BUT HASR'12 EXPERIENCE: ARE DEVELOPERS OF SRE SYSTEMS NAÏVE LISTENERS?

Pavel Matějka [1,2], *Ondřej Glembek* [1], *Oldřich Plchot* [1] *Milan Schwarz* [2], *Tomáš Cipr* [2,1], *Sandro Cumani* [1,3], *Radim Kudla* [2], *Igor Szöke* [1], *Marie Svobodová* [4], *Květoslav Malý* [5], *and Jan Černocký* [1]

(1) Brno University of Technology, BUT Speech@FIT and IT4I Centre of Excellence, Czech Republic,
(2) Phonexia, Czech Republic,
(3) Politecnico di Torino, Italy,
(4) PONTES, Czech Republic,
(5) Ministry of Defense, Czech Republic
matejkap@fit.vutbr.cz

ABSTRACT

Brno University of Technology techniques and results obtained in NIST 2012 Human Assisted Speaker Recognition (HASR) task are described in this paper. The scores of an automatic i-vector based system and 10 naïve listeners were fused. The results show that an automatic system performed generally better than human listeners, with the exception of an engineer that has never done such analysis before. The conclusions of this study should be taken with care due to low number of trials.

Index Terms— automatic speaker recognition, human assisted speaker recognition, HASR, i-vector.

1. INTRODUCTION

Brno University of Technology has a long track of R&D in the area of automatic speaker recognition (SRE). Alone or in consortia, we have been successful in building automatic SRE systems; however, our researchers and engineers heavily suffer from the "IT sickness": they look at the error rates of the systems rather than taking headphones and going to listen to the analyzed signals. This modus operandi is also supported by the request of *no human interaction with the data* in the NIST Speaker Recognition Evaluations [5].

In 2010, NIST organized the first Human Assisted Speaker Recognition (HASR) [1] in which BUT did not take part. The conclusion of participants (see for example [2, 3]) was that with some exceptions, automatic systems outperformed humans.

We have therefore welcomed the 2012 NIST HASR as an excellent opportunity to have BUT engineers (with two external helpers) listen to the data, and compare the results to an automatic system. The main intention of our participation in HASR was not to achieve the best possible results (that's what we are trying in NIST SRE and related projects' evaluations), but rather to finally listen to the data that our systems are supposed to process.

We have submitted a "fusion" of 11 subsystems — a simple average of the scores of 10 naïve listeners and one automatic system based on MFCC i-vector + PLDA. This paper briefly reviews HASR data (Section 2), and describes our automatic system (Section 3). Section 4 contains instructions for and notes on strategy from our naïve listeners, section 5 outlines the calibration and fusion. Results are presented in section 6 and we conclude in section 7.

This work was partly supported by Technology Agency of the Czech Republic grant No. TA01011328, and by European Regional Development Fund in the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070). Sandro Cumani was supported by The European Social Fund (ESF) in the project Support of Interdisciplinary Excellence Research Teams Establishment at BUT (CZ.1.07/2.3.00/30.0005).

2. DATA

BUT participated in HASR1 part of the evaluation involving 20 trials (see [5], section 11, for details). The trials contained two recordings, with sufficient amounts of speech (approximately 1 minute of speech each). There was a mismatch in acoustic channels — while one of the recordings was telephone of generally good quality, the second one was from an interview, with significant channel mismatch compared to the telephone one and lots of noise (60Hz hum). Very low amplitude in the interview recordings was found as the hardest problem by listeners using just standard voice visualization and playback softwares – they were simply not able to amplify such signals enough. UNIX-savvy listeners used `sox` with the `-v` option.

NIST evaluations assume independent processing of trials, with no influence of the other trials. This is hard to fulfill for humans, NIST tried to minimize this dependency by sequentially releasing the trials (new one was released only when the score for the old one was submitted).

The results were submitted to NIST in standard way [5] as hard decisions and scores.

