

An Italian matrix sentence test for the evaluation of speech intelligibility in noise

*Original*

An Italian matrix sentence test for the evaluation of speech intelligibility in noise / Puglisi, GIUSEPPINA EMMA; Warzybok, A.; Hochmuth, S.; Visentin, C.; Astolfi, Arianna; Prodi, N.; Kollmeier, B.. - In: INTERNATIONAL JOURNAL OF AUDIOLOGY. - ISSN 1499-2027. - STAMPA. - 54:Suppl. 2(2015), pp. 44-50. [10.3109/14992027.2015.1061709]

*Availability:*

This version is available at: 11583/2624139 since: 2016-01-11T15:02:11Z

*Publisher:*

Taylor&Francis Group

*Published*

DOI:10.3109/14992027.2015.1061709

*Terms of use:*

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

*Publisher copyright*

(Article begins on next page)

1 An Italian matrix sentence test for the evaluation of speech intelligibility in noise

2

3 Giuseppina Emma Puglisi<sup>a,\*</sup>

4 Anna Warzybok<sup>b</sup>

5 Sabine Hochmuth<sup>b</sup>

6 Chiara Visentin<sup>c</sup>

7 Arianna Astolfi<sup>a</sup>

8 Nicola Prodi<sup>c</sup>

9 Birger Kollmeier<sup>b</sup>

10

11 <sup>a</sup>Politecnico di Torino

12 Department of Energy

13 Corso Duca degli Abruzzi, 24, 10129, Torino, Italy

14

15 <sup>b</sup>University of Oldenburg

16 Medizinische Physik and Cluster of Excellence Hearing4all

17 D-26111, Oldenburg, Germany

18

19 <sup>c</sup>Università di Ferrara

20 Department of Engineering

21 Via Saragat 1, 44122, Ferrara, Italy

22

23 Key words: speech perception, speech audiometry, speech reception threshold, Italian  
24 speech recognition test

25

26 Abbreviations:

27 ANOVA Analysis of Variance

1

2 SRT Speech Reception Threshold

3

4 SNR Signal-to-Noise Ratio

5

6 S50 slope of intelligibility function

7

8 \*) Author to whom correspondence should be addressed.

9

10 Electronic mail: [giuseppina.puglisi@polito.it](mailto:giuseppina.puglisi@polito.it)

11

12

## 1 ABSTRACT

2 Objective: Development of an Italian matrix sentence test for the assessment of speech  
3 intelligibility in noise. Design: The development of the test included the selection, recording,  
4 optimization with level adjustment, and evaluation of speech material. The training effect  
5 was assessed adaptively during the evaluation measurements with 6 lists of 20 sentences,  
6 using open- and closed-set response formats. Reference data were established for normal-  
7 hearing listeners with adaptive measurements. Equivalence of the test lists was investigated  
8 using the open-set response format at three signal-to-noise ratios (SNRs). Study sample: 55  
9 normal-hearing Italian mother-tongue listeners. Results: The evaluation measurements at  
10 fixed SNRs resulted in a mean speech reception threshold (SRT) of  $-7.3 \pm 0.2$  dB SNR and  
11 slope of  $13.3 \pm 1.2$  %/dB. The major training effect of 1.5 dB was observed for the first two  
12 consecutive measurements. Mean SRTs of  $-6.7 \pm 0.7$  dB SNR and  $-7.4 \pm 0.7$  dB SNR were  
13 found from the third till the sixth adaptive measurement for open- and closed-set test  
14 response formats, respectively. Conclusions: A good agreement has been found between the  
15 SRTs and slope and those of other matrix tests. Since sentences are difficult to memorize, the  
16 Italian matrix test is suitable for repeated measurements.

17

18

## 1 INTRODUCTION

2 Even though the Italian language is the 2<sup>nd</sup> most commonly spoken language in Europe  
3 (European Commission, 2012) and the 21<sup>st</sup> most commonly spoken language in the world  
4 (Lewis et al., 2014) only a few speech recognition tests have been developed so far to  
5 evaluate speech intelligibility for both audiology and research purposes (e.g. for intelligibility  
6 measurements in classrooms or in workspaces). Moreover, most of them have not been  
7 optimized for speech intelligibility in noise. This paper therefore describes the construction  
8 and evaluation of an Italian sentence test with the matrix test format in order to be  
9 compatible with an increasing number of tests that implement this format in other languages  
10 (Kollmeier et al., 2015).

11 The most frequently used speech intelligibility test for the Italian language is based on  
12 meaningful mono- or disyllabic words (Bocca & Pellegrini, 1950) distributed over 6 lists  
13 composed of 50 words each. Since meaningful monosyllabic words are rarer in Italian than  
14 disyllabic words, a disyllabic test was proposed by Turrini et al. (1993) which was optimized  
15 with regard to **phonemic balancing** and word familiarity.

16 The other speech audiometry tests that are available in Italian are based on lists of  
17 nonsense logatomes with a CVCV structure (Azzi, 1950), meaningful sentences (Cutugno et  
18 al., 2000) and syntactically fixed but meaningless sentences (Antonelli et al., 1977).

