Taylor & Francis
Taylor & Francis Group

International Journal of Audiology 2015; Early Online: 1–7

International Journal of Audiology

Report

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

Giuseppina Emma Puglisi*, Anna Warzybok†, Sabine Hochmuth†, Chiara Visentin‡, Arianna Astolfi*, Nicola Prodi‡ & Birger Kollmeier#

*Department of Energy, Politecnico di Torino, Torino, Italy, †Medizinische Physik, Universität Oldenburg, and Cluster of Excellence Hearing4all, Oldenburg, Germany ‡Department of Engineering, Università di Ferrara, Ferrara, Italy, #Medizinische Physik, Universität Oldenburg, Cluster of Excellence Hearing4all, and HörTech gGmbH, Oldenburg, Germany

The British Society of Audiology

The International Society of Audiology

NAS NORDIC AUDIOLOGICAL SOCIETY

## Abstract

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

**Key Words:** Speech perception; speech audiometry; speech reception threshold; Italian speech recognition test



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

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

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

The main problem with most of the aforementioned tests is their limited accuracy which is due both to the comparatively small number of test items per test list (that is 5, 10, or 20 items, Prosser & Martini, 2007) and to the variability in intelligibility across test items, which were not controlled during the design and construction of the tests (Antonelli et al, 1977). In addition, the limited accuracy in existing tests is related to the lack of the optimization of speech items in terms of intelligibility. In Cutugno et al (2000) competition noises (babble, traffic, pink, and continuous speech) are recorded and provided together with sentences on two tracks of a CD. Optimization only consists in the equalization of the root mean square (RMS) of all speech items and noise signals. Furthermore, no information is available about perceptual equivalence of the test lists or reliability of the test. The effect due to the availability of only a small number of test items can partially be compensated for by combining test lists, e.g. performing adaptive test procedures and stopping the measurement after, e.g. 8–10 reversals of the adaptive track, or by using test items with several independent elements, such as, e.g. short sentences.

Correspondence: Giuseppina Emma Puglisi, Politecnico di Torino, Department of Energy, Corso Duca degli Abruzzi, 24, 10129, Torino, Italy. E-mail: giuseppina.puglisi@polito.it

(Received 6 February 2015; accepted 7 June 2015)

ISSN 1499-2027 print/ISSN 1708-8186 online 
DOI: 10.3109/14992027.2015.1061709

**Abbreviations:**

| | |
|---|---|
| ANOVA | Analysis of variance |
| SRT | Speech reception threshold |
| SNR | Signal-to-noise ratio |
| S50 | Slope of intelligibility function |

Even then, the difference in redundancy (which is higher for meaningful sentences, lower for semantically correct, but meaningless sentences, and even lower for logatomes) again reduces the number of independent elements and hence the maximum achievable accuracy per time unit. The resulting variance in speech intelligibility makes the comparison of results between different listeners or within the same listener under different conditions difficult (e.g. before and after the application of a hearing aid; Prosser & Martini, 2007).

This work describes the development of a matrix sentence test in the Italian language which was set up in order to have an efficient and valid tool for the testing of speech intelligibility in noise. The developmental steps are compatible to those established for other matrix tests (Kollmeier et al, 2015), and a comparison between languages is therefore possible. Due to using semantically unpredictable sentences with a fixed syntactic structure (name-verb-number-noun-adjective; e.g. *Sofia trascina poche matite utili*, which is Italian for 'Sophie drags a few useful pencils'), the test lists can be used for repeated measurements with the same listener and a high accuracy can thus be achieved with an appropriately high number of concatenated test lists. A further advantage of the matrix test is its possibility of using a closed-set response format: The listener may respond not by repeating the sentence he/she heard but only has to press appropriate buttons in the response matrix. This makes the test suitable for testing a listener in her or his native language, even if the test administrator does not understand the language.

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

## Speech material

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

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

## Recording, cutting, and resynthesis of sentences

The sentences were recorded according to the procedure proposed by Wagener et al (1999 c). One hundred sentences were generated and recorded, so that all the possible combinations of two consecutive words were included to capture the co-articulation between two successive words. The sentences were produced by a native Italian female speaker with standard Italian pronunciation. She was asked to pronounce words with a natural intonation and accentuation, and at a moderate constant speaking rate. The recordings were done in a sound-attenuated booth (fulfilling the requirements of ISO 8253-3, 2012) using a Neumann 184 microphone with a cardioid characteristic and a Fireface UC soundcard with a sampling rate of 44 100 Hz and a resolution of 32 bits. The signals were saved on a PC hard-disc using Adobe Audition 2.0.

