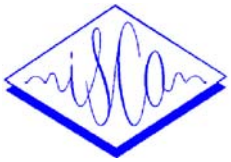

ODYSSEY04 - The Speaker and Language Recognition Workshop
Toledo, Spain
May 31 - June 3, 2004

# Handling Mismatch in Corpus-Based Forensic Speaker Recognition

*Anil Alexander* †, *Filippo Botti* ‡, *Andrzej Drygajlo* †

† Signal Processing Institute, Swiss Federal Institute of Technology (EPFL)
‡Institut de Police Scientifique, University of Lausanne (UNIL)
Lausanne, Switzerland

alexander.anil@epfl.ch, filippo.botti@esc.unil.ch, andrzej.drygajlo@epfl.ch

## Abstract

This paper deals with automatic speaker recognition in forensic applications and handling mismatched technical conditions in a Bayesian framework for evaluating the strength of evidence. Mismatch in recording conditions has to be considered in the estimation of the strength of evidence, i.e., how likely it is that a questioned recording (trace) has been produced by a suspected speaker rather than by any other person from a relevant population. In forensic speaker recognition, in order to estimate such a likelihood ratio, a Bayesian interpretation framework and a corpus based methodology is employed.

Although automatic speaker recognition has shown high performance under controlled conditions, the conditions in which recordings are made by the police (anonymous calls and wiretapping) cannot be controlled and are far from ideal. Differences in the phone handset, in the transmission channel and in the recording tools introduce a variability, over and above the variability of human speech. In this paper we focus on how to estimate and deal with differences in recording conditions of the databases used: detection of whether there is good discrimination between speakers within a database, detection of significant mismatch in recording conditions and statistical compensation in case of mismatch.

## 1. Introduction

The Bayesian approach to forensic speaker recognition relies heavily on the use of databases, in order to establish the strength of evidence. The judicious choice of these databases is crucial to the accurate estimation of the strength of evidence [1] [2].

In forensic speaker recognition the *evidence* ($E$) is the degree of similarity, calculated by the automatic speaker recognition system, between the statistical model of the suspect's voice and the features of the trace. In order to evaluate the *strength* of evidence, the forensic expert has to estimate the likelihood ratio of the evidence given the hypotheses that the suspect could indeed be the source of the trace (also known as hypothesis $H_0$) and that another person, chosen at random from the potential population could be its source (hypothesis $H_1$). In order to do so, it is necessary to choose a potential population, and to record a database that can be used to measure the within-source variability of the suspect's speech. Here the expert is often faced with the problem of being able to decide whether reasonable results can be expected using the databases that he has at his disposal for the case at hand.

When the expert has to deal with a case using the Bayesian interpretation method, he has to decide which databases to choose, which algorithms he should use for modeling and comparing features of speech and how these results should be presented to a judge. He should be able to decide whether the tools he uses perform well with the recordings he plans to use, whether incompatibilities between the databases that he uses can affect the estimation of likelihood ratio and whether normalizations can be performed to reduce the effects of such incompatibilities.

The Bayesian methodology requires, in addition to the trace, the use of three databases: a suspect reference database ($R$), a suspect control ($C$) and a potential population database ($P$). When the performance of the system is being evaluated, it is also usual to use a database of traces ($T$) .

- The $P$ database contains an exhaustive coverage of recordings of all possible voices satisfying the hypothesis: *anyone chosen at random from a relevant population could be the source of the trace*. These recordings are used to create models to evaluate the between sources variability (inter-variability) of the trace in the potential population.
- The $R$ database contains recordings of the suspect that are as close as possible (in recording conditions and linguistically) to the recordings of speakers of $P$ and it is used to create the suspect speaker model, exactly as is done with models of $P$.
- The $C$ database consists of recordings of the suspect that are ideally very similar to the trace and is used to estimate the within-source variability (intra-variability) of his voice.

