SPEAKER VERIFICATION BY INEXPERIENCED AND EXPERIENCED LISTENERS VS. SPEAKER VERIFICATION SYSTEM

Juliette Kahn\textsuperscript{1,2}, Nicolas Audibert\textsuperscript{1,3}, Solange Rossato\textsuperscript{2,4}, Jean-François Bonastre\textsuperscript{1}

\textsuperscript{1}University of Avignon, Laboratoire Informatique d’Avignon (CERI-LIA), France
\textsuperscript{2}University of Grenoble, Laboratoire Informatique de Grenoble (LIG), France
\textsuperscript{3}CNRS, Laboratoire de Phonétique et de Phonologie (LPP), France
\textsuperscript{4}University of Grenoble, Gipsa-Lab, France

ABSTRACT
This paper describes the participation of the LIA in the Human Assisted Speaker Recognition (HASR) task of the NIST-SRE 2010 evaluation campaign and its extension to a larger number of listeners. The human performance in such unfavorable conditions is analyzed in relation to the decision of a speaker recognition automatic system. Results of the perception test showed an important inter-trial variability (from 3\% to 90\% of correct answers for non-target trials) whereas there was no significant difference between the experienced and inexperienced listeners. Some complementarity between speaker verification system and human decisions was also found.

Index Terms— Speaker verification, HASR, perception tests, speaker discrimination task

1. INTRODUCTION
Over the last decade, automatic speaker verification systems have been assessed by the National Institute of Standards and Technology (NIST) [1]. This year, for the first time, the NIST-SRE included a Human Assisted System Recognition task. A small test (15 trials (HASR1) and one of 150 trials (HASR2)) were offered for systems based on human expertise, possibly in conjunction with automatic processing. This pilot effort was designed in order to determine how the human expertise can enhance or supplement the system performance in adverse conditions.

Experiments conducted with speaker verification systems provide reliable results. A performance loss is observed when the length of the training signals changes from 2.5 min to 10s in [2]. Besides, results of such automatic systems are related to the choice of the excerpt used for the training of speaker model [3]. Obviously, the speaker verification system results are sensitive to the recording conditions. The two small sets of trials HASR1 and HASR2 had been selected to maximise the difficulty for automatic systems, involving a challenging set of both target trials and non-target ones [4] in order to see if human listeners could provide additional information.

The human skills for distinguishing voices have been studied for almost a century by linguists and psychologists (since 1937 for [5]). Several studies ([6], [7]) have shown a difference of performance according to the degree of familiarity between listeners and speakers. In [7], the speaker identification error rate was increased from 2\% for the familiar voices to more than 60\% for the unfamiliar voices. Another important factor is the length of the excerpts heard by the listeners. As shown by [8], the discrimination ability with the two sentences ‘get him!’ and ‘face down on the ground and hands behind your back now!’ varies from 52\% with the short sentences to 82\% with the long ones. Moreover, results of [9] underlined that the discrimination performance is not significantly different between sentences with exactly the same words and with different sentences sharing similar phonological content. If the discrimination between human voices is not directly related to the lexical content of the excerpt, language skills are important since the native listeners have shown better results than the listeners who do not know the language spoken [10]. The issue of expertise is still relevant with studies showing conflicting results: [11] concluded on the supremacy of the experts versus the native listeners whereas [12] showed no significant difference between the two groups. As far as we know, the large majority of human discrimination tasks have been using excerpts recorded in the same conditions.

In this paper, we assess the complementarity of the answers given by the ALIZE/spkDet system and those made by humans. Indeed, listeners and speaker verification systems may not rely on the same information to estimate the proximity of two voices [13]. This human discrimination task combines several difficulties: i) the voice samples are recorded in different (and noisy) conditions; ii) the listeners are unfamiliar with the speakers; iii) the listeners are non-native English speakers. We distinguish people who are used to study and annotate speech signals (hereafter called ‘experienced’) and people who are not used to work on speech signals (called ‘inexperienced’). Section 2 describes the Speaker Verification System assessment. Section 3 summarizes the perception test (stimuli, listeners, scoring). The results are presented and discussed in Section 4 and 5 respectively.

2. ALIZE SPK/DET GMM-SVM
The speaker verification system used in this study is the open source ALIZE/spkDet system [14]. It includes a simplified version of the Joint Factor Analysis [15]. The LIA SpkDet toolkit benefits of the LIB-SVM library [16] to estimate SVMs and classify instances. The system is developed on NIST SRE 2008 data. This GMM-SVM system is regularly assessed during the NIST speaker recognition evaluation.

