Merging human and automatic system decisions to improve speaker recognition performance

Rosa González Hautamäki, Ville Hautamäki, Padmanabhan Rajan, Tomi Kinnunen

School of Computing, University of Eastern Finland, Joensuu, Finland

{rgonza,villeh,paddy,tkinnu}@cs.uef.fi

Abstract

Human judgment is the final authority in forensic speaker recognition, but the use of modern speaker verification systems with accurate algorithms to perform the task under various circumstances has a huge potential to help the expert. The ultimate goal is to improve the accuracy of automatic systems when challenging data is provided and find a methodology for human-aided speaker recognition systems. This work presents an evaluation of speaker recognition carried out by human listeners and a gender dependent i-vector recognizer with a strategy for fusion of the decision process. Our experiments with HASR 2010 and HASR 2012 data indicate complementarity in the performance of the automatic system and the naïve listeners decisions.

Index Terms: Speaker recognition, human assisted, NIST HASR 2012, PLDA system

1. Introduction

Recognizing people by their voices in a variety of settings and under unfavorable conditions is a task that the speech community has studied extensively for many decades. Of particular recent research interest is how to combine *human expertise* and state-of-the-art speaker recognition systems. How much can we rely on current automatic systems and how much can we trust on human perception to discriminate speakers? In their two most recent evaluations of speaker recognition technology, NIST[1] included a *human assisted speaker recognition* (HASR) to study how human involvement, experts or naïve listeners, could be effectively incorporated in speaker recognition tasks.

In [1, 2, 3, 4], systems based on human listeners were compared to automatic systems. It was found that automatic systems outperformed listener results but there was evidence of their complementarity [4]. The results of the systems participating in the pilot test for HASR 2010 [5] showed that human-aided systems did not seem to perform as accurate as the best performing automatic systems. The trials selected were too challenging for the listening task, reasons for this could mainly be human perception related task and its limitations, as discussed in [4, 6]: unfamiliarity with the speakers, length of the data, language, presence of additive noise, audio quality, and channel distortions. The most recent NIST evaluation in 2012 also included a HASR test. Even though this test cannot be regarded as similar to forensic related material, it serves the purpose to study the factors that affect evaluations where human judgment is required. HASR material has also enabled the development of tools and methods that could be used to improve speaker recognition systems. For example, in [2], a phonetic annotation tool was upgraded based on the experimental results with their expert based human-aided system. In [7], the authors present a study that uses forensic phonetic approaches to improved automatic speaker recognition. Such methods include modeling of speech events for voice comparison with the objective to include procedures utilized by forensic scientist to identify speaker distinctive speech segments and their phonetic features.

This study presents the performance analysis of human-aided systems with HASR 2010 and HASR 2012 data. A data set from SRE 2008 described in [1] was used to train the fusion method. In contrast to [1] where the automatic system based on *joint factor analysis* (JFA) was used for a HASR like type of trials, we present the performance of a *probabilistic linear discriminant analysis* (PLDA) system using *i-vectors* as features [8, 9] for HASR data, and a strategy to fuse human and automatic system decisions. Three groups of naïve listeners have tested their ability to recognize the speakers with the speech data. The limited amount of trials does not allow a significant statistical analysis but provides grounds to analyze the work of ordinary listeners when challenged to participate in voice discrimination tasks. Unlike [2, 3, 4] where a comparison of human performance and automatic systems with NIST HASR data was done, our goal is to analyze in which kind of audio samples human listeners are able to succeed in relation to state-of-the-art automatic systems and vice versa. The fusion strategy presented (See figure 1) attempts to consider the complementarity in the performance of automatic systems and human decision procedures.

2. Test Setup

Our experiment follows NIST protocol for human assisted speaker recognition tests. The trials contain two audio samples that are processed sequentially, allowing human interaction with the data. The participants could listen to the files as many times as needed, and answer each trial independently. Similarly our automatic system provides a score to define a degree of support for the hypothesis that both audio samples come from the same speaker. In other words, it answers the question, are the speakers present in the trial audio samples same or different?

