Paper presented at the IV Congreso de Fonética Experimental 11-2- 2008 Granada – Spain (in press)

SPEAKER RECOGNITION IN ITALY: EVALUATION OF METHODS USED IN FORENSIC CASES

Luciano Romito* & Vincenzo Galatà**
Phonetics Laboratory - University of Calabria
Italy

IN PRESS IV CFE

1. ABSTRACT

In this paper we present a specific corpus named "*Primula*", recorded and collected at the Phonetics Laboratory at the University of Calabria, reproducing characteristics and instruments usually to be found in legal cases. We then refer to this corpus in a first attempt to evaluate all the Forensic Speaker Recognition (FSR) methods used in Italy using a common data set. Preliminary results demonstrate, however, that much work has yet to be done in order to verify and validate the FSR methods especially when, as happens in Italy, the prosecutions' deductions and conclusions, and subsequently, the verdict, are primarily based on speaker identification.

2. INTRODUCTION

In the 52 pages of a very recent verdict delivered after the trial of a mafia gang in Italy, there are 52 references to wiretapping and only 15 references to direct depositions. This strongly indicates that the prosecutions' deductions and conclusions, and subsequently, the verdict, are primarily based on voice interceptions and on speaker identification.

Today, wiretapping has become one of the most important and widely-used investigation techniques. Nevertheless, there is still no prepared and officially recognized professional protocol for this task, nor does there exist a university or professional formation path that might create such a competence. In addition, there is a notable absence of interest in this matter on the part of the government and associated institutions. The vacuum thus created is often exploited by pseudo-experts, using pseudo-methods, and by anyone who decides take an interest in the subject.

The present work is part of a broad research project initiated at the Phonetics Laboratory at the University of Calabria on Forensic Speaker Recognition (FSR) in legal cases.

The project is subdivided into four phases: the first concerns the development of a registry of experts and of methods in the field; the second is concerned with the establishment of a corpus of voices built *ad hoc* and named "*Primula*"; the third is concerned with the evaluation of the FSR methods used in forensics through our "*Primula*" corpus, while the fourth and last phase is concerned with

* Director of the Phonetics Laboratory at University of Calabria (Italy) and National Coordinator of GFF (Forensic Phonetic Group), an AISV-ISCA speech interest group (luciano.romito@unical.it).
** Ph.D Student in Artificial Intelligence and Psychology of Programming (vgalata@libero.it).

the publication of the outcomes for courts and judicial and legal associations, to be considered the primary "users" of these techniques and methods. The full project covers 9 main points and is organized as follows:

1. development of a registry of Italian corpora for FSR accompanied by a detailed analysis of each single corpus.
2. development of a registry of all methods, and not only those known in scientific circles or known through scientific publications, used in courtrooms.
3. development of a registry of the professionalisms involved, of everyone who uses forensic phonetics and FSR techniques in legal cases.
4. development of a registry of research and papers published in Italy in the last five years on FSR, as well as congresses and projects funded.
5. development of a registry of skills required by Courts and by public prosecutor's offices in order to use FSR techniques.
6. development of a registry of funds invested in the training of law enforcement agencies and for training courses set up by single universities.[1]
7. setting up of a scientific association to provide oversight and to disseminate information about FSR methods, their reliability and their potentialities in legal cases through seminars held in courtrooms and judicial contexts.[2]
8. development of a corpus covering characteristics and instruments normally used for tapping activities with the purpose of evaluating all of the different FSR methods today used in Italy.
9. Evaluation of FSR methods used in forensics in Italy.

The paper presented here will deal with the last two points.

3. FSR IN ITALY

The situation regarding FSR in Italy is far from being acceptable. The first thing to be noted in this field is the absolute absence of either a professional register or university courses or masters courses dealing with FSR. On the one hand, there are very few papers and conferences on FSR and almost no research funds dedicated to it, while, on the other hand, there is an increasing demand for FSR with a vast number of expert reports produced in forensic cases. The result of this is a situation which is entirely dysfunctional. Furthermore, if we consider the educational qualifications of Italian experts, a clearer picture emerges regarding FSR. From results achieved in [1], it is reported that only 57% of experts have a master's degree and more significantly only 13% of them have an

[1] The first 6 points are discussed in [1].

[2] In December 2007, during the fourth Congress of the Italian Association of Voice Sciences (AISV), an Interest Group of Forensic Phonetics (GFF) has been constituted.

