EVALITA 2009 Speaker Identity Verification “Application” Track —Organizer’s Report—

Guido Aversano[1], Niko Brümmer[2], and Mauro Falcone[3]

1. Nuance Communications, Inc.
2. AGNITIO
3. Fondazione Ugo Bordoni

Abstract. This is an analysis and comparison of the accuracy of the seven primary systems submitted to the EVALITA 2009 Speaker Identity Verification “Application” Track. We analyze performance according to test-segment duration, mode of enrollment and telephone types. In particular, we highlight two important problems: the difficulty with mismatched telephone types in enrollment and test; and the difficulty of consistent score calibration across different conditions.

Keywords: Speaker Verification, EVALITA.

1 Introduction

Speaker identity verification (SIV) is the automatic process of recognizing the identity of an individual from the analysis of his voice. The EVALITA 2009 SIV “Application” Track focuses on a specific application area of speaker recognition technology, that is *customer authentication* and automatic *service personalization* for end users (e.g., in domains like banking, e-commerce, messaging systems, customer care, etc.). Inspired by larger world-wide evaluation campaigns, the proposed task is meant to be a first attempt to spread common practices and evaluation protocols for speaker verification through the Italian research community.

Systems submitted for this track have been evaluated on a *remote authentication by telephone* use case scenario. As system performance is affected by the telephone channel, evaluation data included recordings from both fixed and mobile telephone networks and in this report, we place special emphasis on the *cross-channel*, or *mismatched* evaluation tests.

2 Participants and Procedures

The EVALITA SIV-A evaluation campaign, as well as all the other ‘speech tasks’ in EVALITA’09, were announced as an afterthought to the other ‘text tasks’. Indeed, in the original EVALITA’07, there were text tasks only. The introduction of the speech task was thanks to the Italian Speech Interest Group of ISCA, that is AISV (Italian Association of Voice Science). The call for participants had originally been limited to Italian candidates and only later expanded internationally. Although initially about 15 laboratories expressed interest, time constraints and clashes with Interspeech'09 contributed to trimming this number to 7 who finally submitted results on time. These laboratories are listed in Table 1.

Table 1. Participant summary.

Participant	**Abbrev.**	**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
Validsoft, Univ. of Avignon, Univ. of Swansea	VAS	5

No specific problems were encountered in either development or test phases of the SIV-A task. The solution to distribute the data via FTP, and the decision to adhere as much as possible to the well known NIST SRE format and procedures, was found to be a successful strategy. Although this was our first experience in organising an international evaluation, and in spite of the very severe difficulties due to the late starting of the campaign, we believe to have reached the goal. The lessons learned will certainly help us in the organisation of future evaluations, as we have realised there is high interest in such activities internationally.

3 Test Plan

All data distributed for this evaluation was recorded from land-line (PSTN) or mobile (GSM) telephone channels. The spoken language is Italian, with speakers uniformly selected in all regions of Italy.

The evaluation corpus contains *enrollment data* for 100 client speaker models (50 female + 50 male), and a set of 4140 *verification utterances* of which half were of short duration and half were long.

A separate set of *UBM data* consists of 60 other speakers (30 female + 30 male), recorded over 20 sessions (10 PSTN calls + 10 GSM calls). The total duration of the UBM data is 1200 minutes of speech (about 1 minute per call).

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 were not allowed to train large parameter sets (eigenvoices, eigenchannels, UBM, score normalization cohorts, etc.) on this tuning set. Large parameter sets could instead be trained on the UBM data.

Participants submitted results for a prescribed list of verification trials, where the result for each trial contained a decision (acceptance or rejection of the claimed identity) and a confidence score. More details about data, protocols and metrics used can be found in the evaluation guidelines [3].

4 Results

Here we present an analysis of the performance of the primary systems[1] submitted by the 7 participants listed in Table 1.

