POTENTIALS OF A MATRIXED WORD TEST FOR ASSESSING SPEECH RECEPTION IN ROOMS WITH REVERBERATION AND NOISE

Chiara Visentin[1], Nicola Prodi[2]

Engineering Department, University of Ferrara, Ferrara, Italy
[1] chiara.visentin@unife.it
[2] nicola.prodi@unife.it

Abstract

The use of listening tests based on list of sentences is well established in audiology for evaluating hearing loss. The testing material is usually presented diotically under simplified listening conditions (anechoically, in quiet or with stationary noise).

In this work the efficacy of a matrixed word test in the Italian language was assessed for the description of the speech reception performance in complex noisy and reverberant conditions.

The speech material was chosen phonetically balanced; the test sentences were composed by a carrier phrase followed by a sequence of syntactically uncorrelated meaningful target words. Tests were proposed in a closed-set format to 18 native Italian speakers, reporting normal hearing. Speech–intelligibility scores and response times were measured under noise and reverberation, and in a quiet control condition which served to set the performance baseline. In the former condition a speech-shaped steady-state noise was selected as masker and SNR and reverberation were combined to create a listening scenario with an STI=0.48. The tests were administered in a silent room via a three-dimensional audio rendering system. The length of the sequences was varied from two up to six word items, as to select the length able to provide the best performance in resolving the listening conditions.

The results show that optimal results are obtained by the four words test. The outcomes were additionally compared with existing speech recognition tests already available in the Italian language (DRT and matrix sentence test).

Keywords: intelligibility, response time, speech perception, room acoustics.

PACS no. 43.55.Hy, 43.71.Gv

1 Introduction

In real-life listening environments the acoustic cues needed to recognize the speech signal are reduced, due to presence of noise and reverberation and communication becomes harder. A higher degree of attention is required at the perceptual level: the bottom-up representation of the signal is poor and more top-down processing (relating the heard signal to the existing knowledge stored in semantic long-term memory) is required to compensate. The limited cognitive resources are mainly engaged in the lower-level processes of deciphering the signal, leaving few resources for higher-level language processing (elaboration, storing, retrieval and memorization) [1, 2]. The increased allocation of cognitive resources and attention required when listening in unfavorable conditions is perceived as an effortful process [3].

During the years, a variety of listening tests assessing speech reception have been developed, mainly designed for clinical applications and evaluation of the hearing loss of impaired listeners. The speech material is usually presented diotically, in the presence of stable and predictable background noise; the signal is clearly articulated and lacking the changes in modulation caused by the presence of room reverberation. Therefore, in this manner the presented listening situations hardly reflect the acoustic features of real listening environments where spoken communication usually occurs [4] and the test outcomes will not be able to fully account for all the cognitive processing involved in real-word communication.

This work deals with the potentials of a matrixed word test in the Italian language in effectively describing speech reception in complex noisy and reverberant conditions. In order to be an efficient and useful tool, the test should be able to grasp and reflect the perceptual impact of different kind of auditory interferences (background noise, reverberation) typical of real world listening. Such a resource is needed in room acoustics, to understand the perceptual impact of noise and reverberation and control them in the acoustical design of public spaces (e.g., classrooms, conference venue, restaurants…).

In the set-up of the matrixed word test, the following criteria have been followed:

a. *Selection of the test material*. As the test is targeted to real world listening situations, it was decided to avoid the use of nonsense words or the presentation of words in isolation (rather than as part of a sentence, as occurs everyday communication). On the other hand, the presence of a sentence context can help to overcome intelligibility problems: integrating the perceived words into a meaningful sentence facilitates word recall [5, 6]. Recently, the matrix sentence test in the Italian language was developed [7], based on sentences with correct and fixed syntax, but little semantic predictability because of little or no supportive context. In order to increase the cognitive load put on participants and gain a deeper insight on the speech perception mechanisms, the new matrixed word test was based on list of meaningful words not correlated by a syntactic structure.

b. *Measure of the response time to the auditory stimuli*. The matrixed word test outputs are intelligibility scores and response times: speech perception is thus described using both accuracy and speed of the speech processing rate. Response time accounts for top-down processing and reflects the amount of cognitive resources required to interpret and respond to the incoming signal [8]. Furthermore, response time can be considered as a behavioral measure, with a single-task paradigm, of listening effort [3]. The matrixed word test is then administered in a closed-set format, allowing for the retrieval of RT data.

