Modeling Cryptographic Properties of Voice and Voice-Based Entity Authentication

Giovanni Di Crescenzo, Munir Cochinwala, Hyong S. Shim,
Telcordia Technologies
Piscataway, New Jersey, USA
{giovanni,munir,hyongsop}@research.telcordia.com

ABSTRACT

Strong and/or multi-factor entity authentication protocols are of crucial importance in building successful identity management architectures. Popular mechanisms to achieve these types of entity authentication are biometrics, and, in particular, voice, for which there are especially interesting business cases in the telecommunication and financial industries, among others. Despite several studies on the suitability of voice within entity authentication protocols, there has been little or no formal analysis of any such methods.

In this paper we embark into formal modeling of seemingly cryptographic properties of voice. The goal is to define a formal abstraction for voice, in terms of algorithms with certain properties, that are of both combinatorial and cryptographic type. While we certainly do not expect to achieve the perfect mathematical model for a human phenomenon, we do hope that capturing some properties of voice in a formal model would help towards the design and analysis of voice-based cryptographic protocols, as for entity authentication. In particular, in this model we design and formally analyze two voice-based entity authentication schemes, the first being a voice-based analogue of the conventional password-transmission entity authentication scheme. We also design and analyze, in the recently introduced bounded-retrieval model [4], one voice-and-password-based entity authentication scheme that is additionally secure against intrusions and brute-force attacks, including dictionary attacks.

Categories and Subject Descriptors

G.4 [**Mathematical Software**]: Algorithm Design and Analysis; F.2 [**Analysis of Algorithms and Problem Complexity**]: Miscellaneous; J.0 [**Computer Applications**]: General

General Terms

Security, Human Factors, Algorithms, Theory

DIM'07, November 2, 2007, Fairfax, Virginia, USA.

Keywords

Modeling Human Factors, Entity Authentication, Biometrics, Voice

1. INTRODUCTION

One cornerstone in the design of an identity management architecture is a suitable solution to the entity authentication problem. In light of frequent criticisms to both the usability and security inadequacies of simple password schemes, both practitioners and researchers have advocated the use of more powerful authentication mechanisms, frequently called

multi-factor entity authentication (where users combine one factor from any few among different categories, such as "something you know", "something you are", "something you have", etc.) or strong entity authentication (where users include a large enough number of authenticating factors, possibly from the same or different category). As an example, the well-known "single-sign-on" approach to solve the scalability of single-entity multiple-authentication in a multiple-service identity management architecture is naturally realized with the "once and for all" use of strong and multi-factor authentication methods.

In this context the use of biometrics, a typical factor from the "something you are" category, raises a number of technical usability and security challenges, since every form of biometrics achieves its own tradeoff between security and usability, as acknowledged by the current research literature. (See, e.g., [14, 11] for introductory overviews on the various properties of biometrics for user authentication.) Additionally, the use of certain biometrics, such as voice, in the contest of entity authentication opens up especially interesting business cases for identity management in the telecommunication industry and not only, as we later explain in greater detail.

In this paper we focus on voice as it seems the easiest biometric factor to use and deploy and perform comprehensive security modeling and analysis of voice (as a human phenomenon) and voice-based entity-authentication protocols. In other words, we attempt to bypass the usability-security tradeoff of biometrics, by choosing the biometric factor with the highest usability properties, and advancing the state of the art on its security properties.

Previous research work. There is a large amount of research on the solution of various security problems using biometrics. A significant number of papers deals with inherent reproducibility difficulties of biometrics. One active area in the cryptography literature studies key generation, a cornerstone among the various cryptographic tasks, from

fuzzy data (see, e.g. [5]). Another well-studied problem is about safeguarding the privacy of biometric data (see, e.g. [16, 12, 13]). Despite several studies on the suitability of voice within entity authentication protocols, they seem to have escaped rigorous analysis, and we are not even aware of any attempt of defining a formal model in which such methods could be analyzed.

Our technical contribution. In this paper we embark into formal modeling of seemingly cryptographic properties of voice. The goal is to define a formal abstraction for voice, in terms of algorithms with certain properties, that are of both combinatorial (e.g., density, accuracy) and cryptographic type (e.g., hardness of voice impersonation, or voice extrapolation). Naturally, achieving the perfect mathematical model for a human phenomenon is not our goal. However, we do hope that capturing some properties of voice in a formal model would help towards the design and analysis of voice-based cryptographic protocols, as for entity authentication. We show examples for this, by designing and analyzing two voice-based entity authentication schemes, and one voice-and-password scheme.

The first scheme is a natural voice-based adaptation of the UNIX-like password transmission scheme. As the first scheme only allows a number of authentication sessions linear in the amount of voice samples registered during the registration phase, we investigate the problem of removing this drawback.

Our second scheme extends the first scheme using a type of cover-free families, that we call implicitly-samplable. For this scheme, we can prove that any polynomial number of authentication sessions are possible, without violating any correctness or security property.

Finally, we conclude by showing a voice-and-password-based entity authentication scheme that is secure against intrusion and brute-force attacks, including dictionary attacks, in the recently introduced bounded retrieval model [4]. This is obtained by combining our first protocol with a password scheme from [4] that is secure against dictionary attacks in the bounded-retrieval model.

We conclude with a brief description of our company's identity management architecture.

The business case. For almost the entire history of the communication access technologies, information technology and broadcast services and have been separate and voice was provided over fixed or mobile networks, each subject to distinct regulatory requirements. Data was carried over separate data networks and internet access provided by internet service providers, again with their own discrete regulatory regimes. Video services were delivered by broadcasters, with different sets of regulations applying to cable, satellite and terrestrial TV networks.

Advances in technology have changed all this. The introduction of IP networks has transformed service provision, lowering the barriers to entry for new service providers and introducing new service paradigms. Voice services are now being delivered commercially over the internet. Voice over IP (VoIP) is no longer dismissed as a second rate service for geeks. VoIP is the classic example of an Internet application resulting from the separation of transport from services.

