

## Sensory Speaker Verification Technology Overview

This document describes Sensory's speaker verification technology, presents estimates of accuracy of the technology, and offers advice on applying the technology in consumer products.

The task of a speaker verification (SV) system is to listen for and detect a specific passphrase spoken by a particular speaker. SV is used for security applications in which a particular user must be identified. Past applications of SV technology by Sensory include a password diary that prevents access to the diary by anyone but its owner, and a key fob remote for an automobile that unlocks the door only after verifying the identity of the owner. SV can be also used in applications where there are multiple users, and the task is to distinguish authorized from unauthorized users or to distinguish among the authorized users.

### SV Methodology

The SV methodology first involves asking the user to pick a passphrase and to train the device by speaking the passphrase one or more times. These utterances of the passphrase are combined to construct a stored memory of the speaker's voice and passphrase called the *template*. Speaker verification is achieved by challenging the user to provide the passphrase. The user's response, which we'll refer to as the *authentication phrase*, is compared to the template. If the match is sufficiently close, the speaker is accepted; otherwise, the speaker is rejected. The key to a successful SV system is in the details of the matching algorithm. The simple-minded approach of comparing two acoustic waveforms does not achieve the desired result because the relevant aspects of the speech signal—phonetic features and individual differences in speech production—are masked by irrelevant trial-to-trial acoustic variability in the signal. Instead, Sensory approaches SV using nonlinear transformations of the speech signal to obtain a featural encoding of the signal that makes explicit information about phonetics and an individual's speech patterns, and a sophisticated neural network architecture that matches the authentication phrase to the template and produces a *match score*. The neural network extracts features that are reliable predictors of the identity of both speaker and the phrase. The match score can be interpreted as a psychophysical distance between the template and authentication phrase. (By "psychophysical distance," we mean a distance metric that emphasizes differences in human perception and production.) A low match score thus indicates that the template and authentication phrases were likely to have been the same phrase spoken by the same speaker, and a high match score indicates that the two phrases were likely to have been different phrases or to have been produced by different speakers.

The SV neural network produces a match score generally in the range 0–300. For the purpose of speaker verification, a *rejection threshold* must be selected, and all match scores above the threshold are rejected, i.e., speaker verification fails. Consequently, the system can produce two distinct types of errors: the authentication phrase can be rejected when it should have been accepted (a *false rejection* or *FR*), the authentication phrase can be accepted even when it should have been rejected (a *false acceptance* or *FA*). Further, false accepts can be subdivided into cases where (1) the correct authentication phrase is produced by the incorrect speaker (denoted FA-CAIS), (2) the incorrect authentication phrase is produced by the correct speaker (denoted FA-IACS), and (3) the incorrect authentication phrase is produced by the incorrect speaker (denoted FA-IAIS). Because the primary concern of an SV system is to avoid accepting imposters who know the passphrase, the FA-CAIS rate is our key measure of false alarms. Further, the FA-CAIS rate is higher than the

FA-IACS rate, which in turn is higher than the FA-IAIS rate. Thus, if we design an SV system to minimize the FA-CAIS rate, we will also achieve very low FA-IACS and FA-IAIS rates.

To improve the performance of SV, the application can require several successive passphrases, and the sequence is accepted only if each passphrase is accepted. If speaker verification trials are independent of one another, then the false accept and false reject rates for  $n$  passphrases in succession would be  $FA^n$  and  $1-(1-FR)^n$  where FA and FR are the single-trial false accept and reject rates. For example, with a lax rejection threshold set such that  $FA = 25\%$  (25 of 100 trials are falsely accepted) and  $FR = 0.5\%$  (1 of 200 trials are falsely accepted) and  $n = 3$  passphrases in succession, we'd achieve an overall FA rate of 1.6% and an overall FR rate of 1.5%. Thus, with several passphrases in succession, the FA rate can be made arbitrarily low while simultaneously achieving a low FR rate. Unfortunately, speaker verification trials are not independent of one another: if an imposter's voice is able to fool the system for one passphrase, the imposter is more likely to fool the system for another passphrase. Thus, one might not expect the boost in performance with multiple passphrases predicted under the assumption of independence. Nonetheless, with a proprietary decision rule, we obtain the performance predicted under independence—and even better it: Rather than applying a threshold to each of the  $n$  match scores in succession, we achieve superior performance via a decision rule that first aggregates the  $n$  scores. We illustrate this performance below.

