A Statistical Approach to Speaker Identification in Forensic Phonetics

Fabio Leuzzi[1(✉)], Giovanni Tessitore[2], Stefano Delfino[2], Claudio Fusco[2], Massimo Gneo[2], Gianpaolo Zambonini[2], and Stefano Ferilli[1]

[1] Dipartimento di Informatica, Università di Bari, Bari, Italy
{fabio.leuzzi,stefano.ferilli}@uniba.it
[2] Servizio Polizia Scientifica, Polizia di Stato, Rome, Italy
{giovanni.tessitore,stefano.delfino,claudio1.fusco, massimo.gneo}@poliziadistato.it, gianpaolo.zambonini@interno.it

Abstract. Speaker identification can be summarized as the classification task that determines if two voices were spoken by the same person or not. It is a thoroughly studied topic, since it has applications in many fields. One is forensic phonetics, considered very hard since the expert has to face ambient noise, very short recordings, interference, loss of signal, and so on. For decades, these problems have been tackled by experts using their listening abilities, and each of them might represent a research area on its own. The use of semi-automatic techniques may represent a modern alternative to the subjective evaluation of experts, that may enforce fairness of the classification procedure. In a nutshell, we use the differences in speech of a set of different voices to build a population model, and the suspected person's voice to build a speaker model. The classification is carried out evaluating the similarity of a further speech sample (the evidence) with respect to the models. Preliminary evaluations shown that our approach reaches promising results.

1 Introduction

The *speaker identification* problem [11,15,16,35] can be cast as a classification task aimed at determining if two voices were spoken by the same person or not. In this broad sense, it has applications in many fields. In particular, it is a crucial task in forensics, where there is a need to determine the speaker in phone calls. This application domain adds further complexity to the task because calls are typically short in duration with poor quality, ambient noise, interference, loss of signal (in the case of mobile phones), and reduced bandwidth may yield dramatic consequences. Traditionally, the problem has been tackled leveraging abilities of human experts in evaluating the similarities between voices, or in finding peculiarities and defects that allow one to identify the speaker. However, this practice has its drawbacks, among which the limited capabilities of humans in considering complex mixes of parameters and their subjectivity in evaluation.

Nowadays, the most popular methods for speaker identification are the following: (1) listening based methods [24]; (2) spectrograms comparison

A. Appice et al. (Eds.): NFMCP 2016, LNAI 10312, pp. 69–83, 2017.
DOI: 10.1007/978-3-319-61461-8_5

techniques [10,19]; (3) phonetic parameters analysis [1,12,27,37]; (4) automatic techniques [8,9]. In particular, the latter represents a modern alternative to overcome the subjective evaluation of experts, since it relies on algorithmic procedures to predict whether two voices come from the same speaker or not. So, it may ensure more fairness to the classification procedure.

The need of fairness is one of the main motivations for which this research field is so primary for judiciary contexts. Forensics aims to be a fair scientific support to the logical composition of crime events. In this case, such support regards the phone-speaker identification.

From a technical point of view, we can distinguish between *closed* tests (aimed at finding the speaker in a set of voices that surely includes a sample of the speaker's voice) and *open* tests (where this is not ensured). This paper deals with the open case, proposing a technique that uses the differences in speech of a set of different voices to build a population model, and the suspected person's voice to build a speaker model, and then carries out the classification by evaluating the similarity among these models and the anonymous voice. While a preliminary evaluation of this approach was presented in [34], this work aims at a specific analysis of results with respect to the feature selection perspective.

The remainder of this work will present some related work and preprocessing details in Sect. 2, then our approach follows in Sect. 3, after that experimental evaluation is reported in Sect. 4. Forensic results must be understandable to the Court, then Sect. 6 proposes a human understandable translation of the possible classification outcome. Finally, Sect. 7 will conclude with some considerations and future works.

2 Related Works, Background and Preprocessing

Different features may describe the sounds produced by the human vocal apparatus, depending on how it is classified. A first classification is between consonants and vowels. Consonants are produced by forcing air passage in the restricted vocal apparatus. They can be further divided in voiceless, if produced without vibration of vocal cords, or voiced, otherwise. Vowels are produced when the apparatus puts no obstacles, and the sound is determined by the position of tongue and lips. Specifically, they are a periodic signal produced by three factors: the periodic movement of vocal cords that produces the *fundamental frequency* (f_0 – related with the vocal tone of a person); the noise produced by the phonation; the modification of the sound caused by the sound expansion in the mouth. Such components make up the *frequency spectrum*. It is characterized by a sequence of peaks that change depending on the type of sound pronounced, a complex result of the cooperation of tongue, teeth, palate, lips, and so on. The frequency spectrum interacts with the harmonic structure of speech (integer multiples of the fundamental frequency). The harmonics near to the resonance frequency are called *formants*.