However, the internet world is still far from a trusted environment, wrought with security and privacy challenges. It's a world in which end user expectations are that services and information are generally free. In the telecom world, however, end users are accustomed to paying for service. In return they expect, and receive, high quality of service levels. In general the telecom world is also regarded as a more trusted environment. Mobile networks in particular are known for their strong encryption and authentication procedures.

All telecom networks have grown from a base of spam-free, high quality voice. Network operators have the ability to authenticate end users, increasing their advertising worth and reducing fraud for content providers. Mobile network operators with SIM based systems can provide strong authentication without log-in.

This existing climate of trust creates a brief window of opportunity for telecommunications operators. In particular, telecom operators can use their services and capabilities to provide authentication. Telecom operators have the ability to use voice recognition along with SIM card authorization. This allows ubiquitous access via any voice-capable mobile device. The proliferation of mobile devices, the trust in operators by customers as well as the usability provided by voice-enabled communication and subsequent authentication provides a powerful confluence of forces and sound business motivation for operators to provide authentication across diverse networks.

2. MODEL AND FORMAL DEFINITIONS

In this section we present our modeling of voice by formally defining first a 'voice abstraction' and then a 'voice cryptographic primitive'. We then present formal definitions for voice-based entity authentication.

Defining a voice abstraction. We would like to abstract the human process of producing voice samples as a mathematical object. Towards this goal, we focus on a number of parameter dependencies and basic properties of voice, and then capture these in formal definitions.

Input, Output and Parameter dependency. While a voice sample could consist of anywhere from a single letter to a long speech, it will be convenient to restrict to a single word. Given an *alphabet* Σ (e.g., the english alphabet), and a *dictionary* $D \subseteq \Sigma^*$, we define a *word* as an element in dictionary D. Also, let S be a set of voice signals. As a first attempt, one might define voice as a map from words to signals. For any voice sample, the associated voice signal may depend on a large number of factors which we group as follows:

- the human that produces it (this includes all factors involved in how a human vocalizes speech, such as jaw, tongue and vocal cords movement, as well as other human-related factors that affect voice, such as whether the human is female or male, and her/his age);
- which time it is produced at (this incorporates dependency on factors such as possible variations on the human's health condition, as well as variations in a human's voice at different times of the day);
- under which noise conditions it is produced (this includes noise conditions on both the site where the voice sample is produced and the communication line between the speaker and the listener or receiver).

Summarizing these considerations, we can model voice as a map with the following properties: it maps words from a dictionary into voice signals; it is parameterized by a (unique)

identity and a current time; and it is probabilistic (to model noise). Formally, we define the *voice sampling algorithm* as a probabilistic algorithm $VS_{id,ct}$ with inputs in D and outputs in S, where id denotes the identity of the human that produces the voice sample, and ct denotes the current time.

Another important factor is the representation of the voice signal. The area of digital speech processing studies speech signals and the methods to process such signals into digital format. This area is the object of several books, and surveying its results is out of scope for this paper. Here, it suffices to assume the existence of any such method that, given a voice signal, returns a digital measurement of it, that we represent as an r-tuple of elements in some metric space M, with distance function $dist$. Formally, we define the *voice representation algorithm* as a deterministic algorithm VR with inputs in S and outputs in M^r, computable in time polynomial in a parameter n. We also call an element returned by VR a *voice value*.

We also would like to formally capture the following two basic properties of voice:

(a) two voice samples produced by the same human, using different words, and similar times (in fact, even using the same time parameter) result in 'sufficiently distinct' voice values;

(b) two voice samples produced by the same human, using the same word, but at different times, result in 'sufficiently similar' voice values.

We note that both properties are experimentally verified by using algorithms from the speech processing and speech recognition areas. The following definition restates these properties as properties of the defined algorithms VS, VR.

Definition 1. *Let VS and VR be a voice sampling and a voice representation algorithm, respectively, and let δ_d, δ_a be positive real values. We say that pair (VS, VR) is a* voice abstraction *with parameters (δ_d, δ_a) if the following properties hold:*

- *(δ_d-*density*:) for any id, time t and words $w_1, w_2 \in D$ such that $w_1 \neq w_2$, it holds that $dist(\vec{v}_1, \vec{v}_2) \geq \delta_d$, where $\vec{v}_i = (v_{i1}, \ldots, v_{ir}) = VR(VS_{id,t}(w_i))$, for $i = 1, 2$.*

- *(δ_a-*accuracy*:) for any id, word w and times t_1, t_2 such that $t_1 \neq t_2$, it holds that $dist(\vec{v}_1, \vec{v}_2) \leq \delta_a$, where $\vec{v}_i = (v_{i1}, \ldots, v_{ir}) = VR(VS_{id,t_i}(w))$, for $i = 1, 2$.*

Modeling voice as a cryptographic primitive. We now start considering two properties of voice that appear to have some cryptographic nature.

The first property, which we call impersonation hardness, says that it seems hard to find a human producing voice samples that are, in some sense, indistinguishable from another human's voice samples. We note that this property is believed to be true in particular when the indistinguishability notion is interpreted to hold with respect to algorithms from the speech processing and speech recognition areas. In our framework, we formalize the fact that given a voice sampling algorithm with an identity parameter, it is hard for an adversary to find another identity parameter, such that the resulting voice sampling algorithm returns outputs that are similar to the original one. This would be true even if the adversary is given oracle access to the original voice sampling algorithm.

The second property, which we call extrapolation hardness, says that it seems hard to extrapolate the digital representation of a human's voice sample given some digital representations of distinct voice samples from the same human. This would be true even if the adversary is given oracle access to the original voice sampling algorithm.