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



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

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

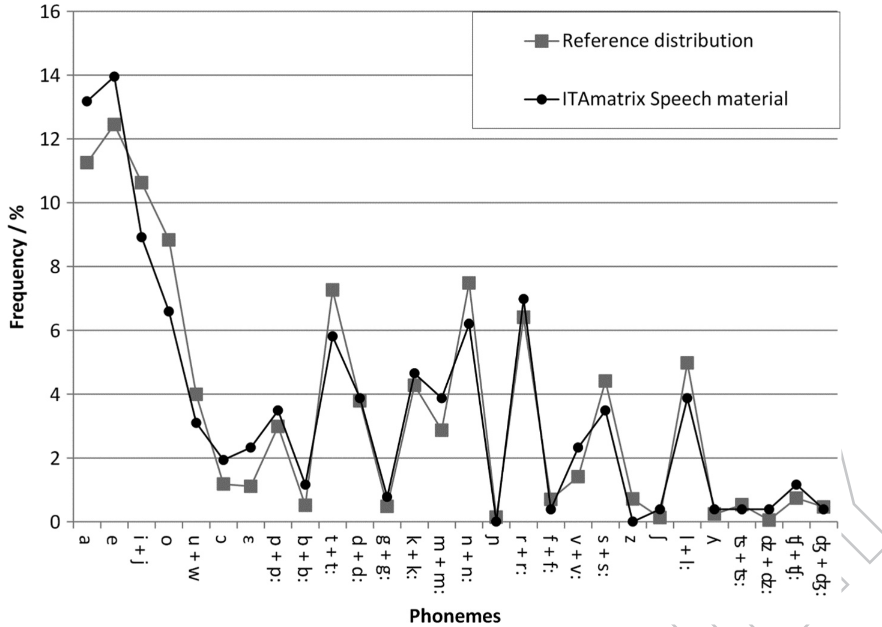


**Figure 1.** Phoneme distribution of the Italian matrix test (gray squares) and the reference phoneme distribution for the Italian language (black circles). The phonemes have been transcribed using the International Phonetic Alphabet symbols.

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

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

## Optimization measurements

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

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

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

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

### *Listeners*

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

### *Procedure and equipment*

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

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

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

$$SI_{word}(SNR) = \frac{100}{1 + e^{4S50(SRT\text{-}SNR)}} \qquad (2)$$

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

### *Results of the optimization procedure*

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

In order to homogenize the intelligibility of the speech material, level adjustments were applied to each remaining word realization (384 out of 500). The level adjustments were limited to ± 3 dB to preserve a natural intonation of the optimized sentences, which was judged by two native Italian listeners. The level adjustments and list exclusions resulted in a mean SRT of − 8.3 ± 1.4 dB SNR and a median slope of 18.0 %/dB over the remaining 384 word realizations included in the 12 remaining lists. In other words, the standard deviation of the word-specific SRT was decreased by 2.3 dB, which resulted in the test-specific slope becoming steeper, that is, from 9.2 %/dB to 15.2 %/dB, according to Equation 1.

Table 2 summarizes the measured and predicted values that were obtained from the optimization procedure.

## Evaluation measurements

The evaluation measurements had various objectives. Besides verifying the characteristics of the optimized speech material, proving the equivalence of the base lists remaining after the optimization procedure, and establishing reference data for normal-hearing listeners, the training effect was addressed, as investigated for other matrix tests.

### *Listeners*

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

### *Procedure*

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

The training effect was evaluated both in a closed- and open-set response format. In the closed-set response format, after the listener listens to the sentence, he/she was given a digital interface that showed a panel containing the 50 words of the base matrix: in this way, the listener can indicate the words they have the understood on the panel. Instead, in the open-set response format the subject has to repeat the words he/she has understood and the experimenter has to

**Table 2.** Mean results from the optimization procedure regarding word-specific SRTs and their standard deviation ($SD_{SRTwords}$), as well as word-specific slopes ($S50_{words}$) and predicted test-specific slopes ($S50_{test}$) according to Equation 1, before and after level adjustment and test list selection. The mean list-specific SRT and slope ($S50_{test}$) measured in the evaluation 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* |
| SRT / dB SNR | − 8.3 | − 8.3 | − 8.3 | − 7.3 |
| $SD_{SRTwords}$ / dB SNR | 3.7 | 3.4 | 1.4 | – |
| $S50_{words}$ / %/dB | 17.7 | 18.0 | 18.0 | – |
| $S50_{test}$ / %/dB | 9.2 | 9.7 | 15.2 | 13.3 |

indicate the correctly repeated words on a display. The SRTs were measured, in order to evaluate the training effect, using an adaptive procedure described by Brand and Kollmeier (2002) with six double lists of 20 sentences (consisting of all the 12 base lists available after optimization). The initial SNR in the adaptive procedure was set at 0 dB, the noise level was fixed at 65 dB SPL and the speech signal was varied to converge to 50% of intelligibility. The answers were stored using word-scoring, i.e. each word in a sentence was scored separately. The SRT was estimated using a maximum likelihood procedure. The order of the test lists was randomized across listeners.