A spectrogram is a plot that represents the components of the sound in three dimensions: time (on the x axis), frequency (on the y axis) and intensity (represented using several color scales, here intensities of gray are exploited. The inner values are usually represented in Hertz. The lower frequency is known as *first formant* f_1, followed by the successive peaks named f_2, f_3, and so on. Generally, vowels are captured by f_1 and f_2, since the first formant indicates the vertical tongue movement (i.e., up or down), and the second indicates the horizontal tongue movement (i.e., back or forth). Furthermore, f_2 and f_3 may provide useful hints for the lips rounding. Formant frequencies are widely accepted features for use in forensic phonetics [18]. Several works are based on the study of f_0 only (e.g., [23]). Unfortunately, to date we cannot assert that voice is like a signature. So, in order to identify the speaker one needs as much information as possible, and it is questionable the fundamental frequency, alone, can be enough.

In order to overcome the uncertainty of the results using the fundamental frequency, [3] investigated the use of the first three formant frequencies and associated bandwidth. They are modeled using a multivariate Gaussian Mixture Model, in order to represent the vocal tract characteristics of the speaker, accounting for within-speaker variability. The results are expressed as a likelihood ratio, and highlight that since formants describe the cavity resonance, they are better suited for application in forensic speaker verification than Mel-Frequency Cepstral Coefficients (MFCC).

In [2], the authors focused on feature selection, investigating several ways to extract Cepstral Coefficients using the two major technologies for mobile communication (GSM and CDMA). Their approach uses the likelihood ratio to quantify the strength of speech evidence. The experiment highlighted the goodness of the MFCC, in spite of the outcomes obtained in [3]. They argue that such results are justified by the removal of the relevant information about the glottal shaping and lip radiation components due to the coding in mobile phone networks (both GSM and CDMA), that should make formant features useless.

In speaker verification task (i.e., the process of verifying the claimed identity of a speaker based on the speech signal), [5,25] create speakers model by measuring the fundamental frequency and formant frequencies of vowels (*a, e, i, o*), and estimating their distributions via Gaussian kernel density estimator. The long-term formant distributions are plotted and examined, accepting or rejecting the speaker. However, the authors pointed out that other information can be extracted from the shape of distributions. Likelihood Ratio [26] is exploited, like in this work, to evaluate the results in [5].

Our approach is text-independent, i.e. it tries to verify the identity without constraint on the speech content. We consider only a real-valued, limited, and continuous signal, i.e., a function that represents the proceeding of a given physical quantity (in our case, sound waves and their spectrum) over time. If a signal has period T (i.e., $x(t+T) = x(t)$), then the function is known when its proceeding in a range of length T is known. The inverse of T is the fundamental frequency $F = \frac{1}{T}$, measured in Hertz if time is expressed in seconds.

Conversely, formants are obtained from the signal spectrum. They are the resonance frequency measured where there is an energy peak in the sound produced by the air passage in the vocal apparatus, keeping into account absorptions due to the sound reflection. The fundamental frequency and the first three formants are the features of our speaker model.

The preprocessing step is carried out by a human operator that considers, for each word, only emphasized vowels, that are less affected by co-articulation and have a more constant signal than others. According to the literature (e.g., [5]), vowel *U* is not considered in this study. The human operator uses *Praat*[1], a software system able to show the graphical trend of the signal energy, allowing one to select the vowel to be analysed and to estimate the power spectrum. He selects the fundamental frequency f_0 using the *CEPSTRUM* method (the result of the Fourier transformation applied to the signal spectrum measured in Decibel), and the formants f_1, f_2 and f_3 (i.e., the first three peaks of the frequency spectrum captured via Fast Fourier Transform). Subsequent formants cannot be detected, due to the poor signal quality.

Figure 1 shows the measurement of the formants of the first vowel *E* in the Italian word *gente* (this particular GUI reports peaks of the spectrum, then the fundamental frequency f_0 is not reported in this Figure).

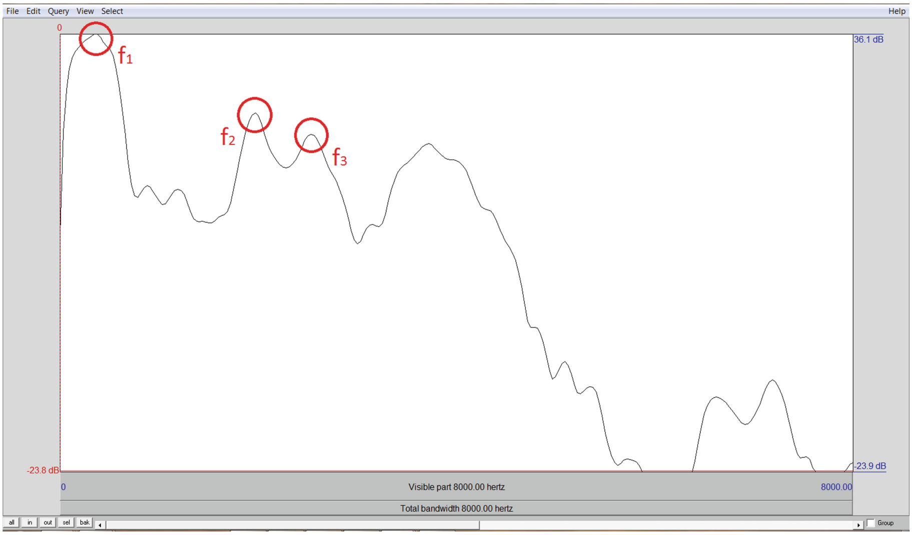
File Edit Query View Select
Help
0
36.1 dB
f1
f2
f3
-23.8 dB
-23.9 dB
0
Visible part 8000.00 hertz
8000.00
Total bandwidth 8000.00 hertz
all in out sel bak
Group

Fig. 1. Formants recognition using Praat.