Definition 2. *Let pair (VS, VR) be a voice abstraction with parameters (δ_d, δ_a), and let ϵ_i, ϵ_e be positive real values. We say that (VS, VR) is a* voice cryptographic primitive *with parameters $(\delta_d, \delta_a, \epsilon_i, \epsilon_e, q)$ if the following properties hold:*

- *((ϵ_i, q)-*impersonation hardness*:) for any identity id and any adversary algorithm A making at most q queries to its oracle, the experiment $ImpExp_{VS,VR,A}$, defined below, returns 1 with probability $\leq \epsilon_i$.*

  $ImpExp_{VS,VR,A}(id, \delta_a)$:

  1. *$(id', t_1, t_2, w_1, w_2) \leftarrow A^{VS_{id,ct}(\cdot)}(id, \delta_a)$*
  2. *let $\vec{v}_1 = VR(VS_{id,t_1}(w_1))$*
  3. *let $\vec{v}_2 = VR(VS_{id',t_2}(w_2))$*
  4. *if $id \neq id'$ and $dist(\vec{v}_1, \vec{v}_2) \leq \delta_a$ then*
     return: *1*
     else **return:** *0.*

- *((ϵ_e, q)-*extrapolation hardness*:) for any identity id and any adversary algorithm A making at most q queries to its oracle, the experiment $ExtExp_{VS,VR,A}$, defined below, returns 1 with probability $\leq \epsilon_e$.*

  $ExtExp_{VS,VR,A}(id, \delta_a)$:

  1. *$\vec{v} \leftarrow A^{VS_{id,ct}(\cdot)}(id, \delta_a)$*
  2. *let $(w_1, \ldots, w_q)$ be the words queried by A to VS*
  3. *let $\vec{v}_i = VR(VS_{id,t}(w_i))$, for $i = 1, \ldots, q$*
  4. *if $dist(\vec{v}, \vec{v}_i) > \delta_a$ for $i = 1, \ldots, q$ then*
     if $\exists w \in D$ s.t. $\vec{v} = VR(VS_{id,t}(w))$ then
     return: *1*
     else **return:** *0.*

Voice-based Entity Authentication in the Standard Model. The task of voice-based entity authentication can be considered analogue to the task of password-based entity authentication, with the only difference that instead of using a password, the entity uses values returned by a voice algorithm. Our model, which we now describe, follows this analogy.

Entities, connectivity, resources. We consider an arbitrary system (or network) containing a number of resources. We would like to allow any of the *users*, denoted as $U_1, U_2, \ldots, U_n$, for some integer n, to access this system locally or remotely through a voice-based authentication protocol verified by a *server* S. The server's storage area contains an *access file*, that we denote as F, with m locations, where we denote as $F[i]$ the content of the i-th location of F.

Subprotocols and Phases. A voice-based entity-authentication protocol can be divided into five main algorithms: a voice sampling algorithm $VS_{id,ct}$ and a voice representation algorithm VR, as defined before, and then a *setup* algorithm, a *registration* algorithm and a *verification* algorithm, which we now define. A setup algorithm, that we denote as SET, is only run by the server. On input a security parameter λ in unary, algorithm SET returns an m-location access file F, for some $m = \text{poly}(\lambda)$ in time at most polynomial in λ. The registration algorithm, denoted as REG, is a possibly probabilistic polynomial time (in λ) algorithm that takes as input a user's identity id, the access file F, and interacts with the voice sampling algorithm $VS_{id,ct}(\cdot)$ as an oracle, returning at the end an updated access file F. A verification algorithm VER is run by the server S, takes as input an identity id and the access file F, and interacts with a voice sampling algorithm $VS_{id,ct}(\cdot)$ as an oracle. At the end, S returns *accept* (briefly, 1) or *reject* (briefly, 0), according to whether the user has been positively identified or not.

We will denote a *voice-based entity-authentication protocol* as the 5-tuple of probabilistic algorithms

$$\mathcal{P} = (VS, VR, \text{SET}, \text{REG}, \text{VER}),$$

and we will assume, for simplicity, that an execution of $\mathcal{P}$ can be divided into three phases: first, an *initialization phase*, where the server runs the setup algorithm; then, a *registration phase*, where each among the n users $U_1, \ldots, U_n$ provides the voice oracle for algorithm REG with server S; finally, an *identification phase*: at any time, any user U_i can provide the voice sampling algorithm $VS_{id_i,ct}$ used by the verification algorithm VER. We denote as $Param$ the list of parameters (represented in unary) associated with $\mathcal{P}$, that can be any subset among: the *number of users* n, the *number of locations* m in the access file, and the *security parameter* λ, and any bound on the adversary's power q (this is defined more precisely later).

Correctness requirement. A basic requirement we expect from a voice-based entity-authentication protocol is that, at any time, a server positively identifies previously registered users.

DEFINITION 3. *Let $\mathcal{P} = (VS, VR, \text{SET}, \text{REG}, \text{VER})$ be a voice-based entity-authentication protocol with parameters $Param = (n, m, \lambda)$. The* correctness *requirement for $\mathcal{P}$ is as follows: for each $i \in \{1, \ldots, n\}$, and any identity id_i, it holds that*

$$Prob \quad [\quad F \leftarrow \text{SET}(Param); \quad F \leftarrow \text{REG}^{V_{id_i,ct}(\cdot)}(Param, id_i, F) : \quad \text{VER}^{V_{id_i,ct}(\cdot)}(Param, id_i, F) = 1 \ ] = 1.$$

Security against entity impersonation. In defining this requirement, for simplicity, we assume that the communication channel between each user and the server is not subject to integrity or tampering attacks. We note that the model in which this link is also subject to adversarial attacks is of orthogonal focus and can be separately studied (but will not be studied in this paper). Thus, we consider the security of a voice-based entity authentication protocol against an active adversary that can monitor and/or play as the verification algorithm in up to a number q of protocol executions and then attempt to impersonate one of the users.

