DOI: 10.17469/O2111AISV000025

FRANCESCO SIGONA, MIRKO GRIMALDI

Calibration and Fusion of Parametric and Non-parametric Statistical Methods for Forensic Semi-Automatic Speaker Recognition Systems: experimental results

In this paper, the most employed Forensic Semi-Automatic Speaker Recognition (FSASR) approaches used in Italy – as emulated in the IMPAVIDO software, and some variants of them – are investigated from the point of view of the accuracy. In doing this, we employed score calibration and fusion with Logistic Regression (LogReg). The metric used to assess the accuracy was the likelihood-ratio cost (Cllr), while the factors of analysis include the database numerosity, different software implementation of LogReg, and the strategy to consider the features available for the different vowels. Our results show that the LogReg calibration has systematically improved the Cllr, suggesting that also the accuracy of the tested FSASR approaches may be improved introducing calibration. Furthermore, our findings suggest that the tested FSASR approaches give more accurate results when the IMPAVIDO database is used.

Keywords: Accuracy, Likelihood-Ratio, Bayesian framework, Forensic Voice Comparison.

1. *Introduction*

In Forensic Speaker Recognition (FSR), the recording of an anonymous speaker is compared with a recording of a speaker whose identity is known. The aim is to help a court of law to decide if the two recorded voices could belong to the same speaker (Morrison, Enzinger, Phil and Zhang, 2017). Many approaches have been proposed to achieve FSR: nonparametric-subjective approaches such as auditory, spectrographic (also known as "voicegram" identification, or "voiceprinting": Kersta, 1962; Tosi, Oyer, Lashbrook, Pedrey, Nicol and Nash, 1972), and parametric-objective approaches, which are based on quantitative measurements and analysis of specific features of the speech signal (Rose, 2002; Rose, 2006) particularly those associated with evidence, are exemplified and critically discussed, and comparisons drawn with generic Speaker Recognition. The centrality of the Likelihood Ratio of Bayes' theorem in correctly evaluating strength of forensic speech evidence is emphasised, as well as the many problems involved in its accurate estimation. It is pointed out that many different types of evidence are of use, both experimentally and forensically, in discriminating same-speaker from different-speaker speech samples, and some examples are given from real forensic case-work to illustrate the Likelihood Ratio-based approach. The extent to which Technical Forensic Speaker Recognition meets the Daubert requirement of testability is

also discussed. © 2005 Elsevier Ltd. All rights reserved.","author":[{"dropping-particle":"","family":"Rose","given":"Phil","non-dropping-particle":"","parse-names":false,"suffix":""}],"container-title":"Computer Speech and Language","id":"ITEM-1","issue":"2-3 SPEC. ISS.","issued":{"date-parts":[["2006"]]},"page":"159-191","title":"Technical forensic speaker recognition: Evaluation, types and testing of evidence","type":"article-journal","volume":"20"},"uris":["http://www.mendeley.com/documents/?uuid=53443a14-a72e-4773-9563-56b19157e8a9"]}],"mendeley":{"formattedCitation":"(Rose, 2006; Jessen, 2008). The objective approaches, in turn, can be divided in acoustic-phonetic, or semi-automatic approaches, and automatic approaches. In the former, the expert plays a key role in the choice of the phonetic units and the extraction of the features to be compared. In the automatic approaches, these steps are mostly accomplished by the software program that also perform the voice comparison.

About the output of automatic and semi-automatic approaches, in recent years the Likelihood Ratio (LR) framework has been proposed as a logical way to report evidence to the court of law (Aitken, Taroni, 2004): furthermore, it has been recommended by the European Network of Forensic Science Institutes (ENFSI)[1] as the ideal method to use in FSR. The LR represents a balance between the probability to get the evidence given the prosecution hypothesis (or same-origin hypothesis, e.g., the questioned material has been generated by the suspect) and the probability to get the same evidence given the defence hypothesis (or different-origin hypothesis, e.g. the questioned material has been generated by someone else). Therefore, two important properties can be recognized to the LR:

1. The LR represents the degree of similarity of a pair of samples (the two recordings), while taking into consideration their typicality with respect to a model of the relevant population, so that a more positive value indicates stronger support for the voices being from the same individual (and vice-versa)
2. The LR value has a meaning by itself, in the sense that a LR value greater than 1 supports the prosecution hypothesis and, more precisely, a LR value equals to 3 (for instance) means that the evidence supports the prosecution hypothesis with a degree of 3 versus 1 (Ramos, Krish, Fierrez and Meuwly, 2017).