In the following, the speech material selected for the listening test is described. Afterwards, the work focuses on the selection of the most suitable number of target words to propose within the test sequence, as to obtain optimized results. Tests were conducted in two listening conditions, varying the length of the test sequences from two up to six words.

2 Selection of the speech material

The test stimuli were created starting from the *corpus* of words of the Diagnostic Rhyme Test (DRT) for the Italian language [9]. The test is based on meaningful disyllabic words, selected to match the language-specific phonemic distribution and optimized as regards word familiarity. The words are organized according to six perceptually distinctive features of the initial consonant (nasal, continuant, strident, coronal, anterior and sonorant); an additional group is present, accounting for the remaining distinctive consonant features of the Italian language. The words can be further gathered according to the combination of the initial consonant with the following vowel, classified according to the tongue position: anterior (i, e), posterior (o, u) and central (a).

Within the DRT *corpus* a subset of 42 words was selected, respecting the perceptual features distribution. It is then ensured that the new speech material is phonetically balanced and representative of the main linguistic feature of the Italian language.

The words were then organized to create the test sequences. Each sequence included a carrier phrase and six target words arranged in a fixed order, a priori establishing the succession of the vowel contexts; specifically, the following series was chosen: a – i/e – o/u – i/e – o/u – a. The choice allows for a decrease in the list length up to 3 words, still keeping an even representation of the vowel contexts.

The test sequences were recorded by an adult native Italian female speaker, who was instructed to speak at the rate of conversational speech. The recordings took place in a silent room with an omnidirectional microphone at a sampling rate of 44.1 kHz.

Each sequence was filtered as to match the long-term spectrum of a female talker suggested in the IEC 60268-16 standard [10]. Sequences with less than six target words were obtained from the recordings by progressively discarding the last items of the sequence while keeping a natural playback.

Finally, the recorded sequences were organized in lists, composed by 13 items each; for each sequence length, two lists were created. Within a test list all the available words were evenly represented.

3 Methods

3.1 Participants

Sixteen subjects (7 male and 9 female) took part in the experiment. Their age ranged from 21 to 35 years (average: 27.0 yr, σ: 5.0 yr). They were all Italian native speaker and reported normal hearing. All participants possessed the same level of qualification and were familiar with other types of listening tests.

3.2 Speech signal and background noise

The test sequences were presented to the participants in two listening conditions: “quiet” (no background noise, no reverberation) and “reverb+noise”.

In both conditions, the speech signal was reproduced at a level of 63 dB(A) at the listening position. In the “reverb+noise” condition a background noise was introduced to produce an energetic masking of the target words; the noise is steady-state and spectrally shaped to match the long term spectrum of a female talker [10]. Furthermore, speech was convolved with the impulse response of a frontal speaker in a simulated room with T_{mid}= 0.94 s. The noise was convolved with the sum of the impulse responses from four source positions in the same room after broadband mixing of the phases in the resulting impulse response in order to loose the directional properties of noise (diffuse noise condition). The revereberated noise was proposed at a fixed SNR of -3 dB; the resulting STI value is 0.47.

3.3 Procedure

The experiment was performed in a quiet sound-treated room; speech stimuli and noise were reproduced via a three-dimensional audio rendering system surrounding the listener. During the test, the participant were seated in the center of the room; in front of them was placed a touch screen, to be used for the items selection.

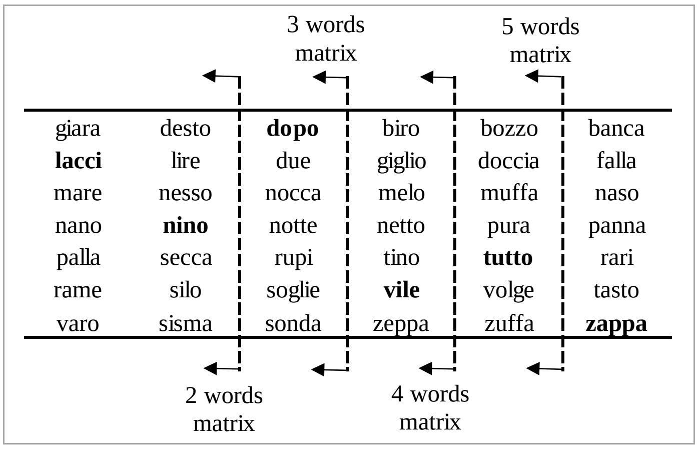
In order to familiarize the participants with the speech material, a training session, presented at a fixed SNR of 10 dB in stationary noise and anechoic conditions, was proposed prior to the experiment. Afterwards, each participant conducted 10 listening tests (5 list length x 2 listening conditions). To minimize the influence of fatigue, sequential and learning effect, acoustic conditions and list lengths were balanced across the participants, using a balanced Latin square.

