ATVS-UAM NIST HASR 2010 SYSTEM DESCRIPTION

Joaquin Gonzalez-Rodriguez, Daniel Ramos, Javier Franco-Pedroso, Javier Gonzalez-Dominguez, Ignacio Lopez-Moreno, and Doroteo T. Toledano

ATVS Biometric Recognition Group, Universidad Autonoma de Madrid, Spain

{joaquin.gonzalez, daniel.ramos, javier.franco, javier.gonzalez, ignacio.lopez, doroteo.torre} @uam.es

1. Abstract

With the objective of knowing the performance floor of naive listeners, the extended 150 trials Human Assisted Speaker Recognition (HASR) test has been performed by a panel of 11 non-native listeners assisted by an automatic system. In order to submit calibrated likelihood ratios, a development set of 32 trials per listener was set up with difficult trials (according to automatic scores) from NIST SRE 2008 evaluation. For each HASR trial, one random person of this panel listened for the train and test speech without time constraints. The listener was then asked for scoring the similarity between both speech segments in a discrete scale (from -3 to 3), and a likelihood ratio was obtained from the calibration rule trained with the development set (ATVS secondary submitted system). The trial was also processed by an automatic system (ATVS tertiary submitted system) whose calibrated likelihood ratio was presented to the user. Finally, the listeners can re-score the speech segments, possibly after new hearing, taking into account the automatic recognition system's score. Once calibrated using development data, those LRs are submitted as ATVS primary system. Development results confirm the expected low performance of non-native naive listeners with unknown speakers.

2. Humans involved in system's decisions

A panel of 11 listeners have participated in the HASR test. Each one has carried out 13 or 14 trials. They used a waveform editor, so they could listen for the speech segments and see additional information such as the waveform, the spectrogram, the pitch contour, etc.

The listener's scores range from -3 up to 3 following the next scheme:

- 3 points: the listener STRONGLY supports that both segments come from the SAME person.
- 2 points: the listener MODERATELY supports that both segments come from the SAME person.
- 1 point: the listener WEAKLY supports that both segments come from the SAME person.
- 0 points: listener equally supports that segments come from the same or different persons.
- -1 point: the listener WEAKLY supports that segments come from DIFFERENT persons.

- -2 points: the listener MODERATELY supports that segments come from DIFFERENT persons.
- -3 points: the listener STRONGLY supports that segments come from DIFFERENT persons.

Listeners must score each trial 2 times:

i) Firstly, BEFORE knowing the automatic system's log likelihood ratio
ii) Secondly, AFTER knowing the automatic system's log likelihood ratio

These scores were then calibrated using a different rule for secondary system (the listener's score BEFORE knowing the automatic system's score) that the one used for primary system (the listener's score AFTER knowing the automatic system's score). These rules are described in section 4.

3. Automatic processing involved

In order to help listeners to produce better information, an automatic speaker recognition system processed each trial and the resulting likelihood ratio was presented to the user. This automatic system is a linearized Factor Analysis – Gaussian Mixture Model system with speaker and channel variability compensation.

For each trial, the channel type was obtained from the Sphere file header. Both speech segments of trials comprising microphone-type channel were Wiener filtered (with the ICSI Wiener filter [1]). For interview-type segments, interviewer's channel (channel B) was used to avoid the interviewer's voice in the interviewee's channel (channel A) by means of merging the VAD labels of each channel, as we do for the SRE. For phonecall-type segments, simply energy-based VAD was used in the channel of interest.

After the Wiener filtering (if it was the case) and the VAD, feature extraction is performed as following:

- 20 ms. Hamming window length, overlapped 10 ms.
- 20 mel-spaced (300-3300 Hz) magnitude filters.
- 38 coefficients per frame (19 MFCC + delta).
- CMN, Rasta and 3-second window Feature Warping.

Then, with the gender information provided by the listener, the core speaker recognition system was executed. This system is based on [2].

The system's scores were ZT-normalized and calibrated. A linear logistic regression scheme has been used for calibration, using the FoCal toolkit [3]. Calibration has been performed in a gender-independent way. Four different calibration rules have been used, depending on the channels specified in the incoming Sphere files:

i) scores generated using microphone-only data (MicMic);
ii) scores generated using microphone data in training and telephone data on testing (MicTel);
iii) scores generated using telephone data in training and microphone data on testing (TelMic);
iv) and scores generated using telephone-only data (TelTel).

The log likelihood ratio obtained (ATVS tertiary submitted system) was that one presented to the user before his final scoring.

