EVALITA'09

Speaker Identity Verification (SIV)

Automatic process of recognizing the identity of an individual from his/her voice.

1

Model

user's voice
word voice
channels
etc.

2

Match

metrics
parameter
statistics
etc.

3

Decide

SIV-A

“Application”

Track -> focus on customer authentication/recognition use case
(e.g. banking, e-commerce, messaging systems, customer care, etc.)

We propose a *remote authentication by telephone* evaluation scenario.

- **major challenge: performance is strongly dependent from the telephone channel**
- **our evaluation data included recordings from both fixed and mobile telephone networks, placing special emphasis on the cross-channel, or mismatched evaluation tests.**

Test Plan

- All data recorded from land-line (PSTN) or mobile (GSM) telephone channels.
- Language is Italian, with speakers uniformly selected in all regions of Italy.
- Datasets:
  - **Enrollment data** for 100 client speakers
  - 4140 **verification utterances**
  - (half were of short duration and half were long).
  - **UBM data** consisting of 60 other speakers, recorded over 20 sessions (total duration 1200 minutes of speech)
  - For 32 of the clients an additional **tuning set** of verification utterances were distributed, for adjusting small parameters sets, like decision thresholds, or score fusion and calibration coefficients.

- **Participants submitted a decision (acceptance or rejection of the claimed identity) and a confidence score, for each verification trial.**

Performance Analysis

We analyse performance for (6x2x2) 24 different trial subsets:
Enrollment={P, G, 3P, 3G, PG, 3P3G} × ***Test duration***={TS1, TS2} x ***Test channel***={P, G}

Enrollment Conditions.		
TC1	P	1 PSTN call
TC2	G	1 GSM call
TC3	3P	3 PSTN calls
TC4	3G	3 GSM calls
TC5	PG	1 call each of PSTN and GSM
TC6	3P3G	3 calls each of PSTN and GSM

Participants

- no Italian labs reply positively to SIV-A call,
- nevertheless we had a very encouraging response from the international side having about 20 labs that expressed interest in participation,
- due to the well known “time” (late call) and “conference overlap” problem (Interspeech2009), only 7 laboratories completed the exercise in time.

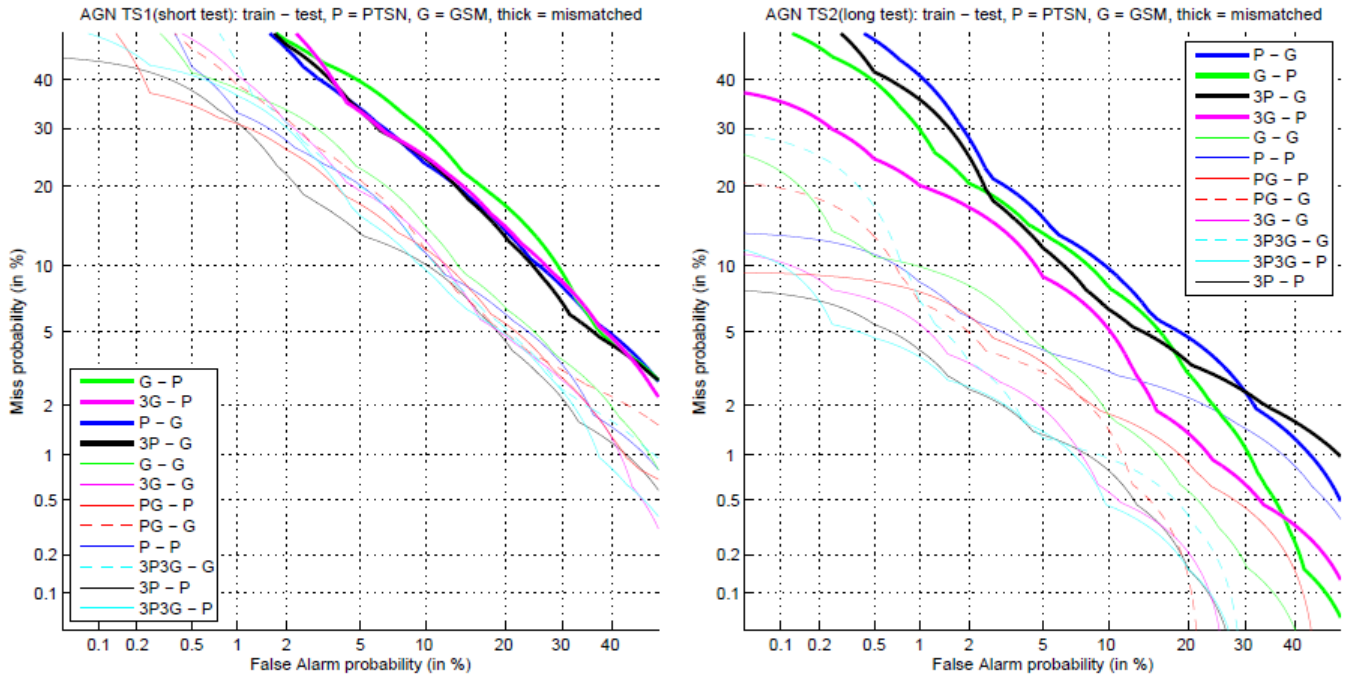
Participant	Abbr.	team size
AGNITIO	AGN	2
Queensland Univ. of Technology, Speech and Audio Research Lab.	QUT	4
Radboud Univ., Nijmegen	RUN	2
Tsinghua Univ., Department of Electronic Engineering	TUE	5
Univ. of West Bohemia	UWB	2
Univ. of Zaragoza, Aragon Inst. for Engineering Research I3A 5	I3A	5
Validsoft, Univ. of Avignon, Univ. of Swansea	VAS	5