3. AUTOMATIC SYSTEM

The automatic system is similar to the i-vector part of BUT/ABC submission for NIST 2010 SRE evaluation [4], but used diagonal (not full) covariance matrices and was gender-independent. BUT's usual experimental systems are based on a relatively messy collection of Matlab and C executables connected by a "glue" of various UNIX shell scripts. In the meantime, our spin-off company Phonexia integrated i-vector extraction and PLDA scoring into its production software [1]. We were therefore happy to use Phonexia's SpeechAPI with command line tools for more convenient and fast i-vector extraction and PLDA scoring.

[1] `http://phonexia.com/download/`

As Phonexia does not disclose the data used for training the models (significant part of this data is provided by its customers), we stuck with the the data-sets designed for NIST SRE10 [6].

The following paragraphs provide brief overview of the system, please consult [6] and references therein for a more detailed description.

3.1. Voice activity detection

Two versions of audio were used:

- original audio.
- editted audio with manually removed long portions of silence, cross-talks, unintelligible speech segments, and the strongest noises.

Voice activity detction (VAD) was performed by our Czech phoneme recognizer (with all phoneme classes linked to the *speech* class) for both versions of audio. The results for the original audio are later denoted as *Automatic-1*, the ones for the editted audio (that were finally fused with human listeners' results) as *Automatic-2*.

3.2. Feature Extraction and UBM

We used short-time gaussianized MFCC 19 + C0 augmented with their delta and double delta coefficients, resulting in 60-dimensional feature vectors. The analysis window is 20 ms long with the shift of 10 ms. Short-time gaussianization uses a window of 300 frames (3 sec).

One gender-independent universal background model was represented as a diagonal covariance, 2048-component GMM. It was trained on the selection of NIST SRE data. The proportion of telephone and interview data were 50:50. Variance flooring was applied in each iteration, where the threshold was computed as an average variance from each previous iteration, scaled by 0.1.

3.3. I-vector system

We used gender-independent i-vector extractor with 600 dimensions. The i-vectors extractor was trained on telephone data from NIST SRE 2004, NIST SRE 2005, NIST SRE 2006, Switchboard II Phases 2 and 3, Switchboard Cellular Parts 1 and 2, Fisher English Parts 1 and 2.

3.4. LDA + PLDA

LDA+PLDA were used to filter out the channel information from the i-vectors. LDA with dimensionality reduction to 200 was trained on the same data-set as the i-vector extractor, however, we only used files from speakers for which we had more than 6 recordings. The same data-set was used for PLDA training.

3.5. Normalization

Our assumption was that PLDA gives well calibrated scores. We have run this system on HASR 2010 with good results, see Table 1 for details. We can count 5 wrong answers for this system and 4 for ABC primary system designed for NIST SRE2010. Interestingly, they agree only in 10 trials, so the systems make different errors.

Table 1. *Results for automatic MFCC i-vector system used in HASR1-2012 (SYS1) and ABC 2010 primary system (SYS2) on 2010 HASR1 evaluation.*

Enroll	Test	SYS1	SYS2	Reference
01_ehzum-a	01_trevr-a	-1.27	1.25	target
02_eibhn-a	02_thftl-b	0.57	-1.66	nontarget
03_ehymn-a	03_tcrte-b	-5.35	2.14	nontarget
04_eqbwg-a	04_trify-a	0.34	-3.51	nontarget
05_ehjmk-a	05_trrkn-a	-0.14	-3.11	target
06_ejntb-a	06_tnbcy-b	-6.23	-3.90	nontarget
07_eftsd-a	07_tzocd-b	5.14	2.40	target
08_elnji-a	08_tdaxi-a	-1.82	-1.05	nontarget
09_eltwa-a	09_tpzez-a	-5.40	-7.48	nontarget
10_euerk-a	10_tzvxk-a	5.12	6.44	target
11_erjdr-a	11_tvndw-a	4.00	8.56	nontarget
12_ezlsi-a	12_thpms-b	-0.32	-2.12	nontarget
13_ekuzl-a	13_tkvay-a	-5.61	-3.30	nontarget
14_ebgbw-a	14_tncns-b	6.35	9.27	target
15_enjvn-a	15_tepkw-b	2.49	-0.14	target

4. NAÏVE LISTENERS

The 10 listeners participating in the evaluation have different levels of phonetic expertise, given mostly by their engineering background. They all have a relatively good command of English, but none is a native speaker[2]. We have been working with audio files during the development of our automatic systems, therefore, we sometimes listen to audio files to verify if, e.g., our VAD is working properly. However, we have never done any forensic expertise. Therefore, all evaluators can be considered as naïve listeners.

