A Proposed Method for Measuring Language Dependency of Narrow Band Voice Coders

Sander J. van Wijngaarden, Herman J.M. Steeneken

TNO Human Factors
PO Box 23
3769 ZG Soesterberg
The Netherlands
{VanWijngaarden,Steeneken}@tm.tno.nl

Abstract

Narrow band voice coders that use vector quantization techniques may suffer from language dependency: the performance of the coder (in terms of speech intelligibility) may depend on the language spoken. For multinational applications, this is undesirable. A test method is proposed that may be used to determine to which extent a vocoder is language dependent. The proposed method, based on a subjective speech intelligibility test in multiple languages, is shown to be feasible by application on known language dependent 'systems': non-native (human) speakers and listeners. The method is shown to be able to significantly prove differences in language dependency, even when using only three languages and nine speaker/listener combinations.

1. Introduction

Speech communication channels are often implicitly assumed to be language independent. Performance measures, such as speech quality and speech intelligibility, are usually assessed without checking the validity of results in languages other than the test language. For many conventional types of communication channels it is fair to expect that this assumption (no language dependency) is valid. Objective speech intelligibility prediction methods, such as the Articulation Index [1] and the Speech Transmission Index [2], have been proven to show a robust relation with subjective speech intelligibility in a host of languages (eg [3]). This may also be explained by observing that many global characteristics of speech are more or less the same across languages, such as the average long-term speech spectrum [4] and the intelligibility-degrading influence of noise [5].

However, when investigating the performance of narrow band voice coders, language dependency may *not* be ruled out. Modern narrow band voice coding algorithms (eg [6,7]) generally use vector quantization (VQ) techniques. VQ helps to achieve lower bit rates without reducing speech quality or speech intelligibility. In order to construct the necessary codebooks for VQ, a corpus of suitable speech utterances is used as 'training material'. In case this corpus contains material from too few (or too similar) languages, the performance of the coder may be language dependent.

When a speech coder is used within a multi-lingual community (such as, for instance, NATO), language dependency is quite undesirable. It seems logical that 'language dependency' should be included as a standard criterion when testing the performance of speech coders for such applications. Unfortunately, measuring language dependency is not easy – it automatically involves carrying out speech 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. Any difference in the implementation between test languages may potentially threaten the validity of the test.

Finally, there is the problem of quantifying the extent to which systems are language-dependent. To make easy interpretation possible, scores on a suitable multi-lingual performance test should be converted into a single language dependency metric. To our knowledge, no such metric has yet been proposed.

2. Available Test Methods

So far, we have used the term 'language dependency' a few times without specifying *which performance characteristic* of a coder is presumed to depend on language. Performance of speech coders is most frequently expressed in terms *of speech quality* or *speech intelligibility*.

Differences in speech quality are usually investigated by having test subjects rate or compare different speech tokens, and having them report on their subjective preference. Although speech quality is a conceptually attractive indicator of performance, it may not be the best characteristic to use when determining language dependency. Speech quality testing paradigms rely on *opinions*; these are known to have a reproducible mean *within* testing populations, but this mean may vary across groups of subjects who speak different languages. A more practical complication is that speech quality tests generally require rather large numbers of subjects (typically 16-40) in order to produce accurate results. Such large subject groups may be hard to recruit in multiple languages.

Speech intelligibility is a more straightforward indicator of performance: if the intelligibility is lower, a lower percentage of messages is understood correctly, and the performance is clearly lower. Speech intelligibility (as opposed to quality) is

not a matter of opinion. Instead, intelligibility tests measure the fraction of speech tokens (sentences, words, or syllables) that are correctly heard. The statistical spread of intelligibility measures between subjects is usually relatively small. This makes speech intelligibility a suitable performance characteristic to use in a language dependency test.
A dozen or more subjective speech intelligibility tests have been reported in literature (eg. DRT[8], CVC[9], SRT[10]). Each test is a compromise between diagnostic power (the ability to identify causes of intelligibility decreases), precision and efficiency. In this case, diagnostic power is not as important as precision and efficiency.
Because of practical considerations, another important factor is the selection of subjects. It may be a non-trivial task to find enough native speakers of multiple (>3) languages, especially if the time needed for each subject is considerable. This rules out lengthy tests, or procedures that require extensive training.
Taking all factors into consideration, a suitable subjective test to base a language dependency test on matches the following profile:

