INTERSPEECH 2009 BRIGHTON

A Human Benchmark for Language Recognition

Rosemary Orr[1,2] *and David A. van Leeuwen*[1,3]

[1]International Computer Science Institute, Berkeley, CA
[2]University College Utrecht, The Netherlands
[3]TNO Human Factors, Soesterberg, The Netherlands

Abstract

In this study, we explore a human benchmark in language recognition, for the purpose of comparing human performance to machine performance in the context of the NIST LRE 2007. Humans are categorised in terms of language proficiency, and performance is presented per proficiency. The main challenge in this work is the design of a test and application of a performance metric which allows a meaningful comparison of humans and machines. The main result of this work is that where subjects have lexical knowledge of a language, even at a low level, they perform as well as the state of the art in language recognition systems in 2007.

10.21437/Interspeech.2009-620

1. Introduction

1.1. Human Benchmarking

For any automated task, we should also ask ourselves what is the quality of performance of *machines* as compared to that of *humans*. This paper looks at human benchmarking for the task of language recognition, that is, establishing the level of human performance in a machine task. With such a benchmark, we can assess the level to which technology can aspire.

There are many issues in conducting human benchmarking of language recognition systems, such as training, test sample duration, learning affect and fatigue. These issues, as well as a review of other work in the field, can be found in our earlier work [1]. This work reports a pilot study, which was carried out in the Netherlands using NIST LRE 2005 evaluation material in 7 languages. One of the findings was that greater proficiency in a language produced better performance in the language identification task. We suspected that there might be a significant improvement in performance where subjects had lexical knowledge of a language, rather than just some exposure.

Where other researchers [2] found a learning effect, we did not. We believed this to be because of the limited number of trials and the lack of feedback during the experimental sessions.

Finally, we observed that there was an imbalance between misses and false alarms, indicating that subjects tended to decide that trial speech segments were not from the target language more than that they were. This happened, despite efforts to help them to balance their answers by telling them that 50 % of the trials *were* from the target language.

In this study, we address these questions using the 14 languages of NIST LRE 2007 [3] in a human benchmark where we compare human performance to machine performance for a number of systems from the NIST LRE 2007 submissions.

2. Human benchmark experimental design

While subjects in the original study were asked to rate their proficiency in each language, the rating was done on a scale of 1 to 5, and gave no information about the type of exposure to the language that they had. For the follow-up design, subjects were classified as to their proficiency in each language. This was done by self-assessment on behalf of the subjects so that they could be classified in one of 5 intuitive groups, ranging from *no exposure* to a language to *native* proficiency. While we still would not have control over the actual training data, we could have some information about relative training in each of the 14 languages. From the results of the pilot study, and as a result of these changes, we expected to find that lexical knowledge of a language would be an important factor in human performance as well as the experienced degree of difficulty.

The design was also changed such that subjects were tested on only one target language per session, whereas in the pilot study, the languages were distributed across subjects. By changing the design in this way, we intended to make it easier to focus on the task, and so to be able to introduce more trials. Furthermore, we wanted to focus on performance as a function of proficiency, which was made easier by this change. We decided not to give feedback during the test as we did not want a learning effect to interfere with the categorisation of language proficiency that the subjects had provided.

Finally, for the current study, more attention was paid to the subjects' awareness of the information about the prior, to see if it affected their decision-making. Not only were they explicitly given the information, and reminded of it throughout, but we also asked afterwards, for every subject, if and how they used this information.

3. Methods and Materials

3.1. Subjects

Subjects were taken from the student and faculty body both in Berkeley, California and in Utrecht, the Netherlands. There were 108 subjects in total, 81 female and 27 male.

3.2. Language Proficiency

In order to test the prediction of performance in the experimental task, as related to training, we asked subjects to classify their proficiency in each of the 14 languages.

Five proficiency categories were chosen, namely *no exposure, some exposure, fair non-native, fluent non-native* and *native*. Subjects were asked, by means of a questionnaire, to estimate their own ability in each of the 14 languages, using these 5 categories.

The last two categories are relatively self-explanatory. The remaining three warrant some explanation.