## SV Evaluation

We present an evaluation of SV for three different Sensory platforms: the RSC 3xx series, the RSC 4xx series, and Voice Activation. The evaluation is based on a data set consisting of 69 speakers, each of whom produced four instances of 40 multisyllable passwords (e.g., AUTOMOBILE, BUMBLEBEE, EFFERVESCENT, LOLLIPOP, MAGNIFICENT, SATELLITE). The data set was roughly half male and half female. The protocol for data collection involved asking each speaker to produce two examples of each password at one point in time, and after several days, the speakers were asked to produce two further examples; we refer to the first two as the *immediate* examples and the second two as the *delayed* examples. The immediate examples were used for training SV and the delayed examples were used as authentication phrases. The purpose of the time lag between the two sets of examples was to simulate the actual use of an SV application, where the user is likely to use train the system initially, and then many days may pass before the system is used for speaker verification. We have found that users have greater difficulty replicating prosody and pitch of a phrase if several days intervene; thus, our data collection procedure is meant to reflect patterns of actual product usage.

The procedure for evaluating the performance of SV is as follows. For each speaker and each passphrase, a template is generated from the two immediate examples. A match score is obtained for each template and each delayed example of the same passphrase, resulting in a total of 3920 match scores involving the correct authentication phrase spoken by the correct speaker (abbreviated CACS) and 188160 match scores involving the correct authentication phrase spoken by the incorrect speaker (CAIS). In addition, match scores involving other combinations of speaker and authentication phrase can be collected for IACS and IAIS cases. Given a particular rejection threshold, the FR and FA rates can be computed from these four sets of match scores: The FR rate is the fraction of CACS match scores above the threshold, the FA rate for incorrect speakers (FA-CAIS) is the fraction of CAIS match scores below the threshold, the FA rate for incorrect authentication phrases (FA-IACS) is the fraction of IACS match scores below the threshold, and the FA

rate for incorrect speakers and incorrect authentication phrases (FA-IAIS) is the fraction of IAIS match scores below the threshold.

By varying the rejection threshold, one obtains a set of FR and FA rates, as illustrated in the following table:

rejection threshold	FR rate	FA-CAIS rate	FA-IACS rate	FA-IAIS rate
95	18.50%	1.08%	0.00%	0.00%
100	8.91%	3.71%	0.00%	0.00%
105	5.56%	6.67%	0.00%	0.00%
110	3.24%	11.07%	0.00%	0.00%
115	1.65%	16.67%	0.02%	0.01%
120	0.89%	23.30%	0.08%	0.02%

As the rejection threshold is raised, the FR rate drops and the FA rates rise. To visualize these numbers, one can plot the FR rate against one of the FA rates. We focus on the FA-CAIS rate, because the FA-IACS and FA-IAIS rates are negligible, suggesting that an imposter is unlikely to fool the system without knowing the passphrase. Figure 1 shows a plot of FR versus FA-CAIS for the Sensory Voice Activation software. The curve is often referred to as an *ROC curve*. (ROC is an abbreviation for *receiver operating characteristic*; the term comes from signal detection theory, which is concerned with the discrimination of targets from lures.) Each point on the curve corresponds to a particular choice of a rejection threshold. The dashed line on the graph is the line for which FR is equal to FA-CAIS. The point of intersection of the dashed line with the curve corresponds to a rejection threshold for which the FR and FA-CAIS rates are equal, called the *equal error rate*.

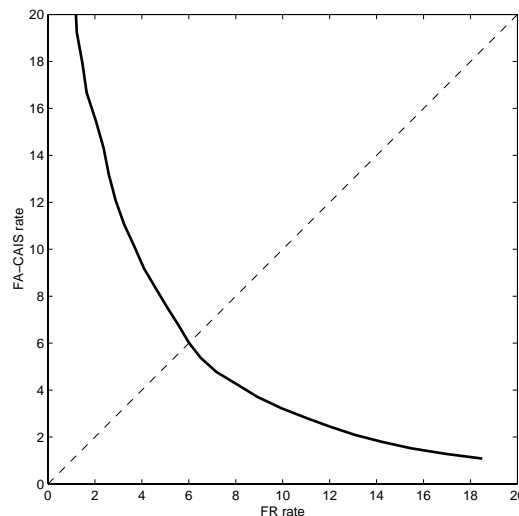


Figure 1

In Figure 2, the ROC curves for Sensory Voice Activation SV that utilizes 1, 2, 3, or 4 successive passphrases are superimposed. The 1-passphrase curve is the same as the previous graph.

