A matrixed speech-in-noise test to discriminate favorable listening conditions by means of intelligibility and response time results.

Journal:	*Journal of Speech, Language, and Hearing Research*
Manuscript ID	JSLHR-H-17-0418.R1
Manuscript Type:	Research Article
Date Submitted by the Author:	24-Jan-2018
Complete List of Authors:	Visentin, Chiara; University of Ferrara, Department of Engineering Prodi, Nicola; Univ Ferrara, Dept. Engineering
Keywords:	Acoustics, Noise, Speech perception

A matrixed speech-in-noise test to discriminate favorable listening conditions by means of intelligibility and response time results.

Chiara Visentin[a] and Nicola Prodi[b]

Department of Engineering

University of Ferrara

Via Saragat 1, 44122 Ferrara, Italy

[a] chiara.visentin@unife.it

[b] nicola.prodi@unife.it

24 January 2018

Abstract

Purpose: The primary aim of this study was to develop and examine the potentials of a new speech-in-noise test in discriminating the favorable listening conditions targeted in the acoustical design of communication spaces. The test is based on the recognition and recall of disyllabic words sequences. A secondary aim was to compare the test with current speech-in-noise tests, assessing its benefits and limitations.

Methods: Young adults (19-40 years), self-reporting normal hearing, were presented with the newly developed words sequence test WST (16 participants, Experiment 1), with a consonant confusion test and with a sentence recognition tests (Experiment 2, 36 participants randomly assigned to the two tests). Participants performing the WST were presented with words sequences of different length (from two up to six words). Two listening conditions were selected: (a) no noise, no reverberation, (b) reverberant, steady-state noise (Speech Transmission Index: 0.47). The tests were presented in a closed-set format; data on the number of words correctly recognized (speech intelligibility, IS) and the response times RT were collected (onset RT, single words RT).

Results: It was found that a sequence composed by four disyllabic words ensured both the full recognition score in quiet conditions and a significant decrease in IS results when noise and reverberation degraded the speech signal. Response times increased with the worsening of the listening conditions and the number of words of the sequence. The greatest onset RT variation was found when using a sequence of four words. In the comparison with current speech-in-noise tests it was found that the WST maximized the IS difference between the selected listening conditions, as well as the RT increase.

Conclusions: Overall, the results suggest that the new speech-in-noise test has good potentials in discriminating conditions with near-ceiling accuracy. As compared with current speech-in-noise tests, it appears that the WST with a four words sequence allows for a finer

mapping of the acoustical design target conditions of public spaces through accuracy and onset RT data.

Introduction

Speech intelligibility can be estimated either with dedicated objective metrics, or by means of subjective methods based on the direct testing of listeners (International Organization of Standardization, 2003; International Electrotechnical Commission, 2011, American National Standards Institute, 2009). During the years, a variety of subjective tests has been developed, and the speech material employed for the scope can vary greatly. For instance, it can consist of items with a fixed pattern (e.g., nonsense words with a CVC or a CVCV structure) or of meaningful words, often with a phonetically balanced distribution across the corpus of items. Aside isolated items the usage of sentences as speech material provides a means of testing which is much closer to realistic communication and was in fact recommended since long in the field of clinical audiology (Plomp, 1986).

Speech-in-noise tests may differentiate for type of presentation, scoring for correct reception and other features, but all provide a relationship between an objective metric and the intelligibility scores, called psychometric function. This function is the base to assess the speech reception and is used in peculiar ways depending on the application. In particular, of outmost importance in clinical audiology is the detection of the speech reception threshold (SRT) that is the signal-to-noise ratio (SNR) at the 50% speech intelligibility. Conversely, in room acoustics design, the psychometric function is used to identify the values of the objective metrics ensuring the targeted speech communication performance. In general, the course of the psychometric function depends on the listening test and is strictly related to the linguistic complexity and redundancy of the stimulus type (Steeneken, 2014). For instance, speech-in-noise tests based on isolated items are characterized by psychometric functions with mild slopes. These tests have the advantage of discriminating speech intelligibility over a wide range of the objective metrics, including the values typically targeted in the acoustical design of rooms where an effective speech communication is required. However, as words rarely occur in isolation, using single items as speech material only partly represents the cognitive processes occurring during real communication, which, besides recognition, also involve the storing and recall of information.

On the other hand, listening tests based on sentences are characterized by steeper recognition functions spanning over limited intervals of the objective metrics. This happens mostly because the sentence test material is typically optimized for the clinical use in audiology, where a steeper slope has the best discrimination potentials for the retrieval of SRT. Unfortunately, 50% scores and SRT are not much representative of everyday listening conditions or of design targets. Overall, it appears that in view of testing speech recognition in favorable acoustical conditions, and especially near ceiling, most of current speech-in-noise sentence tests will not provide an optimal discrimination. Indeed, their psychometric function in this area usually reaches a *plateau*, where a change in the objective indicator is paired with only a slight change in accuracy. The slope of the function can be partly controlled for by reducing the predictability of the speech material: it is known that high or low predictability sentences output substantially different results (Kalikow et al., 1997). However, even when using low-context material, as in the matrix sentence test (Hagerman, 1982; Kollmeier et al. 2015), ceiling-effects show up already starting at low/negative SNRs. For instance, Rennies et al. (2014) found that using the matrix sentence tests in the open-set format as test material, the 100% recognition was scored already starting from Speech Transmission Index (STI) values of 0.43 (corresponding to a SNR of -2 dB, or to the presence of reverberation alone, with T_{60}= 2 s). Therefore, even though sentence tests have the advantage of using stimuli close to what listeners experience in everyday communication, the presence of contextual, semantic and syntactic information supports the speech recognition, helping the listener to fill-in the missing information and compensate for the partial phonetic representation. The 100% recognition score is reached even in suboptimal listening conditions, which are far from those targeted in the room acoustical design. For example, a STI>0.45 is required for a person-to-person speech communication rated just as “Fair” inside a public space (International Organization of Standardization, 2003). Using the results of sentence tests to support the room acoustical design will not be much informative, being probably the speech intelligibility at ceiling or very close to it (Rennies et al., 2014).