4.1 Analysis subsets

We analyse performance for each of 24 different subsets of the verification trials. The subsets are defined as all combinations of the 2 test-segment durations, the 6 enrollment conditions and the 2 test-segment telephone types:

$$\{\mathrm{TS1}, \mathrm{TS2}\} \times \{\mathrm{P}, \mathrm{G}, 3\mathrm{P}, 3\mathrm{G}, \mathrm{PG}, 3\mathrm{P}3\mathrm{G}\} \times \{\mathrm{P}, \mathrm{G}\}$$

where TS1 and TS2 denote the short and the long test durations; P denotes enrollment or test over PSTN (land-line); G denotes GSM (mobile); where 3P or 3G denotes enrollment with 3 calls of one telephone type; and where PG or 3P3G denotes enrollment with multiple calls over different telephone types.

4.2 DET curve analysis

In Figures 1 to 7 we present a detailed DET curve [4] analysis of each system on its own, where performance across the 24 different subsets of verification trials are contrasted.

In order to better display the many overlapping DET curves on the same axes, we used the ROC convex hull (ROCCH) algorithm[2] to obtain smoother curves. Although the ROCCH-DET curve shows lower error-rates in places than those found on the corresponding traditional stepped DET curve, the ROCCH curve is not overoptimistic, in the sense that no point on the ROCCH will give a lower DCF value than that which can be obtained on the traditional ROC/DET curve. This holds [2] for any DCF parametrization (i.e. for any values of C_{miss}, C_{fa} and P_{tar}). Tools for plotting ROCCH-DET curves are available at [1].

Note that the enrollment-test combinations denoted as P-G, G-P, 3P-G and 3G-P are the most challenging ones, since these are *cross-channel*, or *mismatched* conditions and they are represented in our DET-plots as thick curves. The results of all 7 systems show that this is indeed the case.

[1] The VAS system was re-submitted after they had fixed a bug, after the submission deadline and after release of the answer key. They had used NIST's DCF parameters, instead of EVALITA's to set their decision thresholds. This is the only change in their re-submission. This bugfix does not change any DET curves, or EER or minDCF. It does improve actDCF.

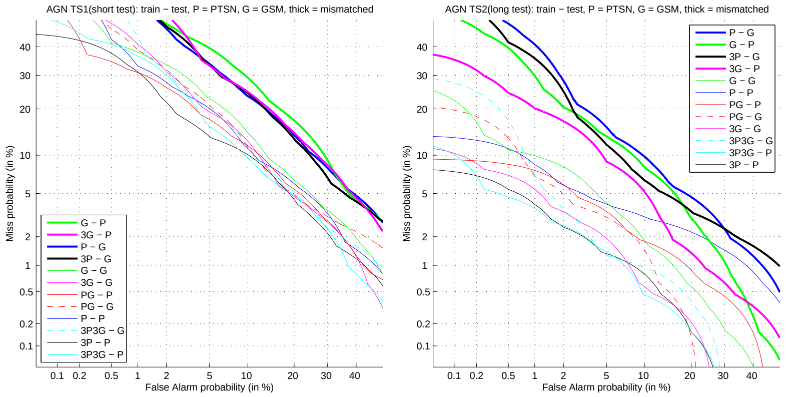

Fig. 1. ROCCH-DET analysis for AGN primary.

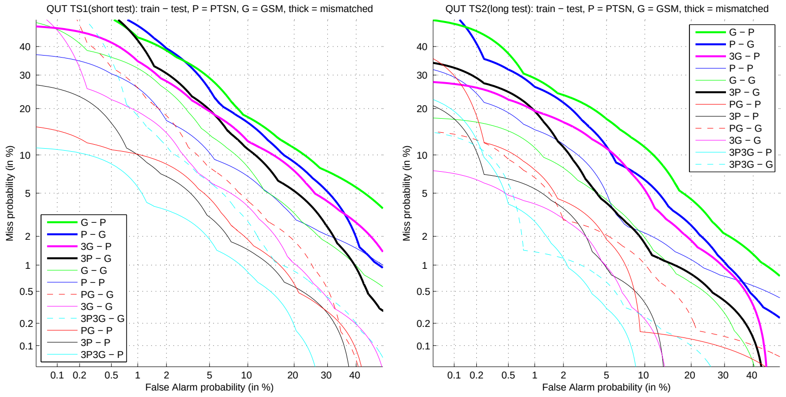

Fig. 2. ROCCH-DET analysis for QUT primary.