3 Statistical Method Applied to the Recognition of the Talker - SMART

Suppose we have a distance measure. Then, we can describe the (possibly) large variability of voices among several speakers, as well as the small variability of

[1] www.fon.hum.uva.nl/praat/.

several dialogues of the same speaker. Such a variability can be represented estimating a distribution of the distances, making a model of the population diversities, together with a distribution of the diversities of the speech produced by a speaker in several contexts. Unfortunately, often there are not enough tracks of the same speaker to evaluate such distributions. So, a method to artificially populate a dataset related to the single speaker is needed. Bootstrap [28] resamples the dataset randomly picking whole records and repeating them. It cannot do otherwise, since the formants are related by complex relationships that impose to keep them together [30].

Missing Data. Often the recordings have poor quality, making hard the detection of some formants. We need to manage missing data. In order to face this necessity we adopted some policies. If the fundamental frequency is missing, the average of the known values is assigned to this cell; if a formant is missing, the cell is filled by its average conditioned by the values of the other formants, obtained via multiple regression; if the missing values are too many in a dimension (i.e. a feature), such dimension is removed completely, since estimations over few values are not reliable; it is noteworthy that first and second formants hardly are lacking. However, it might happen that there are no values for a vowel. In such a case the subject is pulled out from the dataset.

Speaker Representation. Given a generic speaker k, a generic vowel will be represented as $V_k \in \mathbb{R}^{N\times 4}$, where N is the number of that vowel instances, whereas 4 stands for the fundamental frequency and the three formants. Considering that we measured only the instances of vowels A, E, I and O, we have $V_{kj} \in \mathbb{R}^{N_{kj}\times 4}$ such that $j \in \{A, E, I, O\}$, and N_{kj} is the total amount of instances of the vowel j, speaker k. A speaker will be represented averaging the values over the columns, obtaining, for each V_{kj} a row vector $\bar{V}_{kj}$. Then:

$$\bar{S}_k = [\bar{V}_{kA}, \bar{V}_{kE}, \bar{V}_{kI}, \bar{V}_{kO}] \in \mathbb{R}^{1\times 16}$$

where 16 is the total amount of vowels formants.

For the sake of completeness, we can give a fast look to an example of real data. Fixed the speaker k and the instance i_0 of the vowel A, we have:

$$V_{kA}(i_0) = [f_{00}, f_{01}, f_{02}, f_{03}]$$

an example record of which might be:

$$V_{kA}(i_0) = [129, 635, 1288, 2325]$$

Mahalanobis Distance and Statistical Distribution. Several measures have been investigated in [6,7,17,29]. Summing up the results, these works shown the goodness of the Mahalanobis distance, that considers the position of the observations, it weights each observation with a coefficient extracted from the empirical covariance matrix. Such a matrix can be computed over the observed values, it represents the relationship between the features and shows how much a feature changes if related to the other ones.

A covariance matrix $\Sigma \in R^{16\times 16}$ is computed over population matrix S, obtained chaining down the subjects in the population as shown in the following. In particular, recalling that each vowel of a speaker k is $V_{kj} \in \mathbb{R}^{N_{kj}\times 4}$ with $j \in \{A, E, I, O\}$, we expect different values for each N_{kj}, from which we can compute:

$$M_k = max_{j=\{A,E,I,O\}}(N_{kj})$$

The gap of each matrix that does not have M_k rows is filled. The instances are duplicated in the same order starting from the first, until the M_k number of rows is reached. The result, for each vowel, is a new matrix V'_{kj}. The representation of the speaker's data S_k will be:

$$S_k = [V'_{kA}, V'_{kE}, V'_{kI}, V'_{kO}] \in \mathbb{R}^{M_k\times 16}$$

where 16 is the total amount of vowels formants. Putting in a single matrix the set of available speakers, we obtain S.