2.1. Material

Table 1 shows the data used in this study. The original files in SPHERE format were first converted into WAV format and the channel of interest was extracted. Speech segments from the speaker of interest were manually selected (human-based voice activity detection), followed by audio processing depending on the condition of the sample. Interview segments were downsampled from 16 kHz to 8 kHz because the automatic systems are optimized for 8 kHz data. Both samples in the trial were

[1] http://www.nist.gov/itl/iad/mig/sre.cfm

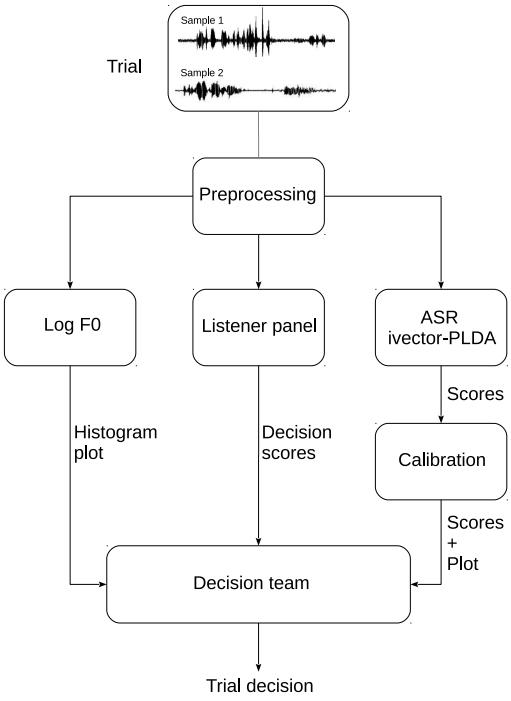


Figure 1: Fusion strategy components.

normalized to set their peak amplitudes to the same level, and in some cases noise reduction was applied: Wiener filter or high pass filter, to reduce the effects of background or additive noise, in specific samples. All trials are English spoken utterances. For the HASR 2010 and HASR 2012 sets, the pair of recordings for each trial were preprocessed as described above before evaluation by the automatic system and by the listeners panel in the case of HASR 2012. However, HASR 2010 files were not segmented manually for VAD before evaluation by the listeners panel.

Table 1: Material used for this study. Telephone and interview type of data.

Corpus	Trials	Listeners
NIST 2008 subset [1]	40	36
NIST HASR 2010	15	36
NIST HASR 2012	20	26

2.2. Listener Panel

The listeners are considered as "naïve", as no formal training in phonetics was required to participate in these experiments. This increased diversity in the panel, from those with little experience in speech analysis to those directly involved in speech related research or applications. None of the speakers were native English speaker but all were fluent or had advanced proficiency in English.

Listener panels for the HASR 2010 and the HASR 2012 overlapped by only 2 finnish speaking listeners. HASR 2010 panel was organized from Singapore and the HASR 2012 panel was organized from Joensuu, Finland. For the HASR 2012 data, the listener panel consisted of 18 male and 8 female with age range of 25 to 60 years and native languages including Finnish, Spanish, German, Turkish, Romanian, Hindi, Farsi, Pushto and Khoekhoe. For the HASR 2010 data, the panel consisted of 36 listeners. Native languages included: Mandarin, Finnish, Farsi, Vietnamese and Malay.

The listeners received a link to play the speech samples as many times as needed. The decision via web-form was one of the following options regarding speaker identity: *definitely same, probably same, definitely different, probably different* and *I cannot decide*. In this way, the decision and confidence of each listener per trial was obtained. Other information requested was related to decision making and evaluation of the trial in question: difficulty of the trial (*easy, moderate, difficult*), time in minutes, methods and cues that the listeners used. The listeners had approximately 24 hours to report their decision. In general, each listener spent between 2 to 20 minutes to evaluate each trial.

