A Human Benchmark for the NIST Language Recognition Evaluation 2005

David A. van Leeuwen[1], *Michaël de Boer*[1,2], *and Rosemary Orr*[2]

[1]TNO Human Factors, Soesterberg, the Netherlands,
[2]University College Utrecht, Utrecht, The Netherlands,
david.vanleeuwen@tno.nl, michaeldeboer@gmail.com, r.orr@uu.nl

Abstract

In this paper we describe a human benchmark experiment for language recognition. We used the same task, data and evaluation measure as in the NIST Language Recognition Evaluation (LRE) 2005. For the primary condition of interest all 10-second trials were used in the experiment. The experiment was conducted by 38 subjects, who each processed part of the trials. For the seven-language closed set condition the human subjects obtained an average C_{DET} of 23.1 %. This result can be compared to machine results of the 2005 submission, for instance that of Brno University of Technology, whose system scored 7.15 % at this task. A detailed statistical analysis is given of the human benchmark results. We argue that the result can best be expressed as the performance of 'naïve subjects.'

1. Introduction

For the evaluation of any system, it is clearly important to have some point of reference for evaluation. For spoken language recognition systems, where a system is required to recognize a language from spoken input, the natural point of reference is human performance for an equivalent recognition task, a human benchmark. The importance of establishing such benchmarks has been emphasized by numerous authors [1, 2, 3], and in particular for recognition systems by, for example, [4, 5, 6].

Generally, it is assumed that humans perform better than machines in recognition tasks. From this point of view, a human benchmark gives system developers a goal towards which they can work, as well as some idea of how to compare systems with the available best. However, it is important to examine this assumption carefully, and to set the criteria and conditions for such a test very clearly.

In terms of human superiority over machines, the evidence from automatic speech recognition (ASR) indeed indicates that machine performance lags behind that of humans, although the gap is closing, if we compare the data from 1997 [5] to more recent data [7]. For language identification tasks, humans have been reported as outperforming machines [4], but there are indications that the machines may be able to outperform humans on short 1–2 second stretches of speech [8]. It seems reasonable to assume that a human will perform better on recognition tasks for languages that they actually speak. A human subject can use not only the acoustic-phonetic models for the task, but also lexical, grammatical and contextual knowledge. However, for recognition tasks involving languages to which a subject has possibly been exposed (though not necessarily) but which they do not speak, this extra knowledge is not available. The same is true of stretches of speech that are too short to include this extra information. It may be that the human subject, for such conditions, is on a level playing field with the machine and it is then foreseeable that the machine performance might come out on top.

Addressing the matter of test criteria, it would seem crucial to set the conditions such that they are as close to those used for evaluation as possible. Furthermore, the evaluation measure should also be same for humans and machines. In his review of human-machine comparisons for recognition systems [5], Lippmann points out that there are a number of factors in previous comparisons of recognition systems which make it difficult to assess these comparisons. For example, the material used to evaluate human performance has not always been the same as that for machine evaluation. Human subjects have been able to compensate for fatigue or inattention by evaluation in groups where majority decisions are reported, as well as or instead of individual decisions. In an excellent research [6], Schmidt-Nielsen and Crystal have applied NIST evaluation rules and data to determine human speaker recognition performance. For language identification tasks, e.g., [8], there has been some effort to equalize the conditions for humans and machines, including keeping the length and quality of the test data similar.

One insurmountable problem is equalizing the experimental condition is that of training time and training data. Estimates have been made for the amount of training data that a human has at various stages of development [9], and the conclusion from that study is that it is not reasonable to require similar training data for humans and machines, at least for the task of *speech* recognition. For both speech and language recognition, more training data is not a feasible solution. Reasoning the other way around, and for the task of *language* recognition, it is not reasonable, for multiple languages, to expect human subjects to submit themselves to the many hours (typically 60 per language in CallFriend [10]) to which a machine is submitted.

In this paper we report the results of a pilot experiment for setting a human benchmark for language recognition. As a basis for comparison, we use the NIST LRE-2005 data [11] to evaluate the performance of human listeners. We have attempted to set up an experiment where the task for the human is as close as possible to that of the machine, taking the above considerations into account. The goal was to make an analysis of the human performance much in the same way as is carried out for machines, such that a direct comparison of performance is possible. Stimuli, decision task and evaluation measure were the same for the human subjects as for the systems evaluated in the NIST LRE-2005 evaluations. The training data was not something over which we had complete control, but for languages which the subjects did not speak and to which they were scarcely or not at all exposed, the training data could be considered similar, although training time was not.

