VOICE BIOMETRICS WHITEPAPER

A Primer for Voice Authentication

Abstract

Many people have had little to no experience with voice biometrics. This document provides essential background information to those interested in learning more or planning their own voice authentication programs.

Steve Hoffman

steve@SayPayTechnologies.com

Overview

Speech recognition services like Apple's Siri and OK Google have become convenient alternatives to the tedious, frustrating and time-consuming effort of keying data into mobile phones. Speech recognition has been around for years and include products like Dragon (Nuance), Cortana, (Microsoft) and Alexa (Amazon). So it's natural for people to think the terms speech and voice recognition are synonymous. Speech recognition is the exercise of using software to recognize sound waves and converting them to a digital representation as for performing searches or text dictation. Speech recognition can be a phenomenal time-savings tool as compared to typed words.

Voice recognition (or sometimes called "speaker recognition") is the exercise of matching a voice utterance to a specific and unique digital representation as a means of identity authentication. Speech recognition engines analyze large samples of voice data containing words spoken by a wide variety of people of different ages, sexes, racial backgrounds, and social and geographic backgrounds. The system creates digital representations with a very high probability of correct interpretation. Each person's voice is uniquely constructed based upon physiological and behavioral characteristics. Physiological aspects are based on the size and shape of each person's mouth, throat, larynx, nasal cavity, weight and other factors; these result in our natural pitch, tone, and timbre. Behavioral properties are those formed based on language, education/influence, and geography, and result in speech cadence, inflection, accent, and dialect.

Voice Recognition Modeling

Voice recognition is much more specific and requires significantly more processing and analysis than speech recognition. Where speech recognition applies broader liberties to converting speech to text, voice recognition must not only convert the speech to text, but also analyze and compare up to 100 unique characteristics of each voice to a master voice print.

Launching a successful voice recognition program first requires collecting a data set of voice samples for the intended geography or location. While English is a common language throughout the world, the sound of each word or sentence sounds differently depending on each person's learned speech attributes. For this reason, English sounds different based upon country and region. Even in the UK, English is noticeably unique when spoken by people from London, Scotland and Ireland. In the USA, accents are distinguishable from those from different parts of the Northeast, South, and West.

How many voice data samples are required for each location is not fixed, but the more voices, the higher the quality of the model. Voice data scientists recommend a data set starting with 500 voices to prove the efficacy of the voice solution.

Enrollment

Voice recognition effectiveness is directly related to following careful and deliberate best practices during enrollment. Enrollment is generally a simple and quick process—requiring the user to speak a passphrase or series of numbers three or four times. Speaking naturally is the most essential best practice, followed by enrolling in an environment without background or ambient noise. Naturally speaking using your normal voice is the best way to recreate each additional voice entry for comparison

to against the master voice print. Naturally speaking is using the same tone, volume, etc. as if you were speaking to an acquaintance right beside you. It's easier to duplicate your natural voice consistently than in any other manner. Many people make the mistake of speaking with increased volume, force or even sounding robotic—try to avoid these pitfalls whenever using voice recognition. As background noise (e.g., traffic, fans, others speaking, music/TV, machinery, etc.) distorts the purity of voice collection during enrollment or comparison, users should take extra care to seek environments with little or no noise. The input device also affects the quality of voice processing; newer mobile phones generally have higher-quality digital microphones and noise-cancellation processing. (If you've ever noticed a small pin-hole on the back of your phone, that is a microphone that collects background noise and generates inverse sound waves for noise cancelation.)

Self-learning

Following voice enrollment, new users sometimes experience lower success rates than their more experienced peers and may need to submit the voice attempts several times before each success. Voice recognition is an imperfect science but can achieve high accuracy rates with usage. New users are sometimes not as relaxed as those with experience who have learned the nuances and idiosyncrasies of voice processing. The learning curve is generally not too difficult with most mastering after several attempts. Further, most voice engines are self-learning and refine each person's voice print with each succeeding entry. Each voice print update adds a new sample that enriches the velocity and expands the breadth of the entire voice model for continuously improved success rates. In time, even entries submitted from environments with modest to moderate noise may be acceptable.

Voice Processing

High-quality voice recognition requires upstream processing on server-class equipment. While some solutions are offered for local on-device authentication, false positive rates (falsely accepting a voice entry than the original owner) dramatically increase. Local authentication is limited to testing far fewer validation conditions as compared to a large online data set capable of analyzing and scoring hundreds of validation conditions. Companies considering deploying voice authentication solutions should target solutions offering the industry norm of False Acceptance Rates (FAR)~ .01% and False Reject Rates (FRR) of ~1%-3%. Bear in mind, most solutions do not rely on voice as the only factor for authentication. With multi-factor authentication, voice recognition is only one of two or more factors, like identifying the user device.

Voice authentication comes in two primary flavors—text dependent, and text independent. Text dependent compares a 6-10 syllable voice "sample" against a master "voice print" and calculates an accuracy score. Text independent captures longer speech input into a voice model and identifies speech mannerisms across a broader spectrum. Text-dependent requires less data but active enrollment by each user (albeit ~30 seconds). Text-independent requires significantly more data, takes longer to process, but enrolls users passively without the need to request any specific utterance. Both have been deployed successfully for call center identification, but text-dependent is the only viable option for functions like website sign-in that must be fast and convenient.