3.3.1 Speech intelligibility measurements

During the test, participants listened to a sequence of words; when background was proposed, it started almost 1000 ms before the carrier phrase and ended simultaneously with the final word.

After the last item had been presented, a panel with the 42 words matrix of the test was shown on the touch screen. In Figure 1, the whole word matrix is reported.

Figure 1 – Word matrix of the listening test; words in bold characters show an example of a randomly built up sequence of six words.


3 words matrix
5 words matrix
2 words matrix
4 words matrix

giara	desto	**dopo**	biro	bozzo	banca
lacci	lire	due	giglio	doccia	falla
mare	nesso	nocca	melo	muffa	naso
nano	**nino**	notte	netto	pura	panna
palla	secca	rupi	tino	**tutto**	rari
rame	silo	soglie	**vile**	volge	tasto
varo	sisma	sonda	zeppa	zuffa	**zappa**

The matrix is organized as follows: each of the six columns contains the seven alternative words of a specific vowel context (e.g., the first column is composed by the words belonging to the “a” vowel context), organized in alphabetical order; the seven words belong to likewise distinctive consonant features. As shown in Figure 1, when the number of words of the test sequence was reduced, the number of matrix columns was varied accordingly.

Participants had to mark the identified words in serial order (the same order as that of the item presentation). It was not possible to change the responses once they had been entered. Once a word was selected in each of the columns, participants confirmed their choice by pressing the “*Continue*” button and the next sequence was automatically presented.

The percentage of correctly recognized words was used as a measure for the speech intelligibility (IS). Results for each participant were averaged across the 13 sequences composing a listening test.

3.3.2 Response time measurements

For the present analysis, the response time (RT) is defined as the time between the end of the waveform of the last word presented and the selection of the first word on the touch screen, therefore corresponding to the response time of the first target word. Response times of the following items were also acquired, defined as the time interval between two successive word selections.
As there was no time constrain on the participants' responses, several unrealistic long RTs were found, probably due to participants' inattention [11]. An absolute cutoff of 8 s was set; higher RT values were discarded during data analysis and considered missing data. On the whole, 68 RTs (1.45% of all the RTs) were rejected.

4 Results

4.1 Percentage of correct responses

Figure 2 illustrates how the different list lengths influence the intelligibility scores, in both "quiet" and "reverb+noise" conditions. As expected, in both conditions a decreasing trend of IS with increasing list length is observed; the presence of noise and reverberation affects the results by decreasing the percentage of correct responses.

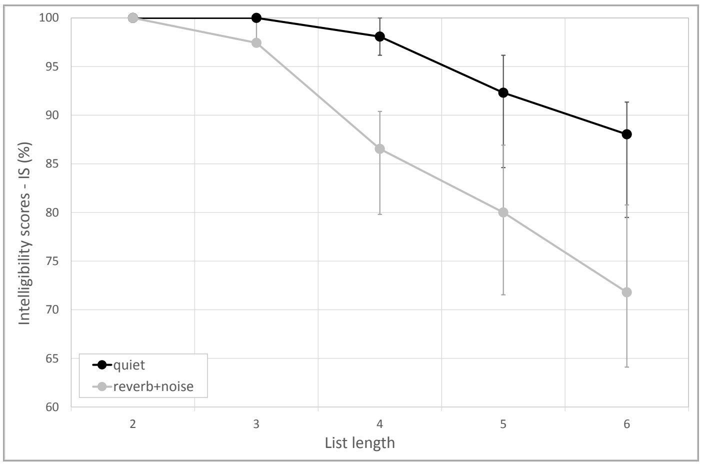


Figure 2 – Median speech-intelligibility scores across participants as a function of the number of words in the sequence. Error bars represent interquartile ranges.