2. Experimental design

The NIST LRE-2005 consisted of three separate tests differing in the nominal duration of the test segments: 3, 10 and 30 seconds. For each duration, 3662 test segments were given. Each segment contained speech in one of the seven given *target languages* English, Hindi, Japanese, Korean, Mandarin, Spanish, and Tamil, or in another non-target language. For each test, two evaluation conditions were defined, an 'open set' and 'closed set' condition. In the closed set conditions, evaluation was limited to include only trials of the seven target languages. Further, a primary condition of interest was defined, where only the data collected by the Oregon Health & Science University (OHSU) were included, accounting a subset of 2505 trials for each duration.

Because processing of trials requires quite some effort on behalf of the subjects, we decided to keep the size of this first experiment limited. We chose to use only the 10 second segments, and only the OHSU primary condition trials. We argued that 10 seconds is a nice balance between having very little linguistic information with 3 seconds and having 30 second trials that are too long to process as a single entity.[1] We further limited the task to the closed-set detection task, which reduced the number of test segments to 2421. Still, merely listening to all trials once would take already over 20 hours. Therefore, we adopted a design where we distributed all trials over many subjects, thus obtaining a human performance figure that is an average over a population of subjects.

2.1. Experimental set up

The available test segments in each of the seven languages l were separated by language, and distributed over $N_s = 38$ subjects. For each subject s, the experiment was divided in seven blocks b. Within a block, a particular language l_{bs} was designated the target detection language. With N_{ls} being the number of test segments in language l assigned to subject s, another N_{ls} segments were randomly chosen from the other languages. These $2N_{ls}$ segments were presented in random order within block b to the subject. The task for the subject was to decide whether or not the test segment was spoken in language l or not. No direct feedback was given about the correctness of the decision. Subjects had been instructed that about half the trials would be in the target language, but number of trials per language was not indicated. Between blocks, the test subjects could take a short break.

In order to facilitate in 'training' the subjects for the several languages, we designed the experimental interface as follows. Subjects were presented information on a computer screen. The top half contained 'training sample buttons.' Theses were seven buttons labeled by any of the target languages. When a button was pressed, a new random 10 second sample from the CallFriend training data partition in that language was played. For a trial decision, test subjects were allowed to play the test segment and any of the training samples as often as they needed for making a good decision. Pressing any of the language/test buttons would stop a possible current playback, and start playing the requested sample. Stop buttons were also provided. The 10-second excerpts were generated from single conversation side speech from the CallFriend database. Silence was removed using an energy-based speech activity detector while retaining about half a second of silence around detected speech, to keep some naturalness. After this, the speech was partitioned in 10 second excerpts. Using the training partition of CallFriend, we obtained a total of 1830 training segments for the seven target languages.

Before the actual experiment, each subject was requested to give information about his/her native language(s), and what the degree of exposure to any of the target languages was on a 6-point scale, ranging from none–very little–little–medium–lots–very much. Further questions included the subject's sex, age, and musical instruments the subject plays. After the collection of trials, some additional 'debriefing' questions addressing the subject's experience of the experiment were asked, from which we hope to learn to improve the design of future experiments.

The seven target language blocks were balanced over the subjects by using a 7×7 Latin square design for each group of 7 subjects. Thus, we compensate for order effects due to learning or fatigue. Subjects were familiarized with the testing procedure and the quality of the speech recordings using a few trials in English.

Apart from the decisions made, we recorded for each subject the sequence of segments played, for both training and test segments, and a time stamp of the decision.

2.2. Recruitment of subjects

The subjects were recruited from students and working staff at University College Utrecht, The Netherlands, where the experiments also took place. The College provides an international liberal arts education, with about 60 % of the students being from the Netherlands and 40 % from the rest of the world. The language of communication is English, and hence the experience and exposure to English is high for all subjects. Subjects were paid for participating in the experiments, either in monetary (students) or gastronomic units (staff). Completion of the experiment varied between 0.5–1.5 hours.

2.3. Comparison to the NIST LRE-2005 task

As indicated above, we tried to set up the experiment such that the task for humans closely resembles that of machines in the NIST LRE context. In this section we will look in more detail at several aspects.

2.3.1. Training material

The training that humans and machines are exposed to, may be the component that is hardest to equalize. While for machines, we have precise control over what speech material is used for modeling, for humans this is not possible. Obviously, for the native language of the test subject, the amount of speech that he/she has been exposed to is much more than the 60 hours available in CallFriend. Moore [9] estimates that the amount of speech to which a 20-year old person is exposed is of the order of 30 000 hours. On the other hand, the exposure to languages other than the native language or English will generally be small for our population[2], and perhaps virtually absent for some languages.

We tried to somehow correct for this large imbalance, both between human and machine and between native and unexposed language, by allowing 'online' training of the different languages. In this way, the detection of a language becomes more like a task of comparison of speech segments. We realize that the amount of speech material available to the subject during the experiments may not come close to the amount necessary for humans to 'build models' [4], and therefore this human

[1] In [6], for the task of speaker recognition, potential boredom and total experiment duration were an argument to choose 3 s segments.