From the 2010 HASR evaluation, we know that an automatic system would do best. We have therefore decided to provide several kinds of information to the listeners:

- audio recording
  - original audio file
  - manually editted audio file.
  - de-noised versions of both above with Wiener filter
- length of speech segments
- SNR implemented according to [7].
- Score from automatic system for the original and edited audio recordings

It was up to the listener whether he or she uses this information or not.

The task of the listener was to assign the score of the trial in the range -5 to 5 with this following instruction provided:

-5: I am 100% sure that the speakers are different
0: I do not know
5: I am 100% sure that the speakers are same

4.1. Different strategies adopted by listeners

Below is a summary of selected listeners' strategies (sorted w.r.t. the accuracy):

[2]Nine are Czech, one is Italian.

- **Listener-10:** My decisions were always based on the output of the automatic system and my knowledge of the architecture of such systems. Rather than trying to decide myself about the trial, I was trying to asses the influence of the quality of the recordings and the channel mismatch to the score produced by the automatic system. I spent 5 to 10 minutes with each trial. In the first pass I listened to both audio files not concentrating on the speech at all. I was concentrated on the quality of the recordings taking into account noise and the volume/gain of the speech. If I found a big channel mismatch and the quality of one recording to be very low, I suspected, that the automatic system would output a negative score even for a target trial. I was also considering the shift in the score of the automatic system when it was provided with automatic and manual VAD. If there was a big shift in the score towards positive value between the automatic and the manual VAD I was inclined to decide the trial as target. At the end of the decision process I selected very short parts of both recordings (usually up to one sentence) and I was repeatedly listening to them. Ideally these sentences contained at least one same word. In this case I concentrated more at this particular word.
- **Listener-1:** I listened to whole audio files, because I was preparing them for automatic system. I was driven mainly by the decision of automatic system and tried to correct the output based on my personal intuition based only on the listening.
- **Listener-8:** I spent about 3 minutes on each trial. I opened both audio files in WaveSurfer and listened to longest speech parts (about 1/2 of each file). I tried to find the same words or laughing or repeating parasite words (ehm, um, you know, ...), and also melody of utterance. Based on this, I produced scores of speaker similarity avoiding too high or too low scoring. In most of the cases, I did not change my score based on automatic system. It was more for comparison of decision for me.
- **Listener-6:** I spent 3 to 5 minutes on each trial. I worked with Wiener-filtered audio-file. I was comparing not only the acoustic similarities (including prosody), but also the content, i.e. true hypotheses were emphasized if the content was found to be related.
- **Listener-7:** I listened either to the entire file or the longest speech segment. I tried to decide especially by comparing the prosody (rhythm, stress, intonation, ...) of the speech. As the two recordings came mostly from different channels I found it hard to compare them based on how it 'sounded' to me only.
- **Listener-9:** I spent about 2-3 minutes on the trial, I listened to segments with speech (about 50%–100%) of one file and then switched to the other file (also 50%–100%). I tried to find similar words or at least emotional (loudest) segments. Then I compared just by listening. No spectrum and other techniques were used. I work on speech recognition, but unlike other participants, I haven't been involved in speaker recognition. I refined my score according to the scores of the others in about 3 cases.
- **Listener-4:** I loaded both files to WaveSurfer. I listened to one speech segments from one recording and to one from the other and I repeated this for about 30sec to 2 minutes till I was sure with my answer and I assigned the confidence. Usually I had to change the loudness of the interview file.

The main problem was that a new trial always arrived only when one had just finished current the one. This caused quite some time-stress: at the beginning, the listeners had a week for a trial, but got delayed and at the end, they had to process two trials a day.

Listeners also struggled with very low audio volume, different channels, non-native language, and noise.