Evaluation (DET)

There are two main errors in SIV performance evaluation

- **False Rejection** (FR), or Missed Detection, when a genuine client is rejected
- **False Acceptance** (FA), or False Alarm, when the system accepts an impostor

The DET (*Detection Error Tradeoff*) curve is the standard performance evaluation representation

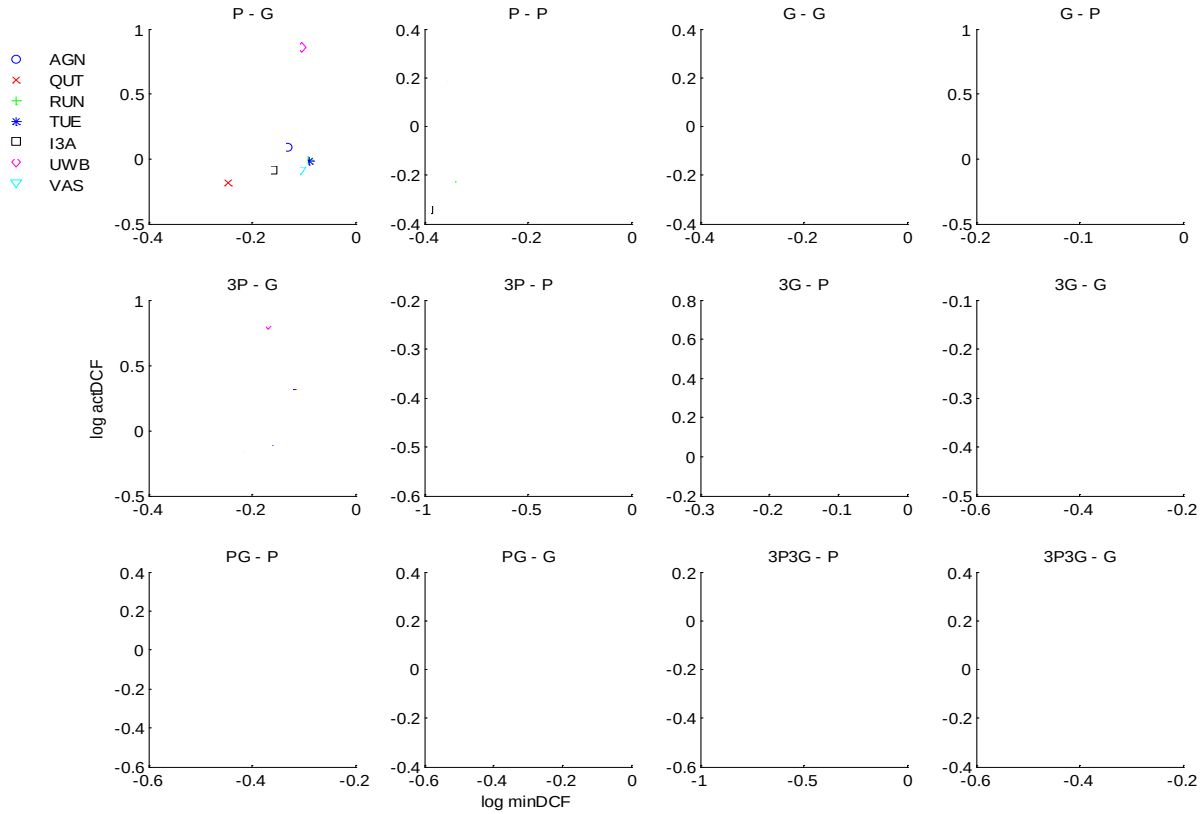


Evaluation (DCF)

Also the DCF (*Detection Cost Function*) or C_{det} has been used in the evaluation, where

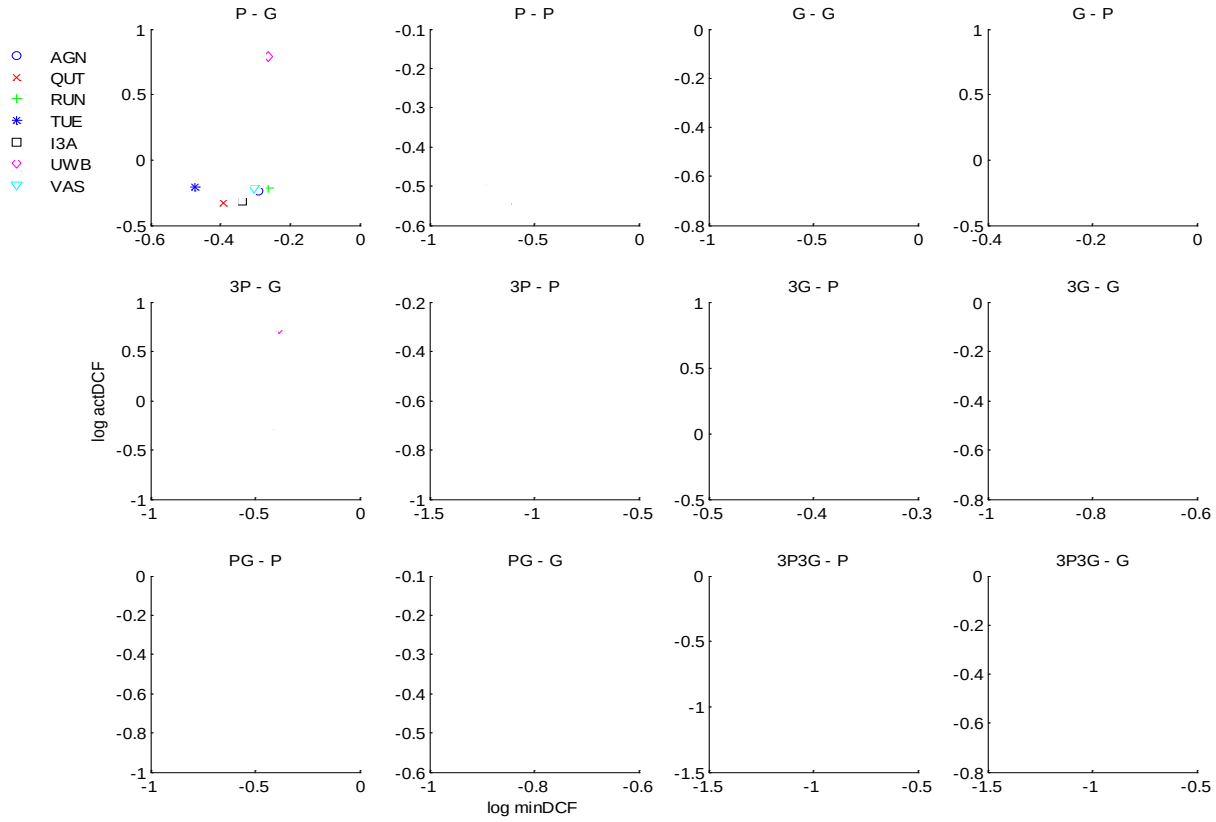
$$C_{Det} = C_{FR} \cdot P_{FR/Client} \cdot P_{Client} + C_{FA} \cdot P_{FA/NonClient} \cdot (1 - P_{Client})$$