[2] Note, however, that it is estimated that the majority of people in the world's population is bi-lingual or multi-lingual.

benchmark experiment can best be described as measuring the performance of 'naïve subjects.'

2.3.2. Detection task

As in the NIST LRE, we have formulated the task primarily as a *detection task*. The main reason for interpreting the NIST task as a detection task is the way the performance is measured, in terms of *detection costs*. In fact, the closed-set task really is a discrimination task, because information about the set of non-target languages is allowed to be used by machines. Similarly, in the human experiment the subjects are aware of the possible alternatives of the target language, and can listen to examples of these at any time before a decision is made.

Another concept that is important to the evaluation measure, is the (synthetic) *prior* of the target language. In NIST LRE this is set to $p = \frac{1}{2}$. We have simulated this prior in two ways. First, we told the subjects that the probability of a test segment being spoken in the target language was $p = \frac{1}{2}$, and second, we chose the evaluation priors of target vs non-target trials to be $1 : 1$. Thus, in the human experiment, we made the evaluation prior equal to the synthetic prior, because we believe that it is very difficult for test subjects to separate the two priors. This is in contrast to the machine, for which the synthetic priors that govern minimum-cost decisions can be chosen independently of the evaluation priors, which were not homogeneous in the case of LRE-2005, resulting in different amounts of test segments per language.

2.3.3. Evaluation Measure

The primary evaluation measure of NIST LRE is the cost of detection, C_{DET}, which since 2005 has the rather complicated definition, here reproduced from [12]

$$C_{\mathrm{DET}} = \frac{1}{N}\sum_{i=1}^{N} C^i_{\mathrm{DET}}, \tag{1}$$

where C^i_{DET} is the detection cost for the subset of trials for which the target language is i

$$C^i_{\mathrm{DET}} = C_{\mathrm{miss}} P^i_{\mathrm{miss}} P_{\mathrm{target}} + C_{\mathrm{FA}} \sum_{j\neq i} P^j_{\mathrm{non}} P^{ij}_{\mathrm{FA}}. \tag{2}$$

Here N is the number of target languages (seven), C_{miss} and C_{FA} normalized cost parameters (set to unity in the evaluation), and P_{target} the prior probability for target language i that must be considered in the decision (set to $\frac{1}{2}$ in the evaluation). Finally P^j_{non} is the prior probability that the test segment is in non-target language j (set by NIST to $(1 - P_{\mathrm{target}})/(M-1)$, where M is the number of test languages, in the primary task $P^j_{\mathrm{non}} = \frac{1}{12}$). The error probabilities P^i_{miss} and P^{ij}_{FA} are determined by the evaluation results, where P^i_{miss} is the proportion of true trials in language i where the system's decision was 'false,' and P^{ij}_{FA} is the proportion of trials with target language i and test segment language j where the decision was 'true.'

In NIST LRE, every test segment is used as target trial for one language and non-target trials for all other languages. In our human benchmark experiment, this is not the case. A test segment is used once as target and—on the average—only once as a non-target trial. Sometimes we calculate C_{DET} over a very small subset of trials, e.g., for a single test subject. In that case P^{ij}_{FA} may not be defined for all test languages j. Then, the prior P^j_{non} is adapted accordingly to be non zero only for the $M' < M$ test languages occurring for target i, $P^j_{\mathrm{non}} = 1/2M'$.

Table 1: Main results of the human benchmark for NIST LRE-2005, analyzed for the different target languages separately.

Language	C_{DET}(%)	P_{FA}(%)	P_{miss}(%)
English	3.63	2.00	5.26
Hindi	34.1	23.5	44.8
Japanese	29.4	18.5	40.3
Korean	31.2	20.4	42.0
Mandarin	25.1	15.6	34.5
Spanish	9.77	8.72	10.8
Tamil	28.6	23.4	33.9
Mean	23.1	16.0	30.2

2.4. Implementation

The experimental protocol was implemented in a Java Virtual Machine running in PC hardware. We used a Sennheiser PC 131 headset in order to play test and training segments to the subjects. The experiments took place in a laboratory room where multiple subjects could run the experiments simultaneously but independently. Attention levels were stimulated by providing enough liquid (in the form of water) and sugar (in the form of sweets). The experimenter was available to the subjects for questions regarding the procedure at all times.

3. Results

The main result of this research is that, for the 38 subjects recruited for this experiment, the average C_{DET} as defined in (1) is 23.1 %. This is for the closed set task using the OHSU subset of 10 second trials of the official NIST LRE-2005 evaluation. In Table 1 we have summarized the human benchmark results per language, and separated the false alarm and miss contributions to C_{DET}.