2.3. Automatic System

The automatic speaker recognition system used is a gender-dependent PLDA system based on the so-called *i-vector* representation of utterances [8, 10]. Two gender-dependent UBMs were first trained using microphone and telephone data from NIST SRE 04, 05, 06, 08 and 08-followup with 1024 Gaussian components. The i-vector extractor, or T-matrix, was then trained using data from SRE 04, 05, 06, Switchboard and Fisher corpora. The i-vector dimensionality was set to 600 and PLDA hyper parameters were trained using Switchboard, NIST SRE 04, 05, 06 and 08-followup corpora. The speaker subspace dimensionality was set to 200 and noise subspace dimensionality to 0. Before using i-vectors in the PLDA, they were whitened and length-normalized to unit norm [8]. We assumed an affine score model for the post-calibrated PLDA scores.

In addition to the human-based strategy, the *long term fundamental frequency* ($F0$) *histogram* [11] was produced for each trial sample of HASR 2012 data. Although not necessarily the most informative speaker cue, $F0$ is known to be robust to channel distortions and it was used as an additional information for decision making and it is also included as a additional score for fusion. $F0$ was estimated using PRAAT [12], that includes a state-of-the-art autocorrelation-based $F0$ tracker. The $F0$ is processed in a logarithmic scale with 60 histogram bins. Kullback-Leibler distance was used to compare the $F0$ distributions of both samples in a trial, which was additional information that could be used to support listeners or automatic system decisions. Table 2 shows the parameter values used for $F0$ extraction.

Table 2: $F0$ estimation parameters using PRAAT

Parameters	Trial gender	
	Female	Male
Pitch floor	100 Hz	75 Hz
Pitch ceiling	600 Hz	400 Hz
Max. number of candidates	15	15
Voiced/unvoiced cost	0.14	0.14

3. Decision and scoring

The trial decision is reached with information from the listeners panel majority vote, automatic system scores and Kullback-Leibler distance of the $\log F0$ histograms.

Fusion of human listeners and automatic system involved two strategies: First, the complementary information described above was analyzed by a discussion panel, which consisted of the authors of the present paper and one extra member, setting a confidence level between -5 to 5, positive confidence referenced to same speaker decision and negative one to different speaker. The listener's decision is considered as majority voting. When the listeners decision and automatic system score were beyond decision boundary, the discussion panel just confirmed the result, with a high confidence. In the case the group of listeners did not arrive to a consensus opinion and the automatic system showed a score clearly identifying with target or non-target, the decision panel considered the automatic system results as decision. For those trials that were technically challenging for the automatic system, listeners opinion and decision panel discussion dictated the final decision. We refer to this decision process as **heuristic**.

In contrast, our second strategy uses **trained weights**. PLDA, listeners panel and $F0$ histogram fusion was modeled as a linear fusion of PLDA score, averaged listeners panel score and KL distance of log$F0$ histogram. Listeners panel score was formed by assigning -1 to votes for *different* speaker and $+1$ to votes for the *same* speaker. Thus, if equal number of listeners voted for same and different, then panel score would be 0. We form a four-dimensional score vector $\mathbf{s} = (s_{\text{panel}}, s_{\text{PLDA}}, s_{\text{KLdist.}}, 1)$ and feed it to logistic regression model,

$$p(H_1|\mathbf{s}) = \frac{1}{1+\exp\{-\mathbf{w}^t\mathbf{s}\}}, \quad (1)$$

where $\mathbf{w}$ is the vector of weights and H_1 is the same-speaker hypothesis. Weights were estimated by minimizing the cross-entropy, as in [13]. For fusion training data, the subset from NIST 2008 was used. Fusion decision is a sign of $\mathbf{w}^t\mathbf{s}$.

The score calibration parameters, scale factor α and offset β, were trained by optimizing a cross-entropy cost function on the NIST 2010 int-tel (det3) scores for the heuristic fusion, and on a NIST 2008 subset of 40 trials for the trained weights fusion (See section 3). The calibration parameters were set to unit cost with a target speaker prior of $p_{\text{target}} = 0.5$. This leads to decision threshold being at origin.