In general, the output of a FSR system is (at least) a number, namely a "score", that may have the first property of LR (in this sense, a score may be LR-like) but does not necessarily also have the second property, i.e. it does not have a clear probabilistic interpretation. Indeed, these scores may contain some bias due to condition mismatch between the data that the system or speaker model was trained on and the conditions of the verification speech. Therefore, after calculating the scores, a further step may be necessary to convert them into true LRs, to be compliant with the LR framework: this step is called *calibration*. It is important to note that the same term, *calibration*, is often used also to indicate a property of a set of LR values, which can be measured, and it is strictly related to the accuracy of a FSR system (Ramos, Gonzalez-Rodriguez, 2013) an incorrect selection of databases, a bad

[1] http://www.enfsi.eu.

choice of statistical models, low quantity and bad quality of the evidence are factors that may lead to likelihood ratios supporting the wrong proposition in a given case. However, measuring performance of LR values is not straightforward, and adequate metrics should be defined and used. With this objective, in this work we describe the concept of calibration, a property of a set of LR values. We highlight that some desirable behavior of LR values happens if they are well calibrated. Moreover, we propose a tool for representing performance, the Empirical Cross-Entropy (ECE.

If the same casework can be processed by two or more FSR systems, their output scores may be combined into a single outcome. This can be achieved by a *fusion* procedure, which aims to produce a compound system that performs better than the single ones.

Score calibration methods based on generative models have recently gained interest (van Leeuwen, Brümmer, 2013; Brümmer, Swart and van Leeuwen, 2014; Plchot, Matějka, Silnova, Novotny, Diez, Rohdin and Glembek, 2017; Cumani, Laface, 2019), but the most widely adopted method is still the score calibration by means of Logistic Regression (Morrison, 2013), a well know and largely applied technique in machine learning, developed for automatic speaker recognition and then become popular in FSR.

The aim of this study is to evaluate the performance of two FSASR approaches traditionally used in Italy, adopting LogReg techniques applied to calibration and fusion. The two systems referred to are IDEM/SPREAD (first appeared in Falcone *et al.*, 1992; see Sigona and Grimaldi, 2015 for a discussion and further references), and SMART (Bove, Giua, Forte and Rossi, 2002; Jona Lasinio, Rossi and Bove, 2004; Romito, Bove, Delfino, Rossi, and Jona Lasinio, 2009). The main difference between the two systems is that IDEM/SPREAD is based on parametric (multivariate Gaussian) statistical models, while SMART is based on non-parametric statistical models.

Recently, both approaches have been integrated into eIDEM and eSMART emulators that are the two main components of a new tool for FSASR, named IMPAVIDO (Sigona, Grimaldi, 2017), and which is used for this study, together with some related variants under development.

The remaining of this paper is organized as follows. Section 2 recalls some fundamentals about LogReg in terms of score calibration and fusion. Section 3 describes the score computation by eIDEM, eSMART and their variants employed for the current study. Sections 4 and 5 describe the experimental setup and present the achieved results. Finally, in Section 6 the main findings are discussed, and future research directions are highlighted.

2. *Logistic Regression and score calibration*

As shown in formulas (1) to (3), if p is the probability of the prosecutor hypothesis, then the LR can be expressed as odds of p (1), the logarithm of LR (let LLR) is the log-odds of p (2), and p is related to LLR by means of the logistic function (3):

$$(1) \quad LR = \frac{p}{1-p}$$

$$(2) \quad LLR = \log LR = \log \frac{p}{1-p}$$

$$(3) \quad p = \frac{1}{1+e^{LLR}}$$

If p>0.5, then necessarily 1-p<0.5, then we say that there is more support for the prosecution hypothesis than the defence one. In this case, is straightforward to observe that LR must be greater than 1, and this is why true LRs have a clear probabilistic interpretation, while raw uncalibrated "scores" do not (a score merely greater than 1 does not necessarily means that p>0.5, i.e. it does not necessarily support the prosecution hypothesis).

What we need now is a function that maps the LR-like scores output by FSR systems to LRs: this is where LogReg comes in. According to LogReg, the following affine relation in the log-odds space gives the score-to-LR transformation:

$$(4) \quad \log(LR) = w_0 + w_1 s$$

where s is the logarithm of the score (log-score), w_0 is a shift coefficient and w_1 is a scale coefficient that must be estimated by means of an appropriate training.

LogReg can also be applied to calibration and fusion of multiples scores coming from different FSR systems, which may process, with their own specificities, the *same* couple of voice segments. In this case, the transformation in (4) must be slightly changed by keeping the shift factor w_0 and including as many w_i scale coefficients as the number of scores (i.e. FSR systems) to fuse.