At this point, given two subjects represented as $\bar{S}_i$ and $\bar{S}_j$, computed averaging column values of the respective matrices, the Mahalanobis distance $d(\cdot,\cdot)$ is:

$$d(\bar{S}_i, \bar{S}_j) = \sqrt{(\bar{S}_i - \bar{S}_j)\Sigma^{-1}(\bar{S}_i - \bar{S}_j)^T}$$

Now, suppose we have a pair of voices and we want to evaluate the possibility that they are produced by the same speaker or not. From a Bayesian point of view, we can introduce two statistical hypotheses to encode these possibilities. Say H_0 is the hypothesis that the two voices come from the same person (*accusatory hypothesis*):

$$P(H_0|d) = \frac{P(H_0)P(d|H_0)}{nf}$$

and H_1 is the hypothesis that the two voices do not come from the same person (*defensive hypothesis*):

$$P(H_1|d) = \frac{P(H_1)P(d|H_1)}{nf}$$

where *nf* is a normalization factor, which can be overlooked. We can combine them, obtaining:

$$\frac{P(H_0|d)}{P(H_1|d)} = \frac{P(H_0)}{P(H_1)}\frac{P(d|H_0)}{P(d|H_1)}\frac{nf}{nf} = \left(\frac{P(H_0)}{P(H_1)}\right)\cdot lr(d)$$

$$lr(d) = \frac{P(d|H_0)}{P(d|H_1)} = \frac{p_B(d)}{p_W(d)}$$

where $lr(d)$ denotes the likelihood ratio over d, $p_B(d)$ is the distribution of distance between the suspected speaker and the population (a.k.a. inter-distance), whereas $p_W(d)$ is the distribution of distance taken within different instances of the suspected speaker (a.k.a. intra-distance). Note that $p_B(d)$ and $p_W(d)$ are real valued functions of d. The strength of evidence is computed in $d(S_i, S_j)$ where S_i is the evidence speaker and S_j is the suspected speaker. Then, in our case, d is:

$$lr\big(d(S_i, S_j)\big) = \frac{P\big(d(S_i, S_j)|H_0\big)}{P\big(d(S_i, S_j)|H_1\big)} = \frac{p_B\big(d(S_i, S_j)\big)}{p_W\big(d(S_i, S_j)\big)}$$

Anyhow, the computation of $p_W(d)$ is not so direct, since often the sample is poor (just a few minutes of recording for the suspected person's voice). We need to refill the gap in order to have a number of simulated suspected-person's recordings comparable to the size of the population dataset. Then we recur to the bootstrap [28] procedure. It builds simulated registration using a random movement of the suspected person's data, generating as many suspected-person's samples as the subjects of the population. The Mahalanobis distance is computed for each pair of samples.

Estimating Speakers Distributions. In order to estimate $p_{B/W}(d)$ we exploit a semi-parametric kernel estimator method. Direct Plug-in Kernel [36] (as used in [7,17]), needs to estimate its smoothing parameter h, using $lsdpi(\cdot)$, shown in Algorithm 1. The semi-parametric kernel [4] is:

$$\tilde{p}_{B/W}(d) = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{h}H\left(\frac{d-d_j}{h}\right)$$

where $\tilde{p}_{B/W}(d)$ denotes the model density of p_B or p_W, N is the size of the population, d_j is the distance between S_k and S_j, h is the smoothing parameter chosen via l-stage Direct Plug-in Kernel, $H(\cdot)$ is the kernel function (Gaussian in our case).

Such parameters ensure the satisfaction of:

$$H(\cdot) \geq 0$$

and

$$\int H(\cdot)du = 1$$

in this way the first formula will satisfy $\tilde{p}_{B/W}(d) \geq 0$ and $\int \tilde{p}_{B/W}(d)dx = 1$, as required for a function to be a probability density function.

4 Evaluation

We considered a dataset of Italian-male phone-call recordings, represented as described in Sect. 2 and made up as follows. $K = \{k_1, \cdots, k_i, \cdots, k_m\}$ is the set of pairs of same-speaker's recordings (in this experimental setting, recording 50 speakers twice). P is the set of single entries (they have not a paired recordings,

Algorithm 1. *lsdpi*$(\cdot)$ – l-stage Direct Plug-in.

```
Input: Number of stages $l$, kernel function $K(\cdot)$ of order 2, a data sample $X$.
Output: Approximation of $\psi_c$.

  $\widehat{\sigma} \leftarrow \sqrt{Var(X)}$
  $c \leftarrow r + 2l$
  $\psi_c \leftarrow \frac{(-1)^{\frac{c}{2}} c!}{(2\widehat{\sigma})^{c+1}(\frac{c}{2})!\sqrt{\pi}}$
  $c \leftarrow c - 1$
  while $c \geq 1$ do
    $g \leftarrow \left[\frac{-2K^c 0}{\mu_2(K)\psi_{c+2}n}\right]^{\frac{1}{2c+5}}$
    $\psi_c \leftarrow n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} K_g^r(X_i - X_j)$
    $c \leftarrow c - 2$
  end while
  return $\psi_c$
```

so they can be used only as negative examples – in this experimental setting, we evaluated 350 single entries). So, we have just K positive test, while the number of negative tests will be:

$$nt(K, P) = \frac{P(P-1)}{2} + 2KP + \left(4\frac{K(K-1)}{2}\right)$$

In this experimental setting, $nt(50, 350) = 100.975$. Our evaluation has a twofold objective: on the one hand, understanding the performance, on the other, finding the set of formants that best represent a voice signature.