Therefore, when one needs to design rooms for speech in compliance with optimal design targets and then to test speech recognition with a percentage correct much higher than 50% or even than 80%, speech material having great discriminating capabilities in the STI region of interest is needed. In particular, the speech material shall be conceived to tap into the recognition, storing and recall processes taking place in everyday communication but also to resolve the limitation due to the ceiling effect of conventional sentence tests already showing up at STI value below the recommended design values. This will allow for a finer detection of differences between suboptimal listening conditions that will support an optimized room acoustic design of indoor spaces specifically tailored to the needs of the occupants.

To this aim, in this work a new speech-in-noise test, named Words Sequence Test (WST) is proposed and its potentials are explored. The concept underlying the development of the test was to create more cognitively demanding speech material, by removing both context and syntactic structure characterizing sentences. In the WST, the listener is presented with a sequence of meaningful words without interleaves, which shall be stored in memory during presentation and subsequently recalled in serial order. As for sentences, the listener is requested to hold information in memory for a period, thus engaging more high-level processing skills than in single word recognition. Additionally, the words sequence lacks the meaning conveyed by the sentence and even the syntactic relationships between the words, so that in the recall phase less cues will be available to the listener. These aspects will affect the functioning of the working memory (WM), which is the cognitive system responsible for the temporary storage of information (Baddeley & Hitch, 1974). In particular, using words sequences the WM will not be backed up by memory for the sentence (Gathercole & Alloway, 2008) and thus, even in favorable listening conditions, it is expected to be depleted faster. So, due to the heavier reliance on cognitive resources, it was expected that using sequences of words as speech material would allow for a finer detection of the effects of listening conditions on speech reception, and thus for a saturation of speech intelligibility at higher values of STI (or SNR) as compared to meaningful sentences.

Beyond accuracy in speech recognition, another dimension of the listening experience that needs to be addressed even when the speech signal is audible and the words are recognized accurately, is the effort perceived by the listeners (Zekveld et al., 2010; McGarrigle et al., 2014). The so-called "listening effort" was lastly defined by Pichora-Fuller et al. (2016) as "the deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a listening task". In order to identify the speech signal listeners deploy a complex interplay of perceptual (bottom-up) and cognitive (top-down) processes (Pichora-Fuller et al., 1995). The relative weight and relevance of both processes depends on the sematic context, on the redundancy of the linguistic structure and on the listening conditions too (Zekveld et al., 2006). In particular, the fact that listening becomes effortful can be explained by the insurgence of *adversity*, which is intended as the mismatch between the external demands posed by the characteristics of the listening and the internal resources that the listener can provide (Lemke & Besser, 2016). Suboptimal acoustical settings (e.g., noisy and reverberant environments) and informationally complex situations are typical examples of conditions eliciting the mismatch. So, whereas in ideal conditions speech is processed automatically, mainly relaying on the perceptual cues (Rönnberg et al., 2008), when the signal is degraded, the reliance on top-down processes increases and a specific processing effort is experienced (Lemke & Besser, 2016). In fact, implicit cognitive resources will not suffice and explicit ones together with and increased attention will be needed to compensate for the poor auditory representation. For this reason changes in the metrics estimating the effort are observed even when speech recognition does not change or is near ceiling (i.e., close to 100%), implying that investigating upon listening effort is useful to provide additional information beyond speech intelligibility measures alone (Suprenant, 1999; Sarampalis et al., 2009). Below near-ceiling conditions, effort-related quantities mirror the changes in accuracy results, strengthening the picture of speech recognition and accounting more closely for the top-down processing.

A variety of measures has been used to study the complex construct of listening effort; they can be divided into three main categories: behavioral measures, physiological measures and self-report ratings (Klink et al., 2012a; Klink et al., 2012b; McGarrigle et al., 2014; Pichora-Fuller et al., 2016). None of the above categories *per se* is thought to gauge listening effort, since different approaches show peculiar pros and cons: in fact, the strategy of integration of measures from separate domains was also envisaged (Pichora-Fuller et al., 2016). Compared to physiological measures (e.g., pupillometry, skin conductance, hear rate) behavioral measures have the advantage of a simpler data collection. Their rationale lays in the model of limited cognitive capacity (Kahnemann, 1973), stating that when more resources are allocated to speech recognition fewer resources will be available for processes related to rehearse and recall of heard information, which is typical of speech processing. The cognitive load changes in critical *versus* uncritical conditions can be traced for instance with response time, which is used in this context as a cognitively-informed outcome variable either in dual-task or single-task experimental paradigms. In the former case, a secondary task other than the speech-related one is monitored and its slowing down (or worsening) is used to estimate the dimension of listening effort (for a review of dual-task paradigms in listening effort evaluations see Gagnè et al., 2017). This approach provides a multi-tasking framework, which is valuable if everyday conditions shall be mimicked, but some open issues are also outlined concerning for instance the policy of allocation of resources between tasks for some categories of listeners, and the effective amount of overall expenditure (Choi et al., 2008; Mc Garrigle et al., 2014). On the other hand, the usage of single-task paradigms provides clear practical advantages over dual-task experiments and dates back to earlier studies addressing the improvement of speech intelligibility tests (Hecker et al., 1966). Later on, several studies using word recognition showed that the response time (RT) to the auditory stimulus decreases with the use of hearing-aids (Gatehouse & Gordon, 1990), with spectral enhancement for hearing impaired subjects (Baer et al., 1993) and was the prevalent effect of digital noise reduction (Gustafson et al., 2014; van den Tillaart-Haverkate et al., 2017). In these later works, the decrease in RT was interpreted as a release of cognitive load and hence was conceptually assimilated to a proxy estimate for the diminishing of processing effort (Lemke & Besser, 2016). Furthermore, using speech recognition single-task paradigms it was shown that the RT increases as the listening conditions become more difficult (Prodi et al., 2010; Houben et al., 2013; Prodi et al., 2013; Mealings et al., 2015) or the stimulus complexity increases (Uslar et al., 2013; Lewis et al., 2016). In addition, Pals et al. (2015) compared the response times in single and dual-task experiments and found that the former showed significant differences between two intelligibility levels whereas the latter did not. Overall, the previous findings suggest that, since the latency before a response includes the time that the listeners need to decode and process the auditory information, more challenging tasks calling for greater cognitive processing (Hällgren et al., 2001) will in general cause longer latencies. In this sense, response time may provide information on the amount of cognitive resources employed in the working memory system to process the incoming signal, and thus be an outcome informative of listening effort. Moreover, McGarrigle et al. (2014) suggested that response time is an important factor to consider when characterizing speech communication: prolonged speech processing may affect speech communication, limiting the amount of information that can be held in memory and affecting the subsequent processing and recall.