5. CALIBRATION AND FUSION

The calibration and fusion is very hard to do in this case, because there is almost no development data available. Our first idea was to use HASR1 2010 data, but there are only 15 trials, which is anyway not enough to train an automatic classifier.

We took simple average of the scores produced by the automatic system on manually edited files and all naïve listeners as the final confidence, and we used the threshold of 0 to get the TRUE or FALSE decision.

We also ran the automatic system on the de-noised versions of the audio files, but since the system was not trained on such data, the scores were shifted and it was hard to derive a threshold for this setup.

6. RESULTS

The complete set of results can be found in Table 2, with the following notation used:

- *Txx* denotes the trial.
- *Listener-x* lines contain the scores provided by listeners. X is used when the listener did not manage to score the trial.
- *Automatic-2* are scores from the automatic system described in section 3 which processed manually edited recordings.
- *Average* is the average score given to the trial by all listeners and Automatic-2 scores.
- *Final confidence* is the average rounded to one decimal point. Note a discrepancy at trial T03: one listener submitted his score late, so it is reflected in the *Average* line. The *Final confidence* line contains the results submitted to NIST. This difference however does not change the final result.
- *Final decision* stands for our hard decision derived by thresholding the final confidence at the value of 0
- *Reference* released by NIST.
- *#Trials* stands for the number of trials a listener or system scored. For systems, obviously *#Trials=20*.
- *Acc* is the accuracy of correctly recognized trials: number of correctly attributed TRUE or FALSE with threshold set to 0 divided by *#Trial*.

For the sake of completeness, the table contains also *Automatic-1*: the results of the automatic system with the original recordings with only VAD (section 3.1) run, without manual editing.

For two trials (T07, T08), the final score was set to 0, and the hard decision to TRUE, as the *Average* was slightly above 0 (on the 3rd decimal point). This turned out to be wrong answer in both cases. In case we set these trials to FALSE, we would decrease the number of our errors from 6 to 4 and reach 80% overall accuracy.

Table 2. *Complete BUT results for 2012 HASR1 evaluation.*

Name&Trial	T01	T02	T03	T04	T05	T06	T07	T08	T09	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	Acc[%]	#Trials
Reference	T	T	F	F	F	T	F	F	T	F	F	F	T	T	T	F	T	T	T	F		
Final Decision	T	T	F	T	F	T	T	T	F	F	T	F	T	T	T	F	T	F	T	F		
Final Confidence	1.6	2.3	-1.5	2.9	-1.5	3	0	0	-1	-2.1	0.4	-1.1	0.2	0.1	0.07	-0.8	2.54	-0.7	2.4	-0.4	0.70	20
Average	1.59	2.30	-1.13	2.90	-1.51	2.97	0.04	0.01	-0.98	-2.08	0.40	-1.13	0.19	0.11	0.07	-0.75	2.54	-0.71	2.38	-0.38	0.75	20
Automatic-2	5.44	3.97	0.2	2.5	-0.1	4.7	-0.6	-1.9	-1.8	-1.9	-3.6	-3.2	-2.1	-2	-2.3	-3	4.4	1.3	4	-2.4	0.70	20
Listener-1	1	2	-2	3	-2	4	-3	-2	-1	-3	-2	-5	-2	2	-1	-1	4	-1	2	-3	0.75	20
Listener-2	1.5	2	-3	4.5	-2	-2	3	3	-1	-1	1	-1	0	2	-1	1	3	1	1	1	0.50	20
Listener-3	3.5	5	1.5	4.5	-2.5	5	2	0	X	-4	2	-2	X	X	X	X	X	X	X	X	0.55	11
Listener-4	1	3	X	X	-2	4	1	3	-3	1	0	X	2	X	-1	-3	1	X	X	0	0.50	14
Listener-5	-2	2	1.5	1.5	-3.5	3	-2	-2	-1	-2	2	-1	0	-2	2	1	2	-1	3	1	0.50	20
Listener-6	2	3	-3	4	-1	1	2	-2	-2	-3	3	X	1	2	-2	X	2	X	X	1	0.63	16
Listener-7	1	-1	-3	3	-2	4	-1	2	-1	-3	2	2	-2	-1	1	-2	2	-3	2	X	0.53	19
Listener-8	-1	2	-0.5	3.5	-1.5	3	-1	3	1	-2	-1	-1	2	1	3	X	4	-2	3	1	0.74	19
Listener-9	3	1	-2	0	2	X	1	-2	X	-2	2	3	2	1	1	2	1	-2	2	1	0.56	18
Listener-10	2	X	-1	2.5	-2	X	-2	-1	1	-2	-1	-2	1	-2	1	-1	2	1	2	-3	0.89	18
Automatic-1	5.87	1.55	-0.3	1.9	-0.6	3.5	-0.2	-2.9	-1	-1.6	-2.5	-3.37	-4.1	-2.4	-3	-3.4	4.2	0.8	4	-0.9	0.75	20