- Highly reproducible intelligibility test
- Equivalent implementation in multiple languages
- Suitable for use with small subject groups (<10 subjects)
- Requires only little time of each subject
- Little or no subject training necessary

3. The SRT Method

Based on the requirements mentioned above, a good candidate intelligibility test method is the Speech Reception Threshold (SRT) method [5,10,11]. This test gives a robust measure of sentence intelligibility in noise, corresponding to the speech-to-noise ratio that gives 50% correct response of short redundant sentences (8 or 9 syllables long).
In the SRT testing procedure, all test sentences are first processed through the tested coders, and stored on a (notebook) computer hard disk. Before each presentation, masking noise is added to the (processed) sentence in order to obtain the required speech-to-noise ratio. The masking noise spectrum is equal to the long-term spectrum of the test sentences. After presentation of each sentence, a subject responds by repeating the sentence as he or she heard it, and the experimenter compares the response with the actual sentence. If the response is completely correct, the noise level for the next sentence is increased by 2 dB; after an incorrect response, the noise level is decreased by 2 dB. The first sentence is repeated until it is responded correctly, using 4 dB steps. This is done to quickly converge to the 50% intelligibility threshold. By taking the average speech-to-noise ratio over the last 10 sentences, the 50% sentence intelligibility threshold (SRT) is obtained.
SRT-results are always reported as an SNR (Speech-to-Noise Ratio) at which the sentence intelligibility is 50%. A high SRT value means that only little noise may be added to reduce the intelligibility to 50%. A low score means that the speech signal can tolerate a fairly large amount of noise and still give 50% sentence intelligibility. Undegraded speech can tolerate more noise before reaching the 50% intelligibility threshold than speech degraded by coding artifacts; the difference in SNR is a measure of the difference in intelligiblity. Similarly, the performance of vocoders in different languages can be compared in terms of speech intelligibility using SRT results.
SRT scores offer high reproducibility when working with only small subject groups, while the time needed for each subject is limited (typically less than 45 minutes). Moreover, the sentences are constructed following a simple set of rules [10]. The original SRT test by Plomp and Mimpen was based on the Dutch language, but the test has proven to be easily implemented in several languages [5], yielding equivalent results in different languages.

4. The Proposed LD Metric

When test scores for different coders in different languages are available, a quick comparison should give some insight into the language dependency of the coders. Clearly, if no differences are found between languages, then apparently none of the coders is language dependent. When some statistical differences *are* observed, the interpretation becomes more difficult.
When inspecting SRT results collected in several languages, the following variables (or sources of variance) in the overall experiment may be identified:

- 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 except ‘coders’, and compare these means. In this case, we are interested in a quantification of the extent to which coders depend on language.
We propose a language dependency-metric that 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:

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

We used $C_{i,j,k}$ to indicate the *critical interval* for statistical significance of the difference $|M_{i,j} - M_{i,k}|$. Hence, if all differences between each pair of tested languages is *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 [13], using (for instance) a straightforward 1-way ANOVA [12].

If there is a significant interaction, we may try to calculate the critical intervals $C_{i,j,k}$. Several statistical methods are available, mostly based on the *studentized range* statistic. For our purposes, we prefer Duncan's Multiple Range test [12,13], but several similar tests (such as the perhaps better known Newman-Keuls test) are also suitable. Assuming a certain alpha-level (in our case always $p<0.05$), critical ranges may be calculated. In the context of Duncan's test, these critical ranges 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, which are identical to the means $M_{i,j}$. This means that the $C_{i,j,k}$ values needed for calculating our LD-metric are identical to the critical ranges determined using Duncan's test. 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.

The practical procedure for calculating L_i will be demonstrated in the next section.

5. Feasibility of the Proposed Method

In order to prove that our proposed method does in principle measure language dependency, we carried out SRT tests in Dutch, English and German. The Dutch sentences were translated (although far from literally) to English and German, observing the same rules that were used for constructing the original Dutch sentences (everyday 8 or 9 syllable sentences, with a maximum of 1 three-syllable word per sentence).