4. Calibration of human listeners' scores

A linear logistic regression scheme has been used for listeners' scores calibration, using the FoCal toolkit [3]. Calibration has been performed in a gender-independent and condition-independent way. Two different calibration rules were trained:

i) one for listeners' scores BEFORE knowing the automatic system's score
ii) one for listeners' scores AFTER knowing the automatic system's score

Listeners' scores used for training the calibration rules have been generated using a development set built from NIST SRE 2008 data. 11 human participants gave scores for 32 different comparisons each, designed to simulate the HASR conditions (i.e., selected form "difficult" comparisons). The number of scores in such 32 comparisons was the same for each gender, condition (MicMic, MicTel, TelMic, TelTel) and half of them were target and half non-target.

Each listener's scores used for training calibration were specially selected considering scores from ATVS automatic speaker recognition systems, and taking into account that the HASR evaluation would encourage "difficult" comparisons. Thus, for each participant, half of the target comparisons were selected from scores having a moderately high value (with log-LR higher than zero), and the rest having a moderately low value (with log-LR lower than zero). The opposite was done with non-target trials. Therefore, the EER of the automatic system in each of the 11 development sets was close to 50%. This means that the score given by the automatic system tends to give no discriminating information to the listener on average.

5. Timing/Computational Resources

Each listener spent about 4 minutes per trial on average to give a score for the primary system. For the secondary system, this time was usually less than that for the primary system (it has not been registered).

6. Development results

As shown in figure 1, the performance of non-native naive listeners with the ATVS generated development set, extracted from difficult trials (for the automatic system) from SRE'08, is poor and highly variable between different participants. Moreover, as the information provided by the system in this case is useless (EER about 50% as trials were selected with 50% of automatic errors), the performance of humans with development data is not improved when they are helped by the automatic system, as shown in figure 2. With actual HASR data, where automatic performance is expected to be about 5% of EER, the performance of humans is expected to improve (if they allow the system to influence them).

As the number of trials per participant is not high (32 trials), results are not statistically significant when analyzed in a participant dependent way, and EERs vary widely from 12.5% to 56%. Even though 8 participants show a clear correlation of performance with time devoted to analysis (figures 4 and 5), there are 3 outlier participants whose performance do not depend so clearly with the time spent in the trials.

One of the main objectives of this work was to show the amount of information provided by naive listeners. In this sense, we tested participant-dependent and participant-independent calibration. Even though participant-dependent calibration seemed promising, due to the wide variety of performances obtained, the data scarcity (32 trials per participant) forced us to opt for a global approach.

In figure 6, ECE (Empirical Cross Entropy) plots for global and participant dependent calibration are shown, where the global approach shows clearly a more robust approach. Figure 7 finally shows the calibrated LRs obtained through both methods, showing that for the global approach, the one used in the submitted systems, the bigger degrees of support that can be provided by a naive listener, supported by the available development data, are lower than 10.

7. HASR results with native and non-native listeners

The extended 150 HASR test has been performed, fulfilling NIST HASR rules, by the abovementioned panel of 11 non native listeners. However, the development set has also being performed by two native Americans (born and grown in the US, residents in Spain after aged 25). They are also doing the HASR test in an off-line mode as very late submission, and results comparing native and non-native performance are expected to be shown at the HASR workshop in Brno.

8. References

[1] http://www.icsi.berkeley.edu/ftp/global/pub/speech/papers/qio/

[2] Albert Strasheim and Niko Brmmer, “SUNSDV system description: NIST SRE 2008”