IN PRESS VCFE

educational qualification in Linguistics. Other educational qualifications are in Economics, Civil or Electronic Engineering etc.

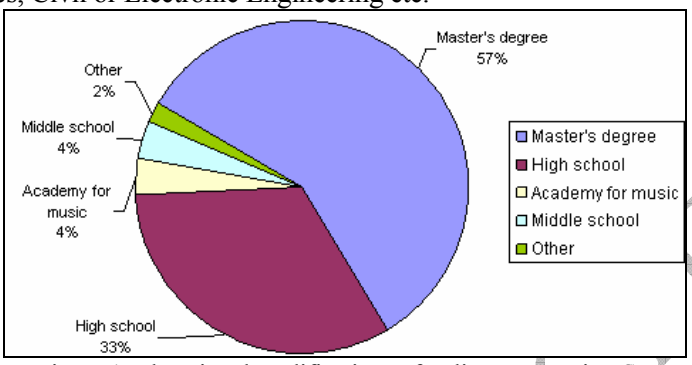


Figure 1 Educational qualifications of Italian Experts in FSR

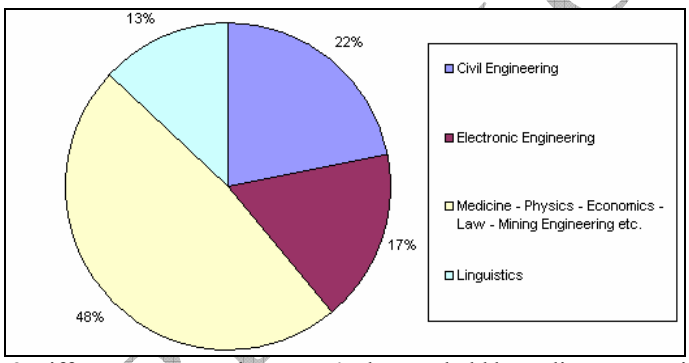


Figure 2 Differences among the master's degrees held by Italian Experts in FSR

Identifying and examining the methods used in FSR we find both subjective and objective methods. For the first of these, auditory and spectrographic methods can be found, while for the second we find parametric methods dealing with *numerical parameters*.

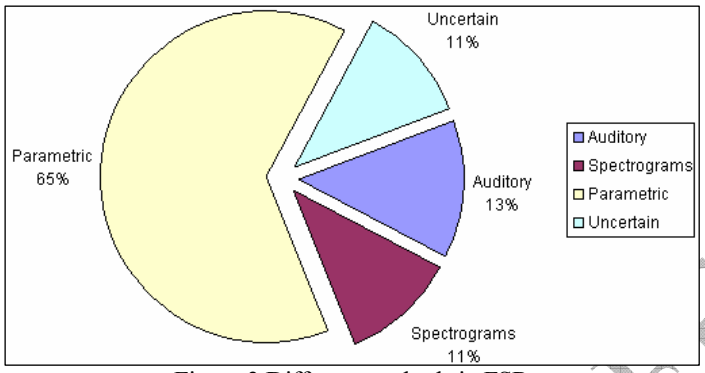


Figure 3 Different methods in FSR

In this context, the expression *numerical parameters* comprises all those methods using acoustic vowel parameters like Fundamental Frequency (F_0), Formant Frequencies (F_1-F_3) or other sets of data such as Articulation Rate (A-rate), Long Term Spectrum (LTS), Historical Formant Frequency or Historical Fundamental Frequency (HFF), comparison of vowels' areas, comparison of mean formants etc.

In this work, we concentrate our attention only on parametric and on automatic methods, that is to say, only on those elements that represent 65% of Italian FSR methods.

As shown in Figure 4, only 26% of experts involved in FSR tasks use a Reference Community for the comparison and in all cases this community is not known.