7. DISCUSSION AND "CONCLUSIONS"

First, it should be clearly stated that the number of trials was very small (BUT did not participate in HASR2 with 200 trials), and the numbers of trials processed by individual listeners were even smaller, so the results are not statistically significant and should be taken with the greatest care.

The automatic systems performed relatively well and actually reached accuracy that is equal to the final performance. It seems that manual editing of the audio leading to better voice activity decisions did not change the behavior of the systems – no conclusion can be made on whether the scores on the edited signals were better or worse than for the original ones. This is in sharp contrast with strong dependency of SRE performance on VAD accuracy seen on many tasks and in many projects.

Concerning the naïve human listeners, it is obvious, that:

- the strategies are defined rather ad-hoc and are well different from standards accepted for forensic comparison.
- the time spent on trials was significantly less than in forensic analysis.
- none of the listeners is native in English, so that little knowledge usually extensively helping human experts (lexical, dialectal, social, ...) could be used.
- on the other hand, listeners could advantageously use the knowledge of automatic SRE system and judge the quality of its output for non-matching conditions.
- listeners were not agnostic of the definition of the task and knew that NIST has selected trials that automatic systems made errors on – this contributed to the tendency to negate the results of the automatic system for severe mismatches between training and test.

The best listener reached 89% accuracy which is far beyond the automatic system and fused results. He is an SRE researcher, never performed this kind of analysis and does not play any musical instrument.

In comparison, the automatic system (state-of-the-art but with no special tuning) reached better results than most of the listeners, which confirms its practical usability.

The experiment would have more value in case three groups of listeners could be defined:

1. real naïve listeners with no special background in phonetics or SRE.
2. SRE developers that understand the problems of automatic systems.
3. listeners trained in sciences requiring careful listening of speech such as phonetics students.

which gives us some "TODOs" for future editions of HASR.

8. REFERENCES

[1] Craig Greenberg, Alvin Martin, et al: Human Assisted Speaker Recognition In NIST SRE10, in *Proc. Odyssey 2010: The Speaker and Language Recognition Workshop*, Brno, 2010.

[2] Reva Schwartz, Joseph P. Campbell et al.: USSS-MITLL 2010 Human assisted speaker recognition, in *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011*, Prague, 2011.

[3] Nicolas Audibert, Anthony Larcher, et al: LIA human-based system description for NIST HASR 2010, in *Proc. NIST 2010 Speaker Recognition Evaluation Workshop*, Brno, 2010.

[4] Niko Brummer, Lukas Burget, et al.: ABC System description for NIST SRE 2010, in Prof. NIST 2010 Speaker Recognition Evaluation Workshop, Brno, Czech Republic, 2010, `http://www.fit.vutbr.cz/research/view_pub.php?id=9346`.

[5] The NIST Year 2012 Speaker Recognition Evaluation Plan, NIST Multimodal Information Group, 2012, `http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf`.

[6] Pavel Matějka, Ondřej Glembek, Fabio Castaldo, Jahangir Alam, Oldřich Plchot, Patrick Kenny, Lukáš Burget, and Jan Černocký, "Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification," in *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011*, Prague, 2011, pp. 4828–4831.

[7] Chanwoo Kim and Richard M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Proceedings of Interspeech 2008*, Brisbane, Australia, Sept. 2008.