DEFINITION 4. *Let $\mathcal{P} = (VS, VR, \text{SET}, \text{REG}, \text{VER})$ be a voice-based entity-authentication protocol with parameters $Param = (n, m, \lambda, q)$, and let $\epsilon \in$ [0,1]. We say that $\mathcal{P}$ is* (q, ϵ)-secure against impersonation *if the experiment $IExp_{\mathcal{P},A}$ returns 1 with probability at most ϵ, where $IExp_{\mathcal{P},A}$ is defined as follows:*

$IExp_{\mathcal{P},A}(Param, \{id_i\}_{i=1}^n)$:

1. *let* $F \leftarrow \text{SET}(Param)$
2. *for* $i = 1, \ldots, n$,
   $F \leftarrow \text{REG}^{VS_{id_i,ct}(\cdot)}(Param, id_i, F)$
3. *set* $qc = 0$ *and* $aux = (id_1, \ldots, id_n)$
4. *repeat*
   let $(j, aux) \leftarrow A(aux)$
   let $aux \leftarrow A^{VS_{id_j,ct}(\cdot)}(aux)$
   let $qc \leftarrow qc + 1$
   until $qc = q$
5. $id \leftarrow A(aux)$ *then*
6. *if* $\text{VER}^{A(aux)}(id, F) = 1$ *then*
   return: *1*
   else **return:** *0.*

Remarks. We note that in this definition we allow the experiment to depend on parameter q for greater generality of definition. If the total number of authentication sessions is at most q, then the adversary is allowed to monitor or actively participate in all sessions. If this number is larger than q, then this definition models the realistic scenario where an adversary can only monitor or actively participate in some of the authentication sessions that a user may be involved in.

3. VOICE-BASED ENTITY AUTHENTICATION PROTOCOLS

The purpose of this section is to present two basic constructions of voice-based entity authentication protocols that can be proved to be secure using our formal definition of a voice cryptographic primitive. The first scheme, which we call the *single voice sample transmission protocol*, can be seen as a simple variation of the conventional password-transmission authentication protocol and uses the ability of choosing a new voice sample that has large enough distance from all previous ones, which follows from the density property of voice sampling algorithms. The second scheme, which we call the *multiple voice sample transmission protocol*, is additionally based on the existence of certain cover-free families which enjoy a special implicit sampling property.

3.1 The single voice sample transmission protocol

We first informally describe this protocol and then formally describe its algorithms and properties.

Informal description. The single voice sample transmission protocol uses as a starting point a voice-analogue of the password-transmission protocol for password-based entity authentication, where the user simply transmits her password to the verifying server, that checks whether the received password coincides with the previously registered password. The voice-analogue version that we present here is especially interesting because of the observation that for a human it is much more convenient to produce multiple voice samples for given words than generating and remembering multiple high-entropy passwords. Accordingly, at registration phase, a user is asked to register multiple voice samples that are sufficiently distant from each other; then, at login phase, the user is asked to produce one voice sample that was not produced before. We can prove that this protocol satisfies correctness, using the density and accuracy properties of voice abstractions, and security against entity impersonation, using the hardness properties of voice cryptographic primitives. We believe that being able to prove the security of such a simple protocol from apparently natural properties of voice, such as those used for our model, gives evidence of the soundness of the introduced voice model.

Formal description. We can then formally describe protocol $\mathcal{P}_1 = (VS, VR, \text{SET}, \text{REG}, \text{VER})$, by using an arbitrary voice cryptographic primitive (VS, VR) with parameters $(\delta_d, \delta_a, \epsilon_i, \epsilon_e, q)$, and by defining the following algorithms.

Algorithm SET. This algorithm just initializes the access file F.

Algorithm REG. The registration algorithm takes several words such that the associated values have sufficiently large distance from each other. Then, it stores them in the access file F, along with a flag for each voice value denoting whether the value has been already used for authentication or not. Formally, on input $Param, id, F$, and given oracle access to algorithm $VS_{id,ct}(\cdot)$, algorithm REG runs the following steps (where ct denotes the current time):

1. For $i = 1, \ldots, \lfloor m/n \rfloor$, randomly choose word $w_{id,i} \in D \setminus \{w_{id,1}, \ldots, w_{id,i-1}\}$ until it holds that
$$dist(VR(VS_{id,ct}(w_{id,i})), VR(VS_{id,ct}(w_{id,j}))) \geq \delta_d$$
for all $j = 1, \ldots, i-1$; if such a word cannot be found within λ attempts then return: $\perp$ and halt.
2. For $i = 1, \ldots, \lfloor m/n \rfloor$, set $uflag_i = 0$ and store record $(id, i, ct, w_{id,i}, v_{id,ct,i}, uflag_i)$ into F.
3. Return: F.

Algorithm VER. The verification algorithm interacts with the voice sampling oracle $VS_{id,ct}(\cdot)$ as follows. It randomly chooses an unused record from F associated with identity id and asks the oracle for a related voice value. Finally, it checks that the received value is sufficiently close to the voice value stored during the registration phase. Formally, on input $Param, id, F$, and given oracle access to the voice algorithm $VS_{id,ct}(\cdot)$, algorithm VER runs the following steps:

1. Randomly choose from F a record
$$(id', i, t, w_{id',i}, v_{id',t,i}, uflag_i)$$
such that $id' = id$ and $uflag_i = 0$.
2. Ask query $w_{id',i}$ to oracle $VS_{id,ct}(\cdot)$, where ct denotes the current time.
3. Let $vs_{id',i}$ be the response and let $v = VR(vs_{id',i})$.
4. Check if $dist(v, v_{id',t,i}) \leq \delta_a$.
5. If yes then set $uflag_i = 1$ and return: 1; else return: 0.

Properties of the above protocol. We would like to prove two properties for $\mathcal{P}_1$: first, a bounded form of correctness based on the existence of a voice abstraction that satisfies δ_d-density and δ_a-accuracy; second, (q', ϵ)-security against impersonation, for some $q' < \lfloor m/n \rfloor$ and some $\epsilon > 0$, under the assumption that there exists a voice cryptographic primitive with parameters $(\delta_d, \delta_a, \epsilon_i, \epsilon_e, q)$ such that $\delta_a < \delta_d$.

Correctness. We would like to prove that $\mathcal{P}_1$ satisfies Definition 3 if every user performs at most $\lfloor m/n \rfloor$ authenticating sessions. First, assume that algorithm REG does not halt in step 1. Then, by the δ_a-accuracy of (VS, VR), it holds that the test in step 4 of algorithm VER is met with probability 1. Then, the probability that $\mathcal{P}_1$ does not satisfy correctness is at most the probability that REG halts in step 1. If every user performs at most $\lfloor m/n \rfloor$ authenticating sessions, then by the δ_d-density of (VS, VR), this happens with probability 0. (Otherwise, algorithm REG halts in step 1 and the test in step 4 of algorithm VER is not met.)