Therefore, based on previous studies, the potentials of the new speech-in-noise test based on sequences of words were explored by measuring, besides speech intelligibility, also response times. A closed-set format was chosen for the presentation of the test stimuli. Consistently with single-task experiments the response time corresponding to the onset time, that is the delay from the end of the auditory stimulus to the choice of the first word, was primarily collected. However, Towse et al. (2008) showed that serial recall tasks also involve memory search and reconstruction processes that take place in inter-words pauses, both in verbal and manual recall. On this basis, also the RT data of the remaining words composing the sequence were retrieved during the experiments, aiming at getting insight on the strategies implemented by the participants to cope with the speech recognition task.

The first experiment of the present study focuses on the development and the optimization of the new speech-in-noise test. Disyllabic word sequences of different length (from 2 to 6 words) were created and proposed in the experiment, with the aim of identifying the sequence ensuring at once the highest accuracy in quiet conditions, and the biggest accuracy decrease once noise and reverberation were added. Following from previous studies on the memory of verbal material (Baddeley, 2000) it was expected that, even in quiet conditions, the amount of words that could be temporarily stored, processed and recalled was restricted by the limited capacity of the listener's WM. Baddeley et al. (1975) found that people are able to remember the number of items that they can pronounce in approximately 2 seconds. Then, assuming a speech rate of 4 to 6 syllables per second as typical for conversational speech in the Italian language (Giordano, 2006; Pellegrino et al., 2011; Koch & Janse, 2016) it was hypothesized that four to five disyllabic words (without any syntactic link) would be correctly recalled in quiet conditions. Beyond this words number, consistently with studies on serial and free recall tasks (Ward et al., 2010) and with a more general behavior pertaining to the capacity of working memory (Cowan, 2010), a decrease in recognition scores with increasing sequence length was expected. Similarly, a decrease in speech intelligibility results was expected with the worsening of the listening conditions. Changes in response times across sequence lengths were also analyzed in the experiment, giving insight into the amount of cognitive resources requested for the task. Following from previous literature results on response time, it was hypothesized that RT would increase when shifting for quiet to noisy conditions, for each sequence length. Furthermore, a slowing down of RT results was expected with the increase of the sequence length, supposing that the cognitive resources would be called for to a greater extent when the stimulus complexity was increased.

In the second experiment, the new WST (with the optimized sequence length) was compared with two current speech-in-noise tests in the Italian language. This was done to set the newly proposed test with respect to conventional and reliable tools used in audiology and in room acoustics. The tests selected for comparison, both implemented in a closed set format, were the Diagnostic Rhyme Test (DRT) (Bonaventura et al., 1986), which is a consonant confusion test based on pairs of rhymed disyllabic words, and the Matrix Sentence Test (MST) (Puglisi et al., 2015) which bases on low-context sentences with a fixed syntactic structure. Owing to the favorable, near-ceiling listening conditions (the same as presented in Experiment 1), no effect on IS for the MST test was expected for both quiet and noisy cases, whereas a significant decrease of intelligibility results in the noisy condition was hypothesized for both DRT and WST. As regards RT, due to the supposedly greater cognitive load, a greater RT difference between the listening conditions was expected for the WST, as compared to the MST and the DRT.

Experiment 1

The aim of Experiment 1 was to investigate the effects of sequence length in a words sequence recognition task (named in the following WST) with reference to both speech intelligibility (IS) and response time (RT).

Methods

Development of the speech material

The word sequences were created starting from the speech material of the Diagnostic Rhyme Test (DRT) in the Italian language (Bonaventura et al., 1986). This intelligibility test bases on a grid of 210 meaningful words, organized in 105 rhymed pairs differing for the initial consonantal feature. All of the words are consonant-vowel-consonant-vowel (CVCV) nouns, being the disyllable the most frequent syllabic structure of the Italian language. The DRT is optimized as regards the language-specific consonant-phonemes distribution, and is based on real words the Italian language. Within a subset of 105 words, the items 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 consonantal features. The words can be further gathered according to the combination of the initial consonant with the following vowel; referring to the tongue position during vowel pronunciation, three groups are identified: anterior (/i/, /e/), posterior (/o/, /u/) and central (/a/). Overall, 21 types of matching between the onset consonant and the following vowel are explored in the DRT test.

In order to define the base matrix of the WST, 42 items were selected among one subset (105 items out of 210) of the DRT *corpus*. The subset was firstly sifted to remove verbs and words with low familiarity, thus retaining a homogeneous group of 84 familiar nouns. The final items were chosen among them respecting the perceptual features distribution (2 x 21 onset consonant-vowel combinations) and matching as closely as possible the phoneme distribution of the Italian language (Tonelli et al., 1998) as regards to onset consonant. The two items selected for each onset consonant-vowel combination were then randomly assigned to one of the two columns of the base matrix with the same vowel context. The base matrix, reported in Table 1, is then organized as follows: each of the six columns contains seven words, differing for the distinctive feature of the initial consonant but with the same context for the first vowel sound (e.g., the first column is composed by words belonging to the /a/ vowel context). The seven words in each column are organized in alphabetical order. The sequence of the vowel contexts (i.e., the succession of the matrix columns) is established *a priori* and it is indicated in the first row of Table 1.

By sequentially selecting the words from the base matrix, the test sequences were created; each one includes a carrier phrase (*Ora diremo le parole…*, which is Italian for “Now we will say the words…”) and six target words. The test sequences were recorded by an adult, native Italian, female speaker, with a trained voice and expertise in stage reading. She was instructed to pronounce the test sequences in a clear manner, speaking at a conversational rate and maintaining a constant vocal effort. Care was taken that, consistently with literature (Giordano, 2006), a speech rate of 4-5 syll./sec was ensured for the target part of the test sequence. In order to limit the presence of factitious positional effects (e.g., lower intensity of the last sequence item, due to the natural decrease at the end of a spoken sequence) the speaker was asked to maintain the same intensity across all the sequence items. The recordings took place in a sound attenuated room, with a 1/2 inch microphone placed about 15 cm in front of the speaker, at a sampling frequency of 44.1 kHz. In Figure 1 the temporal pattern of a typical six words sequence is reported where it is also verified that the co-articulation between successive words was preserved. The maintenance of speaker’s voice intensity across the sequence was verified by means of sound level measures.