To test the hypothesis that the presence of noise and reverberation affects IS scores, pairwise Wilcoxon comparisons of the word scores obtained in "quiet" condition and the scores obtained in "reverb+noise" condition were made for each of the list length. Non parametric test were performed because data are not normally distributed (Shapiro-Wilk test). The tests indicated that IS was significantly worse in unfavorable conditions for the lists of 4 words (p<0.001), 5 words (p<0.001) and 6 words (p<0.001) but not for the lists of 2 (p=0.25) and 3 words (p=0.06).
When the listening condition is kept constant, the presence of differences between different list lengths was tested with a Friedman test, showing in both cases a significant effect (p<0.001). In the "quiet" condition *post-hoc* Wilcoxon test showed that no differences are present in the IS scores between lists of 2, 3 and 4 words, due to a ceiling effect in the responses; significant differences were instead found when the list length is increased to 5 and 6 words (p=0.01 for all comparisons). On the other hand, in

unfavorable listening conditions, significant differences were found between all list lengths, except for the lists of 2 and 3 words (*p*=0.23).

4.2 Response time

Response time results are presented in Figure 3 for the two listening conditions. Shapiro-Wilk tests showed that RT data are not normally distributed; indeed, in general, the RT distribution is positively skewed: it rises rapidly on the left and has a long positive tail on the right [11]. Non parametric statistical test were used to describe the results. It is worth noticing the high interquartile ranges of the RT values, originated by the aggregation of individual data with considerable differences in mean value and dispersion [11]; an additional source of variation is represented by the difference in cognitive strategies implemented by the participants to recall the words list.

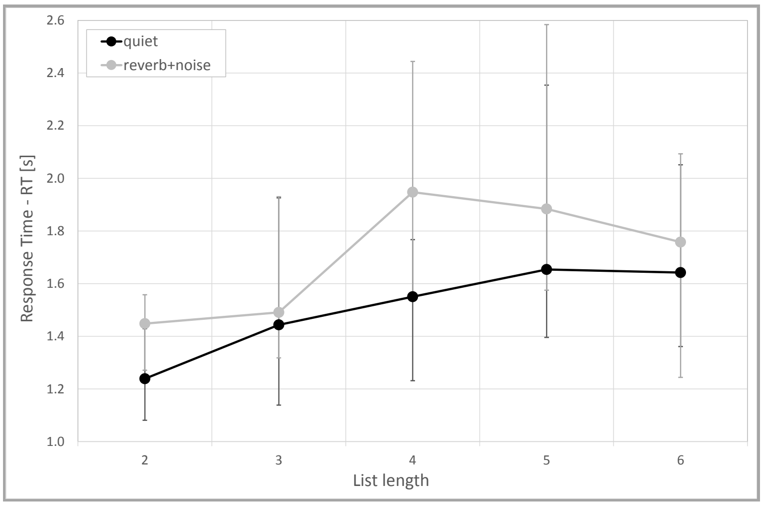


Figure 3 – Median response times across participants as a function of the number of words of the sequence. Error bars represent interquartile ranges.

The significance of the differences between list lengths was tested with a Friedman test followed by Wilcoxon *post-hoc* comparisons. The results of the statistical analysis are presented in Table 1. The statistical analysis shows that in “quiet” conditions RT weakly increases from 2 to 3 words (*p*=0.09), experiment a *plateau* between 3 and 4 words (*p*=1), further increases between 4 and 5 words (*p*=0.02) and finally weakly decreases between 5 and 6 words (*p*=0.09). The pattern modifies with reverberation and noise: RT becomes similar between 2 and 3 words lists (*p*=0.36) and increases between 3 and 4 words (*p*=0.02) where a ceiling effect shows up (*p*=1, between 4 and 5 words lists). Adding a further word leads to a significant decrease of RT (*p*=0.04).

Table 1 – Results of the statistical analysis for RT: differences between sequences length in “quiet” and “reverb+noise” listening conditions. Comparisons with *p* values significant at α=0.05 level are highlighted with a gray background.

	quiet				reverb+noise			
	2 w	3 w	4 w	5 w	2 w	3 w	4 w	5 w
3 words	0.09				0.36			
4 words	0.01	1			0.01	0.02		
5 words	0.01	0.04	0.02		0.01	0.02	1	
6 words	0.01	0.59	0.65	0.09	0.73	1	1	0.04

To examine the effect of reverberation and noise on RT, pairwise Wilcoxon comparisons were made for each of the sequence length. The analysis showed a significant difference between listening conditions for list length 2 (*p*=0.02), 4 (*p*<0.001) and 5 (*p*<0.001); no effects were found for the lists of 3 (*p*=0.6) and 6 words (*p*=0.95).