A brief summary of the methodology proposed in [1] to calculate a likelihood ratio for a given trace is as follows (illustrated in Fig. 1) :

- The trace is compared with the statistical model of the suspect (created using database $R$), and the resulting score is the evidence value ($E$).
- The trace is compared with statistical models of all the speakers in the potential population ($P$). The distribution of log-likelihood scores indicates the between sources variability of the trace with the potential population.
- The control database ($C$) recordings of the suspect are compared with the models created with $R$ for the suspect, and the distribution of the log-likelihood scores gives the suspect's within-source variability.
- The likelihood ratio (i.e., the ratio of support that the evidence ($E$), lends to each of the hypotheses), is given by the ratio of the heights of the *within-source* and *between-sources* distributions at the point $E$ (Fig. 1).

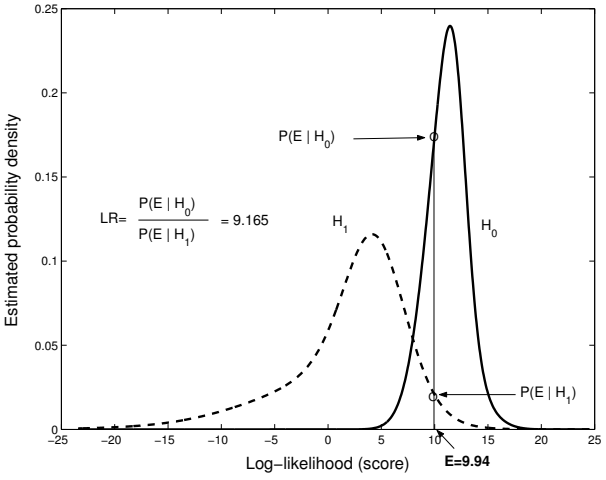


Figure 1: *Graphical representation of the Likelihood Ratio*

### 1.1. Likelihood Ratio and Mismatch

An accurate estimation of the likelihood ratio is possible in the Bayesian framework only if the technical conditions of the reference and potential population databases are identical, and the control recordings are exactly adapted to the conditions of the case. In practice, it can be observed that it is very difficult to satisfy all these requirements. Incompatibilities in the databases used can result in under-estimation or over-estimation of the likelihood ratio.

The likelihood ratio depends heavily on the potential population chosen for the comparison. If an incompatible potential population is chosen, the likelihood ratios derived may be erroneous or misleading. However, recording a potential population is an expensive task, both in terms of resources and time, and the cases that can be correctly analyzed using a given potential population database are limited.

*We propose a method to handle a case under conditions of mismatch by doing the following:*

- *Detect whether accurate discrimination can be performed with each of these databases (Section 2).*
- *Decide whether the databases used are compatible under the Bayesian interpretation framework (Section 3.1).*
- *If the databases are not compatible, estimate statistics for compensation of conditions with a smaller database, and then apply compensation to the potential population database (Section 3.2).*
- *Estimate the new likelihood ratio, with the adapted measures (Section 4).*

## 2. Measuring Discrimination Within a Database

As the first step to any recognition task, it is necessary to check whether the feature extraction and classification algorithms that the forensic expert possesses are capable of discriminating between speakers under the given conditions. Before starting a speaker recognition task with a chosen potential population database, it is necessary to verify whether the system is indeed able to discriminate efficiently between speakers.

The distribution of likelihood scores for cases in which the suspect is truly the source of the test utterance should be approximately in the same range, likelihood scores for cases in which the suspect is not the source of the test utterance should be in another range of values and there should be a good separation between these two ranges. The extent of separation between the scores of the two hypotheses is a measure of the discrimination of the recognition system on the database. In order to quantify this discrimination we introduce the following measure

$$DC = \frac{\mu_{H_0} - \mu_{H_1}}{\sigma_{H_0} + \sigma_{H_1}}, \quad (1)$$

where $\mu_{H_0}$ is mean of the scores when $H_0$ is true, $\mu_{H_1}$ is mean of the scores where $H_1$ is true, $\sigma_{H_0}$ is the standard deviation of the scores if $H_0$ is true and $\sigma_{H_1}$ is the standard deviation of the scores if $H_1$ is true.