3. HUMAN CONTRIBUTIONS
3.1. Samples
Human does not need 2.5 minutes to distinguish voices as showed by [8], and listening to such a long speech excerpt before making...
a comparison is tedious. For these reasons, 6 seconds-long extracts are automatically selected from the model and test segments of each trial and concatenated to build the audio stimulus. The extracts selection is guided by 2 criteria: the selected extracts should include a proportion as large as possible of speech frames and the extracts should include only the interviewee voice. The automatic method succeeded in selecting extracts, except for 3 trials where the selection was manually performed. The selected extracts are then combined into an audio stimulus, generated using the Praat software [17]. A beep is inserted between consecutive segments to signal inter-excerpts switching. All extracts included in the generated stimulus are normalized in intensity. Each file of a model/test pair has a minimum total duration of 42 seconds.

3.2. HASR submission

Three native French listeners (2 female aged 25 and 36, 1 male aged 31) with experience in phonetics and speech analysis, and without any known hearing impairment participate to the HASR submission. For each trial, all the listeners have to decide whether the extracts in the stimulus has been uttered by the same speaker or not. All 150 stimuli from HASR2 database have been evaluated (this set includes the 15 HASR1 stimuli). All the listeners have to judge the trials in the same order, as a requirement of the HASR protocol. When a large amount of noise (especially low-frequency noise) is present in either the model or test segment, the listeners have the possibility to band-filter the signal after visual inspection of the power spectrum, in order to reduce the perceptual heterogeneity caused by the recording channel differences. They can listen to the generated stimulus by selecting parts or as a whole as many times as necessary.

3.3. Perception tests

For this work, the panel of listeners is enlarged in order to check the tendency observed in HASR submission, and the evaluation protocol is randomized in order to compensate for training effects.

3.3.1. Listeners

There are 2 groups of listeners. Group 1 consists of inexperienced listeners and Group 2 consists of experienced listeners. The inexperienced listeners are native French men (20) and women (9) who have been studying English for 9.0 years in average (mean age 29.3). Only one listener has been living in an English-speaking country for a year. The 18 experienced listeners are native French with experience on speech signal annotation. None of them have been living in an English-speaking country for more than 1 year.

3.3.2. Listening

The ratio of target and non-target trials in HASR is unbalanced (51 and 99 respectively for HASR2 and 6 and 9 respectively for HASR1). To end up with a balanced subset ([18]), the 51 HASR2 target trials (including the HASR1 target ones) were balanced with the 9 HASR1 non-target trials and the 42 most difficult HASR2 non-target trials (i.e those with the highest likelihood ratio given by the system, following [1]). The inexperienced listeners ability is evaluated with those trials. The test is divided in two sessions of approximately 40 minutes. Because of the difficulty to find experienced people, for the evaluation by experienced listeners, we select all the trials of HASR1 and the 3 most difficult target trials of HASR2, so as to ending up with a set of 9 target and 9 non-target trials which are included in the selection for the inexperienced ones. The presentation order is randomized and different for each listener. The listeners are allowed to listen to the stimuli once. The evaluation is taken in a quiet environment using closed headphones.

3.4. Scoring

For each trial, the decision of the ‘human system’ is defined by the majority voting among the decisions taken by the listeners. The ratio of the positive answers by the number of listeners defines a confidence score. In order to make comparisons between human and automatic systems possible, the confidence score for each stimulus is mapped to the non target and target score distributions obtained with our system on SRE 2008 data, as summarised in figure 1. Confidence scores below 33% imply that the trials are likely to be non target trials and therefore are mapped with the distribution of the non-target trial scores. Confidence scores above 66% are mapped with the distribution of target trial scores. Confidence scores around 50% are mapped with an interpolation of both distributions.

![Fig. 1. Mapping between positive answers ratio in perception tests and SVM scores distribution. The mean non-target score is attributed to the ratio 1/3, and the mean target score to the ratio 2/3.](image)

4. RESULTS

4.1. HASR submission

The automatic system decision threshold is the one used for the NIST-SRE 2010. With this threshold, the FA ratio is 13% (vs 15.9% on all NIST-SRE trials) and the FR ratio is 12% (vs 4.5% on all NIST-SRE trials) on HASR2 data. The higher FR rate obtained on HASR compared to the one obtained on the whole NIST 2010 protocol confirms that the trials selected for HASR are particularly difficult. Figure 2 shows the results obtained by the LIA and the other participants on HASR2. The listeners took 12 to 180 seconds of listening in order to take their decision (mean: 66 seconds). The listeners FA ratio is 29% for 12.