First of all, a baseline-test was run in order to make sure that the test behaves equivalently in all three languages. The results are given in table 1.

Table 1. Mean SRT and standard deviation (baseline), measured using N subjects

	Dutch	German	English
Mean SRT (dB)	-0.7	-1.1	-1.0
SD	1.6	1.7	1.1
N	9	3	3

None of the results in table 1 differ significantly. The critical interval for statistical significance in table 1 is approx. 1 dB.

Although one will generally aim for a larger number of subjects per language (perhaps 6 to 10), these results give confidence that for the given languages the SRT test is equivalent. Baseline results as given in table 1 may also be used to compensate for small differences between languages by means of a normalization factor.

Now the question arises which (language-dependent) systems may be tested in order to prove that the proposed method does indeed quantify language dependency. The problem here is a lack of existing (quantitative) knowledge on language dependency of coders. We need test conditions that are surely language dependent, as well as conditions that are guaranteed to be language *independent*.

Instead of testing coders (or other systems), a class of test conditions was used for which known language effects exist [5]: non-native speakers and listeners. Instead of testing the performance of a channel, the performance of the participating speakers/listeners is tested. Although the interpretation of results will be different, the principles of the test remain the same.

SRT tests were run in Dutch, English and German, in three conditions: fully native, (non-native) Dutch speakers and (non-native) Dutch listeners. We will refer to these conditions as 'C1', C2' and 'C3', respectively, noting that language dependency is expected for C2 and C3, but *not* for C1. We will refer to C2 and C3 as *non-native* conditions, although one of the languages (Dutch) *is* the native language of the subjects. In all conditions, speech material by three speakers was used and 3 listeners participated.

Only non-native speakers and listeners were selected with good proficiency in the tested languages, both written and orally. Although SRT effects of non-nativeness up to 12 dB have been reported for less proficient speakers [14], such unrealistically low intelligibility scores (when compared to coder performance) did not occur here. All measured SRT values are given in table 2.

The results in table 2 were 'normalized' by subtracting the mean baseline SRT score for each language (table 1) from the actual SRT score. Hence, instead of being true SRT scores, the values in table 1 express the effect of 'non-nativeness' in terms of SRT score. In this case, this normalization is not really necessary, but if larger differences between SRT implementations for the tested languages are found, this may lead to a better language dependency estimate.

Table 2. Full SRT results of the language dependency experiment (9 speaker/listener combinations per condition per language). All SRT results are given in dB.

C1			C2			C3		
NL	EN	GE	NL	EN	GE	NL	EN	GE
0.1	3.3	-0.4	1.3	-0.3	0.4	2.9	7.7	1.2
-3.5	-1.5	-0.4	-1.1	0.9	0.4	0.9	6.1	3.6
-1.9	0.9	0.0	1.3	0.5	0.8	-1.1	4.9	0.8
0.9	-1.5	0.4	0.5	0.1	1.2	2.5	8.9	4.0
0.5	-1.5	2.4	-1.1	-1.9	-0.4	1.3	9.7	4.0
2.5	0.5	-0.4	-1.5	-0.7	1.2	-0.3	8.5	4.0
0.5	-0.7	0.8	0.9	0.5	2.4	0.9	5.3	2.4
-1.9	-0.7	-0.8	-0.3	1.7	2.4	-1.5	5.3	1.2
1.3	0.9	-1.2	-3.1	2.1	3.2	-1.1	2.9	2.8

Since there are three languages (m=3 in equation 1), the number of differences $|M_{i,j} - M_{i,k}|$ to be calculated per 'coder'

is only 3. The values for $|M_{i,j} - M_{i,k}|$, and the associated critical ranges according to Duncan's test, are given in table 3.

Table 3. Differences $|M_{i,j} - M_{i,k}|$ and critical ranges $C_{i,j,k}$.

	$\lvert M_{i,1} - M_{i,2}\rvert$	$\lvert M_{i,1} - M_{i,3}\rvert$	$\lvert M_{i,2} - M_{i,3}\rvert$
C1	0.17	0.17	0.00
C2	1.59	0.70	0.89
C3	2.12	6.12	4.00
	$C_{i,1,2}$	$C_{i,1,3}$	$C_{i,2,3}$
C1	1.53	1.53	1.58
C2	1.65	1.58	1.67
C3	1.68	1.7	1.71

Equation 1 is used to calculate L_i for C1, C2 and C3. Results are given in table 4.