We can compare this result to the results of machines reported for NIST LRE-2005. Unfortunately, even though C_{DET} is NIST's primary evaluation measure, numerical figures have not been reported in open literature [13]. Our own system submission [12], under the name TNO-SDV, achieved $C_{\mathrm{DET}} = 15.6\,\%$ with this task. A more advanced system than our own was that of Brno University of Technology (BUT) [14], who scored $C_{\mathrm{DET}} = 7.15\,\%$. We will use results of these two systems for further analysis.

3.1. Detailed human-system comparison

Looking at our own submitted system results, we can make a trial-by-trial comparison of the errors made by humans and machine. This allows us to use the McNemar statistical test to test whether the difference between humans and our system is significant. The McNemar test is also applied in speech recognition [15] and speaker recognition [16]. The McNemar test in this case tests whether number of unique errors the humans make is significantly different from the number of unique errors the machine makes. Here, the contingency table of trials looks like

	Machine (TNO-SDV)	
humans	Correct	Error
Correct	3361	472
Error	781	233

. (3)

With these numbers, McNemar's χ^2 statistic is 55.2 at 1 degree of freedom, which make the difference clearly significant.

Observing Table 1, we see that the performance of English and Spanish detection is significantly better than that of the

Table 2: Per language comparison of C_{DET} for humans and two systems. McNemar statistics (last two columns) are given only for the TNO-SDV system.

	C_{DET}(%)			Unique errors	
	human	machine		human	machine
Language		TNO	BUT		TNO
English	3.63	17.8	8.96	22	153
Hindi	34.1	20.5	9.79	67	29
Japanese	29.4	14.5	5.70	173	56
Korean	31.2	16.4	8.13	166	60
Mandarin	25.1	11.1	4.45	268	84
Spanish	9.77	14.7	5.64	24	59
Tamil	28.6	14.5	7.38	72	31
Mean	23.1	15.6	7.15	781	472

other languages. Clearly this must be an effect of familiarity of the subjects with these languages. If we do a comparison of human vs machine per target language, the general picture changes. In Table 2 we compare the per-language statistics. The number of unique errors for either humans or machine (TNO-SDV) determine McNemar's χ^2, in all cases shown $p \ll 10^{-3}$. We see that indeed, for the languages English and Spanish, humans perform better than the TNO-SDV system. The BUT system, on the other hand, is only outperformed by humans for English. For Spanish, the McNemar test between humans and machine shows the closest match (30 vs 15 errors), resulting in a $p = 0.037$.

It may be interesting to note that the English language detection in LRE-2005 was considered a hard task because of the occurrance of Indian Enlish speakers among the English trials, while only a little amount of training data for this Enlish accent was distributed. Systems typically showed a large difference in detection capability of the American and Indian accented English trials. We calculated the detection scores for our human benchmark experiment, and found a similar dichotomy of 2.0 % vs 6.7 % for American and Indian accented English, respectively.

3.2. Learning effects

The experimental design allows us to analyze whether there is an effect of the order in which the languages are tested. We can accumulate C_{DET} statistics over the block number b. Averaging over a particular block number involves all subjects and all target languages in a balanced way. In Figure 1 we plotted the development of average C_{DET} over the course of time of the experiment. Visually, there does not appear to be a particular trend of learning (decreasing C_{DET}) or fatigue (increasing C_{DET}). Since the block numbers within the experiment encodes the progression of time more or less linearly, we might test if a linear fit of C_{DET} vs b gives a slope significantly different from 0. Here we need to know the distribution of C_{DET}. This may not be trivial, since C_{DET} is a weighted combination of binomial statistics P^i_{miss} and P^{ij}_{FA}. Assuming this combination makes C_{DET} Gaussian, a normal analysis reveals that the statistics shown in Figure 1 does not indicate a slope significantly different from 0, with t-test statistic $t = -1.42$, and probability $p = 0.214$ that the possible linear effect is chance. Alternatively, we may model C_{DET} having pure binomial distribution, and analyze the dependency on b using the probit link function [17, 16]. This leads to the statistic $z = -1.24$ and probability $p = 0.214$. Note, that although these analyses only say that we didn't *find* an effect, it does not mean that there *is*

