CALIBRATION AND WEIGHT OF THE EVIDENCE BY HUMAN LISTENERS. THE ATVS-UAM SUBMISSION TO NIST HUMAN-AIDED SPEAKER RECOGNITION 2010

Daniel Ramos, Javier Franco-Pedroso and Joaquin Gonzalez-Rodriguez

ATVS - Biometric Recognition Group. Universidad Autonoma de Madrid (UAM). Spain.

ABSTRACT

This work analyzes the performance of speaker recognition when carried out by human lay listeners. In forensics, judges and jurors usually manifest intuition that people is proficient to distinguish other people from their voices, and therefore opinions are easily elicited about speech evidence just by listening to it, or by means of panels of listeners. There is a danger, however, since little attention has been paid to scientifically measure the performance of human listeners, as well as to the strength with which they should elicit their opinions. In this work we perform such a rigorous analysis in the context of NIST Human-Aided Speaker Recognition 2010 (HASR). We have recruited a panel of listeners who have elicited opinions in the form of scores. Then, we have calibrated such scores using a development set, in order to generate calibrated likelihood ratios. Thus, the discriminating power and the strength with which human lay listeners should express their opinions about the speech evidence can be assessed, giving a measure of the amount of information given by human listeners to the speaker recognition process.

Index Terms— Forensic speaker recognition, likelihood ratio, calibration, human listeners, NIST HASR.

1. INTRODUCTION

One frequent characteristic of legal trials where speech evidence is involved is the establishment of opinions about source attribution based on listening to the recordings to compare, typically by a judge, a jury or a panel of listeners. Moreover, there is a common belief that humans are proficient to distinguish people from their voices (even when they are not familiar), which may bias agents in the legal process if such ability is overestimated. However, to our knowledge, the ability of human lay listeners to extract information about whether some speech materials belong to a given suspect (same-speaker hypothesis) or not (different-speaker hypothesis) has not been assessed in depth. Previous studies [1] suggest that human listeners performance is acceptable, outperforming automatic speaker recognition algorithms. Nevertheless, such studies were performed with controlled speech conditions (landline telephone speech), and session variability compensation algorithms far from the current state of the art. Moreover, although discriminating power of listeners was measured in [1], the strength of the support that human listeners should give to the same- or different-speaker hypothesis has not been rigorously assessed to our knowledge, which is critical to avoid overweighting of their opinions.

Given the aforementioned facts, this work aims at assessing the strength of the support that human listeners should yield. We think that such magnitude should be expressed in the form of a likelihood ratio (LR) [2] in accordance to other standards in forensic science such as DNA analysis [3]. Thus, the farther the LR value from 1, the stronger the evidence in favor of the same-speaker (LR$>$ 1) or the different-speaker (LR¡1) hypothesis, and LR=1 represents no support to either hypothesis. For this study, the NIST Human-Assisted Speaker Recognition (HASR) evaluation 2010 has served as a convenient experimental set-up, since it allows the comparison of speaker recognition techniques where human interaction is present, a typical scenario in forensics. We have scientifically tested the performance of the LR values elicited by a panel of 13 listeners, designing a protocol where listeners elicited scores for trials in a development set built using NIST Speaker Recognition Evaluation (SRE) 2008. Such scores have been used to calibrate the scores of the same listeners in the 150-trial task of the NIST HASR 2010. Moreover, we have assessed the performance of such LR values, and we have compared them to the one achieved by the NIST SRE 2010 automatic speaker recognition system over the same data, showing not only that the automatic system clearly outperforms the human lay listener performance for NIST SRE 2010 data, but also that many of the magnitudes of calibrated LR values from human listeners are close to the LR=1 value, indicating weak information given on average to the decision process involved in a trial.

2. NIST HASR 2010 PROTOCOL

This section briefly describes the NIST HASR 2010 protocol in order to understand the motivation of the design of our

This project has been funded by project TEC2009-14719-C02-01 from Spanish Ministerio de Ciencia e Innovacion; project CCG10-UAM/TIC-5792 from Comunidad Autonoma de Madrid and UAM; and the UAM-Telefonica Chair.

submission. The NIST HASR 2010 150 trial condition consists of a set of 150 comparisons (trials), each one considering two speech segments, both of them between 2 and 5 minutes long. Unlike classical NIST SRE rules, in HASR human interaction with the speech data is allowed. The speech in NIST HASR 2010 is a small subset of the NIST SRE 2010 evaluation data, which can be recorded over a telephone or a microphone channel, and from conversational telephonic speech or an interview. Generally, the mismatch among different sessions is severe. In addition to the intrinsic difficulties of the NIST SRE 2010 data, the HASR subset is known to be selected from especially difficult trials, leading to comparisons in extreme conditions. Therefore, this is a challenging test, but also a realistic one, since many common situations in forensic speaker recognition correspond to this scenario.