5 Discussion

The goal of this study was to compare different list lengths and select the one most suitable for efficiently discriminate between the listening conditions. Intelligibility scores were taken into account, as well as response times; the latter quantity, reflecting the cognitive effort put into the listening process, provides complementary information next to IS. Specifically, three requirements were set: achieve the full IS in quiet conditions, maximize the difference between listening conditions with IS and maximize the same difference with RT.

In the "quiet" condition the test with 2, 3 and 4 words are easy enough to guarantee the highest accuracy (IS indistinguishable from 100%). Indeed, in this listening condition, where the absence of external masking ensures a correct hearing of the spoken words, the accuracy result is mainly explained by the (limited) capacity of the verbal short-term memory store [12]. Recalling the results in Figure 1, it can be seen that at short list lengths (up to 4 words) the proportion of words recalled is close to the maximum: all the words of the sequence can be correctly recalled, irrespective of their serial position. At longer list lengths, the recall process starts to heavily tap on the memory store: significant effects of the serial position appear (higher percentage of correct recall for the first and last items of a sequence) and yield a decrease in the overall intelligibility scores.

As regards the decrease of IS when the speech material is proposed in unfavorable conditions, the greater effect is found for the test with four target words. The relative difference between listening conditions is 13%. The reason is twofold: on one hand, the presence of noise and reverberation decreases intelligibility on the acoustic-perceptual level, on the other hand, it affects higher cognitive level of processing [2]. Specifically, the storage of the words to be recalled is influenced, yielding significant differences in the IS of items presented in different serial position [13]. In order to explore the serial position effect in the four-words listening test, IS values were calculated for the single items (Figure 4) and Friedman tests, followed by Wilcoxon paired comparisons, was performed to test the presence of significant differences between the words.

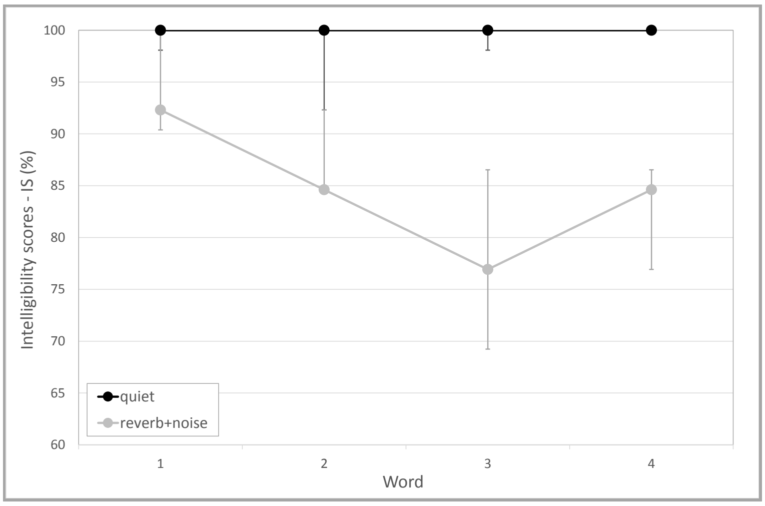


Figure 4 – Median speech-intelligibility scores across participants as a function of the single words in the listening test with four target words. Error bars represent interquartile ranges.

No significant differences were found between the four words in quiet (*p*=0.27). In presence of noise and reverberation differences show up, reflecting significant *primacy* and *recency* effects in words recognition [12]. Pairwise Wilcoxon comparisons of IS values in both conditions were made for each of the words, showing that the presence of noise decrease IS values starting from the second word of the list (word 1: *p*=0.08; word 2: *p*=0.002; word 3: *p*<0.001; word 4: *p*<0.001). Then, the effect of unfavorable conditions on IS seems to be a less efficient retrieval: the increased allocation of cognitive resources for words identification leaves less resources for rehearsal and encoding.

Finally, concerning response time, it is found that the difference between listening conditions is maximized with the four-word sequence, the relative difference being equal to 26%. The response time proves to be a parameter more sensitive than IS, with a relative difference between conditions twice as much the difference calculated with IS. In Figure 5, RTs of the single words are depicted for the test with four words.

