Intelligibility Assessment Method for Semantically Unpredictable Sentences in German

JP Ramirez, A. Raake, D. Reusch
Quality and Usability Lab, Deutsche Telekom Laboratories, TU Berlin
Juan-Pablo.Ramirez@telekom.de

Introduction

Research in perception requires methods to assess the human auditory system performance. To this aim, speech recognition tests are employed on listeners. The task is to retrieve a target message (T) impaired by a masker (M) - typically noise or speech at various intensities - presented in a particular spatial configuration. The output of such tests is the so-called intelligibility, depicting the proportion of the target speech correctly retrieved by the listener in given conditions.

In order to predict speech retrieval abilities among the hearing impaired listeners, audiologists assess the so-called Speech Reception Threshold (SRT). This measurement is the ratio of target and masker levels (TMR) in correspondence with an intelligibility of 50%. Further measurements have proven that the physical characteristics and perceptual nature of the target and masker had crucial consequences on the SRT.

In the task of sentence recognition, the syntax and semantics deliver information which has proven to significantly impact the listener's performance [1]. Several methods were developed in the past in order to quantify the complexity of a given target message [2] [3].They are extensively described by [4].

Boothroyd and Nittrouer (1988) developed linguistic complexity indices that are widely used today. Considering a whole composed of n elements, they link their respective intelligibilities p_w and p_e as follows:

$$p_w = p_e^{\,j} \tag{1}$$

where j quantifies the minimal amount of elements required to retrieve the whole. For a j equal to n, the complexity is at its maximum, and all elements are independent from each other.

A second equation, proposed in the same publication, links the speech recognition scores for items presented both isolated and in their contextual frame. It introduced the parameter k. Since we used recorded sentences, and not the individual words, we did not assess this aspect.

Another measure investigated in previous research was the slope s, as described in [5]. This parameter defines the slope of the psychometric function depicting intelligibility versus TMR at the SRT (yielding 50% intelligibility).

The present paper introduces a corpus of German sentences and measurements of intelligibility in speech-shaped stationary noise conducted with these sentences. Special efforts were paid to gather semantically unrelated words in a correct syntax. Several sentence-retrieval tasks were performed on hearing unimpaired subjects in order to extract the SRT, the slope of the discrimination function s, as well as the complexity indexes j and j' relative to our corpus.

Speech Material

The sentence test proposed in this paper is the transposition into the German language of the work described in [6], which was inspired by [7]. The corpus is phonetically balanced following the average distribution in the language. Each sentence is composed of four keywords with no semantic coherence. In total, 24 lists of twelve sentences each were produced using a corresponding algorithm, yielding 288 sentences. To this aim, four of the five syntactic structures described in [7] were used, with three sentences of each type per list. The keywords actually employed in each list were selected from a corpus extracted from CELEX [8]. The sentence corpus can be characterized as follows:

- Most frequent monosyllabic German words (monosyllabic in canonical form).
- Lexicon of 192 nouns, 109 verbs, 72 adjectives.
- Three repetitions of each word in the 288-sentence corpus.
- Chi-square-based maximization of the agreement between the phoneme-distribution of each list and a phoneme-distribution characteristic of the German language [9].
- Equalization of the word-frequencies per lexical category and per list.

The corpus was recorded by a male speaker with no particular accent at a sample rate of 44.1 kHz in an anechoic environment. To avoid reading effects, the text sentences were randomized for the recordings, and the recorded samples sorted back into the list structure. The electronic levels were normalised to -26 dB relative digital full-scale, excluding pauses by using voice activity detection.

A speech-shaped stationary noise masker was created for the speech in noise listening conditions used across the whole experiments described in this paper. It was produced by overlapping 10 times, in a randomized fashion, the whole set of 288 sentences. Here, the 60 seconds of noise to be produced were filled with subsequent repetitions of each sentence, where the start- and end-points were randomly selected (cosine fade-in and -out). When the 288 seconds were completely filled with a given sentence, the procedure was repeated for the next sentence, until all 288 sentences were used. The entire procedure was carried out 10 times, yielding a stationary noise that presents the same long-term spectrum as the original 288 sentences concatenated one behind the other.

Experimental Results

All the subjects to take part to the following tests were 22 years old in average, presented no hearing impairment (hearing threshold under 15 dB HL for both ear). They were paid on an hour basis. Sentences were presented diotically,

at a level of 70 dB SPL. Intelligibility performance was exclusively based on the retrieval of the four keywords of a sentence, and other words such as articles and pronouns were ignored. The tests were performed in a sound proof booth, in half-hour sessions disrupted by regular pauses. No subject took part in more than one test, to avoid a learning of the employed word corpus. A similar, automated test procedure was used as described in [6]. Hence, no intervention is required from the experimenter once the subject is set alone in the listening cabin and that the test program is launched.

