Forensic Speaker Identity Verification (F-SIV) in Italy First Evaluation Campaign Evalita-2009

Luciano Romito, Vincenzo Galatà

Phonetics Laboratory -Department of Linguistics - University of Calabria
Via P. Bucci cubo 20/A, 87030 Arcavacata di Rende (CS), Italy
luciano.romito@unical.it, vgalata@libero.it
http://www.linguistica.unical.it/labfon/Home.htm

Abstract. We report here the results of a first *timid* attempt to promote an evaluation campaign on a Forensic Speaker Identity Verification task within Evalita 2009. Participants were prompted to test methods and models usually used in forensics on a common corpus collected simulating real forensic characteristics and situations. The Task presented a Training data set including known suspected voices to be compared with voices in two other data sets, namely a Closed-test set of 16 unknown voices and an Open-test set containing different voices to be segmented before the test or comparison. Results achieved by participants are here briefly reported.

Keywords: evaluation, speaker verification and identification methods, forensics.

1 Introduction

The Forensic Speaker Identity Verification (Forensic SIV) is characterized by two main points: the first one is related to the individuals involved in the task consisting of suspected individuals that usually have the aim of not being recognized (and therefore not willing to collaborate); the second one is related to a specific balance of the "decision costs", i.e. between wrong identification scores and failed identification scores (see [1] for a detailed description).

In this first evaluation campaign, participants applying for this track were allowed to use any of the methods/models nowadays available (automatic, semiautomatic or manual ones) and normally used in Italian Courts, but also new methods or new models not tested or verified yet. The aims are essentially to gather state-of-the-art knowledge, to promote research advancement in this area, to stimulate and to test new trends in Forensic SIV by having people/experts working on the same sound material in a situation reproducing a "typical" Forensic case study.

For this track a specific corpus has been made available to all the applicants. The corpus reproduces characteristics and instruments usually found in legal cases.

2 The corpus

The speech corpus contains recordings of Italian male speakers. The recording channels are of three types: high fidelity, environmental and telephonic. The recordings have been captured under five different conditions that determine their quality: 1) silent room condition (this material has been used as *Training* data set); 2) wiretapping in and out of car (made possible with the help of police officers by means of a tapping service); 3) phone-calls in a car; 4) phone-calls in a street; 5) phone-calls in a crowded place (some files of these last four types have been used for the *Closed-set Test* data set). The whole corpus contains the same material recorded in the conditions above listed. For each recording condition, the recorded material contains: a) reading of 10 phonetically balanced sentences; b) reading of 10 repetitions of 3 phonetically balanced sentences. For the environmental recording condition, spontaneous speech material, both inside and outside the car, is also available.

In the same speech corpus another recording session is present and simulates a wiretapping in a noisy place including the four speakers of the speech corpus, together with a large number of other anonymous voices (part of this file has been used for the *Open-set Test* data set).[1]

All sound files have been conformed in terms of quality to the worst recording condition identified in the environmental (car) wire-tapping condition: 8 kHz - 16 bit - mono in *.wav PCM format. Only for the *Training data set* the sound files are distributed in a 44.1kHz - 16 bit - mono *.wav PCM format, too.

All files are labelled following the form [data type]_[xx/xxx]_[a]_[b]_[c(d)].wav, where: [data type] = the letters identifying the *Training* data set with TR, CST for the *Closed-Set Test* and OST for the *Open-Set Test*; [xx/xxx] = two letters [xx] or three digits [xxx] identifying a known or unknown speaker, i.e. S1 or S2 in case of TR data set identifying the two known *suspected* speakers, and 034, 098, n…, three randomly generated digits sequence in the case of CST data set identifying unknown speakers; [a] = type of recording identifying the recording channel and its acoustic quality, i.e. C (silent room recording), A (telephone call in crowded place), S (telephone call in street), I (wiretapping in car), X (recording of a telephone call in a car); [b] = identifies the phonation manner of the speakers, i.e. B (low voice), N (normal voice) and A (loud voice); [c] = type of speech material produced by the speaker, i.e. LR (reading with repetition), LS (one reading), PS (spontaneous speech); (d) = identifies the repeated sentence only if [c] = LR, i.e. 1 (sentence 1), 2 (sentence 2) and 3 (sentence 3).