Security against impersonation. We need to prove that $\mathcal{P}_1$ satisfies Definition 4.

Assume towards contradiction that there exists an adversary algorithm A for which experiment $\text{IExp}_{\mathcal{P}_1,A}$ returns 1 with probability $> \epsilon$. This means that A makes VER return 1, and, in particular, that the test in step 4 of algorithm VER is correctly passed. We distinguish two cases, according to which of these two events happen:

(a) $id = id_j$ for some j returned by A in step 4 of experiment $\text{IExp}_{\mathcal{P}_1,A}$;

(b) $id \neq id_j$ for all j returned by A in step 4 of experiment $\text{IExp}_{\mathcal{P}_1,A}$.

In case (a), we can prove that algorithm A can be used to construct an algorithm B that violates the (ϵ_e, q)-extrapolation hardness of the voice cryptographic primitive (VS, VR), for $\epsilon_e = \epsilon/n$.

More formally, algorithm B, on input id, δ_a and interacting with $VS_{id,ct}(\cdot)$ as an oracle, does the following:

1. randomly choose $\ell \in \{1, \ldots, n\}$ and set $id_\ell = id$;
2. generate algorithm $V_{id_i,ct}(\cdot)$, for $i \in \{1, \ldots, n\} \setminus \{\ell\}$;
3. set $qc = 0$, $aux = (id_1, \ldots, id_n)$ and start running algorithm A on input (aux);
4. repeat
   if A makes a query w to $VS_{id_j,ct}(\cdot)$ then
   set $qc = qc + 1$
   if $j = \ell$ then

make query w to oracle $VS_{id,ct}(\cdot)$

let ans be its answer

otherwise set $ans = VS_{id_j,ct}(w)$

send ans as a reply to A

let aux be A's output

until $qc = q$.

5. let $id' \leftarrow A(aux)$
6. if $id' \neq id$ then return: $\perp$ and halt
7. if $\text{VER}^{A(aux)}(id, F) = 1$ then

let s be the voice signal returned by $A(aux)$

when queried by VER

let $\vec{v} = VR(s)$ and return: $\vec{v}$.

We note that by construction B attempts to perfectly simulate A's view. In particular, we note that B succeeds to perfectly simulate answers to oracle $VS_{id_j,ct}$, for $id_j \neq id$ by generating its own voice sampling algorithms and use them as oracles $VS_{id_j,ct}$, for $id_j \neq id$. We now distinguish two subcases, according to whether A's output after its q queries is equal to B's input id, or not.

In the former subcase, B is successful in violating extrapolation hardness of (VS, VR), as by construction of algorithm VER, the voice signal s returned by A is associated with a previously unused word and thus, by construction of REG, it satisfies

$$dist(VR(s), VR(VS_{id,ct}(w_i))) > \delta_d > \delta_a,$$

for $i = 1, \ldots, q$.

In the latter subcase, B is not successful but this does not happen with probability at least $1/n$, from which we obtain that $\epsilon_e \geq \epsilon/n$.

Case (b) is proved by using algorithm A to construct an algorithm C that violates the (ϵ_i, q)-impersonation hardness of the voice cryptographic primitive (VS, VR), for some $\epsilon_i > 0$.

More formally, algorithm C, on input id, δ_a and interacting with $VS_{id,ct}(\cdot)$ as an oracle, does the following:

1. randomly choose $\ell \in \{1, \ldots, n\}$ and set $id_\ell = id$;
2. generate algorithm $V_{id_i,ct}(\cdot)$, for $i \in \{1, \ldots, n\} \setminus \{\ell\}$;
3. set $c = 0$, $aux = (id_1, \ldots, id_n)$ and start running algorithm A on input (aux);
4. repeat

if A makes the $(c+1)$-th query (w_c, t_c) to $VS_{id_j,t_c}(\cdot)$ then

set $c = c + 1$

if $j = \ell$ then

make query (w_c, t_c) to oracle $VS_{id,t_c}(\cdot)$

let ans be its answer

otherwise set $ans = s_{\ell,c} = VS_{id_j,t_c}(w)$

send ans as a reply to A

let aux be A's output

until $c = q$.

5. let $id' \leftarrow A(aux)$
6. if $id' \in \{id_1, \ldots, id_n\}$ then return: $\perp$ and halt
7. if $\text{VER}^{A(aux)}(id, F) = 1$ then

let (w, t) be VER's query to $VS_{id',t}(\cdot)$

let s be the voice signal returned on this query

if there exists c such that

$$dist(VR(s_{\ell,c}), VR(s)) \leq \delta_a,$$

then return: (id', t_c, t, w_c, w) and halt

8. return: $\perp$ and halt.

As for B, we note that by construction C attempts to perfectly simulate A's view. We now distinguish two subcases, according to whether C's output is $\neq\perp$ or $=\perp$.

In the former subcase, it holds that $\text{VER}^{A(aux)}(id, F) = 1$ and there exists c such that

$$dist(VR(s_{\ell,c}), VR(s)) \leq \delta_a.$$

Thus C is successful in violating impersonation hardness of (VS, VR) if A is successful in breaking the impersonation security of the voice-based authentication protocol $\mathcal{P}_1$.

In the latter subcase, three events might have happened. First, it could happen that $id' \in \{id_1, \ldots, id_n\}$; actually this does not happen by our original definition of the case (b) that we are analyzing. Second, it could happen that $\text{VER}^{A(aux)}(id, F) \neq 1$; however, by our assumption that A violates the impersonation security of the voice-based authentication protocol $\mathcal{P}_1$, this does not happen with probability at least ϵ. This, it could happen that $\text{VER}^{A(aux)}(id, F) \neq 1$ and there is no c such that

$$dist(VR(s_{\ell,c}), VR(s)) \leq \delta_a.$$

However, this does not happen for at least one value of $\ell \in \{1, \ldots, n\}$, and thus does not happen with probability at least $1/n$. Then we obtain that $\epsilon_i \geq \epsilon/n$.

