Measuring Performance in a Biometrics Based Multi-Factor Authentication Dialog

A Nuance Education Paper

2009

Definition of Multi-Factor Authentication Dialog

Many automated authentication applications designed today are built specifically to provide a **Multi-Factor Authentication (MFA)** dialog. To qualify as an MFA a dialog must consist of questions that elicit the customer to provide at least two of the following three components:

- Something the caller knows (ie: PIN, mother's maiden name, secret date)
- Something the caller is (ie: a biometric test such a voice verification)
- Something the caller has (ie: caller provides a code from a provided token such as RSA SecureID)

From a practical standpoint it is difficult for most enterprises to distribute hardware tokens to their entire customer base and expect them to have access to them when calling, so the focus of this paper is on telephony based dialogs based on the first two factors described above. To qualify as Multi-Factor, an authentication dialog must include at least one test that ensures the caller has knowledge that only they should have as well ensuring that the callers voice matches a previously enrolled biometric. In its simplest form, a dialog that asks the user to speak their account number may qualify as a multi-factor authentication dialog since the caller must: (#1) know their account number and (#2) say it with a voice which matches a previously enrolled voiceprint. More complex MFA dialogs may include multiple knowledge and/or biometric challenges.

This paper attempts to define metrics that go beyond typical biometric measurements, such as **Equal Error Rate (EER),** to provide an overall picture of how the system is performing. In particular this paper will attempt to define two key metrics by which Multi-Factor Authentication Dialogs should be measured:

- The **Overall Security Rate** will measure the effectiveness of all the factors in the dialog with regards to preventing would-be impostors with access to the system.
- The **Correctly Authenticated Call Rate** will measure the effectiveness of the dialog to allow cooperative true callers access to the system.

Description of Methodology

Biometric Performance Measurement

The first step in measuring the overall performance of the dialog is to understand the pure biometric performance for any individual biometric test employed in the dialog. Depending on the dialog strategy employed, biometric tests could include text-dependent based identity claims (i.e., account numbers, phone numbers), pass-phrases (i.e., "my voice is my password, " 1-9) or text-independent random phrases (i.e., random names, random digits). **False Accept** (**FA),** which is the likelihood that an impostor would have the ability to successfully pass the biometric test, and **False Reject (FR)**, which is the likelihood that a true caller would fail the biometric test, are typically measured in a controlled environment called an **Impostor Test**. For dialogs that contain multiple biometric tests, the performance of each test should be measured independently.

In an impostor test environment, participants are recruited to enroll in dummy accounts set up especially for the test. Testers are then instructed to make a specific number of calls during a specific time to attempt to access their own account. This data is then stored and measured. Because it is known that all calls during this time period were made by "True Callers," this data can be used to determine the **False Reject Rate (FRR)**. During a second time period, callers are instructed to make a specific number of calls to attempt to access specific accounts that do not belong to them. Care is usually taken at this step to ensure that males attempt to access only accounts belonging to males and that females access accounts only belonging to females. At the end of this period, this data is stored and measured. Because it is known that calls during this time period were made only by "Impostor Callers," this data can then be used to calculate the **False Acceptance Rate (FAR)**.

There is much debate about the correct number of participants needed for an Impostor Test and the methodology under which these tests should be run. While it is beyond the scope of this paper to delve into all of the variables that might make up an impostor test (such as gender, ethnicity, age and channel modeling of the intended calling population), a few best practices are worth noting. In an ideal world statisticians would ask for at least 3000 participants, 1500 of whom would make an enrollment and one true user call and then another 1500 who would make one impostor attempt into one of the enrolled voiceprints. Based on commonly agreed upon confidence threshold calculations, this amount of data is accepted to provide a 95% confidence level and with a 50% variance when targeting a 1% False Accept Rate. In practice it is generally not practical to obtain 3000 participants to make one or two calls each and a smaller set of participants makes multiple calls instead. 300 participants who can make 5-7 true user calls and 10-12 impostor calls (one call each against 10-12 different voiceprints) is generally considered a very good test, although tests of with anything more than 100 participants are useful. Generally tests with less than 100 participants are considered anecdotal at best.

Measuring the Affect of Additional Factors

Biometric performance is just one factor that affects an MFA. The overall **Call Security** and **Correctly Authenticated Call Rate (CACR)** are calculated by considering the impact that additional factors have on overall security of a multi-factor authentication. For example, consider a non-biometric system that requires callers to state their account number to gain access to the system. This dialog has a particular level of security based on the fact that only a percentage of would-be impostors will know a valid account number in the system. Security is improved in this application by adding a second factor, for example requiring callers to also enter a PIN. Security for this type of dialog now requires that a would-be impostor both know a valid account number AND know the associated PIN for this account. From a security perspective, this sets the bar higher for gaining illicit access to the system. The Overall Call Security Rate in this type of dialog would be the likelihood that an impostor knows a valid account in the system combined with the likelihood that the impostor knows the PIN associated with that account.