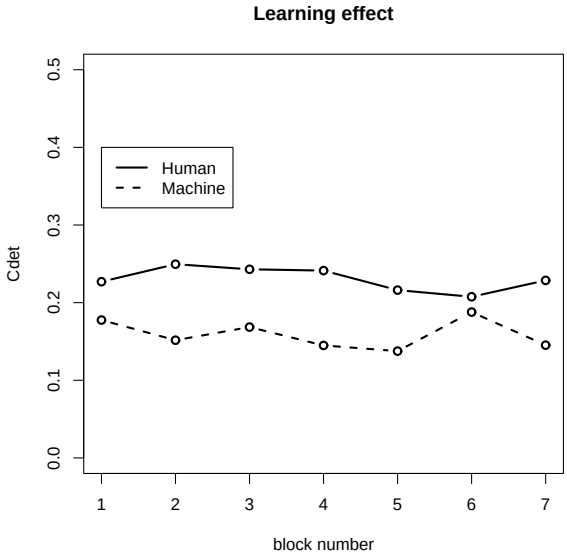

Figure 1: Effect of position in time within the experiment on C_{DET}. The line for the machine should not show a time effect, since all trials are processed independently.

not such an effect. As a comparison we have plotted the C_{DET} values for the TNO-SDV system calculated over the *same* trial subsets in the same figure. This may give an indication of the inherent variability of the trials.

3.3. Test subject variability

So far, we have modeled the human performance as a single 'system' having uniform C_{DET} per language. However, we may assume that individual test subject performance varies widely due to different amounts of exposure to the languages, different linguistic talents and skills, different experimental efforts, etc., despite the fact that the subjects are recruited from a relatively homogeneous population in terms of background and education. The experimental design does not allow to test for this rigidly, because it is a between-subject design.

In Figure 2 we show a histogram of the per-subject C_{DET} for humans (top). For comparison we have included the results over the same trial sets per subject, but then using the TNO-SDV system decisions for each trial set, in the bottom histogram. Except for three outliers, it appears that the variability in C_{DET} by humans does not exceed that measured by the machine. In order to study this between-subject variability better, a more complete design (within-subjects) and more trials per subjects are required.

3.4. Effect of exposure to the target language

One of the questions each test subject answered, was to rate the exposure he/she has had to each target language. Answers were on a 6-point scale, ranging from none to very much. In Table 3 the statistics of these are shown. Clearly, English stands out in having a consistent highest level of exposure. We have also computed per-language C_{DET} aggregated over different exposure levels the subjects indicated, and included these values in the table. Obviously, some C_{DET} values are obtained from very few trials as some cells have very low occupancy numbers, so care must be taken in interpreting the numbers.

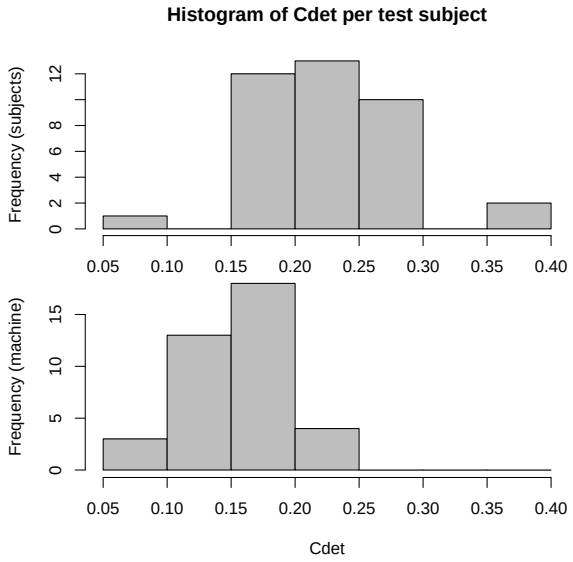

Figure 2: Histogram of per-subject C_{DET} (top). On average, 131 trials are used to determine the per-subject C_{DET}. The lower histogram shows the variability in C_{DET} measured using machine (TNO-SDV) decisions over the *same* trial subsets used in the per-subject analysis.

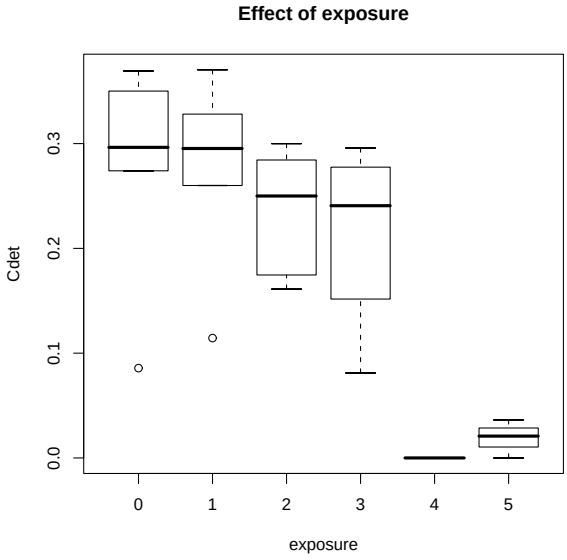

Figure 3: Distribution of per-language C_{DET} as a function of exposure level, as indicated by the test subjects before the experiment.

