**RC25311 (WAT1209-026) September 12, 2012**
**Computer Science**

# IBM Research Report

## Seven Blunders of SIV Testing (and a Guide)

**Jiri Navratil, David Nahamoo**
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

**Research Division**
**Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich**

# Seven Blunders of SIV Testing (and a Guide)

*Jiri Navratil and David Nahamoo*
*IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598*
*{jiri,nahamoo}@us.ibm.com*

## BLUNDER 1

**Assessing text-dependent and text-independent technology using one database**

**Background:** While Text-Dependent (TD) systems require the user to speak a fixed phrase at both enrollment and test time, Text-Independent (TI) ones do not. The TI systems are capable of creating a speaker model from a recording of a user speaking an arbitrary text and matching an arbitrary spoken text against such a model.
Thus, the TI systems can be used in a various ways -- not necessarily pre-determined at time of enrollment, and not necessarily requiring an active cooperation from the user. TI systems typically require a minimum of 30 seconds of enrollment speech (and their sweet spot lies higher). TD systems, on the other hand, require multiple repetitions of the fixed phrase. The TD systems take advantage of the fact that the enrollment and the test speech contain the same words to achieve a higher expected accuracy relative to a TI system.

**Blunder:** There is no way to test TD technology on a TI evaluation task, i.e. "enroll w/arbitrary text, test w/arbitrary text"). Vice versa, it is possible to define an evaluation task as TD, i.e. "enroll w/fixed phrase, test w/fixed phrase" and to test both TD and TI systems on such a task, however, the results do not objectively reflect the strengths of the TI technology and instead deliver results relevant only to the TD technology. Comparing the TD and TI technologies on a common task merely in terms of their accuracy is therefore not very informative. The blunder is committed by reporting results while being ignorant of this fact or by just not disclosing it along with published results.

## BLUNDER 2

**Inadvertently entangling voice and knowledge in text-dependent evaluations**

**Background:** Verification can be achieved by means of a voice match, or of a knowledge match, or of both combined in some way. In Text-Dependent (TD) systems the customers enroll with their private fixed phrase and they must say that phrase at test time. Examiners of TD SIV technology may be tempted to collect a speech database of several tens or hundreds of speakers

each uttering their secret passphrase multiple times. They then use some of these repetitions to create the voice models (enrollment) and use the rest of the utterances for testing.

**Blunder:** Given a target speaker, the examiner proceeds with using the "other" speakers in the collection as imposters to measure the False Accept Rate without realizing that the "other" speakers do NOT say the correct passphrase that matches the target enrollment. Hence, the result does not represent the acoustic precision of the TD technology, but an entanglement of knowledge and voice verification with a (not necessarily reasonable) assumption that imposters do not know the passphrase. In other words the result is over-optimistic.

# BLUNDER 3

**Comparing SIV accuracy obtained from channel-matched and channel-mismatched tasks**

**Background:** A "channel" is a technical term referring to the type of pathway through which the original source signal was transmitted. For instance, in telephony, a channel is determined by the type of handset microphone of the caller and the particular transmission pathway of the call. For instance, an internet user may use a headset microphone and is being connected through VoIP and subsequently routed via a cellular network to the other party. Channel variability and particularly the mismatch from enrollment to test, represents one of the greatest technical challenges to SIV technology. Various channel-mismatch compensation algorithms have been developed and implemented in state-of-the-art systems, however, the adverse behavior of different channels remains a considerable factor in evaluating the SIV accuracy.

**Blunder:** Comparing accuracy of two systems one of which is tested on channel-matched and the other on channel-mismatched databases.

# BLUNDER 4

**Comparing SIV accuracy of two systems each of which was trained and/or tested with different average duration**

**Background:** Speech is a varying process. No same word spoken two times by same person sounds exactly same. Therefore, in creating speaker models and in testing their accuracy, the amount of speech (i.e. the duration of the recording) plays an important role determining the final accuracy. For instance, while testing with 3 seconds of speech, models created from 2 minutes duration of enrollment may perform well, models that were created using only 10 seconds may deliver a dramatically degraded accuracy.

**Blunder:** Comparing accuracy of two systems tested with different average durations in the enrollment and the test stage.