19 The main problem with most of the aforementioned tests is their limited accuracy  
20 which is due both to the comparatively small number of test items per test list (that is 5, 10  
21 or 20 items, Prosser & Martini, 2007) and to the variability in intelligibility across test items,  
22 which were not controlled during the design and construction of the tests (Antonelli et al.,  
23 1977). In addition, the limited accuracy in existing tests is related to the lack **of** the  
24 optimization of speech items in terms of intelligibility. In Cutugno et al. (2000) competition

1 noises (babble, traffic, pink and continuous speech) are recorded and provided together with  
2 sentences on two tracks of a CD. Optimization only consists in the equalization of the root  
3 mean square (RMS) of all speech items and noise signals. Furthermore, no information is  
4 available about perceptual equivalence of the test lists or reliability of the test. The effect  
5 due to the availability of only a small number of test items can partially be compensated for  
6 by combining test lists, e.g. performing adaptive test procedures and stopping the  
7 measurement after, e.g., 8-10 reversals of the adaptive track, or by using test items with  
8 several independent elements, such as, e.g., short sentences. Even then, the difference in  
9 redundancy (which is higher for meaningful sentences, lower for semantically correct, but  
10 meaningless sentences, and even lower for logatomes) again reduces the number of  
11 independent elements and hence the maximum achievable accuracy per time unit. The  
12 resulting variance in speech intelligibility makes the comparison of results between different  
13 listeners or within the same listener under different conditions difficult (e.g. before and after  
14 the application of a hearing aid; Prosser & Martini, 2007).

15 This work describes the development of a matrix sentence test in the Italian language  
16 which was set up in order to have an efficient and valid tool for the testing of speech  
17 intelligibility in noise. The developmental steps are compatible to those established for other  
18 matrix tests (Kollmeier et al., 2015), and a comparison between languages is therefore  
19 possible. Due to using semantically unpredictable sentences with a fixed syntactic structure  
20 (name-verb-number-noun-adjective; e.g. *Sofia trascina poche matite utili*, which is Italian for  
21 “Sophie drags a few useful pencils”), the test lists can be used for repeated measurements  
22 with the same listener and a high accuracy can thus be achieved with an appropriately high  
23 number of concatenated test lists. A further advantage of the matrix test is its possibility of  
24 using a closed-set response format: The listener may respond not by repeating the sentence

1 he/she heard but only has to press appropriate buttons in the response matrix. This makes  
2 the test suitable for testing a listener in her or his native language, even if the test  
3 administrator does not understand the language.

4 The test development procedure consisted selecting 50 words for a base matrix,  
5 recording the words while taking into account co-articulation effects, generating masking  
6 noise, optimizing the speech material by applying level adjustments, and then taking  
7 evaluation measurements. Finally, the Italian matrix test was compared with existing matrix  
8 tests to respond to the three main research aims. First, to understand if the Italian matrix  
9 test shares properties with matrix tests in other Romance languages. Second, to evaluate  
10 whether it is possible to observe the same training effect for open- and closed-set response  
11 formats, as in other languages. Third, to investigate the test-retest reliability of the speech  
12 reception thresholds (SRT), in comparison to other Matrix tests.

13

#### 14 SPEECH MATERIAL

15 In order to establish the 50-word base matrix, which consists of 10 names, 10 verbs, 10  
16 numerals, 10 adjectives, and 10 nouns, two- and three-syllabic words were selected from  
17 among the most frequently used words in the spoken Italian language (see Table 1). Since  
18 commonly used words were chosen (based on frequency dictionary of Bortolini et al., 1972),  
19 the listeners were familiar with the words of base matrix. This minimized the influence of the  
20 listener's linguistic competence on speech intelligibility. The phoneme distribution of the 50  
21 words in the base matrix was compared with a reference phoneme distribution of the Italian  
22 language taken from Tonelli et al. (1998). In the current study, singletone and geminate  
23 consonants were summarized as one phoneme class. The phoneme distribution of the base  
24 matrix was close to the reference distribution, with a maximum deviation of 2.2% for a  
25 phoneme /o/ (see Figure 1).

1 [FIGURE 1 about here]

2 By selecting the words from the sequence provided in the base matrix, grammatically  
3 correct but semantically unpredictable sentences were generated as a random combination  
4 of words from each word group (e.g. *Andrea manda molte tazze normali*, which is Italian for  
5 “Andrea sends many normal mugs”).

6 [TABLE 1 about here]

7

## 8 RECORDINGS, CUTTINGS AND RESYNTHESIS OF SENTENCES

9 The sentences were recorded according to the procedure proposed by Wagener et al. (1999  
10 c). One hundred sentences were generated and recorded, so that all the possible  
11 combinations of two consecutive words were included to capture the co-articulation  
12 between two successive words. The sentences were produced by a native Italian female  
13 speaker with standard Italian pronunciation. She was asked to pronounce words with a  
14 natural intonation and accentuation, and at a moderate constant speaking rate. The  
15 recordings were done in a sound-attenuated booth (fulfilling the requirements of ISO 8253-3,  
16 2012) using a Neumann 184 microphone with a cardioid characteristic and a Fireface UC  
17 soundcard with a sampling rate of 44.1 kHz and a resolution of 32 bits. The signals were  
18 saved on a PC hard-disc using Adobe Audition 2.0.

19 The recorded sentences were filtered with a 40Hz-high-pass filter and each sentence  
20 was set to the same root-mean-square level. Then, the sentences were cut into single words  
21 at a zero-crossing of the waveform, which resulted in 10 different realizations of each word  
22 of the base matrix. The initial cuttings were performed very close to the beginning of each  
23 word, while the final cut close was made to the beginning of the consecutive word in order  
24 to include the co-articulation of the consecutive word at the ending of the words. This means

1 that each realization included a different co-articulation at the end. Thirty test lists of ten  
 2 sentences were resynthesized for each list that contained all of the fifty words of the base  
 3 matrix. Each word realization was included 3 times in these 300 sentences. In order to  
 4 minimize the artefacts due to the resynthesis, individual overlapping times of between 0 and  
 5 20 ms were applied at the transitions between words.