2.1 Training data

The *Training data set* (TR) reproduces the sample voice of two known *suspected* subjects referred to as S1 and S2 contained in the speech corpus (e.g.: TR_S1_C_N_LS.wav; TR_S2_C_N_LR3.wav). The voice samples are clean high fidelity recordings made in a silent room (C). For both known suspected subjects the

[1] For this recording session also four contextual high-fidelity recordings of the four speakers in the *Open-set Test* data set are available and used by the organizers only as a control key.

recorded material contains: a) 1 file containing 10 read phonetically balanced sentences; b) 3 files containing each 10 read repetitions of 1 phonetically balanced sentence. For both known *suspected* subjects, 4 sample files have therefore been provided to the participants for the training task TR.

As above reported these sound files are distributed in a double format: 44.1kHz - 16 bit - mono *.wav PCM and 8kHz - 16 bit - mono *.wav PCM.

2.2 Test data

Two data sets are provided for the test to be carried out with the TR data set: CST data set for the *Closed-set Test* and OST data set for the *Open-set Test*.

For both data sets no answer key has been distributed to participants before the submission of results.

The CST data set is a collection of wiretapping recordings in different environments and in different channels of anonymous speakers. The voices are isolated in 16 different files of different length. The material is composed by read and spontaneous speech and has been distributed to participants in 8 kHz - 16 bit - mono in *.wav PCM format. The files are labelled following the rules above reported: e.g. CST_008_I_N_PS.wav, CST_035_I_N_LS.wav and so on.

The OST data set consists of a single file containing a recording session simulating a wiretapping in a noisy place including the two known *suspected* speakers (S1 and S2) together with other anonymous speakers. Intensity in the file is changing and superimposed voices are possible.

3 Evaluation Measures

Hereafter we present the reading key made available to participants after the submission of the results according to which we evaluated the submitted results following the guidelines for the task.

Table 1. Example of segmentation for the OST (*Open-set Test*) data set with the reading key.

File name	Reading key	In	Out	Duration
OST_0001	S1	0.00.00,264	0.00.04,000	0.00.03,736
OST_0002	S5	0.00.04,160	0.00.15,376	0.00.11,216
...	...	...	...	...
OST_0108	S3	0.12.00,952	0.12.01,800	0.00.00,848
OST_0109	S6	0.12.02,496	0.12.03,383	0.00.00,888
OST_0110	S2	0.12.07,632	0.12.09,583	0.00.01,952
...	...	...	...	...
OST_0270	S1	0.29.58,880	0.29.59,656	0.00.00,776
OST_0271	S4	0.30.00,056	0.30.00,827	0.00.00,772

Table 2. Files available for the CST (*Closed-set Test*) data set with the reading key.

File name	Duration	Reading key
CST_008_S_N_LR3.wav	48 s	S3
CST_019_A_N_LR2.wav	53 s	
CST_030_I_N_LS.wav	45 s	
CST_068_I_N_PS.wav	3 min 43 s	
CST_011_I_N_LS.wav	51 s	S1
CST_028_I_N_PS.wav	2 min 21 s	
CST_049_S_N_LR3.wav	49 s	
CST_054_A_N_LR2.wav	21 s	
CST_013_S_N_LR3.wav	48 s	S4
CST_022_I_N_PS.wav	41 s	
CST_055_I_N_LS.wav	48 s	
CST_081_A_N_LR2.wav	27 s	
CST_018_A_N_LR2.wav	28 s	S2
CST_027_I_N_PS.wav	1 min 22 s	
CST_044_S_N_LR3.wav	40 s	
CST_072_I_N_LS.wav	1 min 2 s	

For the evaluation of the methods used by participants we looked after the results of the comparison between the known *suspected* speakers in the TR data set with the unknown speakers in the CST and OST data sets.

4 Participation Results

We had a total of 11 participants applying for the Forensic SIV task. Each participant, listed in the following table, received the available data according to the guidelines, but unfortunately only two of them submitted the results carrying out the test between TR and CST data set and only one carried out the test with the OST data set, too.

Table 3. Participants in the Forensic SIV task of Evalita 2009.