In both cases of calibration and calibration plus fusion, the required training data consist of a set of scores from comparisons where some are known to be *same-origin comparisons* and others are known to be *different-origin comparisons*, taking the speakers data from the database of the relevant population (at least two different sessions for each speaker are required to make same-speaker comparisons). The estimation of the LogReg coefficients is usually done by means of iterative procedures that minimize a cost function. In the current study, prior probabilities are implicitly set to 0.5 (prior odds = 1, i.e. there is no a-priori preference for the prosecution hypothesis with respect to the defence hypothesis). Hence, the cost function to be minimised, i.e. the deviance statistic, is equivalent to the log-likelihood-ratio cost Cllr given by the following equation (Brümmer, Du Preez, 2006; van Leeuwen, Brümmer, 2007):

$$(5) \quad C_{llr} = \frac{1}{2}\left(\frac{1}{N_{so}}\sum_{i=1}^{N_{so}} \log_2\left(1+\frac{1}{LR_{so_i}}\right) + \frac{1}{N_{do}}\sum_{j=1}^{N_{do}} \log_2(1+LR_{do_j})\right)$$

where N_{SO} and N_{DO} are the number of same-origin and different-origin comparison, $LR_{SO}i$ and $LR_{SO}i$ are the LR values for same-origin and different-origin comparisons.

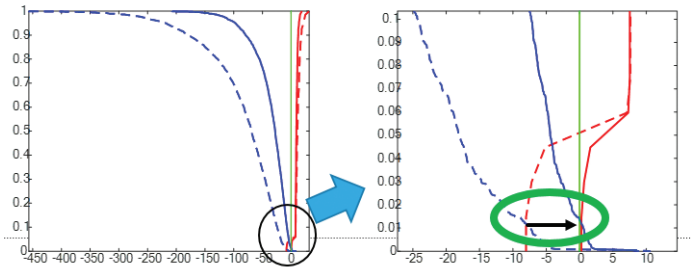
Cllr represents the total probability of error, at any possible prior, that the fact finder would obtain by using the LR value computed by the forensic expert,

in the Bayesian decision framework. Thus, the less the Cllr, the more accurate is the system, and the Cllr computed on calibrated scores, i.e. on LLR, values, is expected to be smaller than the Cllr computed on uncalibrated raw scores. In other words, Cllr can be used as an index to assess the performance of calibration and fusion, i.e. the FSR system, as depicted in Figure *1*.

In doing this, different software implementations of LogReg have been tested:

a) the Focal toolkit[2];
b) the Bosaris toolkit (Brümmer & Villiers, 2013).
c) a variant by Morrison[3] of the LogReg training function of the Focal Toolkit (Morrison, 2011);
d) the *mnrfit* Matlab® function for multinomial regression.

Figure 1 - *Tippet plots illustrate the Probability of false identification (P.f.id., depicted by blue lines on the left side of each plot) and Probability of false rejection (P.f.rej., represented by red lines on the right side of each plot) as functions of log-scores (horizontal axis), for both uncalibrated (dashed lines) and calibrated (solid lines) scores. The plot on the right provides a close-up view of the left plot, clearly demonstrating that, following the calibration process, the point where the two error probabilities (and thus the supports of the two competing hypotheses) are equal approaches log-score = 0, as would be anticipated for a LR value (indeed LR=1 equates to LLR = 0). In this specific instance, the Cllr for uncalibrated scores was 0.226, while the Cllr for the same scores after calibration notably improved to 0.034*



3. *The tested FSASR systems.*

In this study, traditional Italian FSASR approaches have been tested, by means of emulators included in the IMPAVIDO software, eIDEM and eSMART. Both emulators, like the original software, rely on pre-extracted features values, namely, the fundamental and the first three formant frequencies (F0, F1, F2, F3) of the Italian vowels. The IDEM/SPREAD software was designed to handle only 4 Italian vowel /a, ɛ, i, ɔ/, and therefore up to 16 features altogether. The vowel /u/, for example, cannot be input in the IDEM/SPREAD software by design. Due to this limitation, even if the IMPAVIDO software can manage a 7-vowel system

[2] Niko Brummer, "Focal toolkit," available at https://sites.google.com-/site/nikobrummer/focal.
[3] The code of Morrison is available at http://geoff-morrison.net.