Table 1. Feature-subset performances

Tested features	EER	AUC
f_0, f_1, f_2, f_3	0.07692	0.98620
f_1, f_2, f_3	0.07692	0.98108
f_0, f_2, f_3	0.05406	**0.99295**
f_0, f_1, f_3	0.07658	0.98587
f_0, f_1, f_2	0.07182	0.98295
f_0, f_1	0.11538	0.96489
f_0, f_2	0.08800	0.97306
f_0, f_3	**0.03980**	0.98952
f_1, f_2	0.09615	0.97104
f_1, f_3	0.11022	0.96991
f_2, f_3	0.07692	0.97328

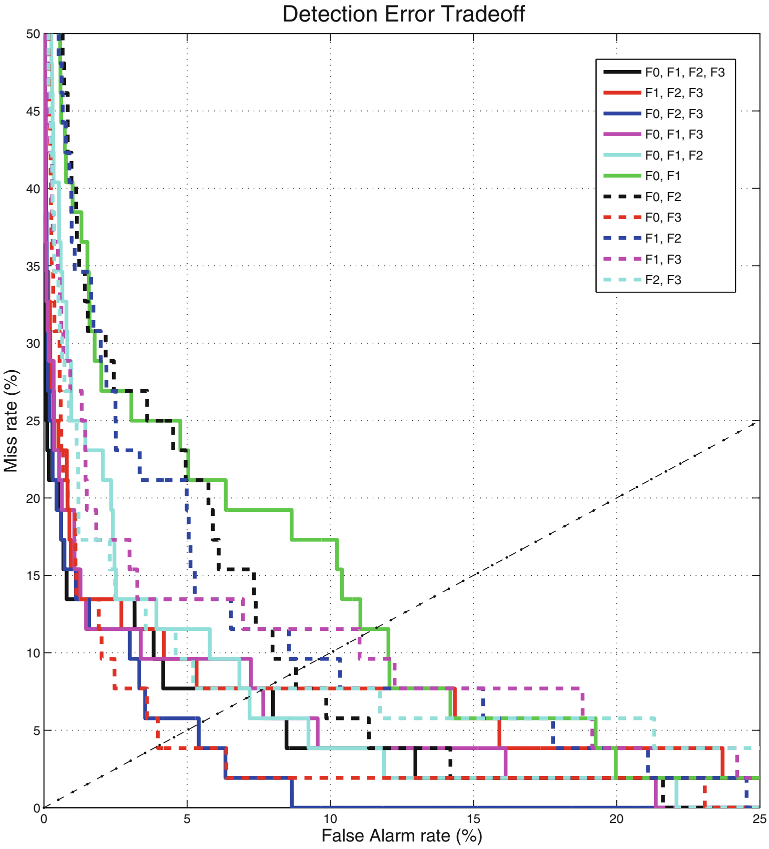

Fig. 2. Diagram of the feature-subset performances

The Likelihood-Ratio $lr(\cdot)$, reported in Sect. 3, expresses how many times more likely we can observe distance d between unknown and evidence voices under the accusatory hypothesis than the defensive hypothesis. It has been used to build the graph in Fig. 2, that shows Detection Error Curve for each subset of features. Table 1 shows the Equal Error Rate (EER) and the Area Under the ROC Curve (AUC) for each subset. The former value is useful to balance the misclassification types, whereas the latter is used to identify the subset that makes less mistakes. The best EER in Table 1 is the curve nearest to zero in Fig. 2, i.e. the subset $\{f_0, f_3\}$, whereas the best AUC in Table 1 is the curve that goes faster to zero in both dimensions in Fig. 2, i.e. $\{f_0, f_2, f_3\}$.

Since both include f_0 and f_3, we should comment the role of f_2. Examining the subset $\{f_0, f_2, f_3\}$ in Fig. 2 we can see that the *false alarm rate* (the worst justice mistake) goes to zero faster than others. Looking at the values in Table 1, the EER of $\{f_0, f_2, f_3\}$ is greater than the EER of $\{f_0, f_3\}$ just a little bit with respect the trend of growth of the EER in general. Furthermore, the subset $\{f_0, f_2, f_3\}$ gives the maximum value of AUC, denoting the smallest error area.

5 SMART: A Particular Case of a Biometric System in the Bayesian Framework

In Sect. 3 we have described step-by-step how the system SMART works. Anyway, it can be seen as a particular case of a more general *Biometric system* used to compute the strength of evidence in terms of Likelihood Ratio in the Bayesian Framework. Let us give a look to Fig. 3. Say E_1 is the crime-scene evidence, and E_0 is the suspected-person's one. For the sake of clarity, example of such pairs could be DNAs, fingerprints, Photos (one from a video-surveillance system that recorded the crime, and the other from suspected-person's); audio tracks, as in our case; and so on.

Whatever is the evidence type, the objective is to establish the strength of evidence that E_0 and E_1 belong to the same person versus the hypothesis that they come from different persons. This objective can be framed in the Bayesian framework as introduced in Sect. 3, in which the Likelihood Ratio is the strength-of-evidence measure.