Nr.	Participant	Company	State	Results
1	Strasheim	Agnitio	Madrid, Spain	Not even one
2	Beritelli	University of Catania	Catania, Italy	Not even one
3	Lindh	University of Gothenburg	Gothenburg, Sweden	Not even one
4	Carfagni & Nunziati	Dept. of Mechanics and Industrial Technologies, Univ. of Firenze	Florence, Italy	Only CST
5	Ciampini	Reparto Carabinieri Investigazioni	Rome, Italy	Only CST
6	Reynolds	MIT - Lincoln Laboratory	Massachusetts, USA	Not even one
7	Prasanna	Indian Inst. of Technology ,Guwahati	Assam India	Not even one
8	Hsu	Delta Electronics Inc.	Taiwan	Not even one
9	Liu	Dept. of Electronic Engineering, Tsinghua University	Beijing, China	Not even one
10	Hemangi Shinde	AISSM's Inst. of Information Technology	Pune, India	Not even one
11	Tucci	Lab. of Phonetics, Univ. of Calabria	Cosenza, Italy	CST and OST

In the following discussion we refer with *Participant 4* to the results submitted by Carfagni & Nunziati from DMTI (University of Florence), with *Participant 5* to Ciampini from Reparto Carabinieri Investigazioni (Rome), and with *Participant 11* to Tucci from the Laboratory of Phonetics (University of Calabria).

5 Discussion

Hereafter we give a very brief presentation of the results submitted and achieved by the three participants in the Forensic SIV task.

Despite the small number of participants the methods used to complete the task are very different: *Participant 4* uses a fully automatic method, while *Participants 5* and *11* use a semiautomatic formant based method but using a different statistical approach: these differences complicated the results comparison. Only the last two methods presented are normally used in Forensics.

For the evaluation we will here only consider those files involving the subjects S1 and S2 of the TR data set and their respective counterparts in the CST or OST data sets presenting the results of their correct or missed identification. However, according to the reading key above listed any further wrong result will be reported.

5.1 Participant 4

Participant 4 (Carfagni & Nunziati, DMTI - University of Florence) carried out the task using an automatic method processing the available data without any human intervention. The system used is based on the Alize/SpkDet software developed and distributed under LGPL by the University of Avignon and using a background population based on a subset of the corpus CSLU adopted to estimate false acceptance (FA) and false rejection (FR) as well as likelihood ratio (LR).[2]

Participant 4 submitted the results only for the first of the two tests demanded by the task: i.e. TR *vs* CST.

According to the results achieved by *Participant 4* none of the comparisons between the voices of the TR data set (i.e. S1 and S2) and the voices present in the 16 files of the CST data set produced positive results. In the following table we report the results and scores reported for the comparison involving the files to be correctly recognized with S1 and S2.

[2] For further details please refer to Carfagni, M. & Nunziati, M., "The Unifi-EV2009-1 Protocol for Evalita 2009" in this volume.

Table 4. Results and scores achieved by *Participant 4* for the comparisons involving S1.

Unknown voice	Known voice	Yes/No	LR	FA	FR
CST_049_S_N_LR3	TR_S1_C_N_LR1	No	0.18	4.1	37
	TR_S1_C_N_LR2	No	0.0018	4.1	37
	TR_S1_C_N_LR3	No	0.026	4.1	37
	TR_S1_C_N_LS	No	2.1e-09	4.1	37
CST_054_A_N_LR2	TR_S1_C_N_LR1	No	0.13	4.1	37
	TR_S1_C_N_LR2	No	0.018	4.1	37
	TR_S1_C_N_LR3	No	0.0031	4.1	37
	TR_S1_C_N_LS	No	6.8e-08	4.1	37
CST_011_I_N_LS	TR_S1_C_N_LR1	No	0.012	4.1	37
	TR_S1_C_N_LR2	No	1.4e-05	4.1	37
	TR_S1_C_N_LR3	No	1.5e-05	4.1	37
	TR_S1_C_N_LS	No	1.9e-13	4.1	37
CST_028_I_N_PS	TR_S1_C_N_LR1	No	0.0044	4.1	37
	TR_S1_C_N_LR2	No	1.1e-06	4.1	37
	TR_S1_C_N_LR3	No	1.2e-06	4.1	37
	TR_S1_C_N_LS	No	1.9e-16	4.1	37

Table 5. Results and scores achieved by *Participant 4* for the comparisons involving S2.