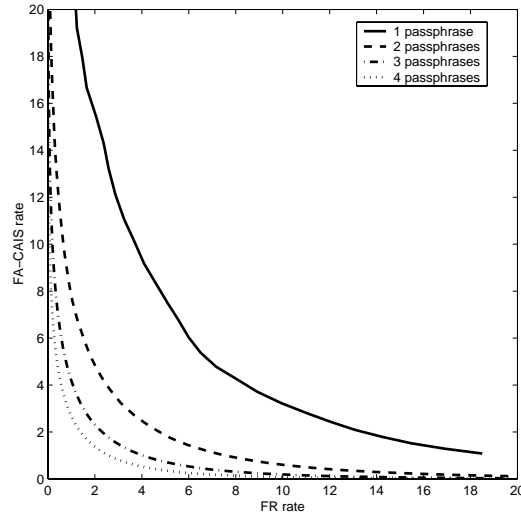


Figure 2

To summarize the performance of Sensory SV technology, the following table shows equal-error rates for each of three different platforms, for 1-4 successive passphrases. An equal error rate of,

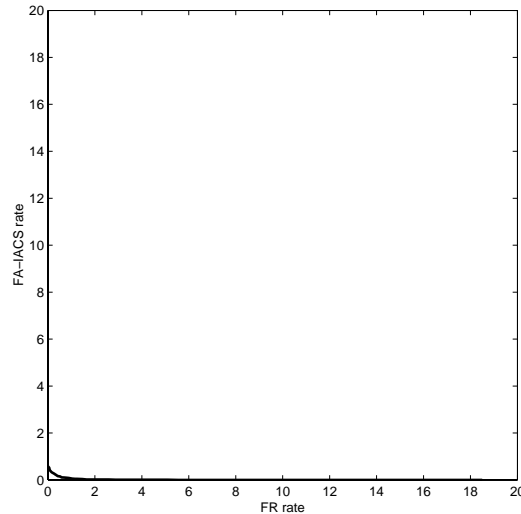
<b>SV Equal Error Rates</b>				
<b>Technology</b>	<b>1 passphrase</b>	<b>2 successive passphrases</b>	<b>3 successive passphrases</b>	<b>4 successive passphrases</b>
RSC 3xx	7.95%	4.12%	2.77%	2.11%
RSC 4xx	7.66%	3.57%	2.10%	1.39%
Voice Activation	6.01%	3.17%	2.15%	1.67%

say, 10% has two interpretations: it could mean that 1 in 10 imposters will fool the SV system if they know the passphrase, or it could mean that 1 out of 10 attempts by any imposter who knows the passphrase will fool the system. Our results indicate that a combination of these two interpretations is correct: some imposters more readily succeed than others, but there is some variability from one attempt to the next. Roughly, the combination of these two factors is 80/20.

We explained earlier that with  $n$  successive passphrases, the  $n$  match scores must be combined to make a final accept/reject decision. Assuming independence of the scores, and assuming a decision rule that accepts the sequence only if each score is individually accepted, the 1-passphrase ROC curve for the RSC 3xx predicts 4.8%, 3.4% and 2.3% equal-error rates for 2-4 passphrases in succession. Despite the fact that the match scores are not independent, we are able to beat out these predictions—as shown in the first row of the table above—via an optimized decision rule that first aggregates the  $n$  match scores (the aggregation procedure being proprietary) and then applies a single threshold.

Figure 3 plots the FR rate against the rate of false accepts for the correct speaker saying the incorrect authentication phrase (FA-IACS). False accepts of this type seldom occur, even when a threshold is chosen such that the FR rate is quite low. From Figure 3, one can see that the equal error rate is about 0.3%. False accepts are even lower when the incorrect authentication phrase is spoken by the incorrect speaker (FA-IAIS), and thus we do not consider them to be a concern. Nonetheless, we emphasize that imposters have almost no chance of fooling the SV system if they

Figure 3



do not know the passphrase, and that the statistics we report in this document concern performance of the SV system if the imposter does know the user's passphrase.

## Practical Advice in Using SV Technology

In the remainder of this document, we address issues that arise in incorporating SV technology into products, and how to best achieve a successful application.