This discrimination coefficient ($DC$) is a measure of the distance between the two distributions. The $DC$ allows us to quantify the separation between the two distributions. It is up to the expert to decide what value of $DC$ would be acceptable for a database. If the $DC$ is one or below, this would imply that several values that belong to the distribution $H_1$ could also have come from the distribution $H_0$, and thus the system shows poor discrimination with this database. Similarly, if $DC$ is between one and two this implies moderate to good discrimination, and above two would imply very good discrimination between speakers.

In order to calculate the $DC$, we performed comparisons using a 194 speaker subset of the Swisscom Polyphone database, using a GMM based classifier with 32 Gaussian mixture components and 12 RASTA-PLP coefficients. The same test was performed using 39 speakers of the NIST 2002 FBI database [3] [4]. If we tabulate these results in a matrix, with the models of different speakers on one axis, and the test utterances from these speakers in the same order on the other axis, the diagonal elements (that represent $H_0$ true scores) should have the largest score in each row. This is because these elements represent scores for $H_0$ true, and higher scores imply higher similarity between the suspect's models and the test features. Figs. 2 and 3 present a likelihood score gray-scaled graphic of the likelihoods returned (i.e., each square represents the score the test utterance of the speaker in that particular column obtained when comparing it with the model of the speaker in row where it was found). The values of $DC$ obtained for this database were 1.7248 and 1.7556 for the FBI and Swisscom databases respectively.

## 3. Handling Mismatch Across Databases

One of the problems in the Bayesian interpretation framework is due to the fact that the assumption that the potential population and suspect reference databases are similar to each other in their technical conditions of recording is not always respected. This means that, although the speech content of each database may be different, the effects of channel distortion, noise and recording conditions should be similar across these two databases. Indeed, in the creation of suspect reference database, every effort is taken to record the database in approximately the same way as the potential population database. However, in practice, it is extremely difficult to reproduce conditions *identical* to that of potential population database.

Generally, in order to create the suspect reference database, the forensic expert is faced with two situations. First, he is supplied with the recordings from the police using their recording equipment, in which case he has little control over the conditions of recording. Secondly, he can perform the acquisition of

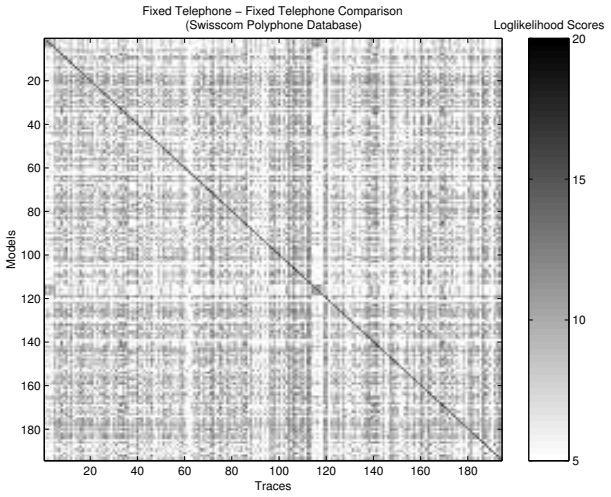


Figure 2: *Illustration of the discrimination on a subset of the Swisscom Polyphone database (Fixed telephone)*

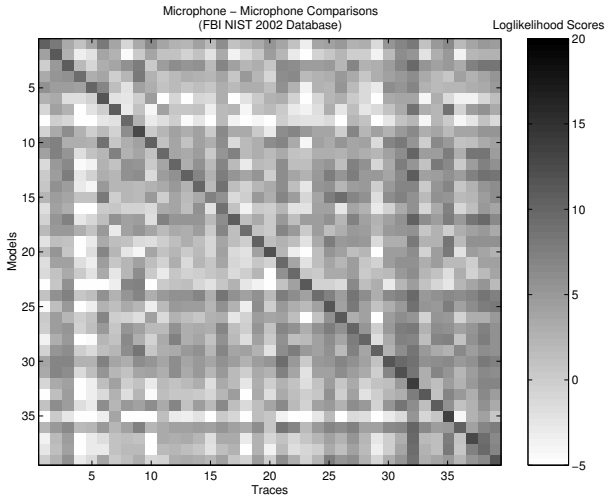


Figure 3: *Illustration of the discrimination on a subset of the FBI NIST 2002 database (Microphone)*

the suspect's voice using his own recording equipment in controlled conditions. However, even this does not ensure complete compatibility with the potential population database, which has a large number of recordings using several different telephones in different transmission and background noise conditions.