4.3 DCF analysis

In Figures 8 and 9 we contrast the relative performance of the 7 submitted systems for each of the 24 subsets. Goodness is measured in terms of 'DCF', (also known as C_{det}), with parameters as defined in the evaluation plan [3]. The 'actDCF' value is computed from the hard *accept/reject* decisions submitted by participants. As a secondary evaluation measure 'minDCF' is computed from the submitted confidence scores, as the point on the DET curve that minimizes the DCF.

The DCF values plotted here are normalized (i.e. scaled), such that $\log \text{DCF} = 0$ for a system that always chooses *accept*. (*Accept* has lower expected cost than *reject*, for the EVALITA DCF parametrization). A system that has ac-

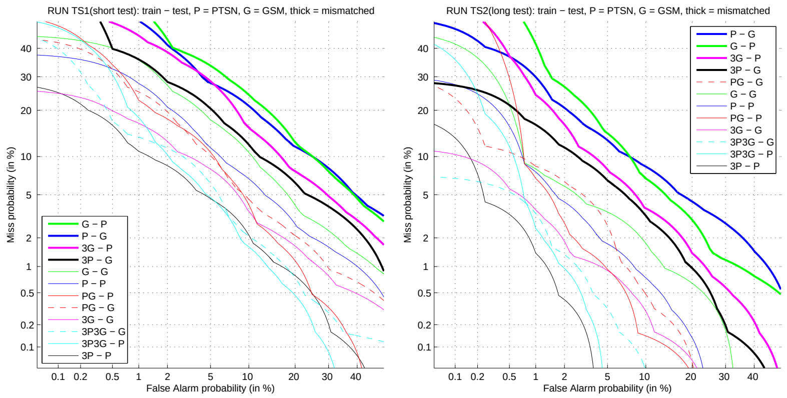

Fig. 3. ROCCH-DET analysis for RUN primary.

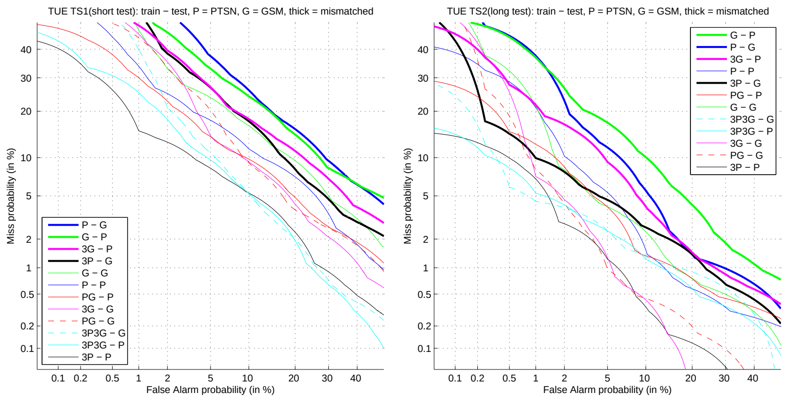

Fig. 4. ROCCH-DET analysis for TUE primary.

tual $\log \text{DCF} > 0$ has bad calibration in the sense that it makes *worse* decisions than the trivial *always accept* strategy.

5 Discussion

The results of this evaluation show some predictable trends and some more unexpected effects. 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 (here most probably the same telephone in enrollment and test).
- Using multiple enrollment calls improves accuracy.