The category *fair non-native* was described to subjects as *a language which is not your native language, in which you*

are not fluent, and which you can use to carry out basic social interactions such as meeting people, eating and drinking with people, grocery shopping. Generally, it might be a language in which you have taken a couple of university-level courses.

The category *some exposure* was described to subjects as *a language that you are familiar with but do not speak. This generally applies to languages that are (sometimes) present in your environment but which you do not need for communication purposes. Often, exposure like this derives from the media, regular holiday destinations, or simply your living environment*.

The category *no exposure* was described to subjects as *a language that you don't think you would recognize easily if you heard someone speaking it, and to which you have had no direct exposure*.

The difference between the categories *fair non-native* and *some exposure* was intended to differentiate between presence and absence of lexical knowledge.

The distribution of subjects across language proficiencies and target languages is shown in Table 1 below. For some languages, for example, English, not all proficiencies could be represented because of the language environments in which the experiments were carried out.

Table 1: Table showing the proficiency of subjects for each of the 14 NIST LRE-2007 languages.

	none	some	fair	fluent	native
arabic	2	4	3	0	2
bengali	3	2	1	0	0
chinese	1	1	2	2	8
english	0	0	3	22	18
farsi	2	1	0	0	2
german	3	4	4	2	2
hindustani	3	3	2	2	1
japanese	2	1	5	1	0
korean	3	5	1	2	0
russian	3	3	2	1	2
spanish	3	2	3	2	3
tamil	2	2	1	0	2
thai	5	3	0	0	4
vietnamese	4	2	2	0	2

Some subjects were proficient at different levels in more than one of the 14 languages. Such subjects could participate in multiple experimental sessions, but then in a different language each time. The order of participation depended on the proficiency levels, and such sessions were run from lowest to highest exposure. In this way, subjects did not get extra acoustic exposure to a language, which could affect the proficiency category that they had assigned themselves.

3.3. Task and material

The task for the humans is that of language *detection*, similar to the machine task in the NIST LRE context, with a closed set of non-target languages. For a given target language, the subjects were presented speech excerpts of around 10 seconds duration, and they had to decide whether each excerpt was spoken in the target language or not. Besides making a decision, subjects had to indicate a confidence level for their decision on a scale of 1 to 5, where 1 was *very uncertain* and 5 was *certain*.

Apart from being able to play the test trial speech, sample speech from any of the 14 languages was also available to the subjects for reference or training, for both male and female speakers. Subjects were allowed to play the trial sample and any of the training samples as often as they wished. They did not have to finish a sample before making a decision or before playing another sample, and each time that sample speech was requested, a new speaker was presented. The training segments were drawn from the CallFriend database augmented with LRE-2007 development test data.

A single experimental session consisted of one combination of subject and target language. Within a session, 160 trials were presented, in random order. The speech was taken from the NIST LRE 2007 evaluation, 10 sec condition. All 80 target trials for a language were included in a session, and 80 non-target trials were drawn randomly from the alternative 13 languages. Thus, the evaluation priors were $\frac{1}{2}$. This was explicitly stated in a briefing before every session.

Before the session started, instructions were given on the screen, and a small test with only English and Spanish samples was run for familiarization with the user interface.

4. Analysis and Results

4.1. Performance metric

In order to make human and machine performance results comparable, we use the same metric as in NIST LRE, i.e., $C^i_{\rm det}$, the cost of detecting language i

$$C^i_{\rm det} = C_{\rm miss} P^i_{\rm miss} P_{\rm tar} + C_{\rm FA} \sum_{j \neq i} P^j_{\rm non} P^{ij}_{\rm FA}. \qquad (1)$$

Here $C_{\rm miss}$ and $C_{\rm FA}$ are normalized cost parameters, set by NIST to unity in the evaluation, and $P_{\rm tar}$ the prior probability for target language i that must be considered in the decision. This has been set by NIST to $\frac{1}{2}$ in the evaluation. Finally $P^j_{\rm non}$ is the prior probability that the test segment is in a non-target language j. This has been set by NIST to $(1 - P_{\rm tar})/(M - 1)$, where M is the number of test languages, here 14. The error probabilities $P^i_{\rm miss}$ and $P^{ij}_{\rm FA}$ are determined from the experiment. $P^i_{\rm miss}$ is the proportion of true trials in language i where the subject's decision was 'no,' and $P^{ij}_{\rm FA}$ is the proportion of trials with the target language i and test segment language j where the decision was 'yes.'