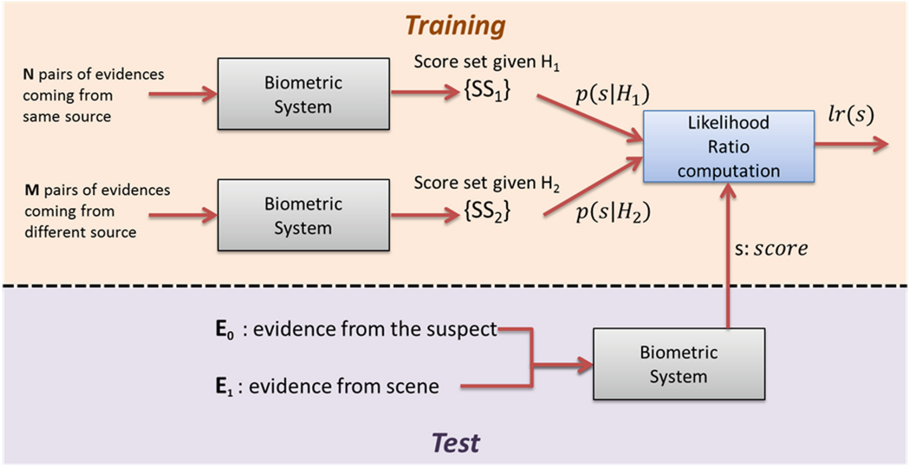
Training
N pairs of evidences coming from same source
Biometric System
Score set given H_1
$\{SS_1\}$
$p(s|H_1)$
Likelihood Ratio computation
$lr(s)$
M pairs of evidences coming from different source
Biometric System
Score set given H_2
$\{SS_2\}$
$p(s|H_2)$
s: *score*
E_0 : evidence from the suspect
E_1 : evidence from scene
Biometric System
Test

Fig. 3. Biometric Bayesian framework

In order to estimate the Likelihood Ratio all we need is a "black box" able to compute the similarity "score", between two evidences and a database containing both pairs of evidences coming from same persons and pairs coming from different ones. Note that in SMART such black box is simply the Mahalanobis distance between a pair of speakers and the score that can be obtained by the inverse of the distance. Going on, the black box *Likelihood Ratio computation* works in two steps:

- **the training phase**, in which two sets of scores are computed from pairs of evidences coming from same (i.e. SS_1) and different (i.e. SS_2) source(s) contained in the database. The sets of scores SS_1 and SS_2 are used to infer the score distributions given the accusatory (i.e. H_1) and the defense (i.e. H_2) hypothesis respectively. In SMART, the set SS_1 is obtained using the bootstrap, whereas the score distributions are computed using a semi-parametric kernel estimator method. After that SMART estimates the Likelihood-Ratio function (i.e. $lr(\cdot)$, or $lr(1/d)$ as defined in Sect. 3). For the computation of SS_2 there are two different approaches: suspect anchored and suspect independent. The former computes SS_2 as the set of scores between the suspected-person's evidence and each other evidence belonging to the database. The latter approach computes SS_2 as the set of scores between all possible (different) pairs of evidences stored in the database. SMART is a suspect-anchored approach;
- **the test phase**, in which the Biometric-system box is used to compute the score between E_0 and E_1. The resulting score value is exploited to obtain the final Likelihood Ratio.

6 Presenting Likelihood Ratio to the Court

Noteworthy, the bootstrap makes our approach non-deterministic, for which the evaluation between speakers (S_1, S_2) is different from (S_2, S_1). There is no theoretical reason to apply the bootstrap to the suspected speaker instead of anonymous one; given that the same classification is expected applying the approach in both directions. Anyway, from a practical point of view, suspected-person's data are often more rich than anonymous, since when the suspects arise, there is enough time to organize the activities in order to record as much dialogues as possible. This is the reason for which the only reliable classification is carried out applying the bootstrap to the suspected-person's data.

We recall that the value of Likelihood Ratio $lr(\cdot)$ quantifies the strength of the evidence. This values must be presented to the Court, then the *European Network of Forensic Science Institutes* (a.k.a. ENFSI) provided detailed guidelines for this purpose[2].

In order to cope with the great amount of different applications of $lr(\cdot)$, its logarithm is commonly used, known as Log-Likelihood Ratio. Given two speakers S_i and S_j, and a distance $d(S_i, S_j)$ on which $lr(d)$ is computed, the Log-Likelihood Ratio function $llr(d)$ is:

$$llr(d) = Log_{10}\big(lr(d)\big)$$

Tables 2 and 3 show the ranges used to evaluate this proposal. For the sake of completeness, corresponding Log-Likelihood Ratio is reported, given that it

[2] With the financial support of the Prevention and Fight against Crime Program of the European Union European Commission - Directorate - General Justice, Freedom and Security. A project funded by the EU ISEC 2010. Agreement number: HOME/2010/ISEC/MO/4000001759.