Table 3: Distribution of exposure to target language over the different subjects. Non-italicized values are the number of subjects that indicated exposure level (column) for language (row), the italic figures below those are C_{DET} values calculated over that subset. The last row is C_{DET} agregated over the exposure level.

	Exposure level					
Language	0	1	2	3	4	5
English						38
						3.63
Hindi	25	11	2			
	35.0	*32.8*	*25.0*			
Japanese	25	12		1		
	31.0	*26.0*		*22.2*		
Korean	5	25	4	4		
	36.9	*30.9*	*28.4*	*29.6*		
Mandarin	8	19	6	3		2
	28.3	*28.2*	*17.5*	*25.9*		*2.1*
Spanish	2	8	10	13	1	4
	8.57	*11.4*	*16.1*	*8.11*	*0*	*0*
Tamil	34	3	1			
	27.4	*37.5*	*30.0*			
mean C_{DET}	*27.9*	*27.7*	*23.4*	*21.6*	*0.0*	*1.9*

For most languages, there appears to be a trend of lower C_{DET} with higher exposure level. This may be better appreciated from Figure 3, where we show the range of these C_{DET} values per exposure level in a box-plot. In Table 3 we also indicated C_{DET} according to (1) averaged over languages with contributions in a particular exposure level.

3.5. Calibration

One of the issues for machines is calibration. It is an art to set the threshold for detection well. In speaker recognition, this is traditionally measured in terms of the difference in C_{DET} and $C_{\mathrm{DET}}^{\mathrm{min}}$, the minimum attainable C_{DET}. For closed-set language detection, the definition of calibration is a lot harder [18]. The ratio of false alarms and misses, given equal priors and costs is indicative of calibration. Returning to the human benchmark experiment, it is interesting to note that there is a consistent imbalance in P_{FA} and P_{miss} (see Table 1). This is despite the fact that evaluation priors were a $1:1$ ratio of target and non-target trials, and an explicit mention of the prior probability of a trial being spoken in the target language being $\frac{1}{2}$. Aggregating decisions over all target and non-target trials, thereby integrating over subjects and target languages, we have $P_{\mathrm{FA}} = 13.6\,\%$ and $P_{\mathrm{miss}} = 28.2\,\%$. With over 2400 trials per category, a statistical test of proportions indicates this difference is significant ($\chi^2 = 154$).

A test subject's internal threshold may be different from others. In psychophysical experiments it is customary to convert individual $(P_{\mathrm{FA}}, P_{\mathrm{miss}})$ coordinates to a threshold-independent parameter d'. In signal-detection theory [19] one assumes that the probability-density functions for target and non-target samples are Gaussian and have equal variance. The detection capability is then completely determined by the difference in means w.r.t. the common variance. This ratio is called d', and relates to a $(P_{\mathrm{FA}}, P_{\mathrm{miss}})$ point as

$$d' = -\Phi^{-1}(P_{\mathrm{FA}}) - \Phi^{-1}(P_{\mathrm{miss}}), \tag{4}$$

where Φ^{-1} is the inverse of the cumulative normal distribution function, also know as the *probit* function. We can calculate d' for each subject separately, thereby effectively equalizing their error rates. We find an average $d' = 1.64$, with a standard deviation of 0.48. This can then be converted back to error rates using the inverse of (4), $p_e = \Phi(d'/2)$, the probability of error, or 'equal error rate.' We thus arrive at a value of $p_e = 20.6\,\%$, which can be compared to the original $C_{\mathrm{DET}} = 23.1\,\%$ thanks to the unit costs and the balanced evaluation priors.[3] This 'im-

[3] A slightly better comparison value is to the average error $\frac{1}{2}(P_{\mathrm{FA}} +$

proved' version of the error rate indicates the expected performance if the human threshold would be uniform and well chosen.

3.6. Debriefing remarks

In this section we itemize some of the remarks that test subjects made during the debriefing questionnaire. One of the purposes of this experiment is to learn from the subjects's experience and adapt the design in subsequent experiments that we plan to conduct.

- Some people found the experiment boring, whilst other found it quite interesting to do.
- There many complaints that some of the training samples for Hindi and Tamil actually were spoken in English.
- This lead to the question whether Indian accented English should be interpreted as English or Hindi/Tamil.
- The quality of the test trial samples was noted as being very low; sometimes only a hum or laughter was audible. People found it hard to listen to telephone-quality audio.
- The task was considered very hard in general.

4. Discussion and Conclusions

The experiment presented in this paper was designed to give an estimate of the human capability of language recognition that can numerically be compared to machine evaluations. To this end, we use the same test data, task and evaluation measure as was used in the NIST Language Recognition Evaluation in 2005. As a first experiment, we have determined the average performance of a population of humans. The main result is that the average C_{DET} for humans is 23.1 %, which is significantly higher than a number of machines that participated in LRE 2005.