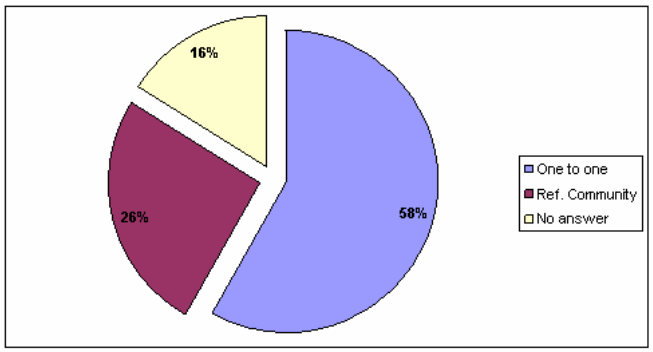


Figure 4 Types of comparisons

INTERSPEECH

4. THE "*PRIMULA*" CORPUS[3]

To evaluate the collected FSR methods used in courtrooms a specific corpus has been recorded in our Laboratory with a view to reproducing characteristics and instruments that are usually to be found in forensic cases. The "*Primula*" speech corpus contains over 900 recordings of 4 Southern Italian male speakers. The recording channels are of three types: high fidelity, environmental and telephone recordings (dark purple boxes in Figure 5 below). The recordings have been captured under five different conditions (green boxes) that determine their quality: silent room, tapping in and out of car (made possible with the help of police officers by means of a tapping service), calls effected in the car, in the street and in the classroom. For the silent room condition we collected loud, low and normal voice recordings (light blue boxes). As the corpus has been conceived, we have the same material for each different condition (light purple boxes). 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 two recording conditions, that is to say for tapping in a car as environmental recording and for telephone call in car uniquely as a telephonic recording, identical and contemporaneous speech material is available for analysis. For the environmental recording condition spontaneous speech material both inside and outside the car is available.

All the recorded material has been conformed in terms of quality to the worst recording condition identified with the tapping in environmental (car) condition: 8 kHz – 16 bit – mono. The whole corpus has been checked manually.

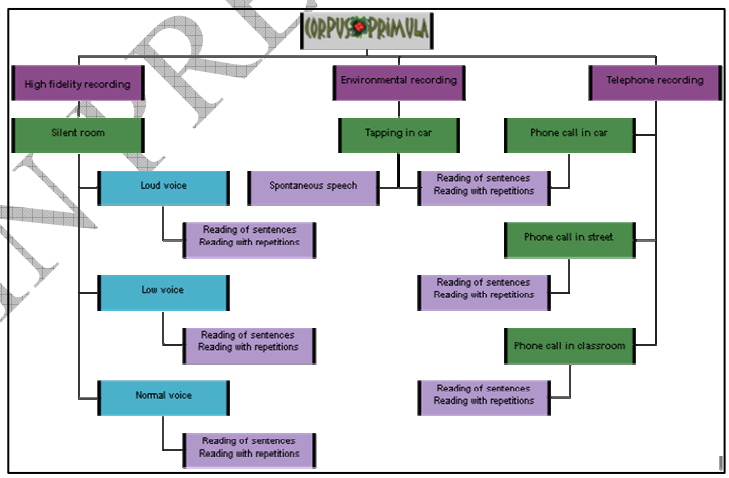


Figure 5 Plan of the "*Primula*" corpus

[3] For further details http://www.linguistica.unical.it/labfon/home_corpus_primula.html

INTERSPEECH

5. EXPERIMENTAL SETTING

In this section we present the experimental setting for the evaluation of the methods used in FSR. The methods investigated in our experiment are:

- *Automatic* method: accomplishes a fully automatic Mel-Cepstrum features extraction;
- *Kernel* method [2, 3, 4]: used by the Forensic Science Police it is based on a statistical analysis of vowels' F_0 and F_1-F_3 frequencies with a Reference Community;
- *Bayes* method [5, 6]: used by the *Carabinieri* it is based on a statistical analysis of F_0 and F_1-F_3 frequencies of four vowels with a National Reference Community;
- Voxys programme [7]: used by some freelancers, it is taken as a method and is based on the use of a software carrying out an analysis only on F_1-F_3 frequencies without F_0;
- Vowels' area overlap method: like the previous one this method is used by some freelancers and like the previous it based on the overlap of F_1-F_2's and F_1-F_3's ellipses in the graphs produced with AutoCAD software;
- Other *parametric* methods: used in other countries (for example by Russian and Canadian communities etc.) and in a certain way imported, but these methods are not better described.