Each recorded sequence was filtered as to match the long-term spectrum of a female talker suggested in the IEC 60268-16 standard (International Electrotechnical Commission, 2011). Afterwards, shorter test sequences with less than six target words were generated from the recordings by progressively discarding the last items of the sequence. Finally, all the test sequences (composed by the same carrier phrase and a variable number of target words) were set to the same root-mean-square level.

The speech material was organized in lists, composed by 13 sequences each; for each base matrix (two up to six columns of seven words), two lists were created. Within a test list, all of the words of the base matrix were evenly represented.

Participants

Sixteen young adults (7 male, 9 female) took part in Experiment 1. Their age ranged from 18 to 35 years (average: 27.0 yr, σ: 5.0 yr). They were either students of the University of Ferrara or PhD students of the local Department of Engineering, recruited by word of mouth. The participants were native Italian speakers and reported the absence of hearing impairments. All of them volunteered for the experiment and gave informed consent before the test took place; they were rewarded with a small fee for participation. No ethics approval was required for the study.

Equipment

The experiment was conducted with the listener seated at the center of a sound attenuated room, treated as to warrant minimal reverberation time (T_{mid}<0.2 s), low background noise (L_{eq}<20 dB(A)) and good sound insulation from the surroundings. Within the room a three-dimensional audio rendering system is installed, which is based on multi trans-aural decoding of binaural signals and whose details and capabilities are reported in Prodi et al. (2010). The stimuli were generated within an Audiomulch® application with the Xvolver VST plug-in for real-time auralization hosted on a control PC placed outside the sound-attenuated room, and were delivered through a SSL Alpha-Link MX® sound card. The audio rendering was driven by the MIDI commands coming from a platform for psychoacoustic testing developed in-house as a LabView® application. The same application managed the tests, recorded participant's choices and stored their respective response times. As the experiment was presented in a closed-set format, a touch screen was used for displaying the base matrix and allowing participant's words selection.

Stimuli

The test sequences were presented in two acoustic conditions. In both of them the speech signal was calibrated to a level of 63.0 dB(A) measured at the listening position using a Brüel and Kjær (B&K) 4165 1/2 inch microphone, a B&K 2639 preamplifier, and a B&K 5935 signal conditioner.

In condition A the signal had neither reverberation nor added noise, resulting in a STI equal to 1.0. Condition A was fixed as an easily reproducible benchmark for optimal speech reception.

In condition B both reverberation and noise were added to degrade the speech signal to some extent. The criterion to set this acoustical condition was to ensure still a good intelligibility, but engaging substantially more cognitive resources than in condition A. On this basis, and considering both the relationship between intelligibility scores and STI, and the STI qualification bands reported in International Electrotechnical Commission (2011), a STI value of 0.47 was fixed. The value was obtained with a signal-to-noise ratio of +3 dB, and a reverberation time (T_{30}, averaged across the octave bands of 500-2000 Hz) of 0.94 s. Reverberation was introduced by convolving the speech signal with the simulated binaural impulse responses of a frontal speaker in a rectangular room of 384 m^3, calculated by means of Odeon®. A long-term speech-shaped (LTSS) noise was used to mask the speech signal. It was obtained starting from a steady-state pink noise signal, which was spectrally shaped in octave-bands to match the long-term spectrum of the speech (International Electrotechnical Commission, 2011). Reverberance from the same simulated room was added to noise by convolving it with the sum of four IRs from four omnidirectional sources located at the lower corners of the room. On the resulting IR, a broadband mixing of the phases was performed by further convolving with a short sample (5 ms) of white noise; this was done to achieve a diffuse noise condition where the directional characteristics of the noise were lost, while its spectral and temporal features had only negligible alterations.

Procedures

During the test, the participants were seated in the center of the sound-attenuated room, with the touch-screen in front of them. The experimental session always started with a training, during which a list of 13 sequences of six target words was presented at a fixed SNR of +10 dB in stationary noise and anechoic conditions. The test sequences were not replicated in the subsequent experiment. The training session was expressly proposed with a more favorable SNR than condition B, as to allow participants to familiarize with the speech material and the test procedure, still being aware of the background noise that would have been later proposed during the experiment. Afterwards, the participants were presented with 10 listening tests (5 lists x 2 listening conditions). Within each test, 13 word sequences of the same length were played back in one of the two listening conditions. To minimize the influence of sequential and learning effect, acoustic conditions and list lengths were randomized across the participants. Furthermore, to avoid participants' fatigue, a small break was proposed halfway through the experiment. The entire experimental session lasted 45-50 minutes, depending on the participant's pace.

During the experiment, participants were presented with a sequence at a time; when the background noise was played back, it started approximately 1000 ms before the carrier phrase and ended simultaneously with the final item of the sequence. After the offset of the sound playback, a panel with the base words matrix of the test was shown on the touch screen; an additional row of question marks was added under the base matrix, allowing for the “no choice” option. The same word matrix was always displayed, that is the words were not re-arranged at every trial, but maintained in alphabetical order within each column. The number of the matrix columns was varied according to the number of words of the test sequence, by gradually discarding the last columns. The 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 all selections were performed, the next sequence was automatically reproduced.

Participants were instructed to pay attention to the task, and asked to respond as accurately as possible without any recommendation as regards response latency (Uslar et al., 2013). Only after the completion of the experiment, the participants were informed that RT data were also acquired.

For each participant, the score (correct/incorrect/no choice) of each word composing a sequence was acquired and used to evaluate the speech intelligibility (IS). Consistently with literature (Uslar et al., 2013; Pals et al., 2015; Lewis et al., 2016), the response time of the first target word (onset RT) was collected. It is defined as the time between the end of the waveform of the last word played back and the selection of the first word on the touchscreen. As regards the choice of the onset RT, it has to be remarked that Cowan et al. (2003) suggested that the processing taking place during this time interval in words serial recall potentially includes rehearsal of the items within the sequence, memory search, and response planning as well as motor programming. On the other hand, in order to minimize the contributions not directly related to acoustic features in the RT measure, also the ΔRT values could be used in the data analysis (Houben et al., 2013; van den Tillan-Haverkate et al., 2017). The quantity expresses the difference between the RT measured in a specific condition and the corresponding RT in quiet, the latter assumed as baseline values (Prodi et al. 2013). As in the present experiments only two listening conditions were directly compared (in quiet and with reverberated noise respectively), the usage of a relative metric for RT was not deemed necessary.