4.2. Overall results

We calculated $C_{\rm det}$ for each session, that is, each combination of subject and target language. Since the subjects' proficiency is known for each language, we can plot the distribution of $C_{\rm det}$ over subjects and languages for each proficiency, i.e., over columns in Table 1. This is illustrated in Fig. 1. It is clear that language proficiency has a strong influence on the detection performance for that language. The values of $C_{\rm det}$ averaged over proficiency are tabulated in Table 2.

Table 2: Average $C_{\rm det}$ over sessions with the same language proficiency.

Proficiency	none	some	fair	fluent	native
$C_{\rm det}$ (%)	28.0	15.8	5.44	3.00	2.05

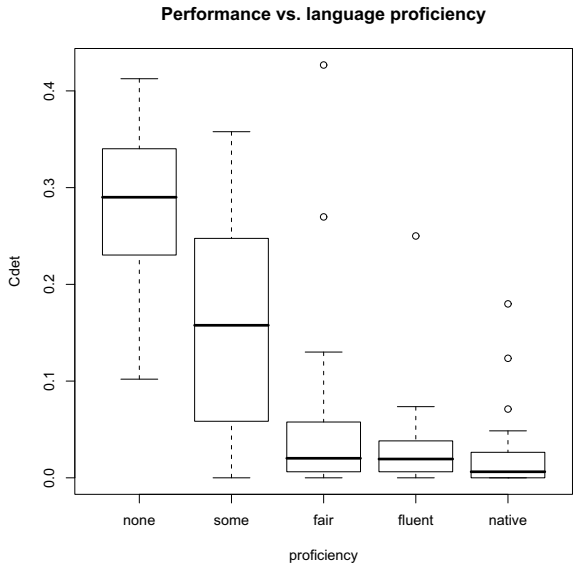

Figure 1: Boxplots for the values of C_{det} per session, pooling target languages and subjects.

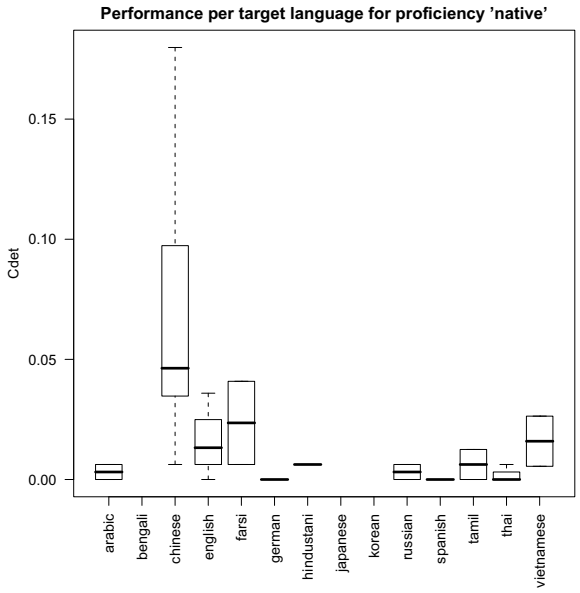

Figure 2: Human performance per target language, for subjects native in the target language. Shown are ranges of C_{det} per subject, using box plots.

4.3. Performance per target language

We can obtain more detailed information by showing the C_{det} per language, for each of the language proficiency levels. In Fig. 2 we have plotted the ranges of C_{det} per language for the native speakers, i.e., from sessions in the last columns of Table 1. Although the number of sessions per language in this condition is small, 0–8 (with an exception of 18 for English) we do get an impression of the variability along languages. Chinese has the relatively high error rates, even for native listeners, which is not unexpected given that we used the NIST LRE 2007 interpretation of "Chinese" being one of Wu, Min, Cantonese and Mandarin—which are really different languages. Similarly, we may expect that errors in English detection for native English listeners may have occurred in the trials containing Indian English. We analyzed the errors made for sessions with English as the target language in Table 3, and indeed, misses typically arise out of Indian English. It is unclear why there are relatively many false alarms for Hindustani, Farsi and Vietnamese, but we know that code-switching in and out of English was reported by some subjects.