We finally have the following

THEOREM 5. *If there exists a voice cryptographic primitive with parameters $(\delta_d, \delta_a, \epsilon_i, \epsilon_e, q)$ such that $\delta_d > \delta_a$, then protocol $\mathcal{P}_1$ is a voice-based entity-authentication protocol that satisfies correctness (if every user performs at most $\lfloor m/n \rfloor$ authenticating sessions) and (q', ϵ)-security against impersonation, where $q' < \lfloor m/n \rfloor$ and $\epsilon < \min\{n \cdot \epsilon_e, n \cdot \epsilon_i\}$.*

3.2 The multiple voice sample transmission protocol

The previous protocol satisfied correctness only if the number of authentication sessions is at most linear in the number of voice samples stored during the registration phase. In this section we show a protocol that satisfies correctness for any polynomial number of authentication sessions, by assuming a known upper bound on the number of authentication sessions in which the adversary is active. We first informally describe this protocol and then formally describe its algorithms and properties.

Informal description. The multiple voice sample transmission protocol is best explained as a direct extension, using a certain variant of cover-free set families [8, 7], of the single voice sample transmission protocol. (Cover-free set families have been used in various areas including information theory, combinatorics, and recently in various areas of

cryptography, including encryption; see, e.g. [1, 9].) Specifically, at login phase, instead of producing a single voice sample, the user will produce a subset of the samples registered during the registration phase. For any given user, the subset chosen is a previously unused subset from the cover-free family, so that a cover-freeness property holds over the set of all voice values registered by this user. It turns out that conventional cover-free families are not sufficient to reach our goal as they would only guarantee a fixed polynomial number of authenticating sessions. Instead, we define an apparently new property for such families, which we call *implicit sampling*, saying that it must be possible to efficiently generate a random subset in the family (as opposed to being able to efficiently generate the entire family, which may not be possible when the family has size not bounded by a given polynomial). We note that not all cover-free families are implicitly-samplable [6], but some known constructions, for instance, from [9], are so.

A useful tool: implicitly-samplable cover-free families. We recall the formal definition of cover-free families, define the notion of implicit sampling, and recall a construction that satisfies this notion and can be used in our protocol.

DEFINITION 6. Implicitly-samplable cover-free families. *Let d, q, k be positive integers, let G be a ground set of size d, and let $F = \{S_1, \ldots, S_k\}$ be a family of subsets of G. We say that subset S_j does not* cover *S_i if it holds that $S_i \not\subseteq S_j$. We say that family F is q-*cover free *over G if each subset in F is not covered by the union of q subsets in F. Moreover, we say that family F is t-*uniform *if all subsets in F have size t. Finally, we say that family F is* implicitly-samplable *if there exists an algorithm Gen running in time polynomial in $|G|$ that generates a random element from F.*

We will use the following fact (this is a simplified version of Lemma 4 from [9], where we additionally observe that their construction is also implicitly-samplable).

LEMMA 7. *[9] For any positive integers q, k, there exists an implicitly-samplable t-uniform family F containing k subsets of a ground set of size d, such that:*

- *F is q-cover-free with probability at least $1 - \epsilon$*
- *$d = \Omega(k^{(q+1)/t}/\epsilon)$*
- *each S_i is generated by randomly and independently choosing t elements from the ground set, for $i = 1, \ldots, k$.*

We note that it is possible to set ϵ to be negligible in λ (i.e., the probability that F is not q-cover-free is negligible), and to set k equal to an arbitrarily large polynomial in λ (i.e., in our scheme this will imply that the number of secure authenticated sessions can be an arbitrary polynomial), and still have feasible values for the remaining parameters. As an example, for any q, we can pick $t = O((q+1)(\log k/\log\log k)(\log 1/\epsilon))$ and have a value of d that is only logarithmic in λ.

Formal description. We can then formally describe protocol $\mathcal{P}_2 = (VS, VR, \text{SET}, \text{REG}, \text{VER})$ by using an arbitrary voice cryptographic primitive (VS, VR) with parameters $(\delta_d, \delta_a, \epsilon_i, \epsilon_e, q)$, by defining algorithms SET and REG precisely as in $\mathcal{P}_1$, and by defining the remaining algorithm VER as follows.

Algorithm VER. The verification algorithm randomly samples a previously unused subset from an implicitly-samplable t-uniform q-cover-free family of subsets, and queries the voice sampling algorithm with the words from the ground set that are elements of this subset. Upon using the voice representation algorithm to compute each value corresponding to each received voice sample, it checks that this value is sufficiently close to the value stored during the registration phase.

Formally, on input $Param, id, F$, and given oracle access to the voice sampling algorithm $VS_{id,ct}(\cdot)$, algorithm VER runs the following steps:

1. Sample a random subset S_i from q-cover-free family F until $uflag_i = 0$;

   if such a subset cannot be found within λ attempts then

   return: $\perp$ and halt.

2. For each index $j \in S_i$,

   read record $(id, j, t, w_{id,j}, v_{id,t,j})$ from F

   compute $v = VR(VS_{id,ct}(w_{id,j}))$,

   where ct denotes the current time,

   check if $dist(v, v_{id,t,j}) \leq \delta_a$.

3. If all checks are satisfied then

   set $uflag_i = 1$ and return: 1;

   else return: 0.

Properties of the above protocol. The proof that $\mathcal{P}_2$ satisfies correctness and security against impersonation under the assumption that there exists a voice cryptographic primitive is obtained as an extension of the analogue proof for $\mathcal{P}_1$. We obtain the following

THEOREM 8. *If there exists a voice cryptographic primitive with parameters $(\delta_d, \delta_a, \epsilon_i, \epsilon_e, q)$ such that $\delta_d > \delta_a$, then protocol $\mathcal{P}_2$ is a voice-based entity-authentication protocol that satisfies correctness and (q', ϵ)-security against impersonation, where $q' \leq q$ and $\epsilon < \min\{n \cdot \epsilon_e, n \cdot \epsilon_i\}$.*

We note that the correctness property for $\mathcal{P}_2$ improves over the same property for $\mathcal{P}_1$ which was only valid for up to $\lfloor m/n \rfloor$ authenticating sessions. Moreover, the security against impersonation of $\mathcal{P}_2$ holds under a bound on the number q' of sessions eavesdropped by the adversary that does not depend on m, n, as in $\mathcal{P}_1$. Perhaps not surprisingly, the usability of $\mathcal{P}_2$ is lower than the usability of $\mathcal{P}_1$, as, for instance, the number of voice samples that need to be produced by the user is considerably larger.