### *Word length*

The passphrases used in our studies were words with 3-5 syllables. Using short 1-2 syllable words/phrases will not provide enough distinctive information to achieve reliable speaker verification, and equal-error rates could jump by a factor of two. Although we have not quantified the effect of using longer 6+ syllable words/phrases, we believe that longer words/phrases will be detrimental to performance due to the fact that the representation of the speech signal loses some resolution for long words/phrases. The 3-5 syllable recommendation is equivalent to a recommendation to choose passphrases of 400-600 milliseconds in duration.

### *Delay between training and usage*

A delay between training and usage of an SV system often affects performance, because the user has increasing difficulty matching the prosody and inflections of the phrases spoken for training as time passes. Thus, when evaluating or testing an SV system, the delay intervening between training and usage is a key variable. Delays occur in naturalistic conditions of usage, and the results we reported earlier incorporate such delays. Equal-error rates are significantly lower (e.g., 6% dropping to 5%) if testing occurs immediately following training. If the reader plans to compare Sensory SV performance to that of competing technologies, the reader should verify that realistic usage conditions were also studied for the competing technologies.

### *Matching training and usage conditions*

In general, one aims to achieve conditions during usage that correspond to the conditions during training. For example, if the SV system is trained in a noise-free environment, it should be used in

a similar environment. Or if users of an SV automobile key fob are walking to the car when speaking the authentication phrase, they may breath heavily, which may contrast with conditions encountered during training, when users were likely stationary.

#### *Variability across users in setting thresholds*

The primary challenge to using SV technology is selecting the rejection threshold. If the threshold is too high, imposters will fool the system when they know the passphrase. If the threshold is too low, authorized users will not pass. To use SV technology appropriately, the following key fact must be understood: The appropriate threshold setting is user dependent. Figure 4 shows a histogram of the equal-error threshold for the 69 speakers in the previously described data base. The x-axis depicts different threshold values, and the y-axis indicates the number of users for whom a given threshold is appropriate. The equal-error threshold varies by about 20% across users. Similar variance in the threshold is observed whether the criterion is to achieve an equal-error rate, or to limit the FR rate, or to limit the FA rate. Thus, regardless of the performance criterion, the appropriate threshold setting will vary from one user to the next.

To elaborate this point further, Figure 5 shows a small real-world test of SV with a single authorized user—call him the *subject*—and a dozen imposters. The graph shows error rate as a function of threshold. The curve that steps down as the threshold is raised is the subject’s FR rate; the curve that steps up is the FA-CAIS rate. The smoother curves are the theoretical FR and FA-CAIS curves based on the large data set described earlier, which we’ll refer to as the *ensemble*. The equal error rate for the subject is around 3% and is achieved for a threshold of 107; the equal-error rate for the ensemble is around 7% and is achieved for a threshold of 117. Although the SV system shows better performance for the subject than for the ensemble, the threshold one would select to achieve an equal-error rate for the subject is significantly lower than the threshold one would select for the ensemble. If the equal-error threshold of the ensemble were used for this particular subject, the subject would have an FR rate near zero but an FA-CAIS rate above 12%.

Unfortunately, variability among speakers is inevitable. The variability is due to differences in distinctiveness of voice and manner of speech, and how precisely speakers replicate their inflections and prosody. Although the SV system is designed to accommodate a range of speakers, the

