Language dependency of candidate and reference voice coders in the STANAG 4591 selection procedure

Sander J. van Wijngaarden[1], Jerzy Lopatka[2] and Ronald A. van Buuren[1]

[1] TNO Human Factors
PO Box 23
3769 ZG Soesterberg, The Netherlands
{vanWijngaarden,vanBuuren}@tm.tno.nl

[2] Military University of Technology
Institute of Communications Systems
00-908 Warszawa, Poland
Jlopatka@wel.wat.waw.pl

ABSTRACT

When developing international standards for voice coding algorithms, it should be kept in mind that many different languages are likely to be used. Narrow band voice coders are sometimes subjectively experienced to suffer from language dependency: the performance of a coder (in terms of speech intelligibility) may depend on the language spoken. In the STANAG 4591 voice coder selection procedure, language dependency was included as a performance criterion. Out of nine candidate and reference vocoders, two were found to be significantly language dependent. By extension of the number of languages in the test from four to five, the sensitivity of the test was increased.

1. INTRODUCTION

Procedures for evaluating speech communication performance are usually based on a single language. The performance in this language is (implicitly) assumed to be representative of performance across various languages. However, when investigating the performance of low-bitrate voice coders, language dependency may very well emerge. In the development phase of any voice coder, optimization efforts are undertaken to improve performance. When these optimization efforts take only one (or too few) languages into consideration, language dependency may be introduced.

When a voice coder is to be used within a multi-lingual community (such as NATO), language dependency is quite undesirable. It seems logical that 'language dependency' should be included as a standard criterion when testing the performance of voice coders for such applications. However, the availability of suitable test methods is limited. A method to assess language dependency of low-bitrate narrow band voice coders was developed specifically for the STANAG 4591 selection procedure [1]. This paper describes the application of this new test method to nine candidate and reference voice coders, and explores the influence of the number of languages used in the test on its sensitivity.

2. THE SRT-LD TEST METHOD

Measuring language dependency must by definition involve carrying out performance tests in multiple languages. Not only does this require speech material in multiple languages to be available; also, experimental subjects who are native speakers of these different test languages will have to be recruited. Besides these practical complications, an important complicating factor is that a single type of multi-lingual performance test is necessary, which *scores performance equivalently* in all tested languages.

The most commonly applied performance tests measure either speech quality or speech intelligibility. Speech intelligibility is a better performance measure to determine language dependency than speech quality. Speech quality estimates, and other measures based on listener opinions and ratings, are more likely to be influenced by cultural and regional factors. This complicates comparison of results across languages. Moreover, speech intelligibility tests generally require fewer subjects to reach a required level of statistical certainty than speech quality tests.

A suitable method to evaluate speech intelligibility in various languages, was found to be the Speech Reception Threshold (SRT) [2]. This method was shown to allow equivalent implementations across many languages [3]. The SRT method was originally developed as an audiological screening tool, but has since been proven effective in quantifying intelligibility for a wide range of applications.

The language dependency test developed for the STANAG 4591 selection procedure is based on the SRT intelligibility test method. We will refer to the language dependency test method as the SRT-LD method [1].

The SRT-LD method is based on a minimum of three languages, preferably more. For the STANAG 4591 selection procedure, four languages were used: English, French (the two official NATO languages), German and Dutch. SRT intelligibility tests were carried out in each of these languages, for speech processed through each of the nine candidate and reference coders (three talkers and ten listeners per language). This resulted in a database of intelligibility scores. The SRT-LD method is based on an evaluation of the following separate sources of variance in this database of intelligibility scores:

- Language
- Coder
- Speaker
- Listener

If we were comparing coders simply in terms of intelligibility, then we might simply average over everything in the list above

0-7803-7549-1/02/$17.00 ©2002 IEEE.

except 'coders', and compare these means. For the SRT-LD test, we are interested in a quantification of the extent to which coders depend on language.

The result of the SRT-LD test is a language dependency-metric L. This metric is calculated from the mean SRT results for n coders in m languages as follows. First, for each coder-language combination the mean SRT value is calculated (across speakers and listeners). We will call this mean $M_{i,j}$ where i is the index for coder and j for language. Our LD-metric L_i will then be defined as:

$$L_i = \frac{2}{m(m-1)} \sum_{j=1}^{m-1} \sum_{k=j+1}^{m} \frac{|M_{i,j} - M_{i,k}|}{C_{i,j,k}} \qquad (1)$$

We used $C_{i,j,k}$ to indicate the *critical interval* for statistical significance (95% confidence) of the difference $|M_{i,j} - M_{i,k}|$. Hence, if all differences between each pair of tested languages are *just* statistically significant for coder i, then L_i will be equal to 1.

