GUARDIA CIVIL SRE 2010 HASR1 SYSTEM DESCRIPTION

Ricardo Nieto
Engineering Department
Criminalistic Service.
General Directorate of the Police and Guardia Civil (Guardia Civil)
Madrid, Spain
rnsalinero@guardiacivil.es

***Abstract*— The Criminalistic Service of Guardia Civil enters for the first time for the HASR1 evaluation of NIST SRE'10. The system used to take part in this evaluation is a forensic tool (IDENTIVOX 2009 / BATVOX 3.0) developed by the company AGNITIO, S.L. Clearly, this system was not developed to operate under the optimal working conditions for this kind of tasks, where systems are supposed to operate at the EER point. Therefore, a shifting work was carried out to compensate in each case the major likelihood risk of false acceptance in this kind of forensic systems. Even with these additional normalisations the expert's subjective opinion was essential, and it was always taken into account when there were suspicions about the system operating under false rejection conditions.**

I. Introduction

The Criminalistic Service enters for this evaluation with an automatic speaker recognition system using MFCC parameters and GMM modelling. It also features approximations to session variability compensation: D-Norm/T-Norm + Z-Norm normalizations in the area of scores, filters for RASTA+Feature Warping+CMN+Channel Factors Compensation in the area of parameters and NAP adaptation in the area of models. It includes the Bayesian evidence model assessment by calculation of robust LRs.

This new version of Identivox features a new universal model (UBM - Universal Background Model) including approximately 1600 hours of speech in Arabic, Chinese, English, Russian and Spanish languages.

The system is considered to be calibrated to strongly penalize the false acceptance likelihood at the expense of the false rejection likelihood. This makes necessary to process the system output data, as well as a subjective assessment based on the experience of the expert responsible for the evaluation.

II. System Description

A. General aspects of the evaluation

When carrying out each trial, we generally take the file named by the organization as "test" in order to generate a reliable speaker statistical model, since it is usually of better quality than those called "model", which despite being microphone files, have a SNR (Signal to Noise Ratio) around 10 db and the ENF (Electrical Network Frequency) signal induction is very prominent.

There is no way to control whether the system is operating correctly; we only have extrapolation of reasoning based on tests carried out on our reference populations and databases.

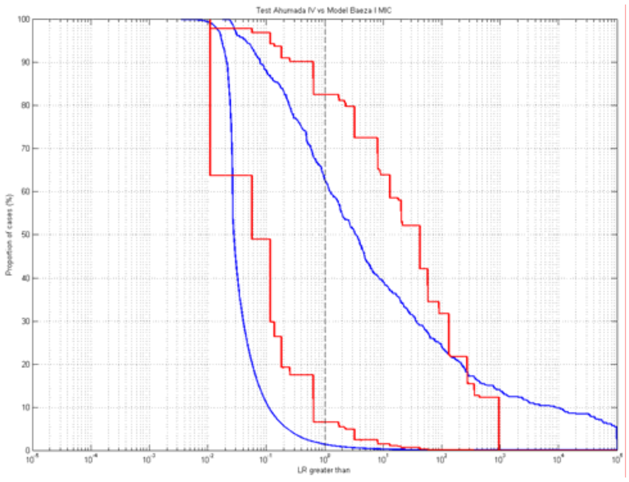
However, at the laboratory we obtain approx. 38% of false rejection likelihood compared with a 1% of false acceptance using our own databases (BDRA – Ahumada IV vs. Baeza I) with variations in time and channel. Assuming that this reasoning can be extrapolated to the NIST HASR1 evaluations, we have compared 100 test files from SRE'08 with each model in telephone channel, obtaining a variable number of impostors with LR values higher than 1.

Using this likelihood value of false acceptance along with our available Tippet plots, the working point has been moved to reduce this likelihood at the expense of increasing the likelihood of false rejection. The aim of the experiment is to work at EER point, which is the most suitable one for the evaluation in question.

Thus, as an example, for false acceptance values around 10% we get a false rejection approximation around 10%, and therefore we would be working closer to the EER.

In case of obtaining a 1% false acceptance, we estimate false rejection around 38%. In order to work at the EER point, we should multiply all LR values by ten and draw conclusions in this new context. See Tippet plot.

This extrapolation can be considered maybe to optimistic, since the quality of audio files in the microphone channel of Baeza I is greater than quality at microphone channels in NIST 2008. Although comparisons are not symmetric ("model" vs "test" differs to "test" vs "model"), a false rejection even greater when false acceptance is reduced is expected.



B. *Feature extraction*

- Score compensation: Normalization (D-Norm/T-Norm + Z-Norm).
- Parameter compensation: Rasta Filters, CMN, Channel Factors Compensation.
- Model compensation: NAP adaptation (use of a channel compensation matrix).
- Coefficients: 19 MFCC+19 Delta.
- Window: Hamming.
- Window length: 20 ms.
- Overlapping (%): 50.
- Sampling frequency: 8000 Hz.

C. *GMM system*

Guardia Civil approach is a likelihood ratio detector with target modelled by Gaussian mixture models (GMMs). We use a Universal Background Model from many speakers as an alternative hypothesis approach. Target models are derived from this UBM by MAP adaptation of the means.

UBMs were trained using 1024 Gaussian mixtures and ML estimation via EM algorithm.

D. *Reference Population*

Reference populations were produced by assessing the averaged spectrum both of test and model files, using the 2008 NIST Speaker Recognition Evaluation Data Collection database and checking the similarity of the fall in low frequency of the files in microphone channel (“model”).

E. *Shifting and normalization*

In order to compensate the effect of decalibration of the working point, the same files obtained to generate the reference populations were used.

This way, when a comparison is made using a model in telephone channel and a test in microphone channel, a telephone reference and a cluster of impostors is used to check the likelihood of false acceptance approximate in channel to the subject of the test, in this case microphone one (depending on the type it can also be MIC07, MIC08 or MIC12).

Once this possibility of false acceptance has been analysed, around twenty files are included (out of the ones used for the checking) in order to proceed with a normalization of means in the area of scores.

F. *Human Aid*

Performing a subjective assessment based on the critical listening has been crucial to support the results drawn by the automatic system.

The profile of the expert chosen to perform the assessment is a person with a Master in Telecommunications Engineering and three year experience producing expert reports on voice comparisons, the vast majority of them in Spanish language.

III. Execution Time

Due to the special characteristics of these tasks, where human intervention is highly marked, the time needed by the system to produce results is considered negligible compared to the hours needed for edition, selection of the reference population, selection of impostors for checking false acceptance and impostors for normalization of the scores means, and especially compared to the time needed by the expert to assess subjectively both the test element and the model.

In general, a single expert would need from five to six working hours for each trial or pair comparison.

IV. Confidence Scale

The assessment scale used to reach a final result in a trial is as follows::

3 – “True” decision / High confidence
2 – “True” decision / Low confidence
1 – “True” decision / Very low confidence
0 – NO confidence
-1 – “False” decision / Very low confidence
-2 – “False” decision / Low confidence
-3 – “False” decision / High confidence

This scale aims at merging both subjective opinions issued by forensic experts and LR calculations produced by the automatic tool Identivox 2009 / Batvox 3.0.

V. CONCLUSIONS

After concluding all the possible tests for every trial using the tool at our disposal, our conclusions are as follows:

Even though the reference populations chosen were reasonable adjusted to the model used both for “test” and “model”, we used mean normalization in the area of scores with impostors, in order to slightly adjust the results.

We know the system performance in terms of false rejection when there is no homogeneity between the test channel and the model, and there is variability related to the lapse of time between the recording times of the test and the model. Additionally, the system working point is set as forensic software according the criteria of innocence presumption, and thus it is optimized to minimize the likelihood of false acceptance at the expense of false rejection. In this context – which might be not the most favourable for this test – we have obtained very conservative LR values.

If this were a real expert report, it could not be performed due to the requirements of our technical proceeding IT-AC-01:

1. The quality of the “model” file is often below 10 dB, and the signal global level is too low.
   This fact along with an encoding of 8 bits per sample (though μ-law) makes its use not recommendable in Identivox 2009 / Batvox 3.0.

2. The “test” file quantification is 8 bits at μ-law and therefore the use of Identivox 2009 / Batvox 3.0 is not recommendable.

Once all impostors have been tested in a channel similar to the used test file (called “model” by the Organization in most cases), false acceptance was found out (LRs above 1) around 2 – 8 %. Extrapolating these data we obtain 38 – 15 % false rejection. In order to work at the EER point, we must properly multiply the LRs.

In some cases, it was subjectively considered that the system was working at the point of false rejection (the statistical data, after a previous shifting process, are 10% of the cases, i.e. between 1 and 2 trials). Therefore, the opinion of the expert was decisive, even if it does not support the objective results produced by the automatic tool.