6 The masking noise was generated through a 30-fold overlapping of all the sentences,  
 7 applying different silent intervals between sentences (for details see Wagener et al., 2003).  
 8 This resulted in a stationary noise with a long-term spectrum that matched the long-term  
 9 spectrum of the speech material.

10

#### 11 OPTIMIZATION MEASUREMENTS

12 Accurate speech intelligibility measurements require a speech recognition test with a steep  
 13 test-specific intelligibility function (e.g. Plomp & Mimpen, 1979; Kollmeier, 1990; Wagener et  
 14 al, 1999b). The slope of a test-specific intelligibility function ( $S50_{test}$ , Equation 1) can be  
 15 considered as the convolution of the mean slope of the word-specific intelligibility functions  
 16 ( $S50_{mean}$ ) and the distribution of the word-specific SRTs ( $\sigma_{SRT}$ ), as shown by Kollmeier (1990,  
 17 2015).

$$18 \quad S50_{test} \approx \frac{S50_{mean}}{\sqrt{1 + \frac{16S50_{mean}^2 \sigma_{SRT}^2}{(\ln(2e^{1/2} - 1 + 2e^{1/4}))^2}}} \quad (1)$$

19 where:  $S50_{mean}$  is the mean slope of the word-specific intelligibility functions and  $\sigma_{SRT}$  is the  
 20 standard deviation across all the word-specific SRTs.

1 The steepness of the test-specific function can be increased by decreasing the spread in the  
2 word-specific SRTs. The spread of the word-specific SRTs can be decreased by applying level  
3 adjustments, i.e. less intelligible words than the average ( $SRT_{\text{word}} > SRT_{\text{mean}}$ ) are increased in  
4 level whereas words of better intelligibility ( $SRT_{\text{word}} < SRT_{\text{mean}}$ ) are decreased in level. The  
5 optimization measurements were aimed at obtaining word-specific intelligibility functions  
6 with their parameters (i.e. word-specific SRT and S50).

7

#### 8 *Listeners*

9 Nineteen native Italian listeners participated in the optimization measurement procedure,  
10 which took place in Oldenburg, Germany. Their ages ranged from between 19 and 31, with a  
11 mean age of 23.9. The pure-tone threshold did not exceed 15 dB HL at octave frequencies of  
12 between 125 and 8000 Hz. The listeners had been in Germany for one year at most at the  
13 time of the measurements. They were all born and raised in Italy. The listeners were paid for  
14 participating in the measurements.

15

#### 16 *Procedure and equipment*

17 The Oldenburg Measurement Applications software (HörTech GmbH, Oldenburg,  
18 [www.hoertech.de](http://www.hoertech.de)) was used for the speech intelligibility measurements. Speech and noise  
19 signals were presented monaurally to the listeners' preferred ear by means of free-field  
20 equalized headphones (Sennheiser model HDA200). The measurement setup was calibrated  
21 to dB SPL using Brüel & Kjær instruments, i.e. artificial ear type 4153, microphone type 4134,  
22 preamplifier type 2669 and amplifier type 2610.

23 Thirty base lists of ten sentences each were constructed, considering that each list  
24 contained all the words of the basic matrix in different combinations and thus was

1 phonetically balanced with respect to the phoneme distribution of the Italian language  
2 reported in Tonelli et al. (1998). Prior to the first measurement session, the listeners were  
3 familiarized with the speech material through a presentation of two test lists. The first  
4 training list was presented without any interferer at 65 dB SPL. The second training list was  
5 presented at a fixed SNR of 0 dB, which resulted in an intelligibility of almost 100%. After the  
6 training session, speech intelligibility was measured at fixed SNRs in the -18 dB to 4 dB range  
7 in 2 dB steps. The sound pressure level of the background noise was kept constant at 65 dB  
8 SPL, and was started and ended 500 ms before and after presentation of the sentence  
9 presentation. Fifty ms rising and falling ramps were applied to the noise signal (using a Hann  
10 window) to prevent abrupt signal onset and offset. The order of the sentences in a list, the  
11 SNR, and the list index were randomized across listeners. The measurements were  
12 conducted with the open-set response format, in which the listener's task was to repeat the  
13 words he/she understood and the test administrator marked the correct responses on a  
14 display. The responses were stored using word-scoring, indicating that each word in a  
15 sentence was scored separately.

16 In order to obtain the word-specific speech intelligibility functions, a logistic model  
17 function (Equation 2) was fitted to the measured data (SI) using a maximum likelihood  
18 procedure:

$$19 \quad SI_{word}(SNR) = \frac{100}{1 + e^{4.550(SRT - SNR)}} \quad (2)$$

20 where  $SI_{word}$  is the intelligibility function of the word.