Table 4. LD-metric L_i for conditions C1, C2 and C3

	C1	C2	C3
L_i	0.07	0.65	2.40

The critical interval for significant difference between the results in table 3 is, according to equation 2, $\delta L_i = 0.58$. Hence, both 'coders' C2 and C3 are significantly language dependent, but C1 (fully native) is not. Furthermore, all differences in language dependency between C1, C2 and C3 are significant.

6. Conclusions and Discussion

The validation of any language dependency test must consist of two parts: first of all, the method must be shown to identify language dependency on systems of which there is an a priori knowledge of language dependency. Secondly, the suitability of the test for practical purposes must be proven by applying it on a variety of voice coding algorithms. In this paper, the first half of this validation process was presented; for lack of knowledge on language dependency of vocoders, the method was tested on 'language dependent humans'. The 'language dependency' of groups of (proficient) non-native speakers and listeners could be proven statistically using the proposed LD-metric. Encouraged by this result, the second part of the validation process, application on vocoders, will now be initiated. Preliminary experiences with application of the SRT method on vocoder-degraded speech give reason to look forward to results of these experiments with confidence.

The limited version of the test used in this paper (only 3 languages, 9 speaker/listener combinations per language) needs a mean SRT difference between languages of approx. 1 dB to obtain a significant result. By using more (or less similar) languages, the test may be made more sensitive to smaller language dependency effects. The same applies for the number of subjects.

7. References

[1] Kryter, K.D., "Methods for calculation and use of the Articulation Index", *Journal of the Acoustical Society of America*, 34: 1689-1697, 1962.

[2] IEC Std. 60268-16 2nd edition, "Sound system Equipment. Part 16: objective rating of speech intelligibility by speech transmission index," Geneva, Switzerland, 1998.

[3] Houtgast, T. and Steeneken, H.J.M.,"A multi-lingual evaluation of the Rasti-method for estimating speech intelligibility in auditoria", *Acustica*, 54:185-199, 1984.

[4] Byrne, D., Dillon, H., Tran, K. *et al.*, "An international comparison of long-term average speech spectra", *Journal of the Acoustical Society of America*, 96: 2108-2120, 1994.

[5] Wijngaarden, S.J. van and Steeneken, H.J.M. "The Intelligibility of German and English Speech to Dutch Listeners". *Proceedings of the International Conference on Spoken Language Processing (ICSLP2000)*, 2000.

[6] Gournay, P and Chartier, F. "A 1200 bits/s speech coder for very low bit rate communciations". *IEEE Workshop on Signal Processing Systems (SiPS'98*), Boston, 1998.

[7] Supplee L., Cohn R., Collura J. and McCree A. "MELP: the new federal standard at 2400 bps", *Proceedings of ICASSP97*, New York, 1997, pages 1591-1594.

[8] Voiers, W.D. "Diagnostic evaluation of speech intelligibility", In *Speech Intelligibility and speaker recognition, Vol.2. Benchmark papers in acoustics*, edited by M.E. Hawley (Dowden, Hutchinson and Ross, Stroudburg), 1977, pages 374-384.

[9] Steeneken, H.J.M., "Diagnostic information of subjective intelligibility tests", *Proceedings ICASSP*, Dallas, pages 131-134, 1987.

[10] Plomp, R. and Mimpen, A.M. "Improving the Reliability of Testing the Speech Reception Threshold for Sentences", *Audiology*, Vol. 18: 43-52, 1979.

[11] Wijngaarden, S.J. van. "Speech intelligibility of native and non-native Dutch speech". Accepted for publication in *Speech Communication*, 2001.

[12] Winer, B.J. "Statistical principles in experimental design", McGraw_Hill, London, 1970.

[13] 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

[14] Buus, S., Florentine, M., Scharf, B. and Canevet, G. "Native, French listeners' perception of American-English in noise." *Proceedings of Internoise 86*, 1986, pages. 895-898.