Unknown voice	Known voice	Yes/No	LR	FA	FR
CST_018_A_N_LR2	TR_S2_C_N_LR1	No	0.5	4.1	37
	TR_S2_C_N_LR2	No	1.5	4.1	37
	TR_S2_C_N_LR3	No	0.33	4.1	37
	TR_S2_C_N_LS	No	0.26	4.1	37
CST_044_S_N_LR3	TR_S2_C_N_LR1	No	0.84	4.1	37
	TR_S2_C_N_LR2	No	0.039	4.1	37
	TR_S2_C_N_LR3	No	1.6	4.1	37
	TR_S2_C_N_LS	No	0.79	4.1	37
CST_027_I_N_PS	TR_S2_C_N_LR1	No	0.27	4.1	37
	TR_S2_C_N_LR2	No	0.0025	4.1	37
	TR_S2_C_N_LR3	No	0.045	4.1	37
	TR_S2_C_N_LS	No	0.055	4.1	37
CST_072_I_N_LS	TR_S2_C_N_LR1	No	0.29	4.1	37
	TR_S2_C_N_LR2	No	0.0026	4.1	37
	TR_S2_C_N_LR3	No	0.036	4.1	37
	TR_S2_C_N_LS	No	0.071	4.1	37

5.2 Participant 5

Participant 5 (Ciampini, Reparto Carabinieri Investigazioni, Rome) carried out only the first part of the Forensic SIV task like *Participant 4*.

Participant 5 adopted a semiautomatic formant based method using the IDEM software distributed by Fondazione Ugo Bordoni. IDEM is a modular system containing a tool for speech analysis and a tool for statistic evaluation [2, 3]. Before

the features extraction all files have been processed by means of a file resampling to 11kHz. The features for F0, F1, F2 and F3 are extracted semi-automatically by an expert for stressed and unstressed vowels /a, e, i, o/ indifferently, available in the sound files by means of extraction algorithms (Cepstrum, LPC and FFT). The statistical tool uses a Bayesian approach to calculate the probability of false identification (P.F.I) thanks to a reference community containing F0 and formant measures of approximately 375 male Italian speakers. Chi-square ($\chi 2$) threshold is set to 32 with lower scores giving *yes* answer and higher scores giving *no* answers.

The four files for S1, as well as those for S2, have been considered by *Participant 5* as single files.

The method applied by *Participant 5* did not produce wrong identifications and all the matching voices have been correctly recognized. Only for CST_054_A_N_LR2[2][3] a missed identification has been registered. The participant also refers that some comparisons could not be executed because of the high noise level in some files or because of the presence of double copies of a same file (problem not found by the organizers or the other participants).

Table 6. Results and scores achieved by *Participant 5* for the comparison involving S1.

Unknown voice	Known voice (S1)		
	$\chi 2$	Yes/No	P.F.I.
CST_011_I_N_LS[1]	29.6	Yes	200
CST_028_I_N_PS[1]	27,8	Yes	400
CST_049_S_N_LR3[1]	25,1	Yes	300
CST_054_A_N_LR2[2]	162,8	No	-

Table 7. Results and scores achieved by *Participant 5* for the comparison involving S2.

Unknown voice	Known voice (S2)		
	$\chi 2$	Si/NO	P.F.I.
CST_018_A_N_LR2[1]	7,8	Yes	15,2
CST_027_I_N_PS[1]	26,1	Yes	17,9
CST_044_S_N_LR3[1]	13,8	Yes	23
CST_072_I_N_LS[1]	10,9	Yes	17,7

In the comparison of S2 with the CST files, *Participant 5* referred to the organizers of the Forensic SIV task that due to high variability the features of S2's fundamental frequency (F0) have been excluded from the comparison: that means the comparison has been carried out using only formant frequencies F1, F2, F3.

[3] The numbers presented in square brackets (i.e. [1]) at the end of the files have not been explained by *Participant 5*.

5.3 Participant 11

Participant 11 (Tucci, Laboratory of Phonetics, University of Calabria) used the same formant based method as well as the features extraction algorithms used by *Participant 5*. *Participant 11* used decisional approach implemented in the SMART III System with a reference population of 305 male Italian speakers containing fundamental frequency and first three formant values for the vowels /a, e, i, o/.[4]

Differently from *Participant 5*, only 5 samples of stressed vowels for /a, e, i, o/ were considered. *Participant 11* used the 8kHz files for the TR data set considering the four files for S1 and S2 as single files (i.e. TR_S1 and TR_S2).

As above reported, only *Participant 11* accomplished to the whole task of the Forensic SIV campaign by processing also the OST data set with a preliminary and mandatory segmentation of the voices present in the file (according to the rules detailed in the guidelines). From the resulting segmentation *Participant 11* perceptually identified six speakers whose productions in the OST file have been collected in six different files (OST_1, OST_2, …, OST_6). The organizers checked out the segmentation carried out by the participant and no wrong attribution was found compared to the segmentation performed by the organizers.