21

22 *Results*

1 The optimization measurements resulted in a mean word-specific SRT of  $-8.3 \pm 3.7$  dB SNR  
2 and a median slope of 17.7 %/dB over all of the 500 word realizations. The test-specific slope  
3 was predicted by means of Equation 1 and resulted in 9.2 %/dB. Eleven realizations of words  
4 were excluded from the final test material. With each excluded realization a whole base list  
5 also had to be excluded, which resulted in 12 base lists remaining at the end of the  
6 optimization procedure. Realizations were excluded for which no adequate fitting was  
7 possible or whose SRTs differed considerably from the general SRT of the respective word.  
8 Included word realizations did not deviate more than 8.5 dB from the average word-specific  
9 SRT.

10 In order to homogenize the intelligibility of the speech material, level adjustments  
11 were applied to each remaining word realization (384 out of 500). The level adjustments  
12 were limited to  $\pm 3$  dB to preserve a natural intonation of the optimized sentences, which  
13 was judged by two native Italian listeners. The level adjustments and list exclusions resulted  
14 in a mean SRT of  $-8.3 \pm 1.4$  dB SNR and a median slope of 18.0 %/dB over the remaining 384  
15 word realizations included in the 12 remaining lists. In other words, the standard deviation of  
16 the word-specific SRT was decreased by 2.3 dB, which resulted in the test-specific slope  
17 becoming steeper, that is, from 9.2 %/dB to 15.2 %/dB, according to Equation 1.  
18 Table 2 summarizes the measured and predicted values that were obtained from the  
19 optimization procedure.

20 [TABLE 2 about here]

21

## 22 EVALUATION MEASUREMENTS

23 The evaluation measurements had various objectives. Besides verifying the characteristics of  
24 the optimized speech material, proving the equivalence of the base lists remaining after the

1 optimization procedure, and establishing reference data for normal-hearing listeners, the  
2 training effect was addressed, as investigated for other matrix tests.

3

#### 4 *Listeners*

5 Fifteen native Italian listeners were tested in Torino (9 female and 6 male subjects, mean age  
6 28) and 11 listeners in Ferrara (5 female and 6 male subjects, mean age 23) using the open-  
7 set response format. The hearing status of the listeners who participated in the  
8 measurements in Ferrara was assessed via self-reporting. Normal hearing of the listeners  
9 measured in Torino was **proven** by means of pure tone audiometry. The pure tone thresholds  
10 did not exceed 20 dB HL at octave frequencies from 125 to 8000 Hz. The training effect, using  
11 the closed-set response format, was evaluated with a separate group of 10 listeners in  
12 Ferrara (5 female and 5 male subjects, mean age 24 years).

13

#### 14 *Procedure*

15 The measurement setup in Torino and Ferrara consisted of a notebook with an earbox 'ear  
16 3.0' sound card (Auritec, Hamburg, Germany) and free-field equalized Sennheiser HDA200  
17 headphones. The measurement setup used in Torino was calibrated in the same way and  
18 with the same equipment as described in the optimization measurements part. In Torino, a  
19 type 2260 amplifier was used instead of a type 2610 amplifier. The measurements in Torino  
20 took place in a **sound-treated booth** that complied with **ANSI S3.1-1999 (R2008)**, while a  
21 room with low background noise ( $L_{eq} = 43.3$  dB) in the University building was selected in  
22 Ferrara. **All the evaluation measurements were conducted monaurally. Each listener could**  
23 **indicate at which of both ears all measurements should be performed.**

1           The training effect was evaluated both in a closed- and open-set response format. In  
2 the closed-set response format, after the listener listens to the sentence, he/she was given a  
3 digital interface that showed a panel containing the 50 words of the base matrix: in this way,  
4 the listener can indicate the words they have understood on the panel. Instead, in the  
5 open-set response format the subject has to repeat the words he/she has understood and  
6 the experimenter has to indicate the correctly repeated words on a display. The SRTs were  
7 measured, in order to evaluate the training effect, using an adaptive procedure described by  
8 Brand and Kollmeier (2002) with six double lists of 20 sentences (consisting of all the 12 base  
9 lists available after optimization). The initial SNR in the adaptive procedure was set at 0 dB,  
10 the noise level was fixed at 65 dB SPL and the speech signal was varied to converge to 50% of  
11 intelligibility. The answers were stored using word-scoring, i.e. each word in a sentence was  
12 scored separately. The SRT was estimated using a maximum likelihood procedure. The order  
13 of the test lists was randomized across listeners.

14           In order to evaluate list equivalence, six double lists (the same as those used in the  
15 training session) were presented to each listener at fixed SNRs of -4.5 dB SNR, -7 dB SNR and  
16 -9.5 dB SNR corresponding to recognition rates of about 80 %, 50 % and 20 %, respectively.  
17 The noise level was kept constant at a level of 65 dB SPL. This part of evaluation  
18 measurements was performed in Torino with 11 out of 15 listeners participating previously in  
19 the training effect measurements. List-specific intelligibility functions for each of 12 base lists  
20 were obtained by fitting the logistic model function (Equation 2) to the mean intelligibility  
21 scores averaged across all listeners.

22

## 23 RESULTS

### 24 *Training effect*

1 The mean SRTs and corresponding standard deviations are shown in Figure 2 as a function of  
2 the sequence of measurements. The SRTs measured with both the open- and closed-  
3 response formats are shown.

