

OC12…)	
18. Coding/compression (e.g., G.711 μ-law, G.711 A-law, GSM-EFR cellular voice coder, CELP voice coder, ACELP voice coder, G.726 ADPCM, G.722 split-band wideband ADPCM, MP2, MP3, AAC, MP4)	a) Questioned: b) Known:
19. Single channel (all talkers recorded on the same monaural channel) or multichannel (e.g., two talkers on separate stereo channels)	a) Questioned: b) Known:
20. Acoustic conditions (background noise and sounds - radio/TV/music, wind noise, background talkers, reverberation)	a) Questioned: b) Known:
21. Environment (e.g., home, office, car, outdoors, subway station, restaurant, booking station, interrogation room)	a) Questioned: b) Known:
22. Number of and durations of known samples and questioned samples	a) Questioned: b) Known:
23. Time span in between samples and range	a) Questioned: b) Known:
24. Additional information *(Note any mismatches between questioned and known samples' audio session information.)*	
C. Speaker Session Information (specify for the *known* and *questioned* samples)	
9. Style (e.g., spontaneous conversational telephone speech, face-to-face conversation, commands, read speech (what material?), question answering, broadcast speech, orated speech)	a) Questioned: b) Known:
10. Language(s)/dialects(s) spoken	a) Questioned:

	b) Known:
11. Speaker state (e.g., stress, emotion, mentally impaired, intoxicated, medicated)	a) Questioned: b) Known:
12. Cooperative or uncooperative	a) Questioned: b) Known:
13. Witting or unwitting	a) Questioned: b) Known:
14. Session data useful for processing this use case (e.g., date and time, telephone number, IP address, geographic location)	a) Questioned: b) Known:
15. Pointers to other sources that are typically linked to this kind of use case	a) Questioned: b) Known:
16. Additional information *(Note any mismatches between questioned and known samples' speaker session information.)*	
D. Speaker Information	
3. Speaker characteristics (e.g., name(s), sex, age/birth date, occupation, place of birth, place raised, race, ethnicity, years of education, native language/dialect, other language(s)/dialect(s), speech impairments/pathologies, social network)	
4. Additional information	