Consider what happens when the suspect is recorded in conditions that do not exactly correspond to the conditions of recording of the potential population database. Following the methodology described in Section 1, the trace is compared to the model of the suspect, and a value for $E$ is calculated. Then, the trace is compared to all the speaker models of the potential population database. The intra-variability of the source is estimated by comparing the suspect control database $C$ with the suspect model of the reference database $R$. It should be remembered that since both $C$ and $R$ come from the suspect, this test should give log-likelihood results that correspond to $H_0$ true. Also, since we can be reasonably sure that the real source of the trace is not part of the potential population, comparison between the trace and the potential population gives results that correspond to $H_1$ true.

In cases where there is a mismatch between databases, it is possible that the technical conditions of the trace and that of the recordings used to create the models of the suspect are more similar to each other than to the potential population. This similarity may correspond not only to the actual similarity in voices but also to similarity in conditions. As shown in Figs. 6 and 7, in certain cases, comparisons within a database can show compatible results, but comparisons across databases may not be compatible.

This may mislead the forensic expert to think that the suspect is the source of the trace because of his better likelihood match score, whereas in reality, this is not only because of the similarity of voices but also because of the similarity in the recording conditions.

This mismatch is illustrated using the Swisscom Polyphone database and the IPSC02 Polyphone database. The Swisscom Polyphone database is used as the $P$ database, and subdatabases of the IPSC02 Polyphone is used to provide $R$, $T$ and $C$ databases in different conditions such as fixed telephone, cellular and analogue tape recordings. In Fig. 4, we see the difference between the distribution of $H_1$ true scores for the potential population ($P$) (Swisscom Polyphone) compared to typical traces, and the $H_1$ true score distribution for tests conducted solely using reference ($R$) and traces ($T$) from the database (IPSC02 Polyphone) in the same conditions.

It can be seen that the average scores for the two $H_1$ distributions are considerably different, hence confirming that the potential population database cannot directly be used to derive the likelihood ratio.

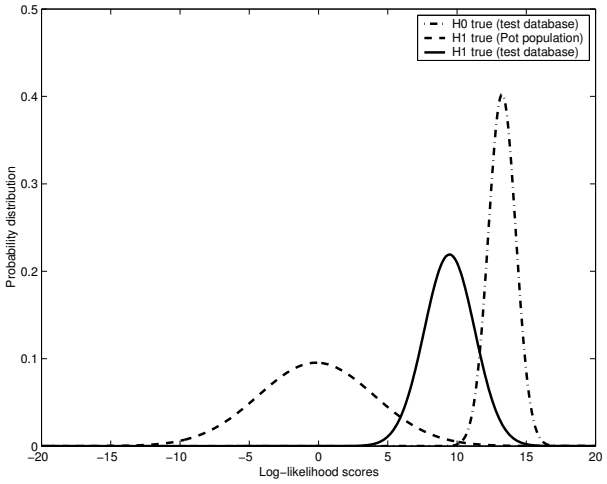


Figure 4: *Log-likelihood score distributions ($H_0$, $H_1$) for potential population $P$ from Swisscom Polyphone and the test $R$ and $T$ databases from IPSC02 Polyphone*

However, it is conceivable that we can build a collection of smaller databases along with the potential population database, corresponding to different case conditions. This collection of databases will contain a considerably smaller number of speakers than the potential population, but will be sufficiently large for us to derive statistics to normalize mismatched conditions.