A. SRT Measurements

Fourteen students were paid to take part to a first test. After a training session where they got acquainted with the speaker's voice and the task, each listened to 5 lists of 12 sentences. The level of the first sentence of a list was increased by the subject himself, until he considered having understood half of the sentence presented to him. Further on, the masker level was set according to the amount of words retrieved in the pervious sentence, following the adaptation procedure proposed by [10]. The aim is to increase the target's level when less than 2 words were understood, and decrease it when more than 2 words were understood. Thus, the TMR converges towards the SRT, which is assessed as the average of the last 8 TMRs presented.
On average, the SRT measured was of -4.1 dB, with a standard deviation of 1.1 dB. Comparison with the Göttingen and Oldenburg tests [10] can be found in the table 1.

Sentence test	SRT dB	Slope *1/dB*	*j/n*, p_e low	*j/n*, p_e high
Göttingen	-6.23	0.192	0.390	0.476
Oldenburg	-7.11	0.171	0.636	0.858
D. Telekom Lab	-4.10	0.120	0.787	0.985

Table 1: SRT and slope of the psychometric function measured via three different sentence tests. Parameter of sentence predictability *j* is given for p_e around 0.2 and 0.8. See text for more details.

In a second test involving 8 subjects, the whole corpus was presented in a randomized order at a constant TMR of -4.1 dB. Each sentence presented an average intelligibility deviating from the 50% expected score. This enabled the computation of a per sentence SRT adjustment based on this deviation.

B. Slope of the Discriminative Function

Since the previous experiment provided intelligibility scores at TMRs that were not set at fixed intervals, but at intervals that depend on the performance of a given subject [10], we performed an additional sentence recognition test at fixed TMR-levels. The aim was to draw the discrimination function covering the relevant interval of TMRs. Ten subjects took part in this test. After a training session, each condition was presented three times in a randomized fashion. Averaged results are shown on the figure 1, with 95% confidence intervals.

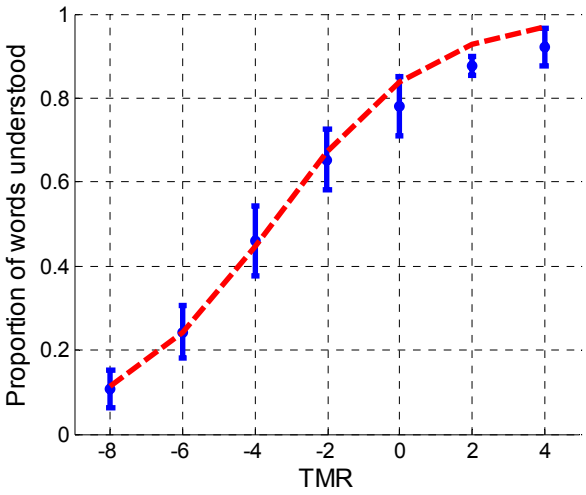

Figure 1: proportion of words understood at 7 fixed TMRs, with 95% confidence interval. The dashed line is the fitting function as depicted in equation 2.

It appears that the SRT measured in this way is higher than the previous one, with SRT = -3.5 dB. But the previously measured value remains in the 95% confidence interval.
Two ways of measuring the slope at 50% intelligibility were considered. One can either consider the function as linear when ranging between 20% and 80% of word retrieval. Using this method, we observe a slope s_{lin} of 10 %.dB^{-1}. Another method, proposed by [11], consists in fitting the following function to the points corresponding to 20% and 80% recognition scores.

$$p_w = \frac{100}{1 + e^{4s\cdot(SRT - TMR)}} \tag{2}$$

The theoretical as well as the experimental measurement lead to $s/s_{lin}=ln(2)/0.6$. Hence the average slope for the corpus is $s=12\%.dB^{-1}$. Values from other sentence test can be compared in table 1.

C. A Glimpse at the per Word Position

One can observe from table 1 that the SRT and slope measurements of the corpus proposed are below what was found for different other sentence tests. This can be explained by the observation that the recognition rate of a word is directly influenced by its position in the sentence (see figure 2). We observe a strict decrease of probability for a word to be understood with its position in the sentence. A first explanation possible could be that human utterances are prone to a strong prosody, and therefore level variations in time along the sentence. The previous word being retrieved, it may be easier to perceive the following word based on the facilitated identification of the word-boundaries, e.g. from co-articulation speech cues.
Assuming i as the position of a word W_i, and $P(W_i)$ its probability to be retrieved, we observe the following correlation:

$$P(W_i) \approx P(W_1)^i \tag{3}$$

Statistically, this is equivalent to state that an element is understood only if each of the preceding ones were. This has

a direct impact on the per-word reception threshold and slopes, and as being their averaged perceived score, on the overall sentence intelligibility.

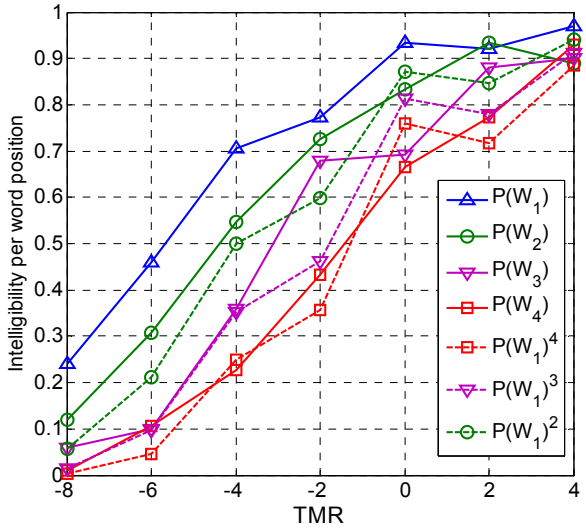

Figure 2: intelligibility derived per word position.