4.4. Comparison to machines

In order to compare the main results to machine performance, we have sampled the set of language recognition systems that participated in LRE-2007 at three points. In Table 4 the performances of 3 systems is given, and because it is not customary to present absolute ranks in NIST evaluations, we've hinted at the system's quality by indicating the year of first participation in the series.

The table reports two values of C_{det}, one obtained using all trials in LRE-2007, and one using the same set of trials that the human subjects judged. The difference between the two values does not exceed 10 %.

The ultimate comparison is that of the systems, generally, to humans. We do that in a graph similar to Fig. 1, but including the three systems as if they represent their own "proficiency levels." We have done this in Fig. 3, where the performance variability within a system is due to the different target languages.

Table 3: Analysis of the errors made for sessions with English as the target language. Languages/accents not mentioned caused no errors

Misses			
Accent	fair	fluent	native
american	1	8	3
indian	9	40	11
False Alarms			
Language	fair	fluent	native
arabic	0	0	1
bengali	0	2	1
farsi	1	14	8
german	0	0	2
hindustani	0	18	14
japanese	0	1	0
korean	0	2	0
russian	1	1	0
spanish	0	1	0
thai	0	1	0
vietnamese	0	5	0

Table 4: System performance for three of the systems that participated in LRE-2007, 10 second trial condition. 'All' indicates inclusion of all trials, i.e., official results, 'hum' means that only the 160 trials used in this human benchmark are considered.

System	First LRE	C_{det}(all)	C_{det}(hum)
MIT	1996	0.0363	0.0391
TSS	2005	0.0702	0.0765
ICSI	2007	0.123	0.120

From the figure, we way conclude that the state of the art of language recognition systems in 2007 is close to a human language proficiency of "fair," as defined in Sect. 3.2. This can be considered the main result of this study.

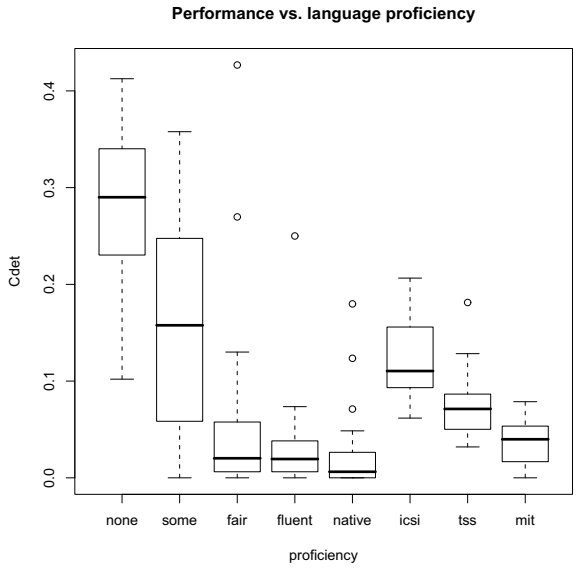

Figure 3: Comparison of Human language recognition performance to Machine performance, for 5 different proficiency levels of humans and 3 different systems.

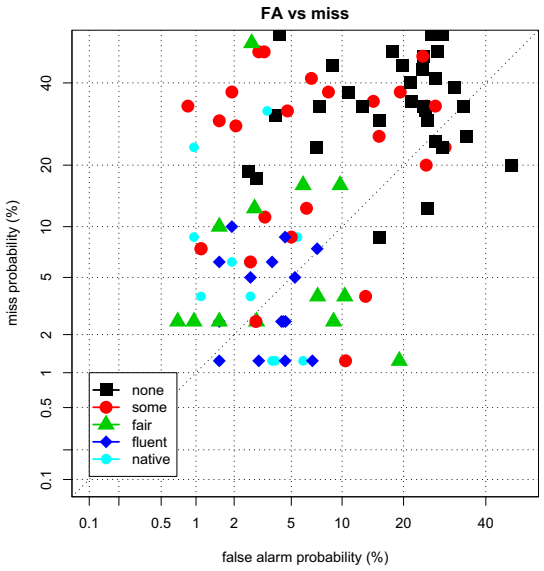

Figure 4: Errors per session, expressed in P_{FA} and P_{miss}, using probit-warped graph axes as is customary for DET plots.