### 3.1. Detection of Mismatch

Often the choice of the potential population database is based on criteria such as language, sex and perceived technical conditions of the recording. However, this kind of choice is not always guaranteed to be sufficient, especially if the recordings of the suspect are very different from that of the potential population.

One of the difficulties that a forensic expert faces is in determining whether a given potential population can be used with a new case, or whether this choice will actually lead to the likelihood ratio being underestimated or overestimated.

We illustrate this data mismatch with the two following experiments:

We choose 12 speakers from the potential population database ($P$) (Swisscom Polyphone database), and 12 speakers from a test database ($R$ and $T$) (IPSC02 Polyphone database) recorded according to the methodology specified in [1]. We then create models for each of the speakers from both the databases and select traces for each of them. In order to test the assumption that the two databases are compatible, we compare each trace with all the models we have trained, for both the databases.

A similar experiment is performed using the FBI NIST 2002 Speaker Recognition database. Here, two sets of 8 speakers are chosen under different conditions (telephone and microphone recordings) as potential population databases ($P$), models are trained for each of these speakers, and compared with test utterances from the same speakers. Comparisons are thus performed within a given recording condition as well as between the two conditions. If we obtain the same range of results of comparisons within a database, and comparisons across databases, we can conclude that the databases are compatible.

Let us consider only cases in which $H_1$ is true. As illustrated in Figs. 6 and 7, there are four possible zones from which we obtain scores corresponding to $H_1$ true. These are the two zones within each database where the suspect is not the source of the trace, and two zones of comparison across the two databases. It can be observed that there is a clear difference between the range of $H_1$ true scores within and across databases.

In ideal conditions where there is no mismatch, all these values have to be in the same range. We propose to use a simple statistic to verify whether all the sets of $H_1$ scores have values that are in the same range [5]. This is the statistic for a *large sample test concerning the difference between two means*.

Suppose we want to see whether we can reject the null hypothesis that there is no difference between the two different distributions.

- Null Hypothesis : $\mu_1$-$\mu_2$= 0
- Alternative Hypothesis : $\mu_1$-$\mu_2 \neq 0$

$$z = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (2)$$

where $\mu_1$ is the mean of the first distribution, $\mu_2$ is the mean of the second distribution, $\sigma_1$ and $\sigma_2$ are the standard deviations of the first and second distributions, and $n_1$, $n_2$ are the number of datapoints in each of the distributions. For example, setting the level of significance at $\alpha = 0.05$, for $z$ we should have $z > 1.96$ if the distributions are not similar or compatible. If this is not the case, we cannot attribute the differences between the two distributions to chance.

The Figs. 6 and 7 illustrate the incompatibility of comparisons across two databases. For this example the values of $z$ were 14.554 for Fig. 6 and 30.7518 for Fig. 7. Thus the two distributions are significantly dissimilar, and are indicative of mismatched conditions.

If this statistic is not satisfied, the expert should try to use other databases that he has at his disposal, and find one that is compatible. Here, he has the option of choosing not to proceed with analysis of the case in the Bayesian interpretation methodology, or to statistically compensate for the mismatched conditions. This would also involve the compensation (presented in Section 3.2) to be applied to the potential population.

In order to measure the extent of change that the acoustic mismatch introduces, we compare the trace not only with the potential population speaker models, but with the smaller database that adequately represents different conditions of recording, with the same set of speakers. When the trace is compared with this smaller database, in two different conditions, we can estimate the probability distribution for each of these sets of scores for each condition. Since the set of speakers is the same, this represents the shift, or biasing due to the conditions of the potential population database.

We have used a 39 speaker subset of the NIST 2002 Speaker Recognition Database (FBI) to simulate a potential population database ($P$). This database contains speakers in three different conditions of recording such as microphone, body-wire and telephone. In order to evaluate the effects of mismatch we extract from the database the same set of speakers in two different conditions.

By comparing the trace with models in these two conditions, we can estimate to what extent the difference in the acoustic conditions of this database would result in a shift in the probability distribution of the scores. For instance, in Fig. 5 the distributions of the scores of the comparisons of traces with two conditions (fixed telephone and microphone) are shown. We see that although these two distributions show similar variances, their means are shifted. If the potential population is in any one of these conditions, a corresponding normalization as described in Section 3.2 can be applied to reduce the bias of the likelihood ratio.