Word position	SRT dB	Slope 1/dB
1	-5.7	0.14
2	-4.3	0.12
3	-2.6	0.12
4	-1.5	0.11

Table 2: SRT and slope of the psychometric function per word position

Probability Analysis

In order to observe the *j* factor depicted in the introduction, equation (1), this last part observes for a given masking level, the average proportion of words that were correctly understood per sentence (p_e) versus the average probability of a sentence being fully retrieved (p_s). *j* was criticised for not showing stable values with p_e varying [4]. The *k* parameter mentioned in the introduction is more stable in this regard, but more complex to assess, since it requires the independent testing of the elements isolated and the elements in the whole. As an interesting complement, [4] introduced the parameter *j'* defined as:

$$p_{w,0} = (1 - p_e)^{j'} \tag{4}$$

where $p_{w,0}$ is the probability of none of the word of the sentence to recognized. Both indices are depicted in figure 3.

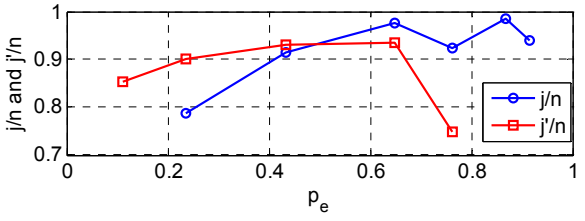

Figure 3: parameters of sentence predictability. Note that experimentally, *j=j'* at p_e=0.5.

For conditions with a high level of noise, it is less likely to find sentences fully understood. The same applies for highly intelligible listening conditions where finding a sentence with none of the words understood is unlikely. Therefore *j* and *j'* are not constant over the whole range of target-to-masker ratio, but show a stable value close to 4 when p_e is over 0.4 for *j* and lower than 0.6 for *j'*.
Considering *j'* and *j* for respectively the lowest and highest half values of p_e, the effective number of statistically independent elements in our sentences are between 3.6 and 4. This consolidates the assumption of low predictability of the corpus' sentences.

Conclusion

This paper introduced a corpus of 288 semantically unpredictable sentences, grouped in 24 lists, close to the average linguistic features of the German language, with correct syntactical structures. Measurements of the speech reception threshold and of the slope of the psychometric function show values below (but comparable to) the literature. This is explained by the non-homogeneous recognition scores among words in a sentence. Observation of parameters of complexity j and j' gave expected values close to 4, reinforcing the unpredictability of the corpus.

Future works will target the case where speech is used as masker, introducing high informational masking [12] as well as voice similarities as a new input to a future model.

References

[1] Boothroyd, A., and Nittrouer, S. (1988). "Mathematical treatment of context effects in phoneme and word recognition", J. Acoust. Soc. Am. 84, 101-114.

[2] Shannon, C. E. (1951). "Prediction and entropy of printed English", Bell Syst. Tech. J. 30, 50-64.

[3] Treisman, A. M. (1965). "Verbal responses and contextual constraint", J. Verbal Learn. Verbal Behav. 4, 118-128.

[4] Bronkhorst, A., Brand, T., and Wagener K. (2002). "Evaluation of context effects in sentence recognition", J. Acoust. Soc. Am. 111, 2874-2886.

[5] Wagener, K., Brand, T., and Kollmeier, B. (1999). "Entwicklung und Evaluaion eines Satztests für die deutsche Sprache. I. Design des Oldenburger Satztests", Z. Audiol- 38(1), 1-32.

[6] Raake, A., and Katz, B.F.G. (2006). "SUS-based Method for Speech Reception Threshold Measurement in French", Proc. LREC (Language Resources and Evaluation Conference).

[7] Benoît, C., Grice M., and Hazan V. (1996). "The SUS test: A method for the assessment of test-to-speech synthesis intelligibility using Semantically Unpredictable Sentences", Speech Comm., 18:381-392.

[8] Max Plank Institute for Psycholinguistics, http://celex.mpi.nl/

[9] Reusch, D., (2009). “Untersuchung zur Sprachverständlichkeit in Bezug auf sprecherspezifische Eigenschaften und semantsicher Vorhersagbarkeit”, Master thesis supervised by Raake, A., and Weinzierl, S.

[10] Brand, T., and Kollmeier, B. (2002). “Efficient adaptive procedure for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests”. J. Acoust. Soc. Am., 111:2801-2810.

[11] Kollmeier, B., and Wesselkamp, M. (1997). “Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment”. J. Acoust. Soc. Am., 102:2412-2421.

[12] Durlach, N.I., Mason, C.R., Kidd, G., Arbogast, T.L., Colburn, H.S., Shinn-Cunningham, B.C. (2003). “Note on informational masking (L)”. J. Acoust. Soc. Am., 113:2984-2987.