Table 2. $lr(\cdot)$ values supporting the prosecution hypothesis

$lr(\cdot)$	$llr(\cdot)$	Typical translation
>10000	>4	Very strong evidence to support
1000 to 10000	3 to 4	Strong evidence to support
100 to 1000	2 to 3	Moderately strong evidence to support
10 to 100	1 to 2	Moderately evidence to support
1 to 10	0 to 1	Limited evidence to support

Table 3. $lr(\cdot)$ values supporting the defense hypothesis

$lr(\cdot)$	$llr(\cdot)$	Typical translation
<0.0001	<−4	Very strong evidence to support
0.001 to 0.0001	−3 to −4	Strong evidence to support
0.01 to 0.001	−2 to −3	Moderately strong evidence to support
0.1 to 0.01	−1 to −2	Moderately evidence to support
1 to 0.1	0 to −1	Limited evidence to support

is easy to use and widely adopted. Since the numeric form of a $lr(\cdot)$ may not be readily interpretable to the Court, the last column reports translations into verbal scale, that prosecutor (Table 2) and defender (Table 3) lawyers can use to present the classification result to the Court.

7 Conclusion

This work presented an approach to Speaker Identification that models the speaker via fundamental frequency and formant features. Distances among these descriptions have been computed using the Mahalanobis distance, in order to model the typical distance in speech among several speakers. Such a model has been obtained estimating the distributions of the differences. In particular, both the set of different speakers and the set of tracks recorded from the same speaker have been modeled, in order to obtain comparable models useful to decide if a novel speaker description is nearest to the unknown speaker model or it is nearest to the population model.

The interpretation of SMART as a general Biometric System, working in the Bayesian framework, provides novel insights for future developments and tests. For instance, one could try to assess how the performance change varying score functions, after that the investigation could follow with the comparison of the outcomes using suspect anchored or suspect independent approach. Moreover, we recall that in forensics it is mandatory to have a system with good discrimination ability (generally verified by AUC and/or EER), but it is mandatory also to have a reliable Likelihood Ratio, making fundamental an investigation about the use of Cost Likelihood Ratio.

Other future works will be focused on clustering speaker description via unsupervised techniques, in order to understand whether formant features are enough to obtain clusters representing italian dialects. Moreover in forensics the discrimination of the model measured by AUC and EER does not suffice to measure the reliability of the computed Likelihood Ratio. For example it is mandatory that the system does not give high positive/negative Log-Likelihood Ratio for the wrong hypothesis. To this aim other error functions, such as Cost-Likelihood Ratio error function, will be investigated in future works.

References

1. Federico, A., Ibba, G., Paoloni, A.: A new automated method for reliable speaker identification and verification over telephone channel. In: ICASSP, p. 1457 (1987)
2. Alzqhoul, E.A.S., Nair, B.B.T., Guillemin, B.J.: Comparison between speech parameters for forensic voice comparison using mobile phone speech. In: Speech Science and Conference 2014 (2014)
3. Becker, T., Jessen, M., Grigoras, C.: Forensic speaker verification using formant features and Gaussian mixture models. In: INTERSPEECH, pp. 1505–1508. ISCA (2008)
4. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press Inc., New York (1995)
5. Grigoras, C.: Forensic voice analysis based on long term formant distributions. In: 4th European Academy of Forensic Science Conference (2006)
6. Calvani, F.: Il problema dell'errore di assegnazione nel riconoscimento del parlatore. Tesi di laurea in Matematica, Universit Tor Vergata di Roma (1996)
7. Calvani, F.: Analisi critica di metodi per la classificazione del parlatore nelle scienze forensi. Tesi di laurea in Matematica, Universit Tor Vergata di Roma (1998)
8. Drygajlo, D.: Forensic automatic speaker recognition. IEEE Sig. Process. Mag. **24**, 132–135 (2007)
9. Drygajlo, A., Meuwly, D., Alexander, A.: Statistical methods and Bayesian interpretation of evidence in forensic automatic speaker recognition. In: EUROSPEECH 2003, Geneva, Switzerland, pp. 689–692 (2003)
10. Koenig, B.E.: Selected topics in forensic voice identification. Crime Lab. Dig. **20**(4), 78–81 (1993)
11. Mathan, L., Bimbot, F., Magrin-Chagnolleau, I.: Second-order statistical measures for text-independent speaker identification. Speech Commun. **17**, 177–192 (1995)
12. Falcone, M., Paoloni, A., De Sario, N.: IDEM: a software tool to study vowel formant in speaker identification. In: Proceedings of the ICPHS 1995, Stockholm, vol. 3, pp. 294–297 (1995)
13. Ferilli, S., Leuzzi, F., Rotella, F.: Cooperating techniques for extracting conceptual taxonomies from text. In: Proceedings of the Workshop on Mining Complex Patterns at AI*IA XIIth Conference (2011)
14. Ferilli, S., Leuzzi, F., Rotella, F.: A run length smoothing-based algorithm for non-Manhattan document segmentation. In: Proceedings of Convegno del Gruppo Italiano Ricercatori in Pattern Recognition (2012)
15. Furui, S.: Digital Speech Processing, Synthesis and Recognition. Marcel Dekker Inc., New York (1989)
16. Paoloni, A., Ibba, G.: Analisi delle voci: il parlatore ignoto. Poste e Telecomunicazioni, pp. 14–25 (1993)