4 [FIGURE 2 about here]

5 A mixed design repeated-measures analysis of variance (ANOVA) was conducted for  
6 the open-set response format with the measurement site as the between group factor (two  
7 levels) and the sequence of measurements as the within group factor (six levels). No  
8 statistical difference was found between the two measurement sites ( $F(1, 24) = 0.41, p =$   
9  $0.526$ ), but a statistically significant main effect of the temporal measurement order ( $F(5,$   
10  $120) = 80.29, p < 0.001$ ) was found as well as a significant interaction between the  
11 measurement order and the test site ( $F(5, 120) = 3.05, p = 0.019$ ). A separate mixed design  
12 repeated-measures ANOVA was conducted, with the response format as the between group  
13 factor (two levels) and the sequence of measurements as the within group factor (six levels).  
14 Mauchly's test was carried out and it indicated that the assumption of sphericity had been  
15 violated for the main effect of the test type,  $\chi^2(2) = 27.20, p = 0.018$ , and the degrees of  
16 freedom were therefore corrected using Greenhouse-Geisser estimates of sphericity  
17 ( $\epsilon = 0.756$ ). A statistical difference was found between the open- and closed-set response  
18 formats ( $F(1, 34) = 14.33; p = 0.001$ ) and for the main effect of the measurement order  
19 ( $F(3.78, 128.46) = 73.17; p < 0.001$ ). ANOVA revealed no significant interaction between the  
20 response format and measurement order ( $F(3.78, 128.46) = 0.48; p = 0.738$ ). The largest  
21 improvement in SRT, that is, of 1.2 dB for the open-set response format and 1.1 dB for the  
22 closed-set response format, was observed between the first and the second measurements.  
23 It decreased to 0.5 dB for the open-set response format and 0.3 dB for the closed-set  
24 response format between the second and third measurement. From the third measurement

1 onwards, only small improvements were found. Therefore, as for other languages, the  
2 reference data was obtained by separately averaging the SRTs of the third measurement to  
3 the last one, , for the open- and closed-set response formats. This resulted in a mean SRT of -  
4  $6.8 \pm 0.8$  dB SNR for the open-set response format and of  $-7.3 \pm 0.8$  dB SNR for the closed-set  
5 response format.

6 The test-retest reliability was calculated in the same way as for the French matrix test  
7 (Jansen et al., 2012), that is, on the basis of repeatable measured SRT with an adaptive  
8 procedure. Only SRTs from the third to sixth measurements were considered to account for  
9 the training effect. Within-subject variabilities of 0.5 dB and 0.6 dB were found for the open-  
10 and the closed-set response formats, respectively.

### 11 *Base list equivalence*

12 The mean intelligibility scores measured with the open-set format at 3 SNRs and the fitted  
13 list-specific intelligibility functions of the 12 base lists of 10 sentences each are summarized  
14 in Table 3. The mean list-specific SRT and slope were  $-7.3 \pm 0.2$  dB SNR and  $13.3 \pm 1.2$  %/dB,  
15 respectively. The lowest and the highest SRTs across lists were  $-7.6$  dB SNR and  $-7.2$  dB SNR,  
16 respectively, while the lowest and highest slopes across the lists were  $11.1$  %/dB and  $15.6$   
17 %/dB, respectively. Although the slope on the intelligibility function for the closed-set format  
18 was not measured in this study, according to Hochmuth et al. (2012) it is expected that no  
19 significant difference is found between the two formats.

20 [TABLE 3 about here]

21 In order to examine the equivalence of the test lists and determine the standard deviation  
22 across listeners, the logistic function was also fitted separately for each listener and each  
23 list. Repeated measures ANOVA with SRT and S50 as the main factors revealed no statistical

1 differences in terms of SRT  $F(11, 110) = 1.6, p = 0.11$  and S50  $F(11, 110) = 1.64, p = 0.098$ . The  
2 mean SRT and S50 averaged across the listeners and lists were  $-7.4 \pm 0.9$  dB SNR and  
3  $14.3 \pm 3.6$  %/dB, respectively.

4

## 5 DISCUSSION

### 6 *Evaluation*

7

8 The optimization of the speech material applied to the word realizations decreased the  
9 variability of the word-specific SRTs by 2.3 dB (from 3.7 dB to 1.4 dB) and thus, according to  
10 Kollmeier's probabilistic model (1990, 2015), increased the predicted test-specific slope by  
11 6 %/dB (from 9.2 %/dB to 15.2 %/dB). The predicted increase in the test-specific slope for the  
12 optimized speech material was confirmed in the evaluation measurements. The measured  
13 mean list-specific slope was 13.3 %/dB, which is 4.1 %/dB higher than the one obtained for  
14 the speech material prior to optimization. This high slope of the Italian matrix test qualifies it  
15 for accurate and efficient speech intelligibility measurements.

16 The mean list-specific slope is highly comparable to those obtained for matrix tests in other  
17 Romance languages, i.e. for Spanish (13.2 %/dB; Hochmuth et al., 2012) and for French  
18 (14.0 %/dB; Jansen et al., 2012), and is close to those of other languages, such as Russian  
19 (13.8 %/dB, Warzybok et al., 2015) or Danish (12.6 %/dB; Wagener et al., 2003). Higher test-  
20 specific intelligibility function slopes were found for the German and Polish matrix tests  
21 (slope of 17.1 %/dB in both cases, Wagener et al., 1999a; Ozimek et al., 2010). The  
22 differences in slope across languages may be related to the specific speaker's characteristics  
23 or to the capability of discrimination of phonemes in noise which may be different from  
24 language to language. Even though the slopes for the Italian, Spanish and French matrix tests

1 are remarkably similar, they are too close to the values of the other languages to be  
2 distinguishable as a separate entity.