The results include identification score *yes/no* as well as *a-posteriori* false acceptance (FA) and false rejection (FR) error. None of the voices in the CST and OST data set has been recognized with S1 or S2 and no cases of wrong identification have been reported.

Table 8. Results achieved by *Participant 11* for the comparison involving S1 in CST.

Unknown voice	Known voice	Identification (yes/no)	False acceptance (FA) error	False rejection (FR) error
CST_011_I_N_LS	TR_S1	NO	9,53%	0%
CST_028_I_N_PS	TR_S1	NO	10,19%	0%
CST_049_S_N_LR3	TR_S1	NO	11,43%	0%
CST_054_A_N_LR2	TR_S1	NO	24,49%	0%

Table 9. Results achieved by *Participant 11* for the comparison involving S2 in CST.

Unknown voice	Known voice	Identification (yes/no)	False acceptance (FA) error	False rejection (FR) error
CST_018_A_N_LR2	TR_S2	NO	0,32%	0%
CST_027_I_N_PS	TR_S2	NO	0,33%	0%
CST_044_S_N_LR3	TR_S2	NO	99,34%	0%
CST_072_I_N_LS	TR_S2	NO	0,33%	0%

[4] The method used and tested by *Participant 11* is implemented in the SMART (Statistical Methods Applied to the Recognition of the Talker) III Project and is exclusively used by the Italian Scientific Police Service. *Participant 11* has been actively involved as research group in all the steps of the SMART I-II-III project (see [4] for a full list of references).

Table 10. Results achieved by *Participant 11* for the comparison involving S1 and S2 in OST.

Unknown voice	Known voice	Identification (yes/no)	False acceptance (FA) error	False rejection (FR) error
OST_1	TR_S1	NO	4,65%	0%
OST_4	TR_S2	NO	0,33%	0%

6 Conclusions

The Forensic Speaker Identity Verification task within the broader evaluation campaign of Evalita 2009 represents a first official attempt for all the participants involved to evaluate methods/models normally used in Italian Courts, and also new ones not tested or verified yet,.

Although the evaluation campaign has been widely promoted by the organizers of the Forensic SIV task, only 11 participants applied it: 4 of them are Italian and only 3 of them completed at least a part of the demanded task. Excluding the methods by *Participant 5* and *Participant 11*, which are exclusively used by Carabinieri or by Italian Police, only *Participant 4* presented and tested a method not even used in forensics.

A simple consideration can be drawn from this experience: considering the Italian participation to the Forensic SIV task here promoted, why is the *island* we know being strongly populated [5, 6] so poorly inhabited (if not unpopulated at all)?

References

1. Hollien, H.: Forensic Voice Identification. Academic Press: SanDiego, CA (2002).
2. Federico, A., Paoloni, A.: Bayesian decision in the speaker recognition by acoustic parametrization of voice samples over telephone lines. In: Proceedings of Eurospeech 93, pp. 2307-2310, Berlin: Germany (1993).
3. Falcone, M., De Sario, N.: A PC speaker identification system for forensic use:IDEM. In: Proceedings of the ISCA Workshop on Automatic Speaker Recognition, Idendification and Verification, pp. 169-172, Martigny, Switzerland (1994).
4. Romito, L., Bove, T., Delfino, S., Rossi, C., Jona Lasinio, G.: Specifiche linguistiche del database utilizzato per lo speaker recognition in S.M.A.R.T. In: Proceedings of the 4th AISV National Conference, "La Fonetica Sperimentale. Metodo e applicazioni", Vol. 4, pp. 632-640, EDK Editore SRL: RN (2009).
5. Romito, L., Galatà, V.: Speaker Recognition: Stato dell'arte in Italia. In: Proceedings of the 3rd AISV National Conference, "Scienze Vocali e del Linguaggio Metodologie di Valutazione e Risorse Linguistiche", Vol. 3, pp. 223-242, EDK Editore SRL: RN (2007).
6. Romito, L., Galatà, V.: Speaker Recognition in Italy: evaluation of methods used in forensic cases. In: Pamies, A., Melguizo E. (Eds.), Language Design, Special Issue (1), pp. 229-240, Método Ediciones, Granada, Spain (2008).