[3] http://sites.google.com/site/nikobrummer/focal

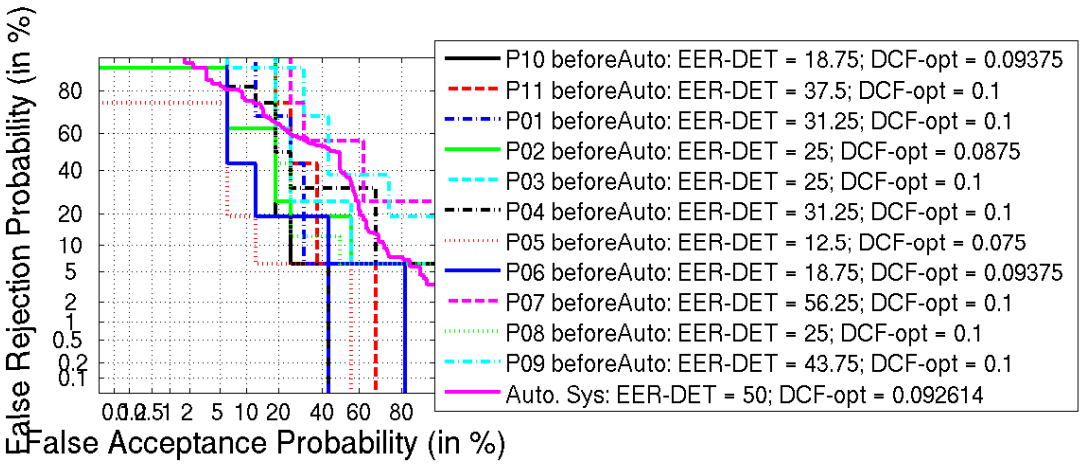

Figure 1.- Development results per participant before knowing automatic system result

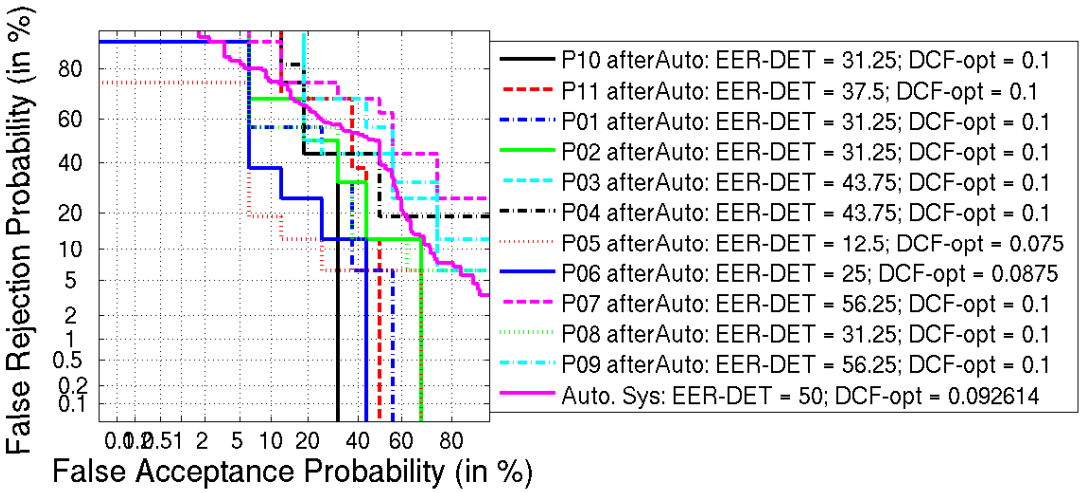

Figure 2.- Development results per participant before knowing automatic system result

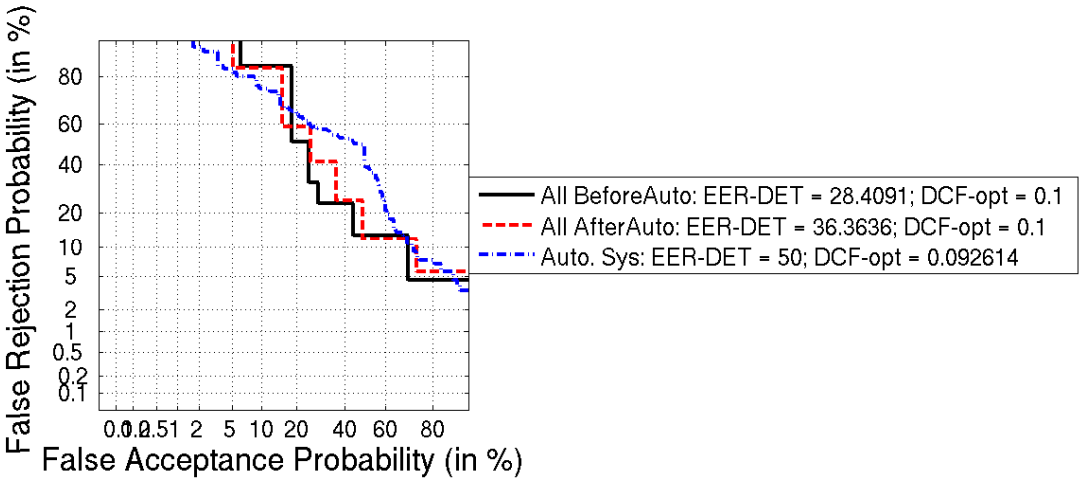

Figure 3.- Average dev results per participants and automatic system in use. As the automatic system was used to select difficult trials for the dev set (50% of erroneous trials), the information provided was randomly misleading.