TC1 cross-site comparison: DCF scatter plots



Evaluation (DCF)

TC2 cross-site comparison: DCF scatter plots



SIV-A: Conclusion

Some of the predictable trends are:

- Shorter test durations give higher error-rates than longer ones.
- Mismatched telephone types give higher error-rates than matched ones.
- Using multiple enrollment calls improves accuracy.

Some of the perhaps more interesting and unexpected effects are:

- The error-rates for the cross-channel conditions were surprisingly high.
- For some systems there seems to be a considerable advantage in the 3P-G compared to the 3G-P condition. *It seems that only in this case was the cross-channel problem `solved' by a few of the systems.*
- For TS2 (long test), calibration `worked' in the sense that for some systems the actual DCF values were better than the trivial always accept strategy, for all of the different conditions. This suggests all of these systems would have real benefit to its users.
  However in the TS1 (short test) case, there are more calibration problems.
- For all of the systems, there is an almost random variation of goodness of calibration (difference between actual and minimum DCF) across the different conditions. Goodness of calibration is strongly condition-dependent, but this dependency is different for different systems.

SIV-F

“Forensic”

Track -> focus forensic real case “simulation”
(e.g. data are real tapped voices, systems are used in real court debate, etc)

We propose evaluation on a reproduction of a "typical" Forensic case study scenario

- **major challenge: performance is strongly dependent from noise and recording mode**
- **The results could be reached with the help of expert and human intervention**
  (i.e. the system should not be necessarily a full automatic system)

SIV-F: corpus

Speech database reflects real forensic conditions:

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)

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
c) for the environmental recording condition, spontaneous speech material, both inside and outside the car, is also available
d) 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)

SIV-F: test & participants

Two test conditions have been evaluated:

1. **Closed-set Test**
   (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.this material has been used as Training data set)
2. **Open-set Test**
   (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 made possible with the help of police officers by means of a tapping service)

Three participants completed the evaluation exercise:

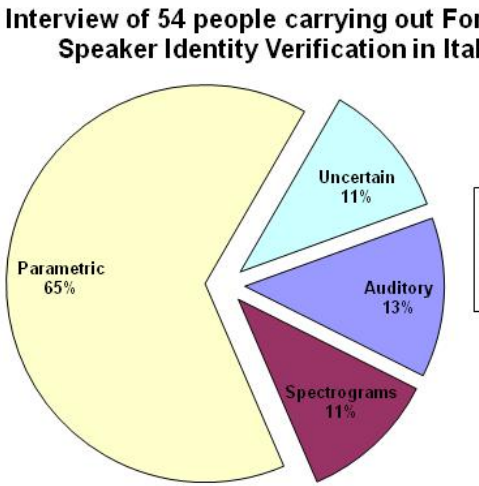
- Reparto Carabinieri Investigazioni (Roma) [IDEM]
- Phonetics Laboratory -Department of Linguistics University of Calabria [SMART]
- Dipartimento di Meccanica e Tecnologie Iindistriali University of Florence [Alize]

SIV-F: results and conclusions

Due to the very different methods and approach used by the three participants it is difficult to compare results
(see organizer's report for further info)

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 so poorly inhabited (if not unpopulated at all)?"



Conclusions

Generally speaking the Italian response to SIV calls was not "enthusiastic".

Nevertheless we find, in spite of all, a good response from the international research community (in SIV-A only).

Resources (i.e. database) could be a problem for future evaluation exercises.