4. VOICE-AND-PASSWORDS ENTITY AUTHENTICATION

The constructions in Section 3 provided 1-factor entity authentication, the factor being voice cryptographic primitives. In practice, these types of constructions have one important vulnerability, analogously to many password-based authentication schemes: they are not resistant to intrusion attacks. An intruder that gains access to one or a few records in the access file is able to compute obtain voice values and use them later to successfully impersonate a honest user.

A solution to this problem for the case of password transmission protocols has been recently given in the bounded-retrieval model [4]. By combining password schemes secure against intrusions in the bounded-retrieval model from this latter paper, and the voice-based authentication protocols described in the previous sections, we obtain 2-factor entity authentication schemes that are secure, in the bounded-retrieval model, against intrusions to the access file (or, password file) and brute-force attacks such as dictionary attacks. In the rest of this section we only briefly sketch one of our construction and its properties, obtained as a direct combination of our voice-based entity authentication protocol $\mathcal{P}_1$ and a password-based protocol from [4]. We start by recalling the definition of the bounded retrieval model.

The bounded retrieval model [4]. In the case of password-based entity authentication, this model assumes a bound q on the number of locations from the password file that can be read by the adversary. This model seems very realistic, as we now explain. On one hand, it seems feasible in practice to make access files very large (since storage is a very cheap resource nowadays); on the other hand, it seems reasonable to assume that any adversary's attempt in retrieving large amounts of data would be easily noticed and thus interrupted. In other words, an attacker needing a large amount of time to retrieve large amounts of sensitive data will most likely be unable to maintain an unauthorized connection for enough time without being detected.

Password-based entity authentication protocols in the bounded retrieval model [4]. In this model, the authors of [4] improve UNIX-like password protocols as follows. When registering a user with a password pw, instead of storing the hash of a password in a single location in the password file, the server stores a combination of this hash value with the random contents of a few carefully chosen locations, this choice being randomized by non-trivial algorithms that take as input the password value itself. When verifying a user claiming a password pw', the same calculations are repeated using pw' and the server checks that the stored value is the same as for some previously registered password. Note that the adversary, not knowing the password, cannot run the same algorithm as the verifying server. In fact, the adversary has provably no significantly better chance of guessing these random location contents than by entirely reading the (very large) password file. But since this model postulates a bound on the number of locations that can be read from the adversary, this will not be possible and the attacker will not be successful in carrying any brute-force attack, such as a dictionary attack.

Password and Voice-based entity authentication in the bounded retrieval model. One of the protocols from [4] instantiating the above paradigm consists of just storing/verifying a tag tg in the access file for each password, the value of this tag depending on the random contents of a few locations in the file, chosen using the password itself. Specifically, the registration algorithm uses dispersers [15] and pairwise-independent hash functions [2] as follows: it computes the output of the dispersers on input the password and rewrites it as a list of locations from the access file; then, it uses the pairwise-independent hash function to hash the contents of all locations in the list and stores its output as a tag. The server's verifying algorithm, on input a potentially valid password, runs the same algorithm as the registration algorithm, and then checks that the resulting tag is the same as the one stored at registration for this user. In our scheme, we suggest to extend this protocol as follows.

Registration phase. The tag tg is computed from the password pw exactly as in the protocol from [4]. Then, the value $v = VR(VS_{id,ct}(pw))$ is computed and the actual value that is stored is $F[j] = v \oplus tg$ (instead of only tg), for some location index $j \in \{1, \ldots, t\}$, if t is the total number of passwords (much smaller than the number of locations of the entire password file).

Verification phase. A tag tg' is computed from the received password pw' exactly as in the registration phase. Then the value $\hat{v}_j$ is computed as $F[j] \oplus tg'$, for all location indices $j \in \{1, \ldots, t\}$, and the authentication is successful if $dist(\hat{v}_j, v') \leq \delta_a$ for at least one $j \in \{1, \ldots, t\}$, where v' is the voice value received by the authenticating user.

Properties. We can prove that the resulting protocol is resistant to bounded retrieval attacks even from adversaries with unrestricted computing power. More precisely, we can first of all extend the formal definition of password protocols secure against intrusion attacks in the bounded-retrieval model from [4] to a formal definition of *voice and password-based entity authentication protocols* that are secure against intrusion attacks in the bounded-retrieval model. Then we can prove that the above protocol satisfies this definition, where the adversary's advantage in impersonating a honest user is bounded by $poly(q, \ell)/2^{|pw|}$, where ℓ is the number of locations used to compute tg (thus, only being slightly larger than the online attack success probability $2^{-|pw|}$). The proof of this fact is obtained using minor modifications to the proof of security from the protocol from [4].

5. OUR VOICE-BASED IDENTITY MANAGEMENT ARCHITECTURE

Telcordia's Identity Management System mainly consists of the following components: Identity Manager, User Interface, and Biometric Provider. Identify Manager coordinates the identity verification process by selecting and posing biometric challenges (including voice-based ones) to end users and collecting and verifying responses from them. It is the Identity Manager that executes the voice-based entity authentication protocol to verify user responses to voice challenges. User Interface presents the front end through which end users register and interact with the system. Biometric Provider evaluates users' biometrics against stored templates and communicates results to Identity Manager.