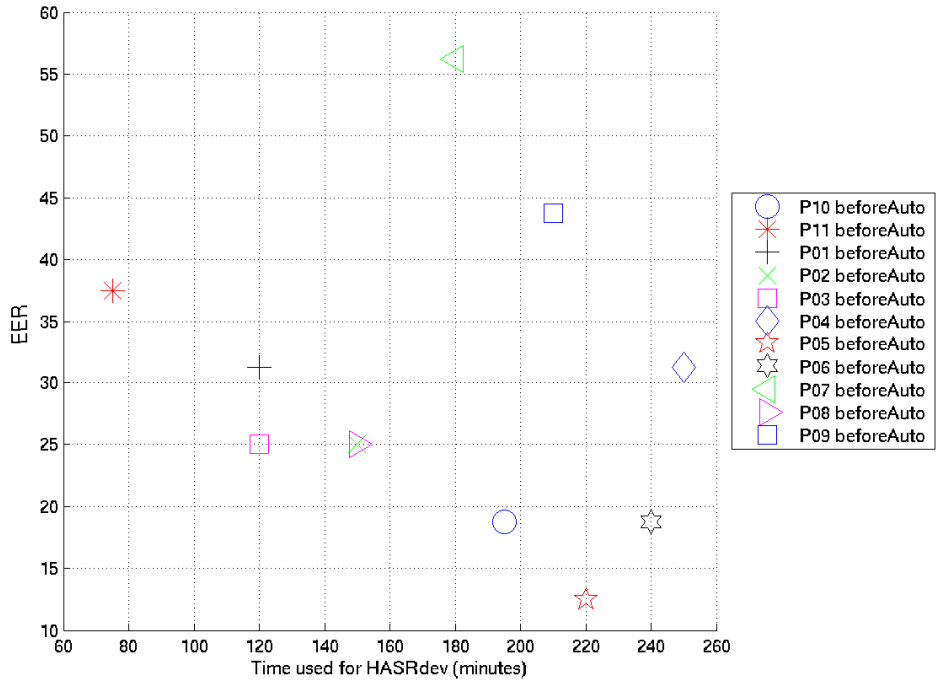

Figure 4.- performance (EER) vs time spent per participant

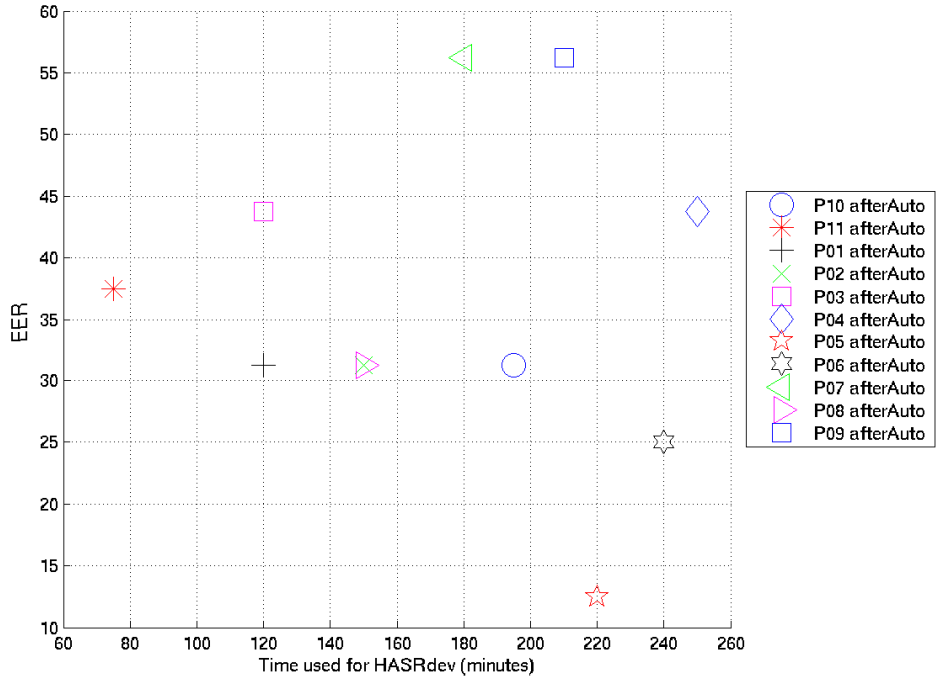

Figure 5.- performance (EER) vs time spent per participant