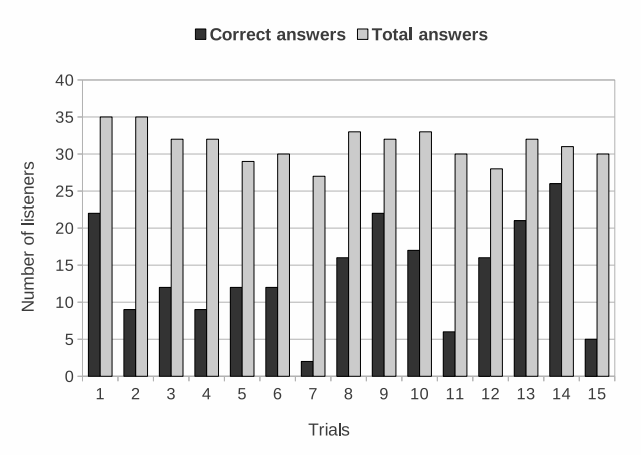
Table 3: System results for NIST HASR 2010 and 2012. Fusion using *trained weights*: PLDA scores, average listeners scores and Kullback-Leibler distance from Log$F0$ histogram distributions

Material	System	Misses	False Alarms	Total
HASR 2010	Listeners	3	6	9
	PLDA	2	0	2
	Fusion	1	4	5
HASR 2012	Listeners	6	3	9
	PLDA	6	2	8
	Fusion	5	1	6

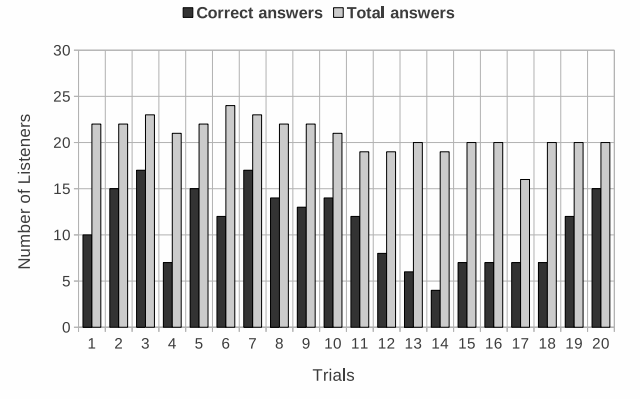
.

4. Results

The results for HASR 2010 and 2012 are reported by hard decisions, *T* if the trial samples belong to the same speaker (target) or *F* if the speaker was different (non-target). Figure 2 shows the performance of the listeners per trial. We observe that some trials were correctly answered just by few listeners. For example, trials 7, 11 and 15 for the HASR 2010 data set, and trials 4, 13 and 14 in HASR 2012. Trial 15 for HASR 2010 and the mentioned trials for 2012 were not recognized successfully by the automatic system.



(a) HASR 2010 [1] without audio preprocessing



(b) HASR 2012 with audio preprocessing (See 2.1)

Figure 2: Listeners performance in each of the trials.

In Table 3, the results for the methods used in this study are shown. Interestingly, the PLDA system performed satisfactorily with HASR 2010 data, outperforming the listeners effort, the trials mistaken were also wrong by the listeners panel. In this case, fusion method did not help to improve the results.

For HASR 2012, five target trials and one nontarget trial were incorrectly classified by both listeners and PLDA system. This suggests that some data will present challenges that current state-of-the-art speaker recognition is unable to process. The characteristics of such trials included additive noise to at least one of the trial samples. The audio quality was also affected by distortion mainly from phone call channel. But not only technical problems characterize these trials, also the speaker could not be discriminated. In the case of nontarget, the speakers sounds so similar that the listeners panel agreed them to be the same speaker, and in the case of target trial, the voice intensity affected by human emotion of the samples, blurred the listeners perception.