4.2. Inexperienced and experienced listeners performance

Overall, the inexperienced listeners give the correct answers for 54% of trials. A binomial test indicates that this ratio is significantly over chance level (p < 0.0001). A separate analysis of target trials (53% success, i.e. 47% FR) and non-target trials (54% success, i.e. 46% FA) indicates that listeners did not perform better for a particular type of trials (t(100) = .074, p = .941). A variability is observed for both target and non target trials: from 10% to 87% of correct
Fig. 2. Results obtained by LIA for HASR2, triangle: SVM result (FA=13%, FR=12%), diamond: humans result (FA=29%, FR=12%), squares: others participants

answers are given for target trials (m = 53%, sigma = 20%), from 3% to 90% for non-target trials (m = 54%, sigma = 21%). The answers for 45 trials out of 102 are significantly different from chance, as checked by binomial tests. However, for 19 of them (10 target trials, 9 non-target trials) the answers of the listeners are incorrect decisions at a rate greater than chance. It is worth noting that listeners show various behavior. The FA+FR ratios fluctuate across listeners from 34% to 56% (20% to 80% for FR, 22% to 73% for FA). The success ratio of listeners is slightly though significantly correlated (r = 0.437, p = 0.016) to the number of years they have been studying English, suggesting that a better knowledge of the speakers’ language helps improving performance. The differences in listeners’ behavior, illustrated by figure 4, are confirmed by the results of Cochran’s Q test (Q(29)=171.219, p<0.0001). Indeed, some listeners show a tendency to give positive answers on most trials, while others give mainly negative answers. Theses behaviors are clearly pointed out with the d-prime ?? that ranges from -0.17 to 0.54. Only 4 listeners perform significantly over chance (d-prime : 0.316031, 0.332524, 0.374449, 0.54355) ; these listeners are specified on figure 4).

Trials built from excerpts pronounced by female speakers are correctly classified in 55% of trials, to be compared with 52% for male speakers. A binomial test indicates a discrimination rate significantly over chance (d-prime : 0.25914, 0.26065, 0.32124, 0.34394) ; the average success rate of experienced listeners on the 18 trials they rated is 39%. They perform only slightly more accurately than inexperienced listeners (33% success on the same data subset, due to an FA rate of 22% vs. 28% for inexperienced listeners). The experienced listeners have taken incorrect decisions on 3 trials and a correct decision on only one trial at a rate greater than chance (binomial test, p < .05).

4.3. Humans vs automatic system

Figure 4 summarizes the comparison of human systems (inexperienced listeners and HASR submission) and automatic system performance on the 102 trials subset, separately for target and non-target trials. Human is considered successful on a given trial when the majority voting on the inexperienced listeners decisions and the HASR submission is successful on this trial. 22% of target trials (14% of non-target trials) are correctly detected by human only, while 14% (25% of non-target trials) are not correctly detected by any system.

As illustrated by figure 5, the relationship between automatic system scores and positive detections rated by inexperienced listeners is weak. Indeed, the correlation coefficient is only r = .208 (p = .036).

5. DISCUSSION AND CONCLUSION

The perceptual task is particularly difficult. We could highlight two main reasons in order to explain this fact. First, the trials selected for inclusion in the HASR task were chosen as the most error-prone for speaker verification systems. Second, the selected listeners are unfamiliar with the speakers, and non-native English speakers. Moreover, the channel differences further complicate the perception tasks when compared to other protocols from the literature. Since the information about recording conditions was limited to a label ‘telephone call’ or ‘interview’, the role of this variability in the success rates could not be assessed. Therefore, it is not surprising that
the performance obtained in this study is far below the best results obtained in the literature.

It is worth noting that the variation of human performance across trials is very large: from 3% to 90% success rate for non-target trials and from 10% to 87% for target trials. Automatic system and humans gave consistent results for about half of the sets. For some trials, a large majority of listeners correctly discriminated the voices while the speaker verification system’s answers were wrong. The success rates of both the experienced group and the inexperienced one were close to each other in the common set of 18 trials. Although the number of trials do not make possible a statistical evaluation of the difference, this result suggests that with the conditions retained for the perception tasks, experienced listeners may not be better than inexperienced ones.

A large part of trials incorrectly detected by the automated system but correctly detected by human systems were given a score close to the decision threshold by the automatic system. Therefore, a rule-based score fusion scheme, in a hybrid system, may be proposed by defining a range of scores centered on the decision threshold, for which human decision would be required to improve accuracy. On the 102 trials evaluated by inexperienced listeners, applying this rule to the 10% scores closest to the decision threshold (8 non-target trials, 3 target trials) leads to a performance improvement on target trials detection. Indeed, the 2 target trials in this range that were incorrectly detected by the automatic system are successfully detected when using such a hybrid system, without impacting the false acceptation rate. Although this performance improvement remains limited, the hybrid systems might be improved by incorporating native listeners in the panel.

### 6. REFERENCES