3. HUMAN LISTENERS IN NIST HASR 2010

The 150 HASR trials have been conducted by a panel of 13 recruited listeners, two of them native. We will call the non-native speakers *Participant01* (or *P01*) to *Participant11* (or *P11*), and *P12* and *P13* will be the English native speakers. Each of non-native participant (*P01* to *P11*) has carried out 12 to 14 trials from the HASR evaluation, completing the 150 trials among all of them. They were assisted by 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. In addition, native listeners (*P12* and *P13*) have performed the full set of 150 HASR trials each one. No other particular rule considering human perception has been used to assign trials to listeners.

The scores elicited by each participant were limited to a range from -3 up to 3 following the scheme as follows. On the one hand, a score of $3/2/1$ means that the listener *strongly/moderately/weakly* supports the same-speaker hypothesis. On the other hand, a score of $-3/-2/-1$ means that the listener *strongly/moderately/weakly* supports that both segments come from *different* people. Finally, a score of 0 means that the listener equally supports both hypotheses. Listeners must score each trial 2 times. First, before knowing the score of the ATVS-UAM automatic speaker recognition system submitted to NIST SRE 2010 (expressed in the form of a LR value); and second, after knowing such score. For space limitations, in this work we will focus on the opinion of listeners before knowing the score from the automatic system, and we will extend the analysis in future contributions.

3.1. Calibration of Human Scores

In order to generate a LR value from the human listener scores, the process known as calibration [2], a set of scores is needed to train the calibration rule. Thus, we constructed a development set from NIST SRE 2008 short2-short3 condition, containing a protocol of 32 trials that each human lay listener must complete before processing their HASR trials. These trials were designed to simulate the HASR conditions (i.e., selected form "difficult" comparisons), which was assessed by the use of the ATVS-UAM automatic system used in NIST SRE 2008. Thus, the trials were selected considering that the automatic system presented an Equal Error Rate (EER) of 50%, an extremely bad detection performance. Each set of 32 comparisons was balanced in gender and channel. Calibration was performed by means of a linear logistic regression model [4], which can be defined as follows:

$$\log(\mathrm{LR}) = \log \frac{P(\mathrm{score}|\mathrm{ss})}{P(\mathrm{score}|\mathrm{ds})} = \alpha \times \mathrm{score} + \beta \qquad (1)$$

where *ss* and *ds* respectively stand for *same-speaker* and *different-speaker* hypotheses. The weights α and β of the linear transformation are obtained from the training scores of the human listeners from the development set [4][1]. In order to train the calibration, two strategies have been followed. First, a global calibration, where the full set of development scores from all listeners have been used to train the linear model, and therefore the same linear transformation is applied to the scores of all listeners. Second, listener-dependent calibration, where the linear model for calibrating HASR scores from a given listener is trained using the development scores of that single listener. The former has the advantage of having more data to train the calibration, but the drawback that if the listeners behave very different the calibration is expected to be sub-optimal; and vice-versa.

3.2. Automatic Speaker Recognition

In this work the ATVS-UAM automatic speaker recognition system used in NIST SRE 2010 is compared to the human listeners. System pre-processing includes Wiener filtering applied to microphone speech segments. Then, feature extraction was performed to all utterances after energy-based Voice Activity Detection (using reference channel provided by NIST if available) with 18-MFCC plus Δ. Matching of speech feature vectors is performed by linearized-Gaussian-Mixture-Models with total-variability session compensation according to [5]. Scores obtained were ZT-normalized and calibrated using linear logistic regression. Background data including calibration was selected from past NIST evaluation databases, and consisted on telephone data for trials involving just telephonic speech and balanced microphone and telephone data for trials including microphone speech.

[1]The FoCal toolkit has been used for training the weights. http://sites.google.com/site/nikobrummer/focal

4. RESULTS

4.1. Discriminating Power

In this section, the discriminating power is measured in terms of DET plots and Equal Error Rates (EER). Figure 1 shows the discrimination performance of the scores from non-native participants for the development and the HASR 150-trial set, compared to the automatic system for the same 150-trial set. It is seen that the development set is only slightly easier for the listeners than the HASR set, and therefore we conclude that the development set was properly designed. Interestingly enough, even when the ATVS-UAM system obtained a EER$=50\%$ in the development set due to design criteria, human listeners can obtain discriminating information from such trials, and also the automatic system is able to outperform humans in the HASR test. This suggest strong complementarity of both information sources.