4.5. False alarm/miss imbalance

One of the outcomes of earlier research [1] was the apparent imbalance in decisions: subjects tend to produce more misses than false alarms. In Fig. 4 we show the individual $(P_{\mathrm{FA}}, P_{\mathrm{miss}})$ pairs for each of the sessions, where the symbol encodes the proficiency level.We find that there are more sessions (38 out of 178) where $P_{\mathrm{FA}} < P_{\mathrm{miss}}$ at rather than vice versa (4 sessions). These results were obtained from a Test of Proportions at $p < 0.05$. Aggregated over languages, $P_{\mathrm{FA}} < P_{\mathrm{miss}}$ for all target languages except English and Spanish. We note that for English, all subjects were proficient at a level that included lexical knowledge. We further note that Spanish is the second language of California, where most of the sessions were conducted. Averaged over all sessions, $P_{\mathrm{FA}} = 0.0804$ vs. $P_{\mathrm{miss}} = 0.180$.

5. Discussion and Conclusions

We have extended our earlier work on a human benchmark in language recognition in 7 languages, using NIST LRE-2005 data [1], to a large scale experiment using the 14 languages in LRE-2007. We have attempted to control for the most important human factor, namely proficiency in the target language, and we observe a performance that increases rapidly as lexical knowledge is introduced into language proficiency. Comparing humans to machines, using the same speech trials and evaluation measure, we can say that state of the art machine performance in 2007, for 10 s speech segments, is comparable to humans with *fair* knowledge of the target language.

We further see that despite our efforts to control the decision threshold of subjects towards a balance in P_{miss} and P_{FA}, there is still a strong tendency for subjects to have $P_{\mathrm{FA}} < P_{\mathrm{miss}}$. One possible explanation is that with 13 non-target languages, it is cognitively difficult for subjects to give these the same prior weight as the single target language. Another explanation is that it is harder to judge a language as "same", when there are obvious difference in test and trial samples in terms of words, speakers and gender.

Given that, as soon as a listener can make use of more than just acoustic cues, the performance increases substantially, it would seem necessary to investigate system improvements which could take linguistic aspects into account. One might think of, for example, language specific voice quality features, fillers and pauses and perhaps syllable and word level features. This would be quite some challenge to system designers, but it seems that such strategies might yield good results, if human performance may be taken as an indicator.

It is difficult to know whether the category *no exposure* can really be tested with this language set. Most of the subjects will have been exposed to the 14 languages via film, music, television, or people that they know who have a background in the cultures in which the languages are spoken. It would be interesting to test this category more thoroughly to see if it is possible to establish a *zero level* of proficiency, using this language set.

6. Acknowledgments

We would like to thank Doug Reynolds from MIT and Christian Müller from ICSI for providing us with the LRE-2007 submitted scores. This work was supported in part by the European Union 6th FWP project AMIDA, 033812.

7. References

[1] David A. van Leeuwen, Michaël de Boer, and Rosemary Orr. A human benchmark for the NIST language recognition evaluation 2005. In *Proc. Speaker and Language Odyssey*, Stellenbosch, South Afrika, 2008. IEEE.

[2] Y. K. Muthusamy, E. Barnard, and R. Cole. Perceptual bechmarks for automatic language identification. In *Proceedings of International Conference Spoken Language Processing*, volume 1, pages 333–336, Adelaide, 1994.

[3] Alvin F. Martin and Audrey N. Le. NIST 2007 language recognition evaluation. In *Proc. Speaker and Language Odyssey*, Stellenbosch, South Afrika, 2008. IEEE.