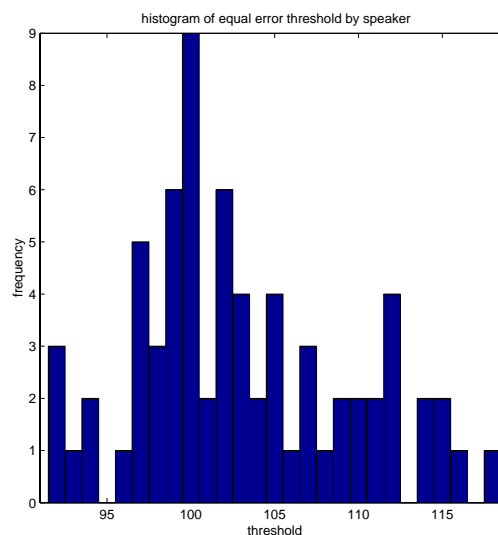


Figure 4

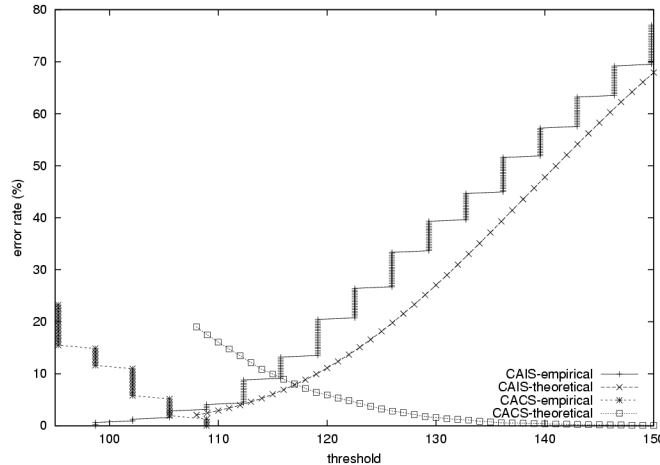


Figure 5

rejection threshold must be tuned for each individual speaker in order to obtain the optimal performance from the system. Thus, rather than selecting one rejection threshold for all users, a user should be allowed to adjust the threshold to accommodate their voice. Control of the threshold should go beyond low and high sensitivity, because such settings correspond to different thresholds for each speaker. We recommend some type of control that allows at least five distinct threshold settings. The particular threshold values will depend on the hardware implementation. For the RSC 3xx, which is the basis of the histogram shown earlier of equal error rates, one might choose thresholds of 95, 100, 105, 110, and 115. An even wider range of thresholds may be desirable if FRs or FAs are to be minimized.

How should a user control the threshold? A simple procedure would involve setting the threshold high initially, and as the user tests the device, the user lowers the threshold by one step if the user finds the FR rate intolerable. This procedure will achieve the lowest possible FA rate given a maximum acceptable FR rate.

#### *Evaluating performance in an application*

Our customers are naturally interested in verifying the performance statistics we've reported in this document. Unfortunately, small-scale real-world tests do not permit strong conclusions. As a simple illustration of this point, suppose a coin is flipped ten times and produces 6 heads and 4 tails. Should one conclude that the coin is biased, producing 50% more heads than tails? One is reluctant to reject the hypothesis that the coin is in fact a fair coin due to the small number of flips. Through statistical testing, we can determine that in fact 10 coin tosses is not a sufficient amount of data to call the coin biased.

Similarly, we can use statistical testing to determine whether tests of SV are indicative of a flaw in the hardware or software. To elaborate this point, we describe one experiment conducted by a customer. The customer evaluated RSC 3xx SV with 12 speakers, assuming each speaker knew the passphrases of the other speakers. A total of 240 speech samples were collected: 120 samples of correct authentication phrases produced by the correct speaker (CACS), and 120 samples of correct authentication phrases produced by the incorrect speaker (CAIS). From these data, an equal-error rate of 10.3% was obtained. (The customer computed 11% for the equal-error rate, but there was an error in their calculation.) Sensory estimates an equal-error rate of about 8% for RSC 3xx SV (see the table above); as a result, the customer was concerned that SV was not performing

properly. However, their result is not indicative of an anomaly, due to the small sample size—small in number of speakers, distinct passphrases, and total samples—used in the experiment.

From our large SV corpus, we selected a subset of data that roughly corresponded to the data collected by our customer (i.e., 10 speakers, about 120 CAIS and 120 CACS utterances), and from these data estimated the equal error rate. We repeated this procedure, each time drawing another random subset of data. Each replication of this experiment produced a slightly different equal-error rate. The distribution of equal-error rates over replications is shown in Figure 6. The mean equal-error rate is around 8%, as stated previously for one-passphrase RSC 3xx SV. However, given the small sample size in each experiment, variability in the estimate of the equal-error rate is observed. Indeed, almost 17% of the experiments resulted in an estimate of the equal-error rate above 10.3%. Thus, even if the equal-error rate in a large population is 8%, small-sample estimates of the equal-error rate will frequently be greater than 10.3%, and one cannot conclude that the 10.3% estimate is inconsistent with the large-sample estimate of 8%.

Testing SV in an application is a worthwhile exercise to check that hardware is functioning. However, Figure 6 points to the limitation of small-scale testing: it can provide a qualitative evaluation, e.g., distinguishing an 8% equal error rate from an 40% equal error rate, but it cannot provide a precise quantitative evaluation, e.g., distinguishing 8% from 10%. Sensory is glad to work with customers to interpret an empirical evaluation and determine whether it is indicative of some anomaly, and to recommend sample sizes that will produce statistically reliable results.

Figure 6