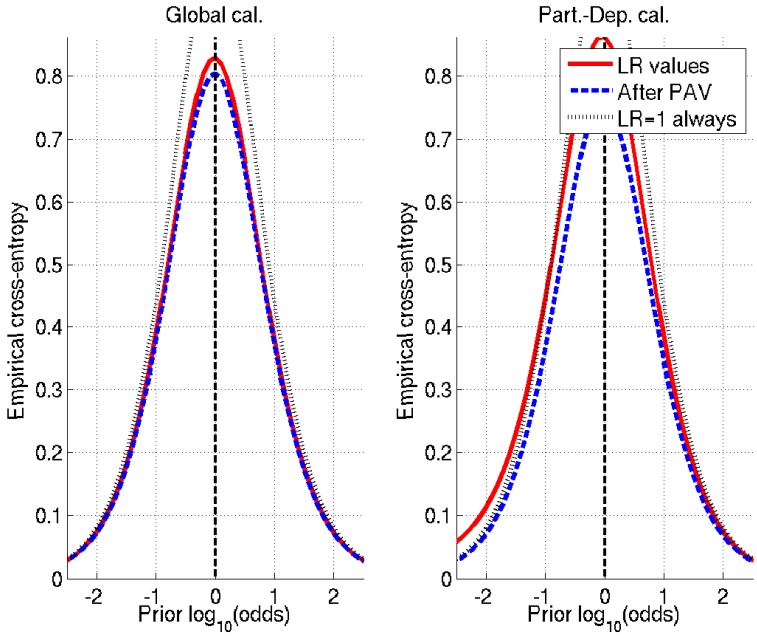

Figure 6.- ECE plots of global and participant-dependent calibration

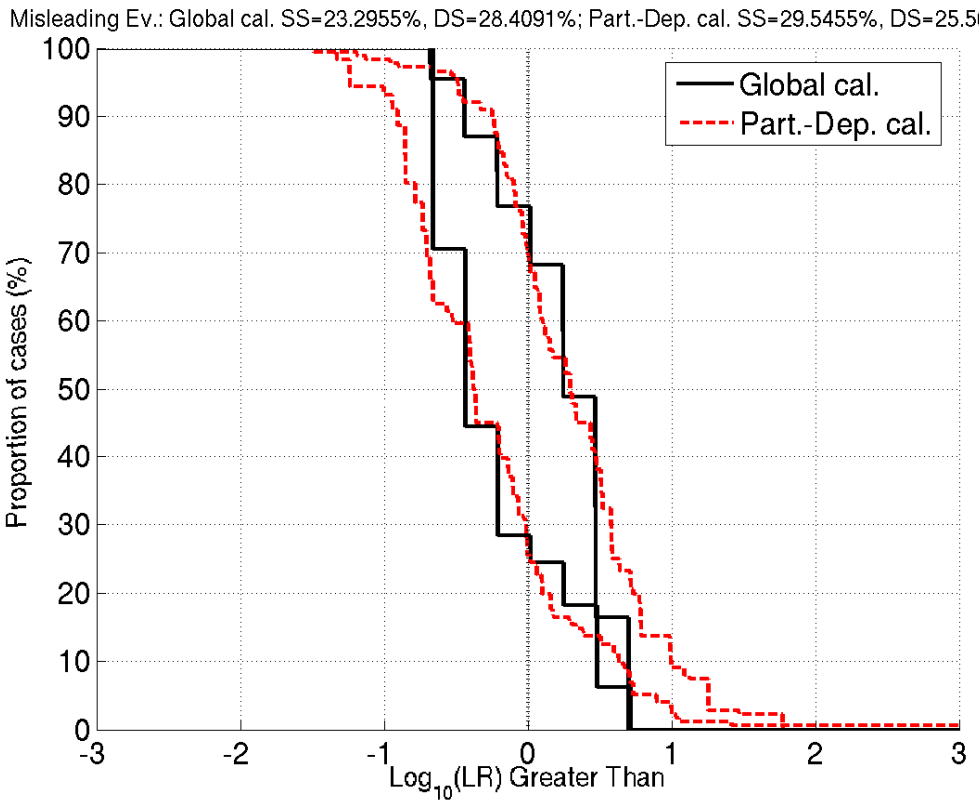

Figure 7.- Tippet plots of calibrated LRs with global vs participant dependent calibration