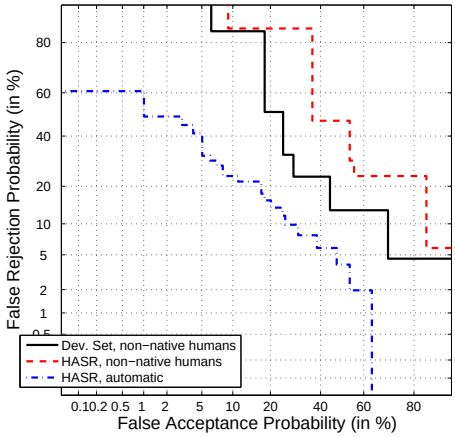

Fig. 1. DET plots showing discriminating power of human listeners compared to the ATVS-UAM automatic speaker recognition system.

Figure 2 shows the EER values of the different participants for development and HASR trials. It seems that, although natives outperform non-natives in development, in HASR their performance is comparable. This can be due to several facts. First, perhaps NIST selection criterion for HASR trials took into account other factors not considered in the development set construction, such as matching contextual information (residence, age, etc.), linguistic similarities, etc. Second, the trials in the development set may include non-native English speakers, which may facilitate the task for English speakers easily identifying non-native accents. Third, native listeners were informed of their high performance on the development trial set before they start the HASR trials, which may result in over-confidence when HASR trials were performed.

4.2. Calibration Strategies

In this section we analyze the different calibration strategies tested, namely global and listener-dependent calibration. Table 1 shows the C_{llr} and C_{llr}^{min} values for development and HASR trials for non-native participants, the lower their value the better. C_{llr} [6] measures the overall performance of a set of LR values, and it is the main figure of merit in table 1, also used in NIST SRE. $C_{llr}-C_{llr}^{min}$ measures the calibration of the set of LR values, which indicates whether the LR values can have a probabilistic interpretation (a key issue in forensics). See [6, 2] for details. For calibrating development scores, we had not a training set in order to test our calibration strategies, and therefore a jackknife procedure was used, where each score was calibrated with scores not coming from the same utterances. Importantly enough, such procedure may lead to overoptimistic results. From Table 1 we see that global calibration is slightly better than listener-dependent (C_{llr}), being calibration ($C_{llr}-C_{llr}^{min}$) also good in both cases. Therefore, we chose global calibration for HASR submission. For HASR trials the C_{llr} value is around 1, which indicates poor performance, much worse than in the development set. This is due to the higher difficulty of the HASR set and to the jackknife procedure, which predicts a slightly better performance in development trials.

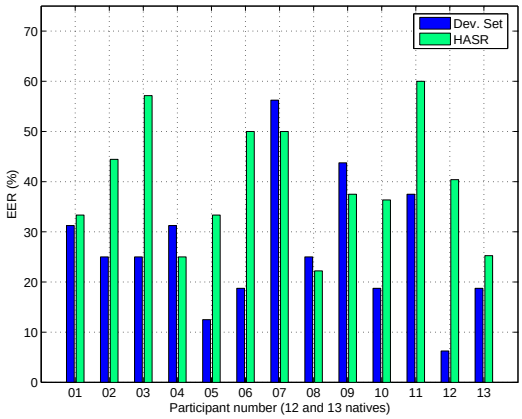

Fig. 2. EER of development and HASR trial sets for different participants. In HASR test, each non-native participant performed between 12 and 14 trials, while native participants (*P12* and *P13*) performed the full set of 150 trials. In development experiment each participant performed 32 trials.

Table 1. C_{llr} and C_{llr}^{min} for different calibration strategies

	Dev. Set (jackknife)		HASR
	Listener-dep.	Global	Global
C_{llr}	0.86	0.83	1.04
C_{llr}^{min}	0.76	0.80	0.96

4.3. Assessing the Strength of the Evidence

The strength of the evidence is related to the magnitude of the LR values, being greater for LR values farther than 1.