In more detail, for the HASR 2012 material, Table 4

Table 4: HASR 2012 trial by trial comparison of results for automatic systems and automatic system assisted by listener panel. Errors, in terms of Misses and False Alarms(FA), are shadowed.

	TRIALS																						
SYSTEM	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**Misses**	**FA**	**TOTAL**
Listener panel (Majority vote)	F	T	F	T	F	T	F	F	T	F	F	T	F	F	F	T	F	F	T	F	**6**	**3**	**9**
PLDA system	F	T	F	T	F	T	T	F	F	F	F	F	F	F	F	F	T	F	T	F	**6**	**2**	**8**
Listener+PLDA+F0 (Heuristic fusion)	T	T	T	T	F	F	F	F	T	F	F	T	F	T	F	T	T	F	T	F	**4**	**4**	**8**
Listener+PLDA+F0 (Trained weights)	F	T	F	T	F	T	F	F	F	F	F	F	F	F	T	F	T	F	T	F	**5**	**1**	**6**
KEY (Ground truth)	**T**	**T**	**F**	**F**	**F**	**T**	**F**	**F**	**T**	**F**	**F**	**F**	**T**	**T**	**T**	**F**	**T**	**T**	**T**	**F**	–	–	–

presents trial-by-trial results for the systems and the majority decision for the listener panel. The correct answers (key) are provided in the last row and the total number of errors in the right column. It is worth noting that five trials (*7,9,12,16,17*) were assigned different answers by the listeners and the automatic system, in these either one was correct. On the other hand, nine trials were correctly answered by both. The $F0$ histogram for trial 2 and trial 4 are presented in figure 3. Trial 2 was correctly classified as same speaker by our fusion system. For trial 4 all systems failed to recognize it as different speaker. The sum of the correctly classified trials by either or both systems gives a total error of 6.

In these experiments, the trained weights fusion was able to correct the mistakes made by either PLDA system or listeners panel using the additional information provided by the log$F0$ histograms distances.

Table 5: Comparison of individual listener performance.

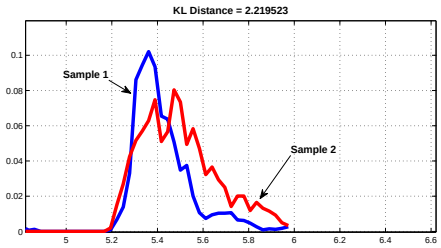
	Listener	**Misses**	**FA**	**Total Trials**	**Device quality**	**Avg. time (min)**
Best	# 19	5	2	20	High headphones	3
	# 26	4	4	20	Low headphones	4
	# 23	2	6	20	Low headphones	9
Worst	# 24	9	1	19	High headphones	2
	# 22	4	7	19	Low speakers	2
	# 5	4	8	20	High headphones	5
Average	# 11	3	5	19	High headphones	4
	# 12	6	3	19	Low speakers & High headphones	2
	# 17	8	0	18	Low speakers & High speakers	2

Individual listener performance for HASR 2012 is presented in Table 5. It shows the best, worst and average performances based on misclassified trials, and the total number of trials answered by the participant. The best subject performed equally to the fusion results. There does not seem to be correlation to the device used during the evaluation nor the time used to make a decision for this specific data.

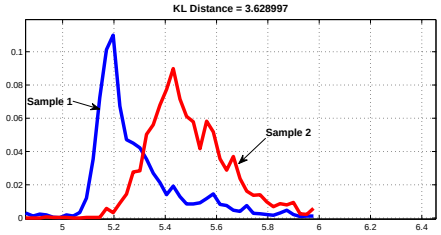
5. Conclusions

This study presented the performance of non native naïve listeners in a speaker verification task using material from NIST human assisted speaker recognition (HASR) test for years 2010 and 2012.

Our results showed that listeners were able to discriminate correctly a few audio samples in which automatic system failed. The listeners decision was based on the perception of the speaker speech style, lexicon, intonation and geographical characteristics (age, gender, dialect). There were also trials



(a)Trial 2. Same speaker histograms. Correctly classified.