Some of the perhaps more interesting and unexpected effects are:

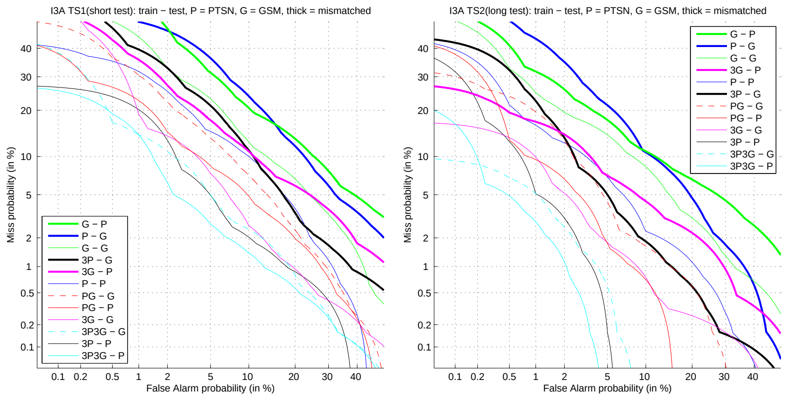

Fig. 5. ROCCH-DET analysis for I3A primary.

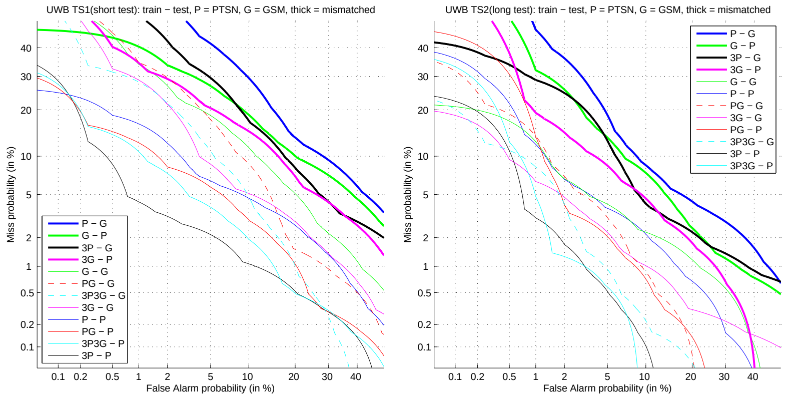

Fig. 6. ROCCH-DET analysis for UWB primary.

- The error-rates for the cross-channel conditions were surprisingly high. The systems submitted in this evaluation were derived from the text-independent technology that has been proven on the Mixer databases, which were used in the recent NIST speaker recognition evaluations between 2004 and 2008, see e.g. [5]. This good performance on the Mixer databases (which contain many different telephones for most of the speakers) created the expectation that performance should have been better on the cross-channel data of this evaluation.
- For some systems there seems to be a considerable advantage in the 3P-G cross-channel condition (fat black curves) compared to the 3G-P condition (fat magenta). It seems that only in this case was the cross-channel problem 'solved' by a few of the systems, albeit at the considerable expense of multiple enrollment calls.

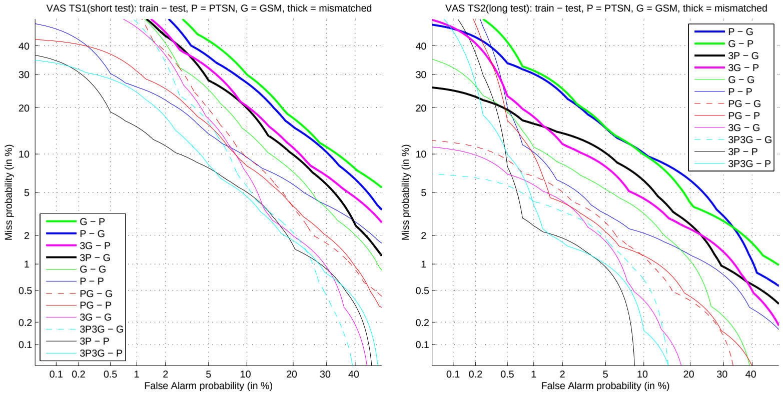

Fig. 7. ROCCH-DET analysis for VAS primary.