/a, ɛ, e, i, ɔ, o, u/ (and can also be easily customized by the authors to manage the formants of any additional vowel sound), in the continuation of this work all the test will be conducted on the 4 vowels listed above.

eIDEM is an emulator of IDEM/SPREAD method. According to this method:

a) the between-speaker variability of the features is modelled by a multivariate Gaussian distribution whose centroid (the vector of means) and variance-covariance matrix are estimated from the database of features that comes with the software (or optionally provided by the user him/herself).
b) the within-speaker variability is modelled by a multivariate Gaussian distribution, whose variance-covariance matrix is estimated from the database, and whose centroid is estimated from the values of the suspect features; this also implies the following:
c) the statistical distribution of the (squared) Mahalanobis distance between the features of the anonymous and the suspect speakers can be modelled by a Chi-squared distribution with as many degrees of freedom as the number of acoustic features used for the test (e.g. 16).

Based on the above assumptions, the software calculates:

1. the value of the (squared) Mahalanobis distance between the feature sets of the anonymous and the suspect (let d^2_{AS}).
2. the probability of false rejection (let $P_{f.rej.}$) as the area of the queue of the above Chi-square distribution starting from the value d^2_{AS}.
3. the value of the probability of false identification (let $P_{f.id.}$), as the integral of the Gaussian probability density function (p.d.f.) that models the between-speakers variability, onto the ellipsoid with squared radius equals to d^2_{AS} and centred to the suspect centroid.
4. a distance threshold (let d^2_{th}) as the smaller between the value of the inverse Chi-square (by default, computed at the assigned value α=0.01) and the value of the squared radius of the ellipsoid which would give a $P_{f.id.}$ equals to a second assigned value (by default, 0.001).
5. the response IDENTIFIED and a LR-like score given by the ratio $(1\text{-}P_{f.rej.})/P_{f.id.}$, if $d^2_{AS} > d^2_{th}$, or the response UNIDENTIFIED and a LR (-like) score given by the ratio $(1\text{-}P_{f.id.})/P_{f.rej,}$ otherwise (for these formulas, see Federico, 2010; but also see Sigona & Grimadi, 2015), about some anomalies in some releases of IDEM/SPREAD).

As far as the calibration and fusion experiments is concerned, two variants of eIDEM were introduced (without any attempt to obtain scores already calibrated at this level):

a) eIDEM_NOTH: it does not care about distance threshold (indeed, the "pure" LR framework does not demand any binary decision depending on thresholds), but simply outputs a score computed as $(1\text{-}P_{f.rej.})/P_{f.id}$, i.e. the same formula that eIDEM uses in case of identification.

b) eIDEM_LK: as eIDEM_NOTH, it ignores the threshold and simply outputs a score given by the ratio between the above mentioned Chi2-squared p.d.f. at the d^2_{AS} point and the between-speaker variability p.d.f. at the point given by the anonymous centroid.

The last FSASR method used in the current study is the one implemented by eSMART, which is based on non-parametric statistics described in (Bove *et al.*, 2002). In this case, the LR (-like) result is given by the ratio of the kernel estimates of the p.d.f. modelling the distance between features of the suspect from himself and the p.d.f. modelling the distance between the suspect ad any other speaker of the database, at a point given by the distance between the suspect and the anonymous speakers.

4. *Experiment setup*

As already mentioned, this study is based on F0, F1, F2 and F3 of the Italian /a, ɛ, i, ɔ/ vowels. Two databases of features representative of the Italian reference population have been used.

The first one is the default database of the IDEM/SPREAD software. This resource essentially consists of a single table containing 26642 rows (records), each of which stores a vector (a quadruple) of values of the parameters F0, F1, F2, F3, together with the indication of the corresponding vowel and an additional text string that indicates (anonymously) the respective speaker. The formants calculation was based on environmental or telephone interceptions. Except for knowing that all the speakers in question, 428, are male, no other information about the socio-linguistic composition of this database is available. Therefore, it is an absolutely unstructured database, which, however, has been used in conjunction with the software for years (and probably still is).

The study would certainly have benefited from a comparison between this database and the SMART database, if only for the sake of fairness. However, it seems that accessing the SMART database is not possible. Nevertheless, it was decided to introduce a second database into the tests, similarly unstructured as the IDEM/SPREAD one. This is the default database distributed with IMPAVIDO, and represents the formants of 310 male voices, in similar recording conditions of the IDEM/SPREAD database.

To assess the capability of an FVC system to recognize two recordings of the same speaker, it is necessary to have a database where each speaker is represented by at least two distinct, non-contemporaneous recordings. However, in both available databases, such a distinction is not present. Therefore, for the purposes of this study, to make same-origin comparison it was necessary to split the single feature matrix of each speaker into two equisized feature matrices, simulating two different sessions for each speaker. In the IMPAVIDO database, each speaker has 20 occurrences for each vowel; thus, 2 sessions of 10 occurrences for each vowel for all speakers were simulated. This number of occurrences is quite adequate to

reproduce a real casework situation. In the IDEM/SPREAD database, only 67 speakers out of 428 have at least 20 occurrences for each vowel. Therefore, for the current study, only those 67 speakers were used from that database.

