Human Assisted Speaker Recognition (HASR) Submission by the I4U Team

Ville Hautamäki, Tomi Kinnunen, Mohaddeseh Nosratighods, Kong-Aik Lee, Bin Ma, and Haizhou Li

Institute for Infocomm Research (I^2R), A*STAR, Singapore
School of Computing, University of Eastern Finland, Joensuu, Finland
School of Electrical Engineering and Telecommunication, University of New South Wales, Australia
{*vishv,kalee,mabin,hli*}*@i2r.a-star.edu.sg*, *tkinnu@cs.joensuu.fi*, *hadis@unsw.edu.au*

Abstract

Being able to recognize persons from their voices is a natural ability that we taken for granted. Recent advances have shown significant improvement in automatic speaker recognition performance. Besides being able to process large amount of data in a fraction of time required by human, automatic systems are now able to deal with diverse channel effects. Though human listeners are more robust to intra-speaker variabilities, they perform badly in the present channel effects. It is with interest in mind that I4U team is participating in the HASR portion of the NIST 2010 SRE. Listening team consisted in total of 43 naive listeners from Singapore, Australia and Finland.

1. System description

The task considered in the HASR is *speaker verification*, that is, deciding whether two given utterances are spoken by the same speaker. A single utterance pair which the decision must be declared for (either by human or automatic method) is called *trial*. For the automatic method, one of the utterances is considered as the enrollment utterance and the other one the test utterance. For human listeners, however, we did not impose such artificial training/test division but the listeners could listen to the samples in any order and as much as they wanted to.

We produced a steroa WAV-file of each trial prior sending it to listeners. Each channel was normalized, so that intelligibility of the speech was as high as possible. Energy difference between speech and (originally) high energy laughter was minimized. Trial was uploaded to the server, where participating listeners could download it whenever they pleased. We set a time limit of few days for each trial, according to the HASR schedule. Each listener could listen trials whenever and wherever they pleased.

Listeners were asked to provide (in web-interface, shown in Fig. 1) a true/false -vote, percieved difficulty (in scale 1-5), time spent (in minutes) and perceptual cues used. Only true/false -vote is used in the HASR portion, other information is used in the post-evaluation study. Final decision was plain majority vote of all the votes received by the deadline. Confidence score for both true and false decifions was set to:

$$\frac{\text{nro. true votes}}{\text{total nro. of votes}}. \quad (1)$$

Number of votes per trial varied between 28 to 38, average being 32.

Each listener was "trained" by 40 trials selected from the NIST 2008 core test. I4U NIST SRE 2008 submission [1] system first scored all core trials. EER point was selected from that set. Then 20 trials were selected by their closeness to the

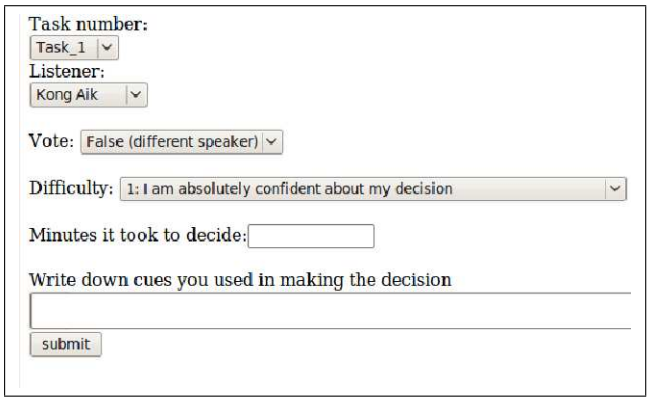

Figure 1: Screenshot of web-interface.

EER point (also trials were selected to be gender and targer/non-target balanced). In this way they correspond to the trials where automatic system will find difficult to make a decision one way or another (not counting calibration cost). This set we denote *hard* set. Automatic system had a classification error of 40% in that set. Other set was selected in a way that using the same threshold automatic system does not make any errors, denoted by the *easy* set. Basic statistics of the two sets are shown in the Table 1.

Table 1: Computing the basic statistics of the two sets of trials using the pool of listeners and automatic system. Statistics include, minimum, maximum and average human error rates.

Data set	Min	Max	Avg	Fusion	Automatic
Hard	20%	60%	40.42%	25%	40%
Easy	10%	45%	26.25%	20%	0%

2. References

[1] Haizhou Li, Bin Ma, Kong-Aik Lee, Hanwu Sun, Donglai Zhu, Khe Chai Sim, Changhuai You, Rong Tong, Ismo Karkkainen, Chien-Lin Huang, Vladimir Pervouchine, Wu Guo, Yijie Li, Lirong Dai, Mohaddeseh Nosratighods, Thiruvaran Tharmarajah, Julien Epps, Eliathamby Ambikairajah, Eng-Siong Chng, Tanja Schultz, and Qin Jin. The i4u system in nist 2008 speaker recognition evaluation. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 0:4201–4204, 2009.