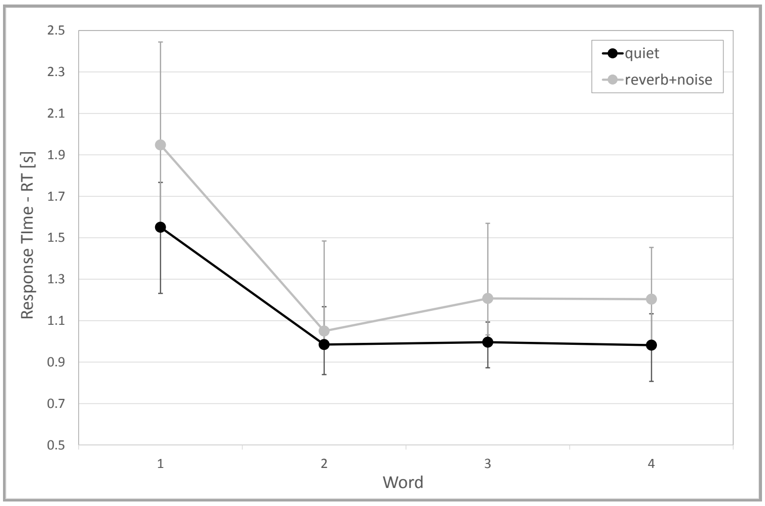


Figure 5 – Median response times (RT) across participants as a function of the single words in the listening test with four target words. Error bars represent interquartile ranges.

The statistical analysis shows that a similar processing strategy is used for words recognition in both listening conditions, being the RT of the first word significantly different from the RT of the remaining words (*p*=0.003 in quiet; *p*=0.005 in unfavorable conditions). Response times of words 2, 3 and 4 are not significantly different. This finding indicates that resolving the words sequence is more or less completed by the time the first word is selected; at least in quiet conditions, RTs of the words 2 to 4 seems to relate to the time needed to find and select the word (already decided) among the proposed alternatives. Paired comparisons Wilcoxon tests showed that noise and reverberation increase the response time of all words, expect the second (word 1: *p*<0.001; word 2: *p*=0.1; word 3: *p*<0.001; word 4: *p*=0.02). Integrating IS and RT results, it can be said that the more the listeners understood correctly, the faster they responded, thus indicating that greater understanding of the presented words facilitates recall [5]. The reason could be related to the fact that processing speech in unfavorable conditions may require the listener to consider more alternative interpretations of the stimuli, prolonging the matching process necessary to reach a decision [14].

5.1 Comparison with other listening tests

The results of the matrixed word test with four words were compared with the results of other speech-in-noise tests currently available in the Italian language, aiming at understanding if the new test allows for a better resolution in discriminating between listening conditions.

Specifically, two listening tests were selected: the DRT test [9] and the matrix sentence test [7]. The listening tests were proposed to two different groups of 18 participants each, in the same acoustic conditions selected for the previous matrixed word test. Non parametric statistics was used to test the significance of the differences between listening conditions. The results are summarized in Table 2

Table 2 – Median IS and RT results for the three listening tests: matrixed word test, DRT test and matrix sentence test. The relative difference between the median values in the proposed listening conditions is reported, together with the *p* values of the Wilcoxon pairwise test.

	IS (%)				**RT [s]**			
	quiet	reverb + noise	*p* value	relative difference	quiet	reverb + noise	*p* value	relative difference
matrixed word	98.1	85.6	<0.001	12.7 %	1.55	1.95	<0.001	25.8 %
DRT	100.0	92.1	<0.001	7.9 %	1.43	1.60	0.002	11.9 %
matrix sentence	100.0	97.5	0.01	2.5 %	1.38	1.71	0.001	23.9 %

All listening tests yield, in the “quiet” condition, intelligibility scores close to the maximum. The decrease due to noise and reverberation is the highest in the matrixed word test and the smallest in the matrix sentence test, where the presence of a syntactic structure backs up the word recognition. Both listening tests involve a combination of an auditory bottom-up processing task and a cognitive (working memory) task but the lack of the syntactic structure calls for a less efficient word retrieval, highlighting subtler differences between acoustic conditions.

Concerning the response time, a direct comparison between the listening tests cannot be performed being the tests based on different cognitive mechanisms. Anyway, looking at the relative differences, it can be noticed that the lowest values are found for the DRT test: this is probably the result of the task being relatively easy and thus requiring a limited engagement of cognitive resources (at least for adult, normal hearing participants). The relative difference is the highest for the matrixed test, even though a large enough value if found also for the matrix sentence test; anyway, when the effect size is calculated, a larger effect appears for the former test (*e.s.*: 0.60 for the matrixed word test, *e.s.*: 0.50 for the matrix sentence test).