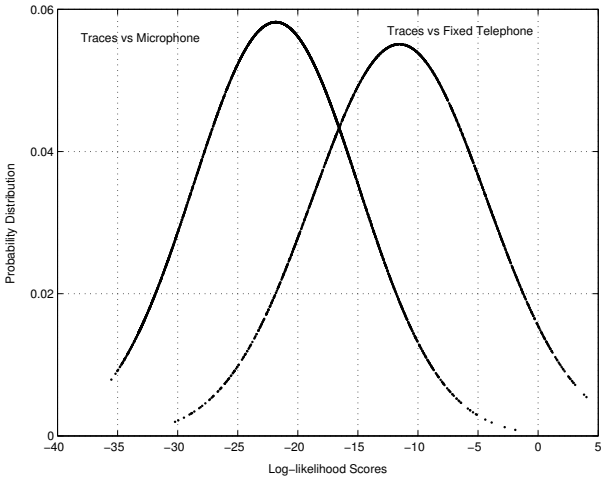


Figure 5: *Distribution of score for the comparison of traces with the population database in two different conditions : fixed telephone and microphone*

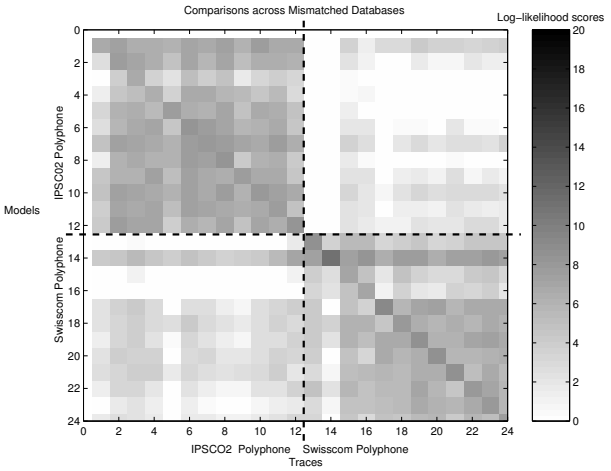


Figure 6: *Comparisons across incompatible databases Swisscom Polyphone and IPSC02*

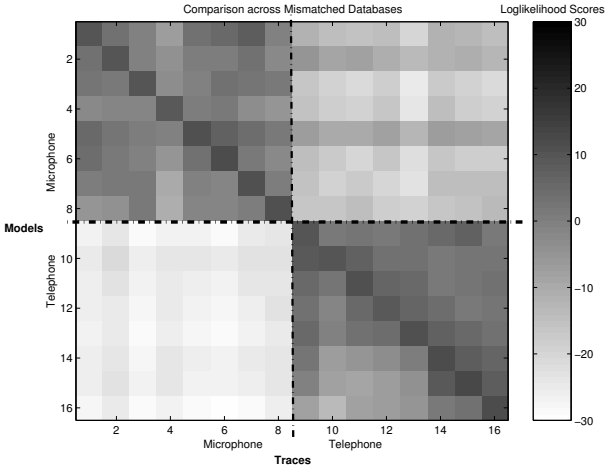


Figure 7: *Comparisons across incompatible databases: FBI NIST 2002 Microphone and Telephone*

### 3.2. Compensation of Potential Population Incompatibility

Recordings obtained from the police do not often match conditions under which the potential population database was recorded, and it is not possible to additionally record the voice of the suspect in matched conditions. As a result, when we compare the trace with the suspect model, and with the potential population models, the results that we obtain may under- or over-estimate the likelihood ratio.

Fig. 4 illustrates this pitfall of applying Bayesian interpretation when the conditions of the potential population database are incompatible with the conditions under which the suspected speaker is recorded. Comparing a trace with two potential population databases in different conditions often gives two distinctly separated distributions of $H_1$ true scores. Since $P(E|H_1)$ directly depends on the distribution of $H_1$ scores, the denominator of the likelihood ratio depends directly on the conditions of the potential population chosen. For the value $E$ given in the figure, the likelihood ratio is greater than one for one condition, and less than one in another.