3         A comparison with the existing Italian speech recognition tests is difficult or even  
4 impossible for several reasons. For example, the development of the speech material with  
5 meaningless sentences (Antonelli, 1977) was based on statistical criteria that only accounted  
6 for usage, frequency, and dispersion of the words; however, the speech material was not  
7 optimized in terms of intelligibility. In addition, Prosser & Martini (2007) argued that the  
8 existing Italian audiometry tests reveal a high variability in intelligibility scores since only a  
9 small number of items are used clinically. Finally, some of the papers about the development  
10 of Italian speech recognition tests are only available in a very limited printed version, and are  
11 therefore difficult to have access to.

12

### 13 *Training effect and base list equivalence*

14 As far as the temporal measurement effect is concerned, denoted in the following as  
15 “training effect”, the present work focused on both open- and closed-set response formats  
16 which resulted in findings comparable to previous matrix tests (Hochmuth et al., 2012;  
17 Warzybok et al., 2015). As for other languages, independent of the response format, the  
18 major improvement in SRT was observed between the first two measurements and it then  
19 decreased to a value below 1 dB after the second measured list (see Kollmeier et al., 2015).  
20 Since no interaction between the temporal order and response format was found, it can be  
21 concluded that the amount of training required to obtain stable results is the same for both  
22 response formats. This is again in agreement with the matrix tests for other languages  
23 (Kollmeier et al., 2015). Furthermore, it can be postulated that the training effect is language  
24 independent and related to the test structure. Following the recommendation for other

1 languages, two test lists of 20 sentences each are recommended to account for training in  
2 order to obtain stable and repeatable results.

3 For the open-set response format, a significant interaction of temporal measurement  
4 and test site was found. It is related to fact that up to the fifth measurement the SRTs  
5 measured in Torino and Ferrara were very close to each other, whereas they slightly differed  
6 in the sixth measurement. For the last training list, listeners measured in Ferrara showed on  
7 average 0.8 dB lower SRT than listeners from Torino. However, this difference is in the range  
8 of the test accuracy (defined by the standard deviation of the reference SRT., It can therefore  
9 be assumed as being irrelevant from an audiological point of view. The mean SRT for the  
10 open-set response format was 0.7 dB higher than the mean SRT for the closed-set response  
11 format. This difference between the two response formats was again in line with previous  
12 findings by Hochmuth et al. (2012) for the Spanish matrix test or by Warzybok et al. (2015)  
13 for the Russian matrix test, which showed differences of 1 dB and 0.6 dB, respectively. These  
14 findings indicate that the visual presentation of word alternatives which is available in the  
15 closed-set response format may help a listener to better recognize the words of a sentence  
16 that were previously presented acoustically, thus lower SRTs can be achieved. **In clinical**  
17 **settings**, the close-set version is mainly recommended for patients of a different native  
18 language than the test instructor. The measurement in open-set format takes usually less  
19 time **than** in the closed-set format. Therefore for a clinical **practice**, when the native language  
20 of a patient and a test instructor is the same, the open-set format is recommended. The  
21 reference data obtained from the adaptive measurements ( $-6.8 \pm 0.8$  dB SNR and  $-7.3 \pm 0.8$   
22 dB SNR for the open and close-set response formats, respectively) are close to those of the  
23 Spanish test in both the open- and closed-set response formats ( $-6.2 \pm 0.8$  and  $-7.2 \pm 0.7$ ,  
24 respectively).

1           The high test-retest reliability of the Italian matrix test (0.5 dB for the open- and  
2 0.6 dB for the closed-set response format) is very close to the reliability of the French matrix  
3 test (0.4 dB for the closed-set response format; Jansen et al., 2012) and of the Russian matrix  
4 (0.6 dB for the open- and 0.5 dB for the closed-set response formats; Warzybok et al., 2015).

5           The evaluation measurements have also confirmed the equivalence of the test lists.  
6 Neither SRT nor S50 differed significantly across test lists. Furthermore, the small difference  
7 in SRTs between the test lists of 0.2 dB is on average comparable with the findings of matrix  
8 tests in the other languages which showed a standard deviation across test lists of between  
9 0.1 dB and 0.2 dB (see Kollmeier et al., 2015 for an overview). Furthermore, the differences  
10 across test lists for SRT and S50 are smaller than the differences across normal-hearing  
11 listeners, which again indicates a high homogeneity of the speech material between the test  
12 lists. This is an effective improvement to the available tests for speech audiometry in Italian,  
13 in which the results are less accurate because of the high variability of the number of items  
14 per list (Prosser & Martini, 2007).

15

## 16 CONCLUSIONS

17 The matrix sentence test has been developed for the Italian language to allow measurements  
18 to be made in an open-set response format with an experimenter present, as well as in a  
19 self-administered closed-set response format. The values obtained from the evaluation  
20 measurements, i.e., reference data for adaptive measurements, parameters of the  
21 psychometric function, the test-list equivalence, training effect and test-retest reliability,  
22 have been shown to be similar to the values obtained in matrix tests in other languages.

23           The adaptive measurements that were introduced resulted in a reference SRT of -  
24  $6.8 \pm 0.8$  dB SNR for the open-set and  $-7.3 \pm 0.8$  dB SNR for the closed-set response formats,

1 respectively. The measurements at fixed SNRs for the determination of the psychometric  
2 function of the Italian matrix test resulted in an SRT of -7.3 dB SNR and a slope of 13.3 %/dB.  
3 It was possible to obtain a high test list equivalence with a standard deviation in SRT across  
4 test lists of 0.2 dB.