In order to bring forth the evaluation task of the methods in the current paper, we selected some recordings from the whole "*Primula*" corpus described in a previous paragraph to create a sub-corpus. For the intra-speaker variability one speaker was selected in all available recording conditions for a total of six voices indicated in the experiment with the following numbers:

- 1 = wiretapping in car of 10 phonetically balanced sentences;
- 5 = wiretapping in car of spontaneous speech;
- 8 = wiretapping in car of spontaneous speech;
- 11 = telephone call in street of 10 phonetically balanced sentences;
- 13 = wiretapping outside car of spontaneous speech;
- 16 = telephone call in car of 10 phonetically balanced sentences;

The voices 1 and 16 contain the identical material recorded at the same time on two different channels. Voice 5 and voice 8 consist of the same recording split up into two different files.

On the other hand, for the inter-speaker variability, four different speakers having same recording condition (telephone call in car of 10 phonetically balanced sentences) were selected for a total of four voices indicated as: 2, 4, 14 and 16.

IN PRESS V CFE

Finally, another voice (with spontaneous speech wiretapped in car) was added. This voice, indicated as 7, belongs to the same speaker as voice 4.
This way conceived all the experiments have been driven on 10 voices as shown in Figure 6.

	Voice1	Voice2	Voice4	Voice5	Voice7	Voice8	Voice11	Voice13	Voice14	Voice16
Voice1		NO	NO	YES	NO	YES	YES	YES	NO	YES
Voice2			NO	NO	NO	NO	NO	NO	NO	NO
Voice4				NO	YES	NO	NO	NO	NO	NO
Voice5					NO	YES	YES	YES	NO	YES
Voice7						NO	NO	NO	NO	NO
Voice8							YES	YES	NO	YES
Voice11								YES	NO	YES
Voice13									NO	YES
Voice14										NO
Voice16										

Figure 6 Reading key showing the relation among the voices compared in each test[4].

We then performed a "parametric" data extraction on the 10 voices following 4 data extraction procedures:

- Automatic, without operator control (i.e. Mel-Cepstrum features extraction);
- Semiautomatic, with operator choice of vowels through a software driven F_0 and F_1, F_2, F_3 frequencies extraction via FFT, LPC and Cepstrum;
- Semiautomatic, with operator manual vowel segmentation and automatic (running a script in Praat) F_0 and F_1, F_2, F_3 frequency extraction;
- Full Manual, operator driven vowel choice and F_0 and F_1-F_3 frequency extraction with Multispeech 3700 v.2.3 by Kay Elemetrics Corp. via FFT and LPC.

According to these described extraction procedures, we obtained 4 different sets of data that we used for the evaluation of the methods above indicated.
The first set is made up by *Only stressed vowels* coming from Semiautomatic procedure with operator choice of vowels and software driven data extraction and from Multispeech (Full Manual procedure): a) 60 and more samples in voice 1 and 2; b) 20 samples in voice 4, 8, 11, 14 and 16; c) less than 20 samples in voice 5, 7 and 13.
The second set is a normalized version of the first set made up of 20 samples of *only stressed vowels* for each voice.
The third set is made up of *all available vowels* (stressed and unstressed) coming from Praat (Semiautomatic procedure) with 100 to 230 samples per voice.

IN PRESS V.C.

[4] With this expression we here refer to a set of comparisons among the given voices as pointed out in Figure 6.

The fourth set is a normalized version of the third set made up of vowels (stressed and unstressed) labelled as *good* (from an auditory feedback) with 90 to 150 samples per voice.

More precisely, we completed 1 test for the *Automatic* method, 8 for the *Kernel* method and 6 for the *Bayes* method[5] for a total of 15 tests with 45 comparisons per test.

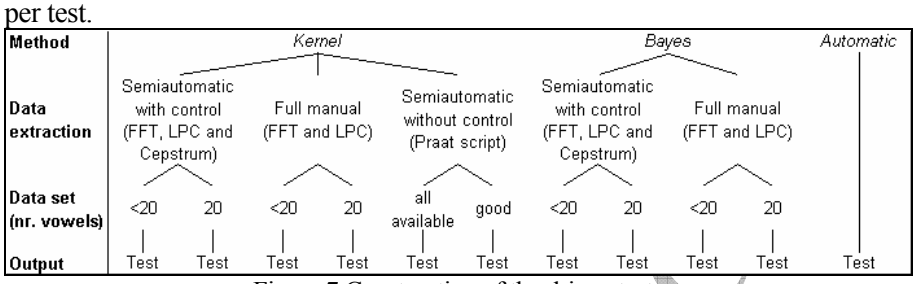


Figure 7 Construction of the driven tests

6. PRELIMINARY RESULTS AND DISCUSSION

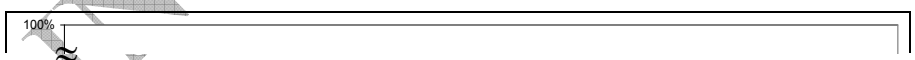
So far we have results only for the first three methods mentioned above with 15 tests completed according to the different data sets.