(b) Trial 4. Different speaker histograms. False accepted trial.

Figure 3: F0 histograms.

which both, the listeners panel and the automatic system were unable to recognize. In the case of HASR 2012 trials, listeners and automatic system were unable to give correct answers for 6 of the 20 trials. Not only the trial selection included speakers with similar voice characteristics, but also channel distortions to phone call samples, background noise or additive noise appeared as possible sources of errors. Also the long term $F0$ histogram distributions was not conclusive in differentiating the speakers in some of the trials even though provided with a visual representation for the speakers samples. Individual listeners performance did not show evidence that the devices used or the time spent for decision affected their performance. Small number of trials make the statistical significance of the obtained results difficult to ascertain. It remains one of the topics we are currently working on.

The fusion strategy of trained weights was able to reduce the errors by two. This indicates that as a future work it is possible to design a probabilistic model where the strengths of automatic systems and human listeners are more explicitly modeled.

6. Acknowledgments

This work was supported by the Academy of Finland projects number 253120 and 253000. The authors would like to thank the participants in the listeners panel.

7. References

[1] V. Hautamäki, T. Kinnunen, M. Nosratighods, K. A. Lee, B. Ma, and H. Li, "Approaching human listener accuracy with modern speaker verification," in *Interspeech 2010*, Makuhari, Japan, September 2010, pp. 1473–1476.

[2] R. Schwartz, J. P. Campbell, W. Shen, D. E. Sturim, W. M. Campbell, F. S. Richardson, R. B. Dunn, and R. Granvill, "USSS-MITLL 2010 human assisted speaker recognition," in *ICASSP 2011*, Prague, Czech Republic, May 2011, pp. 5904 – 5907.

[3] D. Ramos, J. Franco-Pedroso, and J. Gonzalez-Rodriguez, "Calibration and weight of the evidence by human listeners. the ATVS-UAM submission to NIST human-aided speaker recognition 2010," in *ICASSP 2011*, Prague, Czech Republic, May 2011, pp. 5908 – 5911.

[4] J. Kahn, N. Audibert, S. Rossato, and J.-F. Bonastre, "Speaker verification by inexperienced and experienced listeners vs. speaker verification system," in *ICASSP 2011*, Prague, Czech Republic, May 2011, pp. 5912 – 5915.

[5] C. S. Greenberg, A. F. Martin, G. R. Doddington, and J. J. Godfrey, "Including human expertise in speaker recognition systems: report on a pilot evaluation," in *ICASSP 2011*, Prague, Czech Republic, May 2011, p. 5896 5899.

[6] W. Shen, J. Campbell, D. Straub, and R. Schwartz, "Assessing the speaker recognition performance of naive listeners using mechanical turk," in *ICASSP 2011*, Prague, Czech Republic, May 2011, pp. 5916 – 5919.

[7] K. J. Han, M. K. Omar, J. Pelecanos, C. Pendus, S. Yaman, and W. Zhu, "Forensically inspired approaches to automatic speaker recognition," in *ICASSP 2011*, Prague, Czech Republic, May 2011, pp. 5160 – 5163.

[8] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Interspeech 2011*, Florence, Italy, August 2011, pp. 249–252.

[9] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[10] M. Senoussaoui, P. Kenny, N. Brummer, E. de Villiers, and P. Dumouchel, "Mixture of PLDA models in i-vector space for gender-independent speaker recognition," in *Interspeech 2011*, Florence, Italy, August 2011, pp. 25 – 28.

[11] T. Kinnunen and R. González Hautamäki, "Long-term F0 modeling for text-independent speaker recognition," in *Int. Conf. on Speech and Computer (SPECOM'2005)*, Patras, Greece, October 2005, pp. 567–570.

[12] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," Version 5.3.34, retrieved November 2012 from http://www.praat.org.

[13] V. Hautamäki, T. Kinnunen, F. Sedlák, K. A. Lee, B. Ma, and H. Li, "Sparse classifier fusion for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 8, pp. 1622–1631, August 2013.