5           Moreover, the test has yielded a high test-retest reliability of 0.5 dB for the open-set  
6 and 0.6 dB for the closed-set response formats.

7

#### 8 ACKNOWLEDGEMENT

9 This work was supported by the EFRE-project HurDig and by the Cluster of Excellence  
10 Hearing4All of the University of Oldenburg. The authors are grateful to Prof. Giancarlo  
11 Pecorari and Luca Raimondo for their availability in testing the pure tone thresholds of the  
12 listeners in Torino and for the opportunity of conducting measurements in a clinical  
13 environment. We would also like to thank Rossana Carta for the data collecting in  
14 Oldenburg.

15

#### 16 REFERENCES

17 ANSI/ASA S3.1-1999 (R2008): Maximum permissible ambient noise levels for audiometric  
18 test rooms. Washington, D.C.: American National Standards Institute.

19

20 Antonelli, A.R., Barocci, R., Mantovani, M. 1977. Un nuovo materiale vocale in lingua italiana:  
21 le frasi sintetiche. *Nuovo Arch It Otol* 5, 1-13.

22

- 1 Azzi, A. 1950. Prove di acumetria vocale per la lingua italiana. *Arch It Otol* 5, 45-84.
- 2
- 3 Bocca, E. & Pellegrini, A. 1950. Studio statistico sulla composizione della fonetica della lingua  
4 italiana e sua applicazione pratica all'audiometria con la parola. *Arch It Otol* 5, 116-141.
- 5
- 6 Bortolini, U., Tagliavini, C., Zampolli, A. 1972. *Lessico di frequenza della lingua italiana*: First  
7 edition. Milano, Italy: Garzanti
- 8
- 9 Brand, T. & Kollmeier, B. 2002. Efficient adaptive procedures for threshold and concurrent  
10 slope estimates for psychophysics and speech intelligibility tests. *J Acoust Soc Am* 111, 2801-  
11 2810.
- 12
- 13 Cutugno, F., Prosser S. Turrini M. 2000. *Audiometria vocale – vol. I*, ed. GN. ReSound Italia.
- 14
- 15 European Commission, 2012. EUROPEANS AND THEIR LANGUAGES. Special Eurobarometer  
16 386, [http://ec.europa.eu/public\\_opinion/archives/ebs/ebs\\_386\\_en.pdf](http://ec.europa.eu/public_opinion/archives/ebs/ebs_386_en.pdf).
- 17
- 18 Hochmuth, S., Brand, T., Zokoll, M.A., Zenker Castro, F., Wardenga, N., Kollmeier, B. 2012. A  
19 Spanish matrix sentence test for assessing speech reception thresholds in noise. *Int J Audiol*  
20 51, 536–544.

1

2 Hochmuth, S., Jürgens, T., Brand, T., Kollmeier, B. 2014. Multilinguale Cocktailparty – Einfluss  
3 von sprecher- und sprachspezifischen Faktoren auf die Sprachverständlichkeit im Störschall  
4 (Influence of speaker and language on speech intelligibility in noise). *Proceedings of 17th*  
5 *congress of the German Society of Audiology, Oldenburg, Germany,.*

6

7 ISO 8253-3:2012 Acoustics — Audiometric test methods — Part 3: Speech audiometry

8

9 Jansen, S., Luts, H., Wagener, K.C., Kollmeier, B., Del Rio, M., Dauman, R., James, C., Fraysse,  
10 B., Vormès, E., Frachet, B., Wouters, J., van Wieringen, A. 2012. Comparison of three types of  
11 French speech-in-noise tests: A multi-center study. *Int J Audiol* 51, 164–173.

12

13 Kollmeier, B. 1990. Messmethodik, Modellierung und Verbesserung der Verständlichkeit von  
14 Sprache (in German). (Methodology, modeling, and improvement of speech intelligibility  
15 measurmeents). Habilitation, Universität of Göttingen.

16

17 Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M., Uslar, V., Brand, T., Wagener, K.C.  
18 (2015) The multilingual matrix test: principles, applications and comparison across languages  
19 – a review. *Conditionally accepted by Int J Audiol.*

20

1 Lewis, M. Paul, Gary F. Simons, Charles D. Fennig (eds.). 2014. *Ethnologue: Languages of the*  
2 *World, Seventeenth edition*. Dallas, Texas: SIL International. Online version:  
3 <http://www.ethnologue.com>.

4

5 Ozimek, E., Warzybok, A., Kutzner, D. 2010. Polish sentence matrix test for speech  
6 intelligibility measurement in noise. *Int J Audiol* 49, 444–454.

7

8 Plomp, R. & Mimpen, A.M. 1979. Improving the reliability of testing the speech reception  
9 threshold for sentences. *Audiol* 18, 43-53.

10

11 Prosser, S., and Martini, A. 2007. *Argomenti di Audiologia*, Omega Ed. Torino

12

13 Tonelli, L., Panzeri, M., and Fabbro, F. 1998. Un'analisi statistica della lingua italiana parlata.  
14 *Studi Italiani di Linguistica Teorica e Applicata* 3, 501-514.

15

16 Turrini, M., Cutugno, F., Maturi, P., Prosser S., Leoni, F. A., Arslan, E. 1993. Bisyllabic words for  
17 speech audiometry: a new italian material. *Acta Otorhinolaryngol Ital*, 13(1), 63-77.