While biometric performance can be measured, the likelihood of an impostor having information needed to gain access to the system needs to be assumed in most cases. In the example provided above the likelihood of the impostor knowing the Account and PIN must be assumed. In the case of a MFA dialog that uses **Knowledge Verification (KV),** such as a Secret Date, as a second factor, the likelihood that an impostor will be able to correctly provide this information must be assumed.

Calculating Overall Call Performance

While the method for calculating the overall Security Rate and Correctly Authenticated Call Rate proposed by this paper can be applied to any Multi-Factor Authentication dialog, the specific variables to be applied to the formula are dependent on the dialog itself. For the purposes of this paper it is helpful to consider an example MFA dialog. In this example a caller is asked to identify themselves by speaking a self-selected 10 digit phone number. Upon successful identification and biometric authentication, the caller is then asked to provide a self-selected secret date which was supplied to the system at the same time the voiceprint was enrolled. Callers who fail to pass the initial biometric test are not given the opportunity to even hear the question which asks for the secret date. Callers who successfully identify, pass the biometric test, and provide the correct secret date are granted access to the system.

After a round of data collection and analysis, Nuance can provide metrics by which to measure system performance. Once tuning data has been obtained and transcribed, certain calls will be classified as **Un-Automatable Calls**. These calls include hang-ups, agent requests, callers giving incorrect phone or date formats (e.g., 7-digit phone numbers, or month-year dates), providing a different phone number or date than the ones they enrolled with, and other speech that is far off from the question being asked. Essentially these represent calls from people who either didn't make a real attempt at using the system or from people who had so much difficulty that it would become impractical to design an automated system to handle these types of calls.

Using the dialog described above we can develop a specific formula by combining the biometric rates with assumptions about the impostor success rate for a second Knowledge Verification question to measure the system. The Security Rate and the Correctly Authenticated Call rates for the whole application can be calculated as follows:

*Call Security Rate = 100% - (Measured_Biometric_FA% * Impostor_KV_success_assumption%)* [1]

Overall Authenticated Call Rate = 100% - (Un-Automatable_Calls +
(Measured_Biometric_FR% + (100% - %TrueSpeaker_KV_success)))

Measurable Call Rate = 100% - Un-Automatable_Calls%

Correctly Authenticated Call Rate = Measurable_Call_Rate + Un-Automatable_Calls%
(Measured_Biometric_FR% + (100% - %TrueSpeaker_KV_success))

It is worth noting that Nuance does not include Un-Automatable Calls in the calculation of Correctly Authenticated Call Rate since they did not qualify as being possible to authenticate. In general, the Correctly Authenticated Call Rate provides a good metric on the overall usefulness of the system while the Overall Authenticated Call Rate provides the actual percentage of all calls that were authenticated.

During the tuning phase of the project, the business can make decisions about how much time and effort to expend on reducing the Un-Automatable call rate. Business decisions can also be made to attempt to classify the set of un-automatable calls programmatically. For example, hang-ups and specific requests for agents are generally categorizable and can be reported on. Once a steady state is achieved, the Overall Authenticated Call Rate can easily be measured in an IVR system and reports can be generated on a regular basis to ensure that the system maintains the performance goals.

---

1 *For simplicities sake, the Call Security Rate does not use an assumption about the likelihood that an impostor would know the self-selected phone number used by the customer to make an identify claim. A more precise Security Rate calculation would take this into consideration.*

Methodology as Applied to Actual Customer Data

Actual Imposter Test Feedback

The following are the results of an Impostor Test performed for a customer using a dialog that asks callers to speak their 10 digit phone number. The % Client Calls that are Rejected represents the False Reject Rate. The % Impostor Test Callers that were Accepted represents the False Accept Rate. Attempts that fell into the Unsure category were put through an additional biometric test.