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

### *Results of the evaluation procedure*

#### Training effect

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

A mixed design repeated-measures analysis of variance (ANOVA) was conducted for the open-set response format with the measurement site as the between group factor (two levels) and the sequence of measurements as the within group factor (six levels). No statistical difference was found between the two measurement sites ($F(1, 24) = 0.41$, $p = 0.526$), but a statistically significant main effect of the temporal measurement order ($F(5, 120) = 80.29$, $p < 0.001$) was found as well as a significant interaction between the measurement order and the test site ($F(5, 120) = 3.05$, $p = 0.019$). A separate mixed design repeated-measures ANOVA was conducted, with the response format as the between group factor (two levels) and the sequence of measurements as the within group factor (six levels). Mauchly's test was carried out and it indicated that the assumption of sphericity had been violated for the main effect of the test type, $\chi2(2) = 27.20$, $p = 0.018$, and the degrees of freedom were therefore corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.756$). A statistical difference was found between the open- and closed-set response formats ($F(1, 34) = 14.33$; $p = 0.001$), and for the main effect of the measurement order ($F(3.78, 128.46) = 73.17$; $p < 0.001$). ANOVA revealed no significant interaction between the response format and measurement order ($F(3.78, 128.46) = 0.48$; $p = 0.738$). The largest improvement in SRT, that is, of 1.2 dB for the open-set response format and 1.1 dB for the closed-set response format, was observed between the first and the second measurements. It decreased to 0.5 dB for the open-set response format and 0.3 dB for the closed-set response format between the second and third measurement. From the third measurement onwards, only small improvements were found. Therefore, as for other languages, the reference data was obtained by separately averaging the SRTs of the third measurement to the last one, for the open- and closed-set response formats. This resulted in a mean SRT of − 6.8 ± 0.8 dB SNR for the open-set response format and of − 7.3 ± 0.8 dB SNR for the closed-set response format.

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

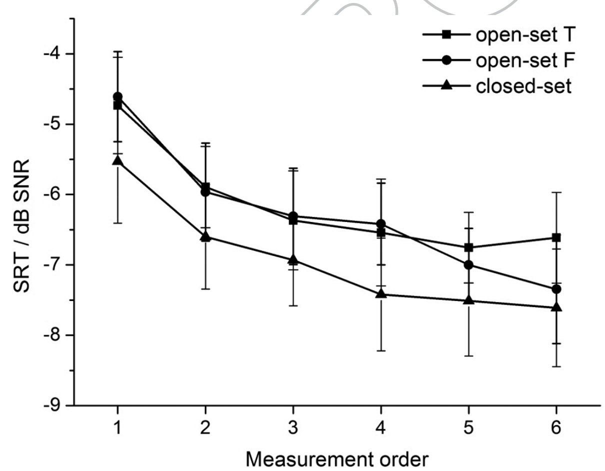


**Figure 2.** Mean SRTs and corresponding standard deviations of the six subsequent training measurements for the open-set response format (measurements from Torino, T, with squares; measurements from Ferrara, F, with circles) and for the closed-set response format (triangles) as a function of the measurements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

#### Base list equivalence

The mean intelligibility scores measured with the open-set format at three SNRs and the fitted list-specific intelligibility functions of the 12 base lists of 10 sentences each are summarized in Table 3. The mean list-specific SRT and slope were − 7.3 ± 0.2 dB SNR and 13.3 ± 1.2 %/dB, respectively. The lowest and the highest SRTs across lists were − 7.6 dB SNR and − 7.2 dB SNR, respectively,

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

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

while the lowest and highest slopes across the lists were 11.1 %/dB and 15.6 %/dB, respectively. Although the slope on the intelligibility function for the closed-set format was not measured in this study, according to Hochmuth et al (2012) it is expected that no significant difference is found between the two formats.

In order to examine the equivalence of the test lists and determine the standard deviation across listeners, the logistic function was also fitted separately for each listener and each list. Repeated measures ANOVA with SRT and S50 as the main factors revealed no statistical differences in terms of SRT $F(11, 110) = 1.6$, $p = 0.11$ and S50 $F(11, 110) = 1.64$, $p = 0.098$. The mean SRT and S50 averaged across the listeners and lists were $-7.4 \pm 0.9$ dB SNR and $14.3 \pm 3.6$ %/dB, respectively.