18

- 1 Wagener, K., Brand, T., Kollmeier, B. 1999a. Entwicklung und Evaluation eines Satztests für die  
2 deutsche Sprache Teil III: Evaluation des Oldenburger Satztests (in German). *Z Audiol* 38, 86–  
3 95.
- 4
- 5 Wagener, K., Kühnel, V., Kollmeier, B. 1999b. Entwicklung und Evaluation eines Satztests in  
6 deutscher Sprache I: Design des Oldenburger Satztests (in German). *Z Audiol* 38, 4–15.
- 7
- 8 Wagener, K., Jøsvassen, J. L., Ardenkjaer, R. 2003. Design, optimization, and evaluation of a  
9 Danish sentence test in noise. *Int J Audiol* 42, 10–17.
- 10
- 11 Warzybok, A., Zokoll, M., Wardenga, N., Ozimek, E., Boboshko, M. et al. (2015). Development  
12 of the Russian Matrix Sentence Test. *Int J Audiol*, doi:10.3109/14992027.2015.1020969.

## 1 List of tables

2 Table 1: Basic word matrix of the Italian matrix sentence test. Words in bold indicate an  
3 example of one randomly built up sentence.

<i>Names</i>	<i>Verbs</i>	<i>Numerals</i>	<i>Nouns</i>	<i>Adjectives</i>	<i>English translation</i>
<b>Sofia</b>	compra	due	<b>scatole</b>	azzurre	<i>Sofia buys two light-blue boxes.</i>
Marco	vuole	poche	matite	piccole	<i>Marco wants a few small pencils.</i>
Anna	prende	quattro	tazze	normali	<i>Anna takes four normal cups.</i>
Sara	<b>dipinge</b>	cinque	pietre	nuove	<i>Sara paints five new stones.</i>
Chiara	vede	<b>molte</b>	tavole	belle	<i>Chiara sees many nice desks.</i>
Maria	cerca	sette	palle	bianche	<i>Maria looks for seven white balls.</i>
Luca	trascina	otto	macchine	grandi	<i>Luca drags eight big cars.</i>
Andrea	regala	nove	sedie	<b>utili</b>	<i>Andrea donates nine useful chairs.</i>
Matteo	possiede	dieci	bottiglie	nera	<i>Matteo owns ten black bottles.</i>
Simone	manda	venti	porte	rosse	<i>Simone sends twenty red doors.</i>

4

5 Table 2: Mean results from the optimization procedure regarding word-specific SRTs and  
6 their standard deviation ( $SD_{SRT\text{words}}$ ), as well as word-specific slopes ( $S50_{\text{words}}$ ) and predicted  
7 test-specific slopes ( $S50_{\text{test}}$ ) according to Equation 1 before and after level adjustment and  
8 test list selection. The mean list-specific SRT and slope ( $S50_{\text{test}}$ ) measured in the evaluation  
9 procedure are also given.

	Optimization			Evaluation
	Before level adjustments (500 word realizations)	Before level adjustments (384 word realizations)	After level adjustment (384 word realizations)	List-specific results
<b>SRT /dB SNR</b>	-8.3	-8.3	-8.3	-7.3
<b><math>SD_{SRT\text{words}}</math> / dB SNR</b>	3.7	3.4	1.4	-
<b><math>S50_{\text{words}}</math> / %/dB</b>	17.7	18.0	18.0	-
<b><math>S50_{\text{test}}</math> / %/dB</b>	9.2	9.7	15.2	13.3

10

11 Table 3: Mean list-specific intelligibility scores measured at SNRs of -9.5, -7, and -4.5 dB SNR  
12 and list-specific SRT and S50 with mean SRT and S50 averaged across 12 base lists.

13

List	Scores [%]			SRT [dB SNR]	S50 [%/dB]
	-9.5 dB SNR	-7.0 dB SNR	-4.5 dB SNR		
<b>1</b>	23.1	62.5	80.9	-7.5	13.4
<b>2</b>	22.9	52.9	78.7	-7.2	12.6
<b>3</b>	21.1	58.9	85.6	-7.5	15.6
<b>4</b>	26.5	56.5	85.8	-7.6	14.0
<b>5</b>	20.5	52.9	82.4	-7.2	14.5
<b>6</b>	22.7	54.9	78.5	-7.2	12.6
<b>7</b>	21.1	53.8	79.8	-7.2	13.5
<b>8</b>	25.8	55.5	82.7	-7.5	13.1
<b>9</b>	27.1	52.5	77.5	-7.3	11.1
<b>10</b>	23.6	54.0	83.5	-7.4	13.9
<b>11</b>	26.4	52.7	79.3	-7.3	11.8
<b>12</b>	26.2	54.9	84.0	-7.5	13.4
			<b>MEAN</b>	<b>-7.3</b>	<b>13.3</b>
			<b>SD</b>	<b>0.2</b>	<b>1.2</b>

1

2

3

4

1 List of figures

2 Figure 1: Phoneme distribution of the Italian matrix test (gray squares) and the reference  
3 phoneme distribution for the Italian language (black triangles). The phonemes have been  
4 transcribed using the International Phonetic Alphabet symbols.

5

6 Figure 2: Mean SRTs and corresponding standard deviations of the six subsequent training  
7 measurements for the open-set response format (measurements from Torino, T, with  
8 squares, measurements from Ferrara, F, with circles) and for the closed-set response format  
9 (triangles) as a function of the measurements sequence.