In the instructions, no information was given to the participants as regards the retrieval of timing data, and thus they were free to implement personal strategies to solve the task, deciding how and when engage in processing. For example, they could assemble the complete sequence before starting the items selection on the touch screen, or, conversely, process each item separately. Since literature results (Towse et al., 2008; Cowan et al., 2003) suggest that for verbal and manual responses to word recall tasks pauses between single items might reflect memory-search and retrieval operations, RT data were also acquired for the items following the first. This aimed at getting more insight in the interpretation of the task performance and exploring the RT variation among the words of a sequence. In these cases, the variable was defined as the time interval between the choices of the words in succession.

Statistical analysis

During the experiment, multiple measures were acquired for each participant, in different experimental conditions; then, even though care was taken in the randomization of the conditions, a non-independence of the measures was expected. In fact, each person potentially has a slightly different individual response time, and this characteristic will affect all the responses from that participant. Furthermore, the responses provided by participants were not normally distributed. Owing to the favorable listening conditions selected for Experiment 1, the IS distribution was concentrated on large values, due to the increased occurrence of results close to the ceiling. As concerns the RT measure, Baayen & Milin (2010) pointed out that a considerable variation in the shape of the response time distribution is to be expected, both at individual level and for the specific experimental condition. In general, anyway, the RT distribution can be considered as positively skewed, raising rapidly on the left and having a long positive tail on the right (Whelan, 2008).

In order to take into account the above mentioned issues and as suggested by (Baayen & Milin, 2010) and already implemented in other studies dealing with auditory response time (Houben et al, 2013; Lewis et al., 2016), a generalized linear mixed effect model (GLMM) with subjects as random-effect factor was implemented to analyze the results of the experiment. The model allowed dealing at once with the random effects introduced by the individual variability and with the not normal distribution of the dependent variables.

All statistical analyses were conducted using the software *R* (R Core Team, 2017) and the *lme4* package (Bates et al., 2015) assuming an α=0.05 significance level. A GLMM with a binomial distribution was used to analyze the IS data. The response variable in the model was coded with a binary score (0/1 corresponding to wrong/correct response); for the analysis, "no choice" responses were considered as incorrect responses. As concerns RT, a Gamma distribution with a log link function was used for the statistical analysis, which allows for a mild a-priori screening for outliers (Baayen & Milin, 2010). Therefore, prior to analysis, only data with excessively long RTs possibly due to participants' inattention were excluded. A cutoff of 8000 ms was set, beyond which RT values were discarded and considered as missing data. Altogether, 68 RTs were rejected (0.82% of the whole sample). Listening condition (A *vs* B), sequence length (2 to 6 words), word position within a given sequence (e.g., first, second…), the interaction between listening conditions and sequence length, and the interaction between word position and listening conditions were included in the statistical models as fixed factors. Both sequence length and word position were treated as factors, because neither the probability of a correct response in the logit scale nor the response time in the logarithmic scale were expected to vary linearly with them. Beside fixed effects, the statistical models always included random effects to take into account participants' specific effects. Model selection was based on a forward procedure using likelihood ratio test. The consistency of the finally selected GLMM models was investigated by checking their assumptions; in particular, this implied a control of the normality of the random effect terms and the residuals as in Everitt and Hothorn (2010). In case of statistically significant effects, pairwise comparisons based on the difference of the means predicted by the GLMM model above were performed using the *lsmeans* package (Lenth, 2016); in order to account for planned multiple comparison, a Benjamini-Hochberg procedure was used.

Results

Speech intelligibility

Concerning the percentage of correctly recognized words, the analysis revealed that both the interactions included in the GLMM were statistically significant.

The significant interaction between listening condition and sequence length ($\chi^2(4)=28.96$, $p<0.001$), indicated that the worsening of the listening condition had a different effect on the estimated probabilities of correct responses, depending on the length of the words sequence. Figure 2 illustrates the intelligibility results in the listening conditions A and B, averaged over the words composing each sequence; in the following, the WST sequences from two up to six words will be named as W2-W6 respectively. Firstly, in order to understand the interaction, the effect of listening condition was investigated for each sequence length. For W2 and W3, IS did not differ between the listening conditions whereas a significant reduction in condition B versus condition A was found for W4, W5 and W6 (p=0.002 for W4; p<0.001 for W5 and W6). Then, for each listening condition, the effect of the sequence length was verified. It was found that in condition A no statistically significant difference in IS results was present between the sequences W2, W3 and W4. A significant reduction in the percentage of correct responses was instead observed when increasing the sequence length to five words (W4 vs W5: z=-1.99, p=0.046); no difference was observed between W5 and W6. Because of the significant interaction between sequence length and listening condition, the pattern of the results in condition B was modified, and a significant IS reduction was found between all test sequences (W2 *vs*W3: z=-2.97, p=0.003; W3 vs W4: z=-7.26, p<0.001; W4 vs W5: z=-5.97, p<0.001; W5 vs W6: z=-6.02, p<0.001).

Then, the statistically significant interaction between listening condition and word position ($\chi^2(15)$=48.58, p<0.001) was analyzed. The interaction points out that the worsening of the listening conditions affects differently the words within the sequence, depending on their serial position. The IS results are displayed for each target word of the sequences in Figure 3, averaged across participants; the corresponding standard deviations are summarized in Table 2. In order to get insight on the statistical result, pairwise comparisons were carried out for the estimated means, separately for each sequence length. No effect of word position was found for W2 and W3. The estimated probabilities of correct responses were similar for all the target words, for both listening conditions, and equal to the full recognition score (i.e., 100%). For W4, no difference was found between the four words of the sequence in condition A but a significant effect of word position was observed in condition B. The probability of a correct response significantly decreased from the first to the third word of the sequence; no difference was found between the last words. For W5, the results of the pairwise comparisons indicated that, even though the IS results significantly differed between the words of the sequence, the pattern of the results was the same in both listening conditions. Two distinct words groups could be identified within the W5 sequence, irrespective of the conditions: a high IS group (consisting of the first, the second and the last item) and a low IS group (formed by the central items of the sequence). Finally, for W6 it was found that in condition A participants had a significantly higher IS for the first, the second and the last word of the sequence. The addition of reverberated noise altered the results pattern of condition A only with reference to the last word of the sequence having now the same IS of the preceding words.