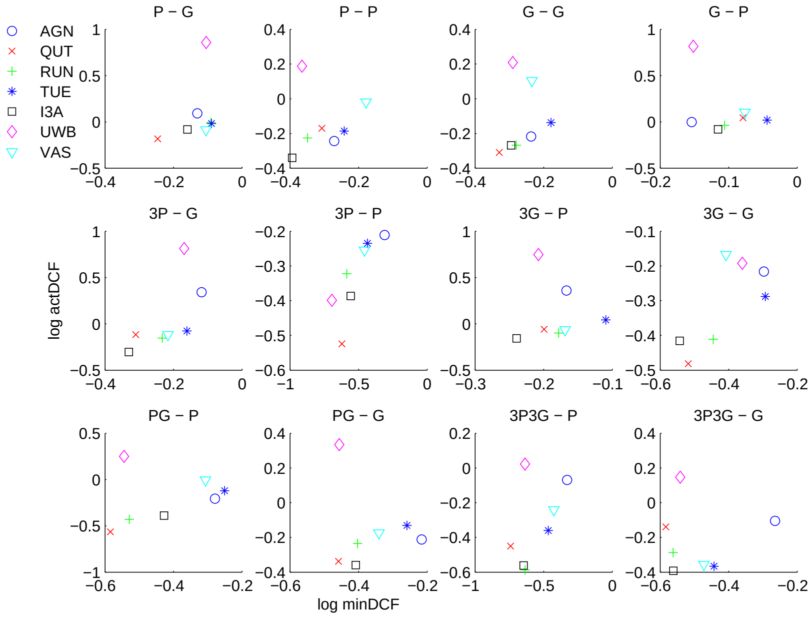

Fig. 8. DCF analysis for TS1. Axes are log base 10 of normalized DCF.

- For TS2 (long test), calibration 'worked' in the sense that for 6 of the 7 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,

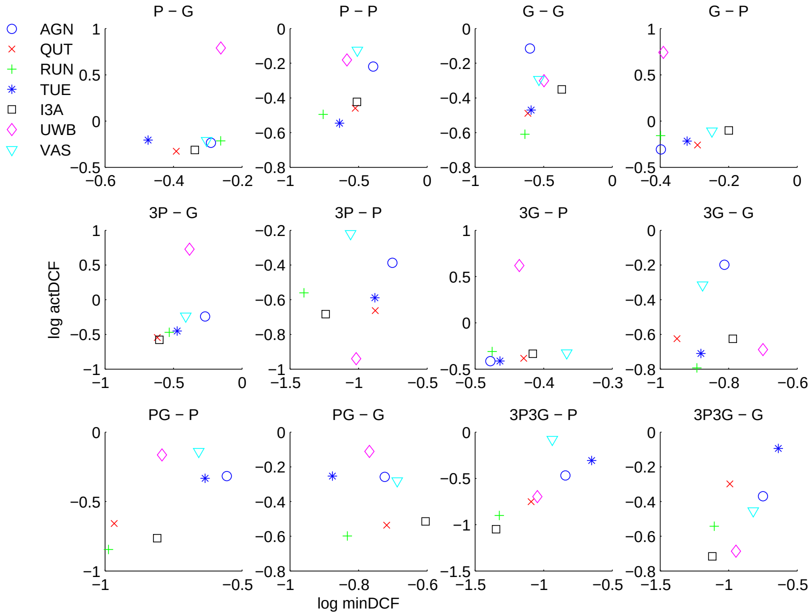

Fig. 9. DCF analysis for TS2. Axes are log base 10 of normalized DCF.

there are more calibration problems, with several actual log DCF values above 0.

– 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.

References

1. ROCCH-DET plotting software, available at `http://focaltoolkit.googlepages.com/rocch`
2. Foster Provost and Tom Fawcett: Robust Classification for Imprecise Environments. In: Machine Learning Journal, vol. 42, no. 3 (2001)
3. Guido Aversano: EVALITA 2009 Speaker Identity Verification—Application Track, `http://evalita.fbk.eu/doc/Guidelines_evalita09_SIV-Application_track.pdf`
4. Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M.: The DET curve in assessment of detection task performance. In: Proc. Eurospeech 1997, Rhodes, Greece (1997)
5. Alvin F. Martin, Craig S. Greenberg: NIST 2008 Speaker Recognition Evaluation: Performance Across Telephone and Room Microphone Channels. In Proc. Interspeech 2009, Brighton (2009)