Voice Scoring

The voice engine scores each voice attempt and responds to the authentication manager with a red, yellow or green light-type status. Green light means the entry passed with a high score; yellow light means the entry passed but with marginal results; red light of course means the entry failed with an unacceptable score. Green light entries are automatically added to the voice print model; yellow light entries are added if a secondary authentication factor is successful like a PIN or password. Red light statuses are never added. Some users with very heavy accents or speech outside from collective norms may become frustrated in the early use of voice recognition. These users may need to override voice rejects with a PIN or password that provides authentication until their voice print profile has been enriched with more voice samples.

Digit-based Voice Authentication

The advantage of using digits presents new opportunities unavailable with passphrases or other biometric methods. Speaking a value allows the user to offer their identification and authentication credentials simultaneously—like combining your username and password in a self-contained package. Current passphrase use cases for signing in require the user to enter or speak their account number, at which time the system prompts them to speak the passphrase. Why does the user need to perform two actions when speaking the account number can identify the user and contains their biometric identity? With customer experience becoming a central theme in all services, any unnecessary action required on the customer's part needs close and careful examination.

The second advantage to digits over passphrases is the ability to identify a specific transaction or "authentication event" which extends voice biometrics beyond limited website sign in. When a transaction receives a digital ID, it enables the user to directly service that function without unnecessary navigation—like paying a bill or approving a wire by simply speaking the transaction ID. The ID may be any of various numbers assigned to the transaction including invoice, account, or customer numbers. At SayPay, we automatically create a transaction identifier using an algorithm that uses the user, amount and parties to the transaction into a guaranteed-unique 8-digit value. When users speak each unique "voice token," they apply their biometric signature which forms in means of nonrepudiation. Even if authentication disputes may be rare with biometrics, the voice token is the digital equivalent of a notary seal that protects and provides all parties with assurance of indisputability.

The third advantage of a digit-based voice solution is the inherent inability for playback. Passphrase providers claim they can detect playback attempts by comparing each voice submissions to all previous submissions. This claim implies that your voice is different enough with each submission but unique enough at the same time. Using a unique value each time eliminates this argument from ever clouding the voice biometric efficacy discussion.

The fourth advantage to digits is three-factor authentication. The default SayPay solution offers 3-factor authentication out of the box with something each user has (the mobile device), something the user is (unique voice) and something the user knows (each unique SayPay "voice token"). Passphrase solutions (e.g., ". . . my voice is my password") by default only offer two-factor authentication with something each user has (the mobile device), something the user is (unique voice). Three-factor authentication—

the holy grail of security teams but deemed too averse to the customer experience—is now possible, feasible and desirable.

The fifth advantage to voice digits is intent anonymity. If a user speaks a bank's generic passphrase in a public location, they are indiscreetly making it known to everyone within ear shot that they are logging into their bank. Perhaps this is not a major concern for most customers, but it should raise a red flag for product managers as it could be an ever-present and irreversible barrier to adoption and usage when making a long-term commitment to using passphrases. Further, research shows that customers would prefer to use a standard approach, like a unique-code for each authentication, than speaking a different passphrase for each of their banking relationships. Onboarding is hard enough without having to go back a few years later and implement a new way that requires customers to re-enroll because a better option was not considered thoroughly enough early on.

Early pioneers of voice passphrases may have been subject to technology limitations like lack of high-quality digital microphones and noise cancelation. Modeling the same passphrase for all users is a much easier exercise as the value remains constant for each institution. However, the added value of a digits-based solution cannot be ignored. A dynamically-generated eight-digit numeric value has 100,000,000 permutations; add alpha characters and the variability increases to 2.8 trillion. Digit-based solutions require analyzing the full value and also parsing the users voice into separate input values for individual analysis and comparison—this in turn achieves a much higher level of user authentication assurance.

Summary

Speech recognition and voice recognition are two separate technologies using speech; the first for application in searches and dictation, and the latter for user authentication. Voice recognition requires a data model of voices that are configured and calibrated for the intended user base based upon local speech attributes like accent and dialect.

The success rate of voice recognition is based upon a solid data model, disciplined enrollment, and ongoing usage where each individual voice print, and the data model at large, continue to improve.

Voice validation involves analyzing a voice input record based upon up to 100 unique characteristics and comparing the results to the stored voice print; the outcome is a score signaling if the voice comparison was highly accurate (green), possibly accurate (yellow) or highly inaccurate (red).

Voice authentication comes in two primary flavors—text dependent, and text independent. Text dependent compares a voice "sample" against a master "voice print" while text independent performs longer speech input across a broader spectrum.

Companies considering deploying voice authentication solutions should target solutions offering the industry norm FAR ~ .01% and FRR of ~1%-3%. A digits-based approach is generally preferable over a passphrase to due increased flexibility, security, standardization, and user experience.