Before proceeding, it is important to remember some points:

1. we are here dealing only and solely with objective methods;
2. we processed only F_0 and F_1, F_2, F_3 frequencies of different sets of data (stressed vowels, all available vowels etc.) extracted with different algorithms (LPC, Cepstrum, FFT etc.);
3. we applied the different methods, and therefore their specific statistic (Kernel, Bayes etc.), maintaining their *default* settings.

For questions related to space and until the fulfilment of the evaluation task, any discussion about single comparison scores[6] in each test will be voluntarily avoided considering for the moment only global results.

As reported below in Figure 8 **Errore. L'origine riferimento non è stata trovata.**there is so far a 32% *Mistakes* rate among all the tests realized.



IN PRESS IVC

[5] In both the *Kernel* and the *Bayes* method we employed a total number of 4 external collaborators who completed the tests and to whom we gave only F_0 and F_1, F_2, F_3 data to process.

[6] This would imply a detailed discussion regarding different channels and speaking styles involved and probably influencing the extracted and processed parameters.

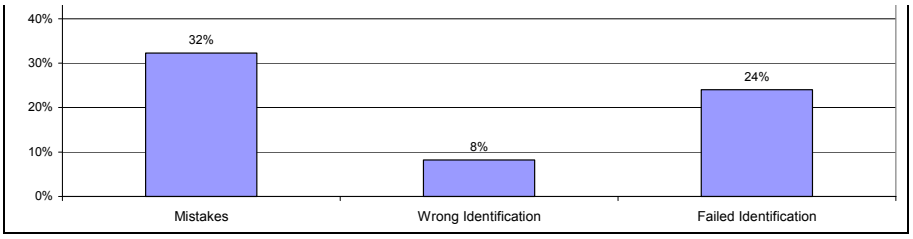


Figure 8 Global mistakes' types

We will focus here mainly on *Wrong Identification* scores (8%) more than on *Failed Identification* ones (24%)[7].

Leaving aside any discussion of different data sets, recording channels etc. it will be interesting to note the significant difference between semiautomatic and automatic methods examining the *Wrong Identification* scores among the statistical methods used as shown in Figure 9 .

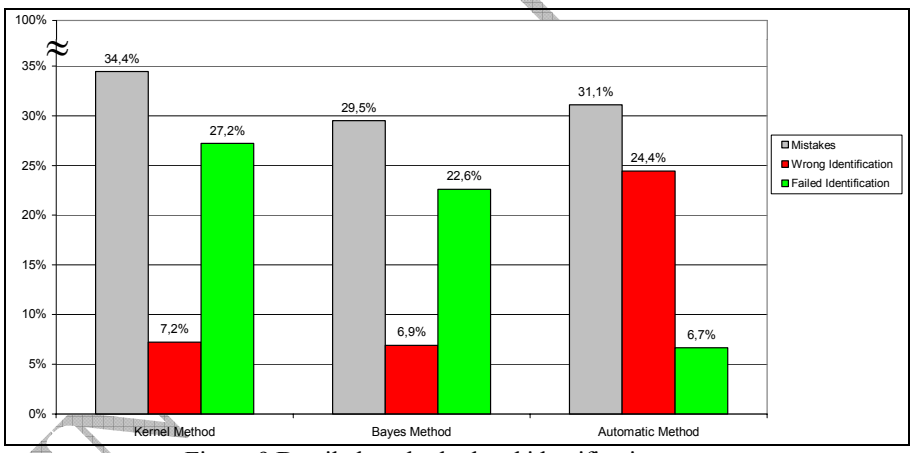


Figure 9 Detailed method related identification scores

Beside mistake scores of 34,4%, 29,5% and 31,1% respectively for *Kernel*, *Bayes* and *Automatic* methods, we find lower *Wrong Identification* scores for semiautomatic methods, i.e. 7,2% for *Kernel* and 6,9% for *Bayes*.