Phone Number Verification Decision	% Client Calls	% Impostor Test Callers
Reject	**0.6**	91.9
Unsure	0.6	4.1
Accept	98.8	**4.0**

These performance characteristics represent the operating point requested by the customer. Callers who fall into the unsure range were given an additional biometric test. Because thresholds are adjustable (ie: by increasing the pass threshold, the False Accept rate can be lowered at the cost of increasing the False Reject rate), customers make a decision to choose the operating point which makes the most sense for their business and security needs. The following table represents the range of operating points available based on the data in this Impostor Test:

Threshold Delta	PN FR (%)	FA, PN only (%)	Projected FA after KV (%)
-0.50	0.3	8.1	0.8
-0.45	0.3	7.5	0.8
-0.40	0.3	6.9	0.7
-0.35	0.3	6.3	0.6
-0.30	0.4	5.8	0.6
-0.25	0.4	5.4	0.5
-0.20	0.5	5.0	0.5
-0.15	0.5	4.8	0.5
-0.10	0.6	4.4	0.4
-0.05	0.6	4.1	0.4
0.00	**0.6**	**4.0**	**0.4**
0.05	0.7	3.9	0.4
0.10	0.7	3.9	0.4
0.15	0.8	3.9	0.4
0.20	0.8	3.8	0.4
0.25	0.9	3.7	0.4
0.30	1.0	3.4	0.3
0.35	1.0	2.9	0.3
0.40	1.1	2.6	0.3
0.45	1.2	2.3	0.2
0.50	1.3	2.1	0.2
0.55	1.3	1.9	0.2
0.60	1.3	1.8	0.2
0.65	1.4	1.6	0.2
0.70	1.4	1.5	0.2
0.75	1.6	1.5	0.2

Assumptions for Biometric Based MFA Dialog

While the Biometric FA and FR rates were calculated for this dialog based on Impostor Test data, assumptions were needed in order to calculate the Call Security and Correctly Authenticated Call Rates. For this recently deployed customer application, the second factor used was a Secret Date that was self-selected by the user during the enrolment period. The following assumptions were initially set:

1. 97% of true speakers will successfully pass the KV step (secret date)
2. 10% of impostors will successfully pass the KV step
    - This was considered to be a generous assumption as an impostor must first defeat the biometric test before even hearing the hint for the secret date

Based on a tuning with real customer data, it was possible to determine the percentage of true speakers who successfully passed the KV dialog. After tuning, the failure rate at the secret date dialog was determined to be 2.9%. As such, only the assumption about the proportion of impostors who will successfully pass the KV step is still necessary. Armed with this information, the revised numbers used in Call Security and Correctly Authenticated Call Rates were as follows:

1. 97.1% of true speakers will successfully pass the KV step
2. 10% of impostors will successfully pass the KV step

Overall Field Performance

After a tuning of this application, the Un-Automatable Call rate was determined to be 12.7%. As described above, calls were categorized as Un-Automatable due to hang-ups, agent requests, callers giving incorrect phone or date formats (e.g., 7-digit phone numbers, or month-year dates), providing a different phone number or date than the ones they enrolled with, and other speech that is far off from the question being asked.

By combining the measured biometric rates with the 2nd factor security assumptions the Security Rate and the Correctly Authenticated Call rates for the whole application can be calculated as follows:

*Call Security Rate = 100% - (Measured_Biometric_FA% * Impostor_KV_success_assumption%)*
*Call Security Rate = 100% - 4% * 10%*
Call Security Rate = 100% - .4%
Call Security Rate = 99.6%

Overall Authenticated Call Rate = 100% - (Un-Automatable_Calls +
(Measured_Biometric_FR% + (100% - %TrueSpeaker_KV_success)))
Overall Authenticated Call Rate = 100% - (12.7% + (1.3% + (100% - 97.1%)))
Overall Authenticated Call Rate = 100% - (12.7% + (1.3% + (2.9%)))
Overall Authenticated Call Rate = 83.1%

Measurable Call Rate = 100% - Un-Automatable_Calls%
Measurable Call Rate = 100% - 12.7%
Measurable Call Rate = 87.3%

Correctly Authenticated Call Rate = Measurable_Call_Rate + Un-Automatable_Call_Rate -
(Measured_Biometric_FR% + (100% - %TrueSpeaker_KV_success))
Correctly Authenticated Call Rate = 87.3%+12.7% - (1.3% + (100% - 97.1%))
Correctly Authenticated Call Rate = 100% - (1.3% + ((2.9%))
Correctly Authenticated Call Rate = 100% - (4.2%)
Correctly Authenticated Call Rate = 95.8%

It is important to note that while these results are actual measured results from a real deployment, the specific measured results of a deployed system will vary. Variances can be caused by a large variety of

factors such as audio quality issues, dialog design issues, number or tunings, etc. These numbers do however, provide a good rough level estimate of what a similarly designed multi-factor authentication dialog should be able to provide.

Conclusions

While the specific formulas used in this paper do not apply to all Multi-Factor Authentication Dialogs, it is believed that the Overall Security Rate and Correctly Authenticated Call Rate are much more effective ways of measuring system performance, than by looking at the individual performance of biometric gates without considering the impact of the overall dialog.