A detailed breakdown reveals that the human detection errors for English and Spanish were lowest. The C_{DET} for English as target language (3.6 %) is lower than that of the best machine result we could find, that of BUT (9.0 %).[4] The high variability of human performance for the target languages (cf. Table 1) is interesting, and requires some investigation. One apparent reason for such variability might be the level of exposure to the target language, following the very clear effect of high exposure levels in Figure 3. However, the highest exposure level category is dominated by English (cf. Table 3), and there might be other factors that explain the high variability of performance for the different target languages. We suspect that one of these is the difference between *exposure* to a language and *ability* to speak the language. The type of knowledge available to a *speaker* of a language includes syntactic, lexical, morphological, semantic and sometimes even cultural information, as well as the acoustic-phonetic information which is available to listening *non-speakers*. We know all subjects in this experiment had knowledge of English. Many of them had knowledge of Spanish, having taken this language for the language requirement at the College. While they may have noted their exposure as being only medium, that does not reflect the fact that they have followed a course in this language for at least one semester at university level. In order to investigate the difference between exposure and ability as a possible reason for performance variability, it would be necessary to explicitly separate subjects. One could envisage two groups, one with only acoustic exposure and some limited lexical knowledge, for example people who spend holidays in a particular linguistic environment, and those with some language ability. The line between exposure and ability may be difficult to draw, but it seems clear that exposure should be defined such that the information gained from it is not superior to that available to the machine.

Contrary to what was found in [4], we did not find a learning effect. There are several reasons for this. First, we did not give feedback about trial decisions, and second, the size of the blocks was quite small. Although our experimental design was such that possible learning effects would not confound per-language results, the fact that we did not observe a learning effect over the whole experiment indicates that this is really a 'snap shot' measurement of this population's language recognition abilities. We might conclude that our results best describe that of 'naïve' people—subjects that have hardly any time to learn the acoustic-phonetic characteristics of a new language.

We observed that there was an imbalance in misses and false alarms, typically $P_{\mathrm{miss}} > P_{\mathrm{FA}}$, despite efforts to convince subjects to balance their answers. Something which might explain this is the fact that, for non-targets, there were six possible languages versus only one target language, and maybe, unconsciously, humans try to equalize the prior probabilities to some extent. It is also possible that the subjects were not too aware of the fact that the target priors were $\frac{1}{2}$ within each experimental bock. The experimental design did not allow for subjects to balance answers in retrospect, and the varying and unknown number of trials per block made on-line adjustment of decision thresholds very hard. In a way, this supports the idea that humans implicitly adhere to the paradigm of independent trial decisions—which is quite essential to machine evaluation.

In [6], for the task of speaker recognition, the same priors of $\frac{1}{2}$ were used for similar reasons, but in that experiment apparently no striking imbalance between misses and false alarms was observed. Further, they could obtain better human performance by simulating a group decision, because the same trials were judged by a panel of typically 16 subjects. The group decision methods were not based on individual decisions, but on 10-point scale confidence ratings. Our experimental set-up did not include more than one subject judging the same trial, and we chose not to collect confidence ratings. For future research, it might be interesting to include these in the experimental design. Still, however, we did find a way to compensate for variability in each subject's threshold in Section 3.5.

From the remarks made by the subjects at the end of the experiment, it emerged that they found the comparison method difficult. In part, this is because of the time that would be required to do a truly thorough comparison, listening to samples of each language until certain of the decision about the test language. We would note that, even with the relatively small amount of trials, as compared to other work [4] some subjects spent as much as one and a half hours on the test. It seems likely that better results would be obtained for human performance if feedback were provided after each decision. Providing feedback would introduce a learning effect, and the analysis of the results would then have to focus on the final blocks of testing. More blocks would have to be introduced. However, continuous feedback would also have the advantage of eventually speeding up the decision making process. Since the test is already experienced as being difficult, prolonging it should be carefully

$P_{\mathrm{miss}}) = 22.8\,\%$, because the averaging over test subjects weights languages in their evaluation proportions, rather than equally like C_{DET} does.

[4] Note that for the machines, the hardest English trials were Indian English, for which little training data was available.

managed. Subjects could sit the experiment in several sessions, as in [4]. Where there is a learning effect, this is likely to be experienced as a reward for effort on behalf of the subject, and may go some way towards alleviating the perceived difficulty of the task. The subjects were only asked to rate the difficulty of the task, and not to explain what was difficult about it. Informal questions reveal that the difficulty is for languages where the subject has no *ability*, but possibly some exposure.