17. Ghizzoni, A.: Il problema dell'identificazione del parlatore nelle scienze forensi: modelli, metodi di classificazione e analisi dei dati. Tesi di laurea in Matematica, Universit Tor Vergata di Roma (1999)
18. Grimaldi, M., dApolito, S., Gili Fivela, B., Sigona, F.: Illusione e scienza nella fonetica forense: una sintesi. Mondo digitale (2014)
19. Kersta, L.J.: Voiceprint identification. Nature **196**, 1253–1257 (1962)
20. Leuzzi, F., Ferilli, S., Rotella, F.: Improving robustness and flexibility of concept taxonomy learning from text. In: Appice, A., Ceci, M., Loglisci, C., Manco, G., Masciari, E., Ras, Z.W. (eds.) NFMCP 2012. LNCS, vol. 7765, pp. 170–184. Springer, Heidelberg (2013). doi:10.1007/978-3-642-37382-4_12
21. Leuzzi, F., Ferilli, S., Rotella, F.: ConNeKTion: a tool for handling conceptual graphs automatically extracted from text. In: Catarci, T., Ferro, N., Poggi, A. (eds.) IRCDL 2013. CCIS, vol. 385, pp. 93–104. Springer, Heidelberg (2014). doi:10.1007/978-3-642-54347-0_11
22. Leuzzi, F., Ferilli, S., Rotella, F.: A relational unsupervised approach to author identification. In: Appice, A., Ceci, M., Loglisci, C., Manco, G., Masciari, E., Ras, Z.W. (eds.) NFMCP 2013. LNCS, vol. 8399, pp. 214–228. Springer, Cham (2014). doi:10.1007/978-3-319-08407-7_14
23. Lindh, J.: Preliminary F0 statistics and forensic phonetics. In: Lindh, J., Eriksson, A. (eds.) Annual Conference of IAFPA, Department of Linguistics, Gteborg University (2006)
24. Nolan, F.: Speaker recognition and forensic phonetics. In: The Handbook of Phonetic Sciences (1997)
25. Nolan, F., Grigoras, C.: A case for formant analysis in forensic speaker identification. Int. J. Speech Lang. Law **12**(2), 143 (2005)
26. Rose, P.: Forensic Speaker Identification. Taylor & Francis London, New York (2002)
27. Paoloni, A., Falcone, M., Federico, A.: The parametric approach in forensic speaker recognition. In: Proceedings of COST 250 Workshop on Speaker Recognition by Man and Machine: Directions for Forensic Applications, pp. 45–51 (1998)
28. Rosati, F.: Sperimentazione del metodo bootstrap nel problema del riconoscimento del parlatore. Tesi di laurea in Matematica, Universit Tor Vergata di Roma (2001)
29. Rossi, C.: Il problema di decisione dell'identificazione del parlatore. Caratterizzazione del parlatore, pp. 173–176 (1996)
30. Rossi, C.: Classification and decision making in forensic sciences: the speaker identification problem. In: Rizzi, A., Vichi, M., Bock, H. (eds.) Advances in Data Sciences and Calssification, pp. 647–654. Springer, Heidelberg (1998). doi:10.1007/978-3-642-72253-0_88
31. Rotella, F., Ferilli, S., Leuzzi, F.: An approach to automated learning of conceptual graphs from text. In: Ali, M., Bosse, T., Hindriks, K.V., Hoogendoorn, M., Jonker, C.M., Treur, J. (eds.) IEA/AIE 2013. LNCS, vol. 7906, pp. 341–350. Springer, Heidelberg (2013). doi:10.1007/978-3-642-38577-3_35
32. Rotella, F., Ferilli, S., Leuzzi, F.: A domain based approach to information retrieval in digital libraries. In: Agosti, M., Esposito, F., Ferilli, S., Ferro, N. (eds.) IRCDL 2012. CCIS, vol. 354, pp. 129–140. Springer, Heidelberg (2013). doi:10.1007/978-3-642-35834-0_14
33. Rotella, F., Leuzzi, F., Ferilli, S.: Learning and exploiting concept networks with connektion. Appl. Intell. **42**(1), 87–111 (2015)
34. Forte, A., Rossi, C., Bove, T., Giua, P.E.: Un metodo statistico per il riconoscimento del parlatore basato sull'analisi delle formanti. Statistica LXII, 177–192

35. Furui, S., Matsui, T.: Adaptation of tied mixture based phoneme models for text-prompted speaker verification. In: ICASSP, pp. 125–128 (1994)
36. Wand, M.P., Jones, M.C.: Kernel Smoothing. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, London, New York (1995)
37. Wolf, J.J.: Efficient acoustic parameters for speaker recognition. J. Acoust. Soc. Am. **51**(6), 2044–2056 (1972)