[7] We consider *Wrong Identification* scores to be more important because methods used in FSR should offer the highest possible guarantee for the protection of individual civil liberties.

We believe that better results can be obtained and we are certain that inside each test there are even lower *Wrong Identification* scores.
In the near future, once the evaluation has been completed, we will analyse the results with a "magnifying glass" looking at the single tests and at the single comparisons to find the best combination of algorithm, parameters and so on to lower *Wrong Identification* scores. Much work is also needed with regard to Speakers' Reference Communities that should better be defined and organized in order to guarantee more consistent representativeness.
Concerning *Automatic* methods a just balance between *default* thresholds has to be found to lower as much as possible the *Wrong Identification* scores.

7. CONCLUSIONS

As already remarked, the results here presented are still incomplete. They represent, however, a first attempt to evaluate and test the different methods on a common reference corpus.
Although the experimental setting needs to be improved, and although many other methods have not yet been evaluated from a scientific and impartial point of view, it is very important to verify the extent to which one or other method can give consistent and faithful results in FSR. Further and future work will be therefore carried out by including more information in the comparison, like for example articulation and speech rate, historical fundamental frequency etc. to verify if better scores can be achieved.
Once this step is completed it will be necessary to involve the experts who usually operate FSR providing them with the *.wav files and asking them to collaborate and participate in what we could call "an evaluation campaign" in FSR using their own method to validate its scientific weight, maybe after a preliminary subjective comparison.
Any FSR task could, or should, therefore become a set of different steps instead of a simple and impartial operation, a series of proofs and counterproofs to validate or to deny evidence transforming this task into a sum of objective and subjective procedures. Any subjective procedure that is defined "subjective" can of course not be evaluated because it is the result of the expert's opinions and considerations. His competence is very important because, based on this competence, an evaluation is feasible.
In conclusion, a statistical method, or better, a combined algorithm, could, for example, put together different outcomes and information (whether subjective or objective) with diverse scores to obtain a single conclusion.
Last but not least, the scientific outcomes and results should be translated into a linguistically acceptable scale for FSR's end-users in the wake of [9]'s proposed

IN PRESS IN CFE

verbal scales of likelihood ratio in order to express the strength of a piece of evidence.

8. REFERENCES

[1] Romito L., Galatà V. (2006), Speaker Recognition: Stato dell'arte in Italia. Valutazione dei corpora, dei metodi e delle professionalità coinvolte, in: *Atti del 3° Convegno Nazionale AISV, "Scienze Vocali e del Linguaggio Metodologie di Valutazione e Risorse Linguistiche",* Trento, 29/11-1/12, 2006, EDK Editore SRL: RN, 2007, Vol. 3.

[2] Brutti P., Fabi F., Jona Lasinio G. (2002), Una proposta di meta-analisi basata sulla combinazione di classificatori per il problema del riconoscimento del parlatore, *Statistica*, vol. 3, pp. 455-473.

[3] Bove T., Brutti P., Fabi F., Jona Lasinio G., Giua P.E., Forte A., Rossi C. (2003), *Three approaches to the speaker identification problem for forensic use*, Convegno Cladag 22-24 settembre 2003 (invited relation).

[4] Bove T., Jona Lasinio G., Rossi C. (2004), The speaker recognition problem, *XLII Riunione Scientifica della Società Italiana di Statistica*, pp. 429-440.

[5] Federico A., Paoloni A. (1993), Bayesian decision in the speaker recognition by acoustic parametrization of voice samples over telephone lines, in: *Proceedings of EUROSPEECH 93*, Berlin, Germany, pp. 2307-2310.

[6] Falcone M., De Sario N. (1994), A PC speaker identification system for forensic use: IDEM, in: *Proceedings of the ISCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, Switzerland, pp. 169-172.

[7] Candelo P. G. (2004), Voxys 5.0, *Peg Informatica*, Torino.

[8] Romito L., Lio R. (2008), Stabilità dei parametri nello Speaker Recognition: la variabilità intra e inter parlatore, in: *Atti del 4° Convegno Nazionale AISV, "La Fonetica Sperimentale: Metodo e Applicazioni",* Cosenza, 3-5/12/2007 (in press).

[9] Champod C., Evett I. (2000), Commentary on Broeders (1999), *Forensic Linguistics 7/2*, pp. 238-243.

IN PRESS VCFE