The likelihood ratio can shift from a position of supporting the $H_0$ hypothesis to a position of supporting the $H_1$ hypothesis as a consequence of mismatch. This can have significant consequences if the likelihood ratios are reported without indicating the possible effects of a mismatched potential population database.

We propose the normalization of potential population scores to reflect the conditions of the case, and to reduce the effects of underestimation or overestimation of the likelihood ratio. In order to calculate the shift brought about due to incompatible potential populations, we select two small databases containing an identical set of speakers, but recorded in two different conditions. One of these databases should be in the conditions of recording of the potential population in question. This database should ideally be a sub-database of the potential population database, or at least one that is in the same conditions as the potential population. We estimate the changes due to the difference in conditions using this sub-database, and apply it to the entire potential population.

In speaker verification tasks, normalizations like *T-norm* [6], *H-norm* and *Z-norm* [7] are used to normalize the imposter scores. In this paper, a normalization from a mismatched condition to the conditions corresponding to the case is applied to the potential population scores. All potential population values are normalized by the measure in Eq. 3. With this normalization we shift the mean score of the potential population (compared to the trace) towards the mean score $H_1$ of the sub-database, and scale its standard deviation to reflect the standard deviation of the sub-database.

$$f(X) = \left(X + \mu_{H_{1C_2}} - \mu_{H_{1C_1}}\right) \cdot \frac{\sigma_{H_{1C_2}}}{\sigma_{H_{1C_1}}} \qquad (3)$$

where $\mu_{H_{1C_1}}$ and $\sigma_{H_{1C_1}}$ are the mean of and standard deviation of the $H_1$ score distribution in condition 1, $\mu_{H_{1C_2}}$ and $\sigma_{H_{1C_2}}$ are the mean and standard deviation of the $H_1$ score distribution in condition 2.

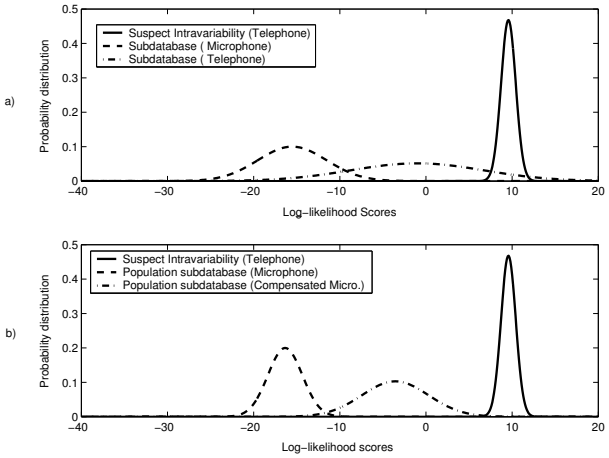


Figure 8: *Compensation of mismatched conditions*

This normalization has the effect of moving the potential population scores ($H_1$ distribution) closer to the estimated mean of the $H_1$ target sub-database distribution as well as adapting its variance to that of the target sub-database distribution. If we apply this normalization to the scores, we observe that the potential population scores are normalized to that of the smaller database in the conditions similar to the suspect database ($R$).

In Fig.8.a we observe the distribution of the $H_0$ true scores (suspect intra-variability), and the distribution of the $H_1$ true scores obtained comparing the trace to the models sub-database in two conditions (telephone and microphone). In Fig.8.b we observe the same $H_0$ true score distribution in telephone condition, the distribution of $H_1$ true scores for a subset of the potential population database in one of the conditions (microphone) and the compensated $H_1$ true score distribution for the microphone condition.

## 4. Handling a Case

Consider that a forensic expert has received a recording of a suspect and a questioned recording for which the court would like to determine the source. Let us also assume that for this case the expert has a cooperative suspect of whose voice a sufficiently long recording can be made. This allows the expert to create models and estimate the intravariability in the suspect's voice. This assumption is consistent with those required by the Bayesian interpretation framework proposed in [1].