Now we are left with the problem of calculating $C_{i,j,k}$. Critical intervals may be obtained by carrying out an appropriate statistical analysis. First of all, we need to know if we can prove an overall *interaction* between 'coder' and 'language' from our SRT data: we wish to find out if the relation between intelligibility and 'coder' is modified by 'language'. This is easily done using off-the-shelf statistics software packages, such as Statistica [4], using (for instance) a straightforward 1-way ANOVA [5].

If there is a significant interaction, we calculate the critical intervals $C_{i,j,k}$. using Duncan's Multiple Range test [4,5]. These critical intervals indicate which difference between two *marginal means* is just significant. In this case, the marginal means are the mean SRT values across speakers and listeners, $M_{i,j}$. We can now calculate L_i for each coder from equation 1.

The metric L_i has some attractive features, particularly due to the use of critical ranges. Because of this normalization, a value of "1" has an intuitive interpretation; the difference in performance between two languages is (on average) just significant if $L_i = 1$.

Another attractive feature of L_i is the statistical interpretation of differences. The 95% confidence range of each of the terms in equation 1 is, because of the normalization term Ci,j,k , equal to 1. By using the basic error propagation rules (or by examining the sampling distribution of L_i) the critical interval for differences between values of L_i is is easily derived: this only depends on the number of statistically independent observations m according to

$$\delta L_i = \sqrt{\frac{1}{m}} \qquad (2)$$

This also means that any L_i differing more from zero than this value, indicates that the coder may be assumed language dependent with 95% confidence.

3. SRT-LD RESULTS USED IN THE SELECTION PROCEDURE (FOUR LANGUAGES)

As stated above, the SRT-LD test as applied to the selection procedure was based on four languages. The results are summarized in table I.

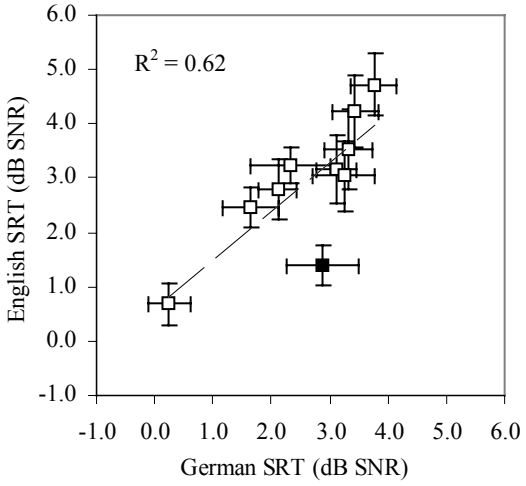
Table I. Language dependency metric L_i (four languages) for each of the nine candidate and reference coders (labeled c1-c9). The 95% confidence interval according to Eq. 2 equals 0.5 for each coder.

c1	c2	c3	c4	c5	c6	c7	c8	c9
0.23	0.52	0.25	0.26	0.97	0.17	0.47	0.38	0.40

A coder is proven to be language dependent if its L_i differs significantly from zero. This is the case for coders c2 and c5.

It is worthwhile to explore if the values of L_i are in agreement with our intuitive notions of language dependency. A relatively large value of L_i (as for coder c5) should imply relatively large variations of the speech intelligibility (SRT scores) across languages. Figure 1 shows the relation between SRT scores in two of the test languages (German and English).

Figure 1. Relation between speech intelligibility (SRT) in German and English. The solid point represents c5



There is a significant overall correlation (across coders) between intelligibility in these two languages; the same is true for all other combinations of languages. However, as figure 1 shows, the SRT for coder c5 is relatively higher for German than for English (meaning that the intelligibility for German is lower than for English). This trend is confirmed by the relations between other languages in the test. Hence, the conclusion that c5 is a language dependent coder could have been expected intuitively, purely by inspecting the database of SRT scores by eye.

The differences between values of L_i for the various coders are relatively small in relation to the 95% confidence interval. This means that it is likely that language dependency may be proven

for other coders as well, by taking measures to improve the sensitivity of the SRT-LD test. This can be achieved in a number of ways:

- by identifying and eliminating sources of variance
- by increasing the number of individual speech intelligibility measurements (more talkers and listeners)
- by increasing the number of languages in the test

Typical sources of variance to eliminate are: within-language differences in regional accents, differences in talker gender, and differences in gender, educational level and social status of the listeners. In short, the subject population (across languages) should be as uniform as possible.