We claim that we have succeeded in conducting a human benchmark of language recognition which is very close to the way machines are evaluated. The most important area where we could not equalize human and machine evaluation is in the control of the training material. We have chosen a design where during trial decision a virtually unlimited amount of training trials in any of the languages under evaluation could be consulted, making the task one of comparison of speech segments. This is not the same as learning to know a language or, as a machine would do, building a 'model' of the language using of the order of 60 hours of speech. One could raise the question whether it is 'fair' to compare the human benchmark results to the machine. Perhaps a thorough instruction of the subjects in the characteristic differences in the language by linguists—in a course[5] of, say, 60 hours—can be considered a better comparison in training condition. An alternative would be to *lower* the amount of training time for the machine.

There are still many aspects, apart from training condition, that we have not addressed in a systematic way, yet. Among these are the trial duration, which seems to have practical experimental limitations, and between-subject variability, which will not only lead to better insights, but also allow for performing simulated group decisions. We hope to address these and other questions in future research.

5. Acknowledgements

This work was supported in part by the European Union 6th FWP project AMIDA, 033812. The authors would like to thank Lukáš Burget from Brno University of Technology for sharing the BUT NIST LRE-2005 results with us.

6. References

[1] Harold F. O'Neil Jr., Yujing Ni, Eva L. Baker, and Merlin C. Wittrock, "Assessing problem solving in expert systems using human benchmarking," *Computers in Human Behaviour*, vol. 18, no. 6, pp. 745–759, November 2002.

[2] Frances A. Butler, "Benchmarking text understanding systems to human performance: An exploration," Tech. Rep., Centre for Technology Asessment, Graduate School of Education, University of California, Los Angeles, 1990.

[3] Eva L. Baker and Elaine L. Lindheim, "A contrast between computer and human language understanding," Tech. Rep., UCLA Centre for Technology Assessment, 1988.

[4] Y. K. Muthusamy, E. Barnard, and R. Cole, "Perceptual bechmarks for automatic language identification," in *Proceedings of International Conference Spoken Language Processing*, Adelaide, 1994, vol. 1, pp. 333–336.

[5] Richard P. Lippmann, "Speech recognition by machines and humans," *Speech Communicatrion*, vol. 22, pp. 1–15, 1997.

[6] 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.

[7] David Pallett, "A look at NIST's benchmark ASR tests: Past, present, and future," `http://www.nist.gov/speech/history/`, 2003.

[8] Martine Adda-Decker, Fabien Antoine, Philippe Boula de Mareuil, Ioana Vasilescu, Lori Lamel, Jacqueline Vaissiere, Edouard Geoffrois, and Jean-Sylvain Linard, "Phonetic knowledge, phonotactics and perceptual validation for automatic language identification," in *Proc. ICPhS*, Barcelona, 2003, pp. 747–750.

[9] Roger K. Moore, "A comparison of the data requirements of automatic speech recognition systems and human listeners," in *Proc. Eurospeech*, Geneva, 2003, pp. 2581–2584.

[10] Linguistic Data Consortium, "Callfriend corpus," `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC96S46`, 1996.

[11] "The 2005 NIST language recognition evaluation plan," `http://www.nist.gov/speech/tests/lang/2005/`, 2005.

[12] David A. van Leeuwen and Niko Brümmer, "Channel-dependent GMM and multi-class logistic regression models for language recognition," in *Proc. Odyssey 2006 Speaker and Language recognition workshop*, 2006.

[13] Alvin F. Martin and Audery N. Le, "Current state of language recognition: NIST 2005 evaluation results," in *Proc. Odyssey 2006 Speaker and Language Recognition Workshop*, San Juan, June 2006.

[14] Pavel Matějka, Lukáš Burget, Petr Schwarz, and Jan Černocký, "Brno University of Technology system for NIST 2005 language recognition evaluation," in *Proceedings of Odyssey 2006: The Speaker and Language Recognition Workshop*, 2006, pp. 57–64.

[15] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *IEEE Proc. ICASSP*, Glasgow, 1989, pp. 532–535.

[16] David A. van Leeuwen, Alvin F. Martin, Mark A. Przybocki, and Jos S. Bouten, "NIST and TNO-NFI evaluations of automatic speaker recognition," *Computer Speech and Language*, vol. 20, pp. 128–158, 2006.

[17] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, Springer, New York, fourth edition, 2002, ISBN 0-387-95457-0.

[18] Niko Brümmer and David A. van Leeuwen, "On calibration of language recognition scores," in *Proc. Odyssey 2006 Speaker and Language recognition workshop*, San Juan, June 2006.

[19] J. A. Swets, *Signal detection and recognition by human observers; contemporary readings*, Wiley, New York, 1964.

[5] The course might also include some instruction about detection priors.