Response time

For the analysis of RT data, a preliminary model was setup, including listening condition, response type (correct/wrong/no choice) and their interaction as fixed factors. Following the significant interaction ($\chi^2(2)=28.03$, $p<0.001$), pairwise comparisons showed that no difference was present in RT of the two listening conditions for "no choice" responses. On the contrary, the two conditions were discriminated by both correct ($z=7.89$, $p<0.001$) and wrong ($z=2.11$, $p=0.031$) responses. Furthermore, RT data associated to "no choice" responses were significantly greater than RT data associated to both wrong and correct responses ($p<0.001$ for all comparisons). Keeping in mind the role of inter-items pauses in words recall tasks (Towse et al., 2008; Cowan et al., 2003), this finding could be interpreted as an expenditure of a prolonged amount of time in attempting to process the stimuli until a "no choice" response was finally selected. As these results would distort the RT data set, values corresponding to "no choice" responses were removed from the analysis and considered as missing data; in total, 245 RTs were discarded, corresponding to the 2.9% of the dataset.

The statistical analysis showed that the interaction between listening condition and word position was significant ($\chi^2(15)=47.80$, $p<0.001$). In order to understand how the presence of reverberated noise affected the pattern of the RT results, pairwise comparisons of the predicted logarithmic RT were performed.

In the analysis of the results, the onset RT was firstly considered, intended as a qualifier of the whole sequence (Uslar et al., 2013) and associated to the average accuracy of the words sequence. Onset RT results for the sequences from S2 to S6 and for the listening conditions are presented in Figure 4. When examining the pairwise comparisons between onset RT results in the two listening conditions within each sequence length, it was found that participants were slower to respond in the worst versus the best listening condition for W2 (z=6.99, p<0.001), W4 (z=4.94, p<0.001) and W5 (z=2.16, p=0.031). The effect of listening condition on onset RT was greater for W4 than for W2 and W5, being the mean ratio (ratio between the predicted mean in condition B over the predicted mean in condition A) respectively equal to 1.29, 1.11 and 1.12. No effect of the listening condition was instead found for the onset RT of W3 and W6. Concerning the effect of sequence length on onset RT, a significant increasing trend was observed when increasing the sequence length up to W5. In both listening conditions, all comparisons between adjacent sequence length (e.g., W2 and W3) were significant with p<0.001, except for W3 *vs* W4 in condition A which was not statistically significant. A significant onset RT decrease was observed when the sequence length was set to six words, with the onset RT of W6 being significantly faster than the onset RT of W5 (z=-3.05, p=0.002) in condition A, and of both W5 (z=-6.69, p<0.001) and W4 (z=-3.78, p<0.001) in condition B.

Then, the analysis was extended as to consider the serial position effects, with the aim of understanding the RT pattern within the words of the sequences and its change across the listening conditions. In the following, where not differently stated, the comparisons were significant at p<0.001. Figure 5 shows the RT results for the five sequence lengths, detailed for each target word; in Table 3 the corresponding standard deviations are reported. When considering W4, it was found that the pattern of the RT of the words was the same in both listening conditions. Specifically, a significant decrease was found between the RT of the first word and the RTs of the following target words whereas no differences were present between the other words of the sequence. When reverberated noise was added, a significant RT increase was found for all the words of the sequence. Differently, for W2 and W5 the same RT pattern was found in both listening conditions, but the presence of reverberated noise significantly affected only certain words of the sequence. For W2 a significant decrease was found between the RTs of the first and the second word; in condition B only the RT of the first word increased. In W5, participants were found again to respond significantly slower on the first item of the sequence, but differences were observed between the remaining words, with a significant RT increase on the third and the fourth word of the sequence. The shortest RTs were found on the second and on the last word of the sequence. In condition B, a significant RT increase was observed for the first and the last words alone (A *vs* B – w1: *p*=0.031; w5: *p*=0.002). Lastly, for W3 and W6 the statistical analysis revealed that the pattern of the RTs of the words changed when reverberated noise was added. For W3, in condition A, a significant difference was observed between the RT of the first word and the subsequent two, which instead had the same RT. The presence of reverberated noise altered the RT pattern, and a significant increase was found between the second and the third word of the sequence (w2 *vs* w3: *p*=0.024). For W6 a more complex scheme was already observed in condition A, with the RTs of the first and the third word being similar and significantly higher than the RTs of the remaining words. In condition B, the RTs of the first three words remained unaltered, whereas it significantly increased in the remaining words of the sequence (A *vs* B – w4: *p*=0.006; w5: *p*=0.01; w6: *p*<0.001).

Discussion

Effects of sequence length on speech intelligibility

As reported above, it was hypothesized that in ideal listening conditions only sequences with a number of words comprised between four to six, corresponding to what can be voiced in approximately 2 seconds, would have been correctly recalled. The results of Experiment 1 showed that the sequences ensuring full percentage of words recognition in anechoic and no-noise conditions were only those up to W4. While for W2 and W3 the full IS was exactly scored by all participants, the presence of a small amount of wrong answers in W4 yielded a mean IS score of 97.8%. However, the result was not statistically different from the shorter W2 and W3 where the full score was achieved. When the number of words in the sequence increased further, memory errors were expected, due to the limited capacity of the working memory. Consistently, it was observed that the overall IS results in W5 and W6 were significantly lower than the optimal performance. The result was driven by a significant effect of the words serial position implying that, even when the correct hearing of the words was ensured, the probability of correct recall depended on the word position within the sequence. Specifically, for W5 and W6, IS was significantly higher for both the first and second words and the last words, pointing out a "primacy" and a "recency" effect of similar magnitude. One classic interpretation of serial position effects (Murdock, 1962) is that the words earlier in the list are put into long-term memory: they undergo a bigger amount of processing, having more opportunity to be rehearsed. Conversely, the words from the end of the list go to the limited-capacity short-term store and are still present there when recall starts. The central items of the sequence that benefit from neither of the two effects are recalled most poorly.