The steps involved in handling mismatch in the case are as follows:

- Select a potential population database ($P$) that is similar to the recordings of the suspect reference ($R$).
- Estimate statistics for $H_0$ true and $H_1$ true distributions for this test database as described in Section 3.1. If a mismatch is detected between the $P$ database and the $R$ database, then
  - Select another, more compatible potential population database, or
  - Investigate the possibility of recording the suspect in conditions compatible with the databases available, and
  - If neither of the above options are possible, either decide not to analyse the case using the Bayesian interpretation framework or apply statistical compensation of the mismatched conditions to potential population scores.
- Choose small sub-databases which contain the same speakers in the two conditions that are mismatched between the $P$ and $R$ databases.
- Compare the trace ($T$) with these two sub-databases, and estimate the shift in the distribution of $H_1$ true score distributions in these two conditions.
- Using the statistics of $H_1$ true distributions in the database in the two conditions, apply the normalization suggested in Eq. 3, to the potential population scores in the Bayesian interpretation framework, and derive the compensated likelihood ratios.

## 5. Discussion

If an existing mismatch is undetected, it is likely that the uncompensated usage of the Bayesian interpretation framework gives erroneous results. After detecting a mismatch, the expert has a choice of selecting another compatible database if possible, deciding not to analyze the case in the Bayesian interpretation framework or performing statistical compensation for the mismatched conditions. Detecting and compensating mismatches between databases helps to reflect more accurately the similarity or dissimilarity of the real speech contained in the recordings.

Although primarily the Bayesian interpretation framework is used for the evaluation of evidence in court, it is also a valuable tool for investigatory purposes. The methodology of compensating mismatch is equally important for both these purposes in order to avoid results that are affected by differences in recording conditions.

When it is not possible to use reference recordings that are made in exactly the same way as those of the potential population, this compensation reduces the effect of mismatch, making the potential population comparisons close to that of the suspect. This is very important in forensic science, as the suspect has to be treated just like the speakers of the potential population, and it is necessary not to favor *a priori* one or the other hypothesis.

## 6. Conclusions

In this paper we have proposed detection of mismatch and statistical compensation of forensic speaker recognition results biased due to this mismatch between databases in the Bayesian interpretation framework. We have described how to detect incompatibilities between suspect and potential population databases, deriving statistics for databases similar to that of a given case and then if necessary, statistically compensating the differences which come about because of the incompatibility of databases.

## 7. Acknowledgments

The authors would like to thank Quentin Rossy for his help in creation of the IPSC02 Polyphone database which has been used in the experiments in this paper.

## 8. References

[1] D. Meuwly and A. Drygajlo, "Forensic Speaker Recognition Based on a Bayesian Framework and Gaussian Mixture Modeling (GMM)," in *2001, A Speaker Odyssey: The Speaker Recognition Workshop*, 2001, pp. 145–150.

[2] A. Drygajlo, D. Meuwly, and A. Alexander, "Statistical Methods and Bayesian Interpretation of Evidence in Forensic Automatic Speaker Recognition," in *Proc. Eurospeech 2003*, Geneva, Switzerland, 2003, pp. 689–692.

[3] S. D. Beck, *FBI Voice Database For Automated Speaker Recognition Systems, User's Manual for Training and Testing*, Federal Bureau of Investigation, March 1998.

[4] H. Nakasone and S. D. Beck, "Forensic Automatic Speaker Recognition," in *2001: A Speaker Odyssey*, Crete, Greece, 2001, pp. 139–144.

[5] C. Aitken, *Statistics and the Evaluation of Evidence for Forensic Scientists*. John Wiley & Sons, 1997.

[6] R. Auckenthaler, M. J. Carey, and H. Lloyd-Thomas, "Score normalisation for text-independent speaker verification system," *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 Speaker Recognition Workshop*, vol. 10 (1-3), pp. 42–54, 2000.

[7] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.