The system is designed to be flexible in order to allow use of different biometrics on as needed. Currently, the prototype utilizes voice as the main source of biometrics. However, depending on applications and operations environments, different biometrics, e.g., iris scan, face recognition, and fingerprints, may be required. In addition, it is conceivable, or even desirable, to use multiple biometrics at the same time. To this end, the system defines an interface, by which the evaluations of users' responses to biometric identifiers and confidence values can be submitted, regardless of the actual biometrics used in the system. In the current prototype, the voice biometrics is provided by a commercial product, i.e., ScanSoft SpeechSecure, which functions as a Biometric Provider.

Voice also provides the main means by which users interact with the system. Specifically, users dial a phone number and answer a series of challenges by voice. In the current prototype, a commercial Voice XML (VXML) platform, VoiceGenie, is used to execute, on command by the Identify Manager, a set of custom VXML files. These VXML files form the User Interface component of the system and are responsible for sending challenges and collecting answers from users. The exact set of challenges played to users varies from session to session and is decided by the Identity Manager. Processing of collected user answers is also determined by the Identity Manager and is dependent on the types of questions. For example, for (voice) biometric questions, the Identity Manager specifies that user responses first be sent to the Biometric Provider for evaluation and then sent to it the evaluation results. The evaluation process determines how closely a user's response to a biometric question matches the stored template of who the user claims to be. The result is a numerical value, referred to as a confidence value. This way, the Identity Manager is decoupled from the tasks of processing and evaluating biometric data, which further enhances the flexibility of the system.

In the current prototype, registering new users is a two-step process. First, the user visits a web site, where s/he enters the registration information, including the user's organization number, office number, phone number, office address and a set of personal questions and answers. The user is then asked to call a number to create his/her voice template. Subsequently, the user goes back to the web site where s/he finalizes the registration. The described registration process is meant to simulate a secure registration environment, in which the user's identifying information can physically be verified by a live person; imagine the process of opening a new bank account at a branch office.

Java Servlets are used to implement Web pages for user registration. They also implement the communications and control interfaces between the Identify Manager and User Interface component (i.e., the VoiceGenie platform). Identify Manager and Voice-Based Entity Authentication Algorithms are implemented as "Plain Old Java Objects" (POJO), which execute in response to user actions at the registration Web pages and during voice sessions via the VoiceGenie platform. The prototype system utilizes commercial database systems to store user registration information. The communication between the VoiceGenie and Scansoft SpeechSecure, used for evaluation of users' voice responses, uses HTTP.

6. CONCLUSIONS

Human voice has a great but perhaps unrealized yet potential as a simultaneously very usable and very secure biometric factor for entity authentication. As an attempt to avoid the well-known "usability vs. security tradeoff" in entity-authentication solutions, we have put forward a model for the design and quantitative analysis of protocols based on human voice. Our goal being the design of cryptographic protocols, and specifically voice-based entity authentication schemes (a cornerstone of most identity management systems), our modeling has focused on certain cryptographic properties that appear to be intrinsic features of voice. In this model, we have designed and proved the security properties of three schemes. The first scheme is essentially analogue to a standard password transmission protocol, which is arguably the simplest entity-authentication scheme by purely digital tools. The second scheme improves the first scheme on security but naturally not on usability. The third scheme combines voice-based schemes (such as one of the first two in this paper) with password transmission schemes to achieve further increased security against intruders that manage to gain access to the verification server's storage.

7. REFERENCES

[1] D. Boneh, G. Di Crescenzo, R. Ostrovsky, and G. Persiano, *Public-Key Encryption with Keyword Search*, in Proceedings of Eurocrypt 2004, LNCS, Springer-Verlag.

[2] J. L. Carter and M. N. Wegman, *Universal Classes of Hash Functions*, Journal of Computer and System Sciences, vol. 18, 1979, pp. 143-154.

[3] G. Davida, Y. Frankel, and B. Matt, *On Enabling Secure Application through Off-Line Biometric Identification,* in Proceedings of 1998 IEEE Symposium on Research in Security and Privacy.

[4] G. Di Crescenzo, R. Lipton, and S. Walfish, *Perfectly-Secure Password Protocols in the Bounded Retrieval Model*, in Proceedings of TCC 2006, LNCS, Springer-Verlag.

[5] Y. Dodis, L. Reyzin, and A. Smith, *Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy Data*, in Proceedings of Eurocrypt 2004, LNCS, Springer-Verlag.

[6] D. Z. Du and F. K. Hwang, *Combinatorial Group Testing and its Applications,* World Scientific, Singapore, 1993.

[7] P. Erdos, P. Frankl and Z. Furedi, *Families of finite sets in which no set is covered by the union of r others,* in Israeli Journal of Mathematics, 51: 79–89, 1985.

[8] W. Kautz and R. Singleton, *Nonrandom binary superimposed codes,* in IEEE Transactions of Information Theory, vol. 10, pp. 363-377, 1964.

[9] J. Garay, J. Staddon, and A. Wool, *Long-live Broadcast Encryption*, in Proceedings of Crypto 2000, LNCS, Springer-Verlag.

[10] R. Morris and K. Thompson. *Password Security: A Case History*, in *Communications of the ACM*, Vol. 22, no. 11, 1979, pp. 594-597.

[11] L. Myers, *An Exploration of Voice Biometrics*, http://www.sans.org/reading_room/whitepapers/authentication/1436.php, April 2004.

[12] N. Ratha, J. Connell, and R. Bolle, *Enhancing Security and Privacy in Biometric Authentication Systems*, in IBM Systems Journal, vol. 40, n. 3, 2001.

[13] S. Prabhakar, S. Pankanti, and A. Jain, *Biometric Recognition: Security and Privacy Concerns,* in IEEE Security and Privacy Magazine, vol. 1, n. 2, March 2003.

[14] B. Schneier, *Inside Risks: The Uses and Abuses of Biometrics,* in Communications of the ACM, vol. 42, no. 8, pp. 136, Aug. 1999.

[15] M. Sipser, *Expanders, Randomness or Time vs. Space*, in Journal of Computer and System Sciences, vol. 36 (3), June 1988.

[16] Y. Sutcu, Q. Li, and N. Memon, *Protecting Biometric Templates with Sketches: Theory and Practice*, in IEEE Transactions on Information Forensics and Security, 2007