When the speech recognition task was performed in less favorable listening conditions (i.e., the speech was processed with the simultaneous presence of reverberation and background noise) a decrease in IS results was observed. Interestingly, a significant interaction was found between sequence lengths and listening condition. Indeed, in the STI interval here considered, the accuracy of normal-hearing young adults on the shortest sequences (W2 and W3) was not impaired by the presence of reverberated noise. Thus in a listening condition with STI=0.47, that can be generally rated as "Fair" (International Organization of Standardization, 2003), normal-hearing young adults were able to cope successfully with a recognition-and-recall task for sequences up to three disyllabic words. On the contrary, for longer sequences (W4, W5 and W6), the participants' performance was significantly impaired by the presence of reverberated noise. The smallest IS differences from the ideal listening condition were found for the initial words of the sequence; a high amount of processing is generally devoted to the first word, which thus benefited of the primacy effect even in more challenging acoustic conditions. The speech reception instead greatly worsened on the subsequent words of the sequence, in line with the results of Kjellberg et al. (2008). As more resources were needed for the phonological coding of the degraded perceptual stimulus, fewer resources were available for encoding and rehearsing the sequence items, yielding an accelerate loss of information.

Effects of sequence length on response time

In condition A, the lowest RTs were expected, to serve as a reference against the values measured in more challenging acoustic conditions. It is worth noticing the specific pattern of the single words RTs in condition A (Figure 5). Up to W4, participants were significantly slower to select the first item then the following ones. The contraction of inter-items delays that occurred up to W4 compared to onset RT indicates the prevalence of processing during the onset RT (Towse et al., 2008). In this cases, the onset RT seems to be appropriate to represent the whole sequence. Moving from W4 to W5 (or to W6), significantly alters the RTs pattern, which now shows the presence of a prolonged RT on the third word, corresponding to a long pause before the selection of the third item of the sequence. Thus, the increased sequence length calls for increased processing resources, which is manifested in a salience of the RT of the third word. In these cases, the onset RT does not seem appropriate to represent the whole sequence.

In condition B, slower RTs were generally measured, reflecting the increased processing required for listening to a degraded speech signal. This finding is consistent with previous literature results, for single task listening experiments (Gatehouse and Gordon, 1990; Houben et al., 2013; Gustafson et al., 2014; Pals et al., 2015), and can be explained by the WM model for ease of language understanding (Rönnberg et al., 2008). The model explains changes in cognitive engagement in terms of matching between the phonological information extracted from the speech signal and the phonological information represented in long-term memory. When a mismatch arises, explicit processing resources have to be invoked: more alternative interpretations of the stimuli are elicited, prolonging the matching process necessary to reach a decision.

It is interesting to notice that, when onset RT was considered, no difference was observed between the two listening conditions for certain sequence lengths (W3 and W6). The finding can be explained by looking at the RT of the words composing the sequence, and specifically at the significant interaction between word position and listening condition for W3 and W6. Whereas the listening condition did not affect the RT of the first two words of the sequence, a significant increase was observed starting respectively from the third and the fourth word of the sequence. It might suggest that modifying the acoustic conditions yields a change in the pattern of RT of the words and hence on the allocation of processing resources during the entire recall.

Thus, when using words sequences as target material for speech-in-noise tests in closed-set format, the measure of onset RT not always describes the main processing time associated to the sequence. In fact, depending on the sequence length or on the listening condition, additional memory search and retrieval processes might be needed before selecting specific items of the sequence (Cowan et al., 2003). It could be hypothesized that a more comprehensive RT measure, taking into account the total RT (sum of the RT of the single words) or a weighted combination of the single words RT, might provide a better description of the entire processing. Specific and more detailed investigations are needed to explore this aspect further. As a first attempt, a GLMM statistical model with total RT as response variable and listening condition as factor was set up for each sequence length. In this case, the effect of listening condition was always significant, indicating a statistically significant increase of total RT in condition B with respect to condition A (W2: z=3.88, W3: z=4.78, W4: z=8.95, W5: z=5.58, W6: z=5.44; p<0.001 for all comparisons). Interestingly, and consistently with the results outlined for onset RT, the effect of the listening conditions was the greatest for W4. The mean ratios between the conditions were 1.09, 1.09, 1.27, 1.14 and 1.14 for the increasing sequence length (W2 to W6).

Overall, the results outlined suggest that RT reflects the deployment of cognitive processing during the recall task and hence provides complementary information to IS. However, aiming at using RT in speech-in-noise tests, the repeatability of the RT absolute results across experiments should be explored. In fact, being RT a behavioral measure, its absolute value is bound to be modulated by several factors besides the response to the stimulus alone. Whereas the relative differences in RT values between conditions are driven by the stimulus complexity (as defined by the type of listening test and by the acoustical conditions), other factors besides attention, such as mode of presentation (e.g., in laboratory or in field), environmental comfort, and subset of listening conditions could somehow affect the absolute values of the metric from one experiment to another. All of these factors were controlled in the present experiments but specific investigations on their influence, which are still lacking, are needed to enforce the methodology.

Choice of the optimal sequence length

The main aim of the Experiment 1 was to examine the IS results across the sequence lengths in order to select the most suitable one for describing the accuracy changes in favorable listening conditions. For the purpose, two requirements were set: a) the full score (100% word recognition) in condition A, b) the highest variation in the IS results of the two conditions. Whereas the former requirement was set as to provide a reference for the accuracy measure, the latter was set as to ensure the highest detail in discriminating listening conditions in the STI interval (> 0.47). Following from the finding previously outlined, the only sequences meeting the first requirement were W2, W3 and W4. Among the three sequences, W4 alone underwent a significant IS reduction between condition A and B and cold be thus selected as the optimal one for describing IS in favorable listening conditions. The IS difference between conditions A and B was equal to 15%, which is wide enough to potentially allow for a meaningful discrimination of intermediate listening conditions.

As concerns the behavior of RT for W4, the same RT pattern across the words was found for both listening conditions. Thus an assumption could be made that the same pattern is preserved in the range from STI=1 down to STI=0.47 for the present listening conditions and hence the onset RT could be appropriate to represent the main processing time associated to the W4 sequence. However, with the aim of extending the results of the present experiment to lower STI values or different conditions (e.g., other background noise types), additional investigations should be carried out, specifically focused on the presence of an interaction between the word position and the listening condition. Under such circumstances, changes in the RT pattern could occur for W4 too.

Experiment 2

In Experiment 2, the results obtained in Experiment 1 for the four words sequence (WST-W4), selected as the optimal for testing speech reception in near-ceiling conditions, were compared with current speech-in-noise tests.