Involving more talkers and listeners in the test is very straightforward, but the benefit is expected to be relatively small in relation to the additional effort needed. This is especially true if the test is to be carried out at one physical location; native listeners of multiple (foreign) languages may be scarce. It is possible to spread the test across multiple test facilities, in different countries. This greatly increases the availability of native talkers of various languages, making it much easier to perform the SRT-LD test from a practical point of view. Unfortunately, this also introduces new sources of variance (differences in acoustic conditions, instructions by different test leaders, etc.)

Increasing the number of languages beyond four is probably more effective than increasing the number of talkers and listeners beyond three and ten. However, this is virtually impossible to achieve at a single test facility, let alone by a single test leader (who would need to have sufficient command of five different languages in order to be able to supervise the tests). The question is, whether the benefit of adding additional languages is greater than the adverse effects of adding new sources of variance. This is explored in the following section.

4. SRT-LD RESULTS BASED ON FIVE LANGUAGES

SRT intelligibility tests were also carried out in Polish. The Polish part of the experiment was carried out at a different test facility, by a different experimenter, than the rest of the experiment. The rest of the experiment was completely carried out in the Netherlands (although with native talkers of each of the test languages, who happened to be living in the Netherlands).

The Polish part of the language dependency test was not used for the selection procedure, but merely performed out of scientific interest. After extension to five languages, the values of L_I given in table II were obtained.

Table II. Language dependency metric L_i (five languages) for each of the nine candidate and reference coders (labeled c1-c9). The 95% confidence interval according to Eq. 2 equals 0.45 for each coder.

c1	c2	c3	c4	c5	c6	c7	c8	c9
0.37	0.46	0.22	0.47	0.84	0.34	0.43	0.59	0.31

Not surprisingly, coders c2 and c5 are still significantly language dependent. By extension to five languages, language dependency can now also be proven for c4 and c8. It is noteworthy that c8 is based on the same vocoding algorithm as c5, but at a different bitrate.

Despite the inclusion of data from a different test facility, the test method was found to improve, in terms of statistical discrimination, by the addition of the Polish data.

5. CONCLUSIONS AND DISCUSSION

The SRT-LD method is effective in measuring language dependency of low-bitrate narrow band voice coders. As applied in the STANAG 4591 selection procedure, its sensitivity (statistical discrimination power) is relatively small. This can be improved by extending the number of languages in the test. It appears that the use of data obtained at multiple test sites is an acceptable solution to cope with the practical problems associated with recruiting native talkers of multiple languages as test subjects.

The fact that several voice coders, including state-of-the-art candidate algorithms, were found to be language dependent, shows the need to include language dependency as a performance measure in voice coder selection procedures.

It should be noted that the SRT-LD method can be used to evaluate the use of a voice coder for *multi-lingual* applications, but not necessarily for *cross-lingual* applications. Non-native speech tends to be less intelligible even under undegraded conditions (e.g. [6]), but it is also likely that adverse interactions with voice coding algorithms reduce the intelligibility of non-native speech even further. This effect is not included in the SRT-LD test results.

6. REFERENCES

1. Wijngaarden, S.J. van, Steeneken, H.J.M, "A Proposed Method for Measuring Language Dependency of Narrow Band Voice Coders". *Proc.. Eurospeech 2001-Scandinavia*, pp. 2495-2498, 2001.
2. Plomp, R. and Mimpen, A.M. "Improving the Reliability of Testing the Speech Reception Threshold for Sentences", *Audiology*, Vol. 18: 43-52, 1979.
3. Wijngaarden, S.J. van, Steeneken, H.J.M., Houtgast, T., "Methods and models for quantitative assessment of speech intelligibility in cross-language communication". *Proc.. RTO Workshop on Multi-lingual Speech and Language Processing*, Aalborg, Denmark, 2001.
4. StatSoft, Inc. (2000). STATISTICA for Windows [Computer program manual]. Tulsa, OK: StatSoft, Inc., 2300 East 14th Street, Tulsa, OK 74104, phone: (918) 749-1119, fax: (918) 749-2217, email: info@statsoft.com, WEB: http://www.statsoft.com
5. Winer, B.J. "Statistical principles in experimental design", McGraw_Hill, London, 1970.
6. Wijngaarden, S.J. van. "The intelligibility of Non-native Dutch speech". *Speech Commun.* 35, pp. 103-113, 2001