## Discussion

### *Evaluation*

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

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

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

### *Training effect and base list equivalence*

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

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

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

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

## Conclusions

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

The adaptive measurements that were introduced resulted in a reference SRT of $-6.8 \pm 0.8$ dB SNR for the open-set and $-7.3 \pm 0.8$ dB SNR for the closed-set response formats, respectively. The measurements at fixed SNRs for the determination of the psychometric function of the Italian matrix test resulted in an SRT of $-7.3$ dB SNR and a slope of 13.3 %/dB. It was possible to obtain a high test list equivalence with a standard deviation in SRT across test lists of 0.2 dB.

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


## Acknowledgements

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

***Declaration of interest:*** The authors report no declarations of interest. Birger Kollmeier serves as the scientific director of HörTech gGmbH (www.hoertech.de), a non-profit organization owned in majority by Universität Oldenburg. Copyright of the speech material is held by HörTech gGmbH.


## References


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

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

Azzi A. 1950. Prove di acumetria vocale per la lingua italiana. *Arch It Otol*, 5, 45–84.

Bocca E. & Pellegrini A. 1950. Studio statistico sulla composizione della fonetica della lingua italiana e sua applicazione pratica all'audiometria con la parola. *Arch It Otol*, 5, 116–141.

Bortolini U., Tagliavini C. & Zampolli A. 1972. *Lessico di frequenza della lingua italiana*: First edition. Milano, Italy: Garzanti.

Brand T. & Kollmeier B. 2002. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *J Acoust Soc Am*, 111, 2801–2810.

Cutugno F., Prosser S. & Turrini M. 2000. Audiometria vocale - vol. I, ed. GN ReSound Italia.

European Commission, 2012. Europeans and their languages. Special Eurobarometer 386, http://ec.europa.eu/public_opinion/archives/ebs/ebs_386_en.pdf.

Hochmuth S., Brand T., Zokoll M.A., Zenker Castro F., Wardenga N. et al. 2012. A Spanish matrix sentence test for assessing speech reception thresholds in noise. *Int J Audiol*, 51, 536–544.

Hochmuth S., Jürgens T., Brand T. & Kollmeier B. 2014. Multilingual cocktail party: Einfluss von sprecher- und sprachspezifischen Faktoren auf die Sprachverständlichkeit im Störschall (influence of speaker and language on speech intelligibility in noise). *Proceedings* of 17th congress of the German Society of Audiology, Oldenburg, Germany.

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

Jansen S., Luts H., Wagener K.C., Kollmeier B., Del Rio M. et al. 2012. Comparison of three types of French speech-in-noise tests: A multi-center study. *Int J Audiol*, 51, 164–173.

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

Kollmeier B., Warzybok A., Hochmuth S., Zokoll M., Uslar V. et al. 2015. The multilingual matrix test: Principles, applications, and comparison across languages: A review. *Conditionally accepted by Int J Audiol*.

Lewis M.P., Simons G.F. & Fennig C.D. (eds.). 2014. *Ethnologue: Languages of the World, Seventeenth edition.* Dallas, USA: SIL International. Online version: http://www.ethnologue.com.

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

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

Prosser S. & Martini A. 2007. *Argomenti di Audiologia*, Torino: Omega Ed.

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

Turrini M., Cutugno F., Maturi P., Prosser S., Leoni F.A. et al. 1993. Bisyllabic words for speech audiometry: A new Italian material. *Acta Otorhinolaryngol Ital*, 13(1), 63–77.

Wagener K., Brand T. & Kollmeier B. 1999a. Entwicklung und Evaluation eines Satztests für die deutsche Sprache Teil III: Evaluation des Oldenburger Satztests (in German). *Z Audiol* 38, 86–95.

Wagener K., Kühnel V. & Kollmeier B. 1999b. Entwicklung und Evaluation eines Satztests in deutscher Sprache I: Design des Oldenburger Satztests (in German). *Z Audiol*, 38, 4–15.

Wagener K., Josvassen J.L. & Ardenkjaer R. 2003. Design, optimization, and evaluation of a Danish sentence test in noise. *Int J Audiol*, 42, 10–17.

Warzybok A., Zokoll M., Wardenga N., Ozimek E., Boboshko M. et al. 2015. Development of the Russian matrix sentence test. *Int J Audiol*, doi:10.3109/14992027.2015.1020969.