All the required same-origin and different-origin comparison were computed with a cross-validation approach, separately for each database.

In the case of the IDEM/SPREAD database, the uncalibrated scores were computed by doing all the possible comparisons between a session of a speaker and the other session of the same (for same-origin) or different (for different-origin) speaker. For each comparison, reference population model has been trained using all database content except the speaker (for same-origin) or the two speakers (for different-origin) used for the comparison: in this way, we avoided probable bias. The calibrated scores were computed by calibrating each single score, while the others have been used to train the LogReg procedure.

The above approach was not feasible for the IMPAVIDO database, due to its much larger number of speakers. In this case, a K-fold cross-validation has been used, in which the database content was split into K equi-sized groups of speakers. Each group was used to make all the possible comparisons among its members, while the remaining K-1 groups were used to train the model of the reference population for those comparisons. The calibrated scores were computed in a similar way.

All the results were calculated considering both the full dimensional problem, i.e., the four vowels together, and separately vowel-by-vowel, so that the behaviour of LogReg calibration was checked in both cases.

The fusion was performed for each couple of available emulators, for single-vowel and all-vowels conditions, fusing uncalibrated scores and fusing already calibrated scores.

5. *Results and discussion*

In the Table 1, the achieved Cllr values for uncalibrated scores, and for those calibrated with the different LogReg implementation, for each emulator, in all-vowel and single-vowel condition, for both the selected IDEM/SPREAD and IMPAVIDO database are showed.

In any case, it is confirmed that LogReg calibration always improves the Cllr. The best calibration variant, in the sense that leads to the lowest mean Cllr with the lowest standard deviation, is not very clear because it seems to be dependent on the emulator and the database. The best result has been achieved with the Bosaris toolkit with the eIDEM_LK emulator, using all the vowels, for the IMPAVIDO database.

Table 1 - *Results of calibration of single method (eIDEM.aeio means that all vowels have been used; eIDEM.a means that one vowel only has been used). Each cell contains the mean value of the log-likelihood-ratio cost (Cllr) and its standard deviation (in parentheses). "Raw" means "uncalibrated"*

	Selected IDEM/SPREAD database					*IMPAVIDO database*				
System id.	Raw	Focal	Morr.	Bosaris	Mnrfit	Raw	Focal	Morr.	Bosaris	Mnrfit
eIDEM .aeio	0,30 (0,1)	0,05 (0,02)	0,04 (0,01)	0,04 (0,02)	0,12 (0,03)	0,04 (0,02)	0,01 (0,01)	0,012 (0,001)	0,02 (0,04)	0,011 (0,009)
eIDEM .a	0,80 (0,1)	0,8 (0,1)	0,23 (0,04)	0,23 (0,04)	0,70 (0,1)	0,74 (0,04)	0,14 (0,02)	0,120 (0,009)	0,74 (0,04)	0,29 (0,02)
eIDEM _NOTH. aeio	0,98 (0,01)	0,04 (0,02)	0,035 (0,01)	0,04 (0,02)	0,12 (0,03)	1,000 (0,009)	0,01 (0,02)	0,006 (0,002)	0,03 (0,06)	0,04 (0,11)
eIDEM _NOTH.a	0,99 (0,01)	0,27 (0,03)	0,27 (0,03)	0,27 (0,03)	0,91 (0,06)	1,000 (0,009)	0,14 (0,01)	0,14 (0,01)	0,14 (0,01)	0,34 (0,02)
eIDEM _LK.aeio	15 (0,5)	0,11 (0,02)	0,11 (0,02)	0,11 (0,02)	0,52 (0,09)	4,6 (0,1)	0,1 (0,3)	0,020 (0,001)	0,004 (0,003)	0,009 (0,003)
eIDEM _LK.a	3,86 (0,09)	0,24 (0,04)	0,23 (0,04)	0,24 (0,04)	0,65 (0,07)	2,05 (0,05)	0,11 (0,01)	0,113 (0,008)	0,109 (0,009)	0,27 (0,02)
eSMART .aeio	3,6 (2,1)	0,29 (0,10)	0,3 (0,1)	0,29 (0,10)	0,80 (0,2)	3.1 (1.5)	0,15 (0,04)	0,146 (0,036)	0,15 (0,04)	0,33 (0,07)
eSMART .a	2,0 (1,2)	0,35 (0,15)	0,3 (0,1)	0,34 (0,14)	1,1 (1,0)	1.5 (0,5)	0,21 (0,03)	0,20 (0,03)	0,21 (0,03)	0,50 (0,06)

Another research question of this study was to verify whether it is more convenient, in terms of Cllr, to combine emulators, each working on a different vowel, than having a single emulator that works on all the vowels simultaneously. The answer, as shown in the Table 2, is that it seems to depend on the FSASR emulator and the database. Anyway, the results clearly show that for the eSMART emulator is always better to let it work on each vowel separately and then fusing the scores. This means that the current implementation of eSMART (inspired to Bove *et al.*, 2002), that uses all the 4 vowels together, can be improved.