Thus, we represent in Figure 3 the proportion of cases in the experimental set where the LR is greater than a given value ($\log(\mathrm{LR})$ *greater than...*) for same- and different-speaker trials (Tippett plots). This representation allows to see the proportion of LR values much bigger or much lower than 1. For HASR human scores, it can be clearly seen that there are not LR values greater than 10 or smaller than 0.1, which means that each calibrated LR value given by human scores is giving little support to the same- or different-speaker hypotheses. This weakness is explained by a nice and fairly intuitive property of calibration: if the discriminating power of a set of scores is very low, the calibrated LR values generated from such scores will tend to be close to 1. In other words, if someone (or some system) is not proficient at discriminating people from their voices, calibration encourages the strength of their opinions to be moderate. Finally, Figure 3 shows that many of the LR values given by the automatic system are much farther than 1 (there is a proportion LR values greater than 10^3 or lower than 10^{-3}), indicating higher strength of the evidence than for human listeners.

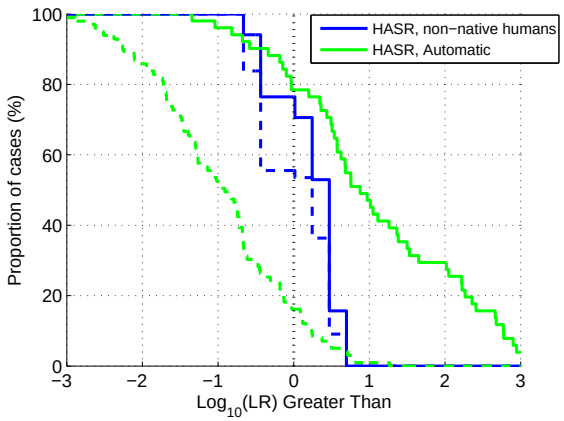

Fig. 3. Tippett plots: proportion of LR in the experimental set being higher than a given value, both for same- (solid) and different-speaker (dashed) trials.

5. CONCLUSIONS

This work has presented a rigorous study about the performance of human lay listeners, not only regarding their discriminating power for speaker recognition tasks (DET plots), but also with respect to the strength of the opinions that they elicit about speech evidence. The study has been carried out in the context of the NIST HASR 2010, a challenging environment with severe session variability and unfavorable conditions. We have recruited a panel of 13 listeners, who have elicited opinions in the form of scores, both for the HASR evaluation trials and for a development set build from NIST SRE 2008 and intended to mimic HASR 2010 conditions. Later, we have calibrated HASR scores thanks to the development set scores. This yields calibrated likelihood ratios (LR), which numerically represent the degree of support of the listeners for the same-speaker or different-speaker hypothesis in each trial. Calibrated LR values allow us not only to measure the discriminating power of human listeners, but also the strength of the evidence evaluated by them.

The main conclusion of this study is that the strength of calibrated LR values elicited by human listeners is significantly low, mainly due to their poor discriminating power in the HASR conditions. In fact, calibrated LR values from human listeners are not greater than 10 or 0.1, indicating an extremely weak support to the same-speaker or different-speaker hypotheses. In conclusion, such opinions will add little information about whether the speakers in both speech materials are or not the same. Moreover, automatic speaker recognition technology clearly outperforms human listeners. These conclusions are in contrast of those found in previous work [1], where the conditions of the speech was much more controlled and the state of the art of the technology was far from the performance of current session variability compensation techniques.

Due to space limitations, this study only shows a small part of all the analysis to be performed on the available scores, including the comparison of native and non-native speakers; the use of different calibration strategies; the measurement of correlation and complementarity of human listeners and automatic speaker recognition; and the fusion of the opinions from both sources.

6. REFERENCES

[1] Astrid Schmidt-Nielsen and Thomas H. Crystal, "Speaker verification by human listeners: Experiments comparing human and machine performance using the nist 1998 speaker evaluation data," *Digital Signal Processing*, vol. 10, pp. 249–266, 2000.

[2] J. Gonzalez-Rodriguez, Phil Rose, D. Ramos, Doroteo T. Toledano, and J. Ortega-Garcia, "Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[3] David Balding, "Interpretation of DNA evidence as a paradigm for speaker recognition," in *Proc. of Odyssey 2010*, Brno, Czeck Republic, 2010.

[4] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[5] Najim Dehak, Reda Dehak, Patrick Kenny, Niko Brummer, Pierre Ouellet, and Pierre Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proc. of Interspeech 2009*, Brighton, UK, 2009, pp. 1559–1562.

[6] N. Brümmer and J. du Preez, "Application independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 230–275, 2006.