6 Conclusions

The purpose of this work was to evaluate the efficacy of a matrixed word test in the Italian language for testing speech reception in noisy and reverberant environments.

A series of listening tests was performed, in order to determine the optimal number of target words, according to the following criteria: maximization of IS in quiet conditions and maximization of the difference between quiet and realistic conditions for both IS and RT. The only words sequence simultaneously satisfying all the requirements, was the one with four target words, that was therefore selected for further applications. An additional comparison of the matrixed word test with existing speech recognition tests available in the Italian language, showed that the new test allows for a better resolution in discriminating listening conditions, in terms of both intelligibility scores and response time.

The results here presented are encouraging; the new test promises to be a reliable and effective way to assess speech perception in complex listening situations for normal hearing adults. Furthermore, its flexible structure (that can be composed by different sequences of target words) potentially allows its use with different categories of listeners, accounting for their cognitive development (e.g., children under 14 years), language knowledge (e.g., foreign language speakers) or hearing impairment.

Further research is underway where the matrixed word test with four target words is applied to a variety of real-life listening environments.

References

[1] Rudner, M.; Lunner, T. Cognitive spare capacity and speech communication: a narrative overview. *BioMed research international*, 2014.

[2] Ljung, R.; Israelsson, K.; Hygge, S. Speech Intelligibility and Recall of Spoken Material Heard at Different Signal-to-noise Ratios and the Role Played by Working Memory Capacity. *Applied Cognitive Psychology*, Vol. 27, 2013, pp. 198-203.

[3] McGarrigle, R.; Munro, K.J.; Dawes, P.; Stewart, A.J.; Moore, D.R.; Barry, J.G.; Amitay, S. Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper'. *International journal of audiology*, 2014.

[4] Brungart, D.S.; Sheffield, B.M.; Kubli, L.R. Development of a test battery for evaluating speech perception in complex listening environments. *Journal of the Acoustical Society of America*, Vol. 136, 2014, pp. 777-790.

[5] Uslar, V.N.; Carroll, R.; Hanke, M.; Hamann, C.; Ruigendijk, E.; Brand, T.; Kollmeier, B. Development and evaluation of a linguistically and audiologically controlled sentence intelligibility test. *Journal of the Acoustical Society of America*, Vol. 134, 2013, pp. 3039-3056.

[6] Carroll, R.; Ruigendijk, E. The effects of syntactic complexity on processing sentences in noise. *Journal of psycholinguistic research*, Vol. 42, 2013, pp. 139-159.

[7] Puglisi, G. E., Warzybok, A., Hochmuth, S., Visentin, C., Astolfi, A., Prodi, N., Kollmeier, B. An Italian matrix sentence test for the evaluation of speech intelligibility in noise. *International journal of audiology*, Vol. 54, 2015, 44-50.

[8] Houben, R.; van Doorn-Bierman, M.; Dreschler, W.A. Using response time to speech as a measure for listening effort. *International journal of audiology*, Vol. 52, 2013, pp. 753-761.

[9] Bonaventura, P.; Paoloni, F.; Canavesio, F.; Usai, P. Realizzazione di un test diagnostico di intelligibilità per la lingua italiana ("Development of a diagnostic intelligibility test in the Italian language"), *Internal Technical Report No. 3C1286*, Fondazione Ugo Bordoni, Rome, 1986.

[10] IEC 60268-16: 2011: *Sound system equipment – Part 16: Objective rating of speech intelligibility by speech transmission index*, Belgium, 2011.

[11] Whelan, R. Effective analysis of reaction time data. *Psychol. Rec.*, Vol 58, 2008, pp 475-482.

[12] Ward, G.; Tan, L.; Greenfell-Essam, R. Examining the relationship between free recall and immediate serial recall: the effects of list length and output order. *Journal of experimental psychology*, Vol. 36, 2010, pp. 1207-1241.

[13] Kjellberg, A.; Ljung R.; Hallman, D. Recall of words heard in noise. *Applied cognitive psychology*, Vol. 22, 2008, pp. 1088-1098.

[14] Ljung, R. Room acoustics and cognitive load when listening to speech. Ph.D. Thesis, University of Gävle, Sweden, 2010.