The same listening conditions of Experiment 1 were presented to a different panel of listeners by using the original Diagnostic Rhyme Test (DRT) and the Matrix Sentence Test (MST) in the Italian language. The main reasons leading to the choice of the DRT and the MST instead of other available speech-in-noise tests as benchmarks against the new WST-W4 were the following: (1) both share some basic features with the WST, either the speech corpus (DRT) or the matrixed structure (MST); (2) their use as speech intelligibility tests is established for the Italian language as regards respectively both room acoustic design and clinical practice; (3) they can be easily implemented in a closed-set format, allowing for the same experimental paradigm across the tests; (4) being based on stimuli of different linguistic complexity, the two tests were expected to provide both IS and RT results with specific behavior in the considered STI interval (Lewis et al., 2016).

Methods

Speech material

The DRT bases on single, meaningful words organized in rhymed pairs ("*nido/lido*", phonetic translation: /'nido/, /'lido/). One item of each pair composing the test was recorded, embedded in a carrier phrase (*La prossima parola che leggeremo è nido*, which is Italian for "The next word we will read is nest"). The recordings took place in the same sound-attenuated room as used for the WST, with a different female speaker. She was instructed to speak in a natural way, at the rate of conversational speak, avoiding any emphasis on the final, target word. The recorded material was filtered, as to match the long-term spectrum of a female talker (International Electrotechnical Commission, 2011) and set at the same rms-level. Three test lists of 18 words each were then created.

The speech material of the Matrix Sentence Test is composed by sentences with correct and fixed syntax (name-verb-number-noun-adjective) but no semantic predictability. Digital recordings of the test sentences were acquired under agreement from the producer Hoertech Gmbh and were spectrally shaped as to match the long-term spectrum of a female talker (International Electrotechnical Commission, 2011). Forty-eight sentences were randomly selected among the test corpus and evenly divided into three test lists.

Participants and experimental procedure

Thirty-six normal-hearing, native Italian speakers were recruited with the same modality as described in Experiment 1. All of them self-reported no hearing impairment. Participants were randomly assigned to one of the two types of speech-in-noise tests (DRT or MST); the characteristics of the two groups are summarized in Table 4, where the group of listeners presented with Experiment 1 is also described. A Friedman test showed that the age distributions were not significantly different among the three groups.

The DRT and the MST were presented using the same equipment and in the same listening conditions as for Experiment 1. During the experimental session, each listener firstly was presented with a training session; similar to Experiment 1, one test list (18 words for the DRT and 16 sentences for the MST) was presented in anechoic conditions, with a stationary noise and a SNR equal to +10 dB. The training sequences were not replicated in the subsequent experiment. Afterwards, the test lists were presented in the selected listening conditions; for both tests, listening conditions and test lists were counterbalanced across participants. During the DRT, participants listened to a target word with the carrier phrase; their task was to select the correct alternative between the three options appearing on the screen at the signal offset (the two rhymed words and the “None of the two” choice). During the MST, the participants’ task was to select sequentially the five words composing a sentence; the (5x10) base matrix appeared on the screen at the end of the audio reproduction. Within each column, the words were arranged in alphabetical order; the same matrix (without changes in the words order) was displayed during the experiment. As for the WST, a row of question marks was added under the base matrix, to be selected when a choice could not be made (“no choice” option).

For both listening test the advancement was self-paced, as the next target word/sentence was reproduced only after the words selection was completed. The experiment (either DRT or MST) took about 15 minutes for its completion, including the instructions and the training session.

Speech intelligibility scores (IS) and RT data were acquired for each participant. For the DRT, IS was defined with a binary coding (0/1 corresponding to wrong/correct choice); when the "none of the two" alternative was selected, it was considered as a wrong response. The RT was defined as the time elapsed between the end of the audio reproduction and the selection of one of the three alternatives. For the MST, just like for the WST, the score (correct/incorrect/no choice) of each word composing the sentence was acquired, and used to evaluate IS. Similarly, onset RT corresponded at the time between the stimulus offset and the choice of the first item; the response latencies of the following four words were also acquired.

Statistical analysis

Following Experiment 1, a generalized linear mixed model procedure was used for the analysis of the results of Experiment 2. The statistical model for IS data included as fixed factors the test type (DRT, MST, or WST-W4), the listening condition (A or B), the word position, and the interaction between test type and listening condition. As regards RT data, a model was set up with the onset RT as response variable, including test type, listening condition, and their interaction as fixed factors. The serial position effects of the MST response time results were explored with a dedicated model (fixed factors: word position, listening condition, and their interaction). Again, extremely long RTs (> 8 s) and RT data corresponding to a "no choice"/ "none of the two" selection were a priori excluded from the analysis, resulting in the removal of 10 RTs (0.3% of the sample) for the MST and 8 RTs (1.2% of the sample) for the DRT.

Results

Speech intelligibility

Figure 6 displays the IS results for the three speech-in-noise tests, across the two listening conditions. The statistical analysis revealed a significant interaction between the test type and the listening condition ($\chi^2(2)=18.44$, $p<0.001$).

In condition A the IS results of the three speech-in-noise tests were not statistically different and not distinguishable from the ceiling of 100% correct responses. Then, for DRT and WST-W4 a statistically significant decrease was found in IS between condition A and condition B (WST-W4: $z=-8.76$, $p<0.001$; DRT: $z=-2.33$, $p=0.03$) whereas the IS results in the two conditions were not statistically different for the MST. The largest difference between the two listening conditions was found for WST-W4 (IS: 98% - 85%), and the smallest for DRT (100%-92%).

Response time

The comparison of the onset RT of the three tests in the two listening conditions is displayed in Figure 7. The statistical analysis showed a significant interaction effect between the considered factors ($\chi^2(2)=8.79$, $p=0.012$), indicating a different effect of listening condition on the onset RT across the three tests. No statistically significant differences were found in the onset RT of the three tests in condition A. For all tests, onset RT increased significantly with the worsening of listening conditions (DRT: $z=3.96$; MST: $z=6.07$; WST-W4: $z=8.24$; $p<0.001$ for all comparisons) but to a different extent. The mean ratios between condition B and condition A were respectively 1.12, 1.15 and 1.26, indicating that the increase was the greatest for WST-W4 and the lowest for the DRT.

Finally, the effect of serial position on RT results was analyzed for MST. Figure 