# BLUNDER 5

**Blunder 5 – Publishing selected "lab" results**

**Background:** Purely from the theoretical standpoint, the "No Free Lunch" and the "Ugly Duckling" theorems (Theorems 9.1 & 9.2, p. 456 in Duda et al. "Pattern Classification," Wiley, 2nd Ed., 2001) tell us that there is no classifier inherently better than others in their generalization properties. Simply put, they imply that there is always a classifier "of choice" that performs best given a particular data set, or, conversely, there is a data set that gives nice results for a given classifier. While this rather extreme theoretical construct luckily does not apply in full to our practical scenario of SIV testing, one should keep in mind that it is possible to obtain a dataset which CAN produce over-optimistic results. The crux of a fair selection process is randomness. Samples must be chosen randomly, rather than systematically. Imagine a large data set from which we systematically selected only those samples that had produced no error - we will obtain a set that may have a 0% EER, however, that measurement will not generalize to practice --it is over-optimistic.
A second component of the trouble with lab results is over-tuning. This is a process in which the examiner uses perhaps a randomly selected data set, but she tweaks the configuration parameters of the SIV engine iteratively to improve the results on this set. The larger the number of parameters tweaked, and the larger the number of tuning iterations, the higher the odds that the final result will be a blunder:

**Blunder:** Publishing over-optimistic lab results. It is the examiner's responsibility to maintain good experimenting procedures and assess the obtained results with respect to their practical meaning. Best practice to avoid this blunder is to participate in blind evaluations, such the NIST Speaker Recognition Evaluations (www.**nist**.gov/speech/tests/spk ), which use randomly selected, unseen data and do now allow any tweaking.

# BLUNDER 6

**Quoting False Accept Rate without quoting False Reject Rate, and vice versa**

Background: A SIV engine may make two types of mistakes: False Accepts (FA) and False Rejects (FR). The two types may or may not be equally important, however, both are sought to be reduced simultaneously across an entire operating range by the engine developers. After the development is completed and the engine is ready to be deployed, typically, the application developers are given access to a threshold parameter. Changing the value of the threshold allows them to trade off FA for FR, and vice versa, dependent on the application. With any SIV, raising the acceptance threshold will reduce the FA rate but, at the same time, will increase the FR rate.

Blunder: A quote of either FA, or FR, without the accompanying FR, or FA, respectively, is meaningless. It is trivial to achieve a small error rate of one type by increasing the error of the other.

# BLUNDER 7

**Lacking statistical significance**

Background: SIV, as many other speech technologies, are inherently of statistical nature. Their evaluations should also be seen as *statistical* experiments producing merely averages (or expectations) of accuracies, rather than deterministic statements. During the evaluation process, sanity-check methods must be in place to guarantee a statistical significance of the results. It is a fact from the theory of statistical testing that a smaller number of samples contained in an error rate entails a smaller confidence in that result and produces a wider band of "blurriness" (noise)

around it (this is similar to predicting the outcome of an election from a small sample of voters). As a consequence, as we measure smaller error rates we need to increase the number of data samples in order to maintain significance of the results.

Blunder: Quoting a statistically unreliable number, particularly at low EER, without realizing it or without disclosure. For instance, the published number is, say, 1% and was obtained on a corpus whose size is, say, 100 samples of each type (true users and imposters), i.e. 1 mistake out of 100 of each type. Based on standard statistical significance tests, the error could actually lie by chance anywhere up to 3.5% EER.

# RECOMMENDATIONS

**Learning from blunders**

*Avoid comparing apples and oranges I:*
Be aware of the TI/TD distinction of an evaluation task. We may not be able to avoid our TI technology being tested against TD technology on a TD task, however, we should make our customers aware of the issues involved. The customer should be guided in terms of the overall solution such that they gain the insight that TD will not provide a secure and flexible solution to them as it is prone to attacks by recording and is constrained to text fixed at enrollment time.

*Avoid comparing apples and oranges II:*
Knowledge and speaker model entanglement - In comparing two different systems, make sure that they are compared fairly with respect to the use of knowledge and speaker model. Specifically, comparison should only be based on common criteria for both systems: either speaker (voice) model capability, or knowledge capability, or a combination

*Avoid comparing apples and oranges III:*
Results should only be reported with accompanying information about the length of training and test, and about channel match/mismatch conditions. Results quoted without this information are meaningless and should not be used for comparison.

*Describe apples properly I:*
The result of a SIV system should be stated in terms of the FA and FR pair and not just one of them alone. The commonly accepted measure EER is just a special case of this pair.

*Describe apples properly II:*
Since every statistical evaluation really identifies a range and not an absolute number, e.g. predicting the outcome of an election based on a small sample, reporting results should be accompanied by the size of the tested population. This allows for simple formulas to be used to qualify the range of the performance of a system.

*Finally… eat apples with caution:*
We trust our lab results because we know all the experimental details. The less one knows about someone else's lab results the more caution one should exercise. The best way to assess technologies is to look at their relative performance as measured on an identical data set, under fair conditions, ideally performed by a third party.