The next question is to decide for the best emulator, in terms of Cllr. The answer is achieved by means of a paired samples t-test (1 tail), $\alpha = 5\,\%$ on the Cllr values, and it is depicted in the Table 3. Using the selected IDEM/SPREAD database the best Cllr is achieved with eIDEM_NOTH emulator (working on all the vowels) and Morrison's LogReg variant for calibration, while using the IMPAVIDO database

the best result is given by eIDEM_LK and Bosaris toolkit for calibration. Even in this experiment, the result achieved using IMPAVIDO database was far better.

Table 2 - *Experimental results about the fusion of the single-vowel systems. In each cell of the table, the best fusion strategy, and the achieved mean log-likelihood-ratio cost (Cllr) are indicated*

	Selected IDEM/SPREAD database	*IMPAVIDO database*
Focal		eSMART: fusing the RAW single-vowel systems (Cllr=0,6)
Morrison	eIDEM_LK: fusing the RAW single-vowel systems (Cllr=0,04) eSMART: fusing the RAW single-vowel systems (Cllr=0,04)	eIDEM_LK: fusing the RAW single-vowel systems (Cllr=0,017) eSMART: calibration of the single-vowel systems, then fusion (Cllr=0,06)
Bosaris	eIDEM_LK: fusing the RAW single-vowel systems (Cllr =0,07) eSMART: fusing the RAW single-vowel systems (Cllr =0,254	eSMART: fusing the RAW single-vowel systems (Cllr=0,05)
Mnrfit	eIDEM_LK: fusing the RAW single-vowel systems (Cllr=0,14) eSMART: fusing the RAW single-vowel systems (Cllr=0,30)	eSMART: calibration of the single-vowel systems, then fusion (Cllr=0,06)

Table 3 - *The performance of the single emulators. Log-likelihood-ratio cost (Cllr) vales are expressed as mean value and standard deviation (in parentheses)*

	Selected IDEM/SPREAD database	*IMPAVIDO database*
Focal	eIDEM Full dim.calibrated: Cllr= 0.05 (0.02) eIDEM_NOTH Full dim.calibrated: Cllr= 0.04 (0.02)	eIDEM Full dim.calibrated: Cllr=0,01 (0,01)
Morrison	**eIDEM_NOTH Full dim. calibrated: Cllr = 0.03 (0.01)**	eIDEM_NOTH Full dim. calibrated: Cllr=0.006 (0.002)
Bosaris	eIDEM Full dim.calibrated: Cllr= 0.04 (0.02) eIDEM_NOTH Full dim.calibrated: Cllr= 0.04 (0.02)	**eIDEMLK Full dim. calibrated: Cllr = 0.004 (0.003)**
Mnrfit	eIDEM Full dim.calibrated: Cllr= 0.12 (0.03) eIDEM_NOTH Full dim.calibrated Cllr= 0,12 (0.03)	eIDEM_LK Full dim. calibrated: Cllr=0,009 (0,003)

The last experiment concerned with the fusing process of two different emulators. This is necessary in order to try to get a better new system. The findings show that in a very few cases we found a fusion which was better than any other "single" emulator in terms of Cllr (paired samples t-test (1 tail), α =5 %). Conversely, the

improvement was always very small, as highlighted in the Table 4. Anyway, among the experimental results we obtained, also the best-fused system (i.e. the one highlighted in Table *4*) was not better than the best single system, i.e. the eIDEM_LK full dimensional with Bosaris calibration. (Cllr= 0.004 (0.003)).

Table 4 - *results about the fusion of emulators. Log-likelihood-ratio cost (Cllr) values are expressed as mean value and standard deviation (in parentheses)*

Database	*Calib. method*	*Best achieved fusion*	*...better than the best single system, which is*
selected IDEM	Focal	eIDEM Full dim. uncalib. + eIDEM_NOTH Full dim. uncalib.: Cllr= 0,03 (0,02)	eIDEM_NOTH Full dim calib.: Cllr= 0.04 (0.02)
IMPAVIDO	Morrison	eIDEM_NOTHFull dim. uncalib + eIDEM_LK with Calibrated fused vowels: Cllr= 0,005 (0,002)	eIDEM_NOTH Full dim. Calib.: Cllr=0.006 (0.002)

6. *Conclusions and further remarks*

In the present study, the performances of two FSASR approaches traditionally used in Italy, namely IDEM/SPREAD and SMART, were evaluated through emulations included in the IMPAVIDO software, using LogReg techniques applied to calibration and fusion.

For this purpose, tests were conducted using the default database of IDEM/SPREAD, which does not contain socio-linguistic information about speakers. Since the database of SMART is not accessible, similar tests were also conducted on the default database of IMPAVIDO, which, like that of IDEM/SPREAD, does not contain socio-linguistic information about speakers. Furthermore, the tests were conducted on the vowels /a, ɛ, i, ɔ/, not on the complete vowel system of the Italian language, to maintain the same limitation as the IDEM/SPREAD software.

Despite these limitations, the work presents significant strengths. As far as is known, fusion procedures have never been applied (nor published) to either IDEM/SPREAD or SMART, neither previously documented in literature, nor by the authors of the respective systems, nor by independent researchers. Furthermore, it does not seem that the two systems, with calibration and fusion applied, have been compared in this manner before. Additionally, there is also no previous indication of a software tool, similar to the one developed and employed by the authors, being available to conduct such a comparative study.

The first finding is that, according to the experimental hypothesis, the LogReg calibration has improved the Cllr in every circumstance, suggesting that also the accuracy of the tested FSASR approaches may be improved introducing calibration.

The second finding is that the tested FSASR approaches give more accurate results when the IMPAVIDO database is used. This may be due to the larger

number of speakers and the required number of vowel occurrences with respect to the selected IDEM/SPREAD database.

These preliminary results suggest that a more detailed study would be necessary in order to explore the behaviour of such approaches depending on other factors such as the numerosity of the feature sets to compare (i.e. the number of occurrences for each vowel) and possibly the noise conditions. Reducing the requirements on vowel quantity, at least for anonymous sample, can help to simulate those caseworks where very limited data are available: this is a quite simple operation. On the other hand, increasing that prerequisite, or even replicate the same experiment with introducing a sufficiently large number of noisy speech recordings, is a challenge task because **one** need to accurately label a very large number of vowels, which is very time-consuming.

Also, both the tested databases do not distinguish between different speaking sessions of a same speaker; neither have any other meta-data such as noise level. On the other hand, calibration can be influenced by the actual conditions in which the speech recordings have been produced. In the present study, we have implicitly assumed that the features values in both the tested databases where representative of typical casework conditions, where the signal-to-noise ratio (SNR) of the speech was above 10 dB, which is a threshold commonly accepted to extract spectro-acoustic measurements on vowels. Probably, a deeper study of the impact of calibration on the tested FSASR would add more details to the picture.

Another important conclusion may be drawn specifically for the eSMART emulator. As already mentioned, it is a component of the software IMPAVIDO which was inspired to the work of Bove *et al.* (2002), which in turn can be considered the basis for the modern SMART software. The achieved results show that the way in which it uses the vowel information is not effective: a better system can be achieved by using eSMART on each single vowel, and then use LogReg to fuse the four scores into a calibrated and fused LR. This result appears to be of considerable importance. In fact, it suggests that even the original SMART software might function more effectively than it presumably does today. In such a case, it could reasonably be suggested to conduct partial tests on each vowel separately from the others, and then obtain the final comparison result through the fusion of the individual partial results. A specific study on that software is therefore highly suggested.

The tested fused systems were sometimes able to outperform each single system that participated to fusion, but it was not able to perform the best single system we achieved. This may be explained by the fact that, despite the difference in the statistical approaches between eIDEM and eSMART, both systems work on the same features, which may be somehow redundant. From this point of view, a future direction of investigation might be to fuse a FSASR system based on a different kind of features, or, even better, to fuse a semi-automatic with fully automatic approach.

References

Aitken, C.G.G., Taroni, F. (2004). *Statistics and the evaluation of evidence for forensic scientists*. Chichester, Wiley.

Bove, T., Giua, P.E., Forte, A. & Rossi, C. (2002). Un metodo statistico per il riconoscimento del parlatore basato sull'analisi delle formanti. In *Statistica*, 62(3), 475-490.

Brümmer, N. & Du Preez, J. (2006). Application-independent evaluation of speaker detection. In *Computer Speech and Language*, 20(2-3), 230–275.

Brümmer, N., & Villiers, E.D. (2013). The BOSARIS Toolkit: Theory, Algorithms and Code for Surviving the New DCF. ArXiv, abs/1304.2865.

Brümmer, N., Swart, A. & van Leeuwen, D. (2014). A comparison of linear and nonlinear calibrations for speaker recognition. In *Odyssey 2014: The Speaker and language Recognition Workshop*, Tokyo, Japan, 1-5- November, 14–18.

Cumani S., Laface, P. (2019). Tied Normal Variance–Mean Mixtures for Linear Score Calibration. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 12-17 May, 6121-6125.

Falcone, M., Paoloni, A., De Sario, N., Saverione, B. (1992). IDEM: un sistema per l'analisi e la rappresentazione vocale. In *Associazione Italiana di Acustica, XX Congresso Nazionale*, Rome, 20-24 Aprile, 417-422.

Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M & Ortega-Garcia, J., (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. In *Computer Speech and Language*, 20(2-3 SPEC. ISS.), 331–355.

Jessen, M. (2008). Forensic phonetics. In *Language and Linguistics Compass*, 2, 671–711.

Jona Lasinio, G., C. Rossi & T. Bove (2004). The Speaker Recognition problem. In *Proceedings of the XLII Scientific Meeting of the Italian Statistical Society (SIS)*, Palermo, 20-22 June, 429–440.

Kersta, L.G. (1962). Voiceprint identification. In *Nature*, 196, 1253–1257.

Morrison, G.S. (2011). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model-universal background model (GMM-UBM). In *Speech Communication*, 53(2), 242–256.

Morrison, G.S., (2013). Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. In *Australian Journal of Forensic Sciences*, 45(2), 173-197.

Morrison, G.S., Enzinger M., Phil, E., Zhang, C. (2017). Forensic Speech Science 2017. 183–197. In Freckelton I., Selby H. (a cura di), *Expert Evidence* (Ch. 99). Sydney, Australia: Thomson Reuters, 1-140.

Plchot O., P. Matějka, Silnova, A., Novotny, O., Diez, M., Rohdin, J., Glembek, O., *et al.* (2017). Analysis and description of ABC submission to NIST SRE 2016. In *Interspeech Proceedings*. Stockholm, Sweden. 20-24 August, 1348-1352.

Ramos D., Krish R.P., Fierrez J., Meuwly D. (2017). From Biometric Scores to Forensic Likelihood Ratios. In Tistarelli, M., Champod, C. (a cura di), *Handbook of Biometrics for Forensic Science. Advances in Computer Vision and Pattern Recognition*. Berlin: Springer, 305-327.

Ramos, D. & Gonzalez-Rodriguez, J. (2013). Reliable support: Measuring calibration of likelihood ratios. In *Forensic Science International*, 230(1–3), 156–169.

Romito, L., Bove, T., Delfino, S., Rossi, C., Jona-Lasinio, G.J. (2009). Specifiche linguistiche del database utilizzato per lo Speaker Recognition in S.M.A.R.T. In Romito, L., Lio, R., Galatà, V. (a cura di), *La Fonetica Sperimentale: metodo e applicazioni*, Torriana: EDK Editore, 632–640.

Rose, P. (2002). *Forensic speaker identification*. London: Taylor and Francis.

Rose, P. (2006). Technical forensic speaker recognition: Evaluation, types and testing of evidence. In *Computer Speech and Language*, 20(2-3 SPEC. ISS.), 159–191.

Sigona, F., Grimaldi M. (2015). Il riconoscimento del parlante in ambito forense: uno studio indipendente sul software IDEM/SPREAD in uso ai Carabinieri. In *Sicurezza e Giustizia*, IV, 16–21.

Sigona, F., Grimaldi M. (2017). Tools for Forensic Speaker Recognition. In Orletti, F., Mariottini, L. (a cura di), Forensic Communication: Theory and Practice, Cambridge: Cambridge Scholars Publishing, 127-150.

Tosi, O., Oyer, H., Lashbrook, W., Pedrey, C., Nicol, J., & Nash, E. (1972). Experiment on Voice Identification. In *Journal of the Acoustical Society of America*, 51(6B), 2030–2043.

van Leeuwen, D.A., Brümmer, N. (2007). An introduction to application independent evaluation of speaker recognition systems. In Müller, C. (a cura di), *Speaker Classification I: Fundamentals, Features, and Methods*., Heidelberg: Springer-Verlag, 330–353.

van Leeuwen, D.A., Brümmer, N. (2013). The distribution of calibrated likelihood-ratios in speaker recognition. In *Interspeech Proceedings* (ISCA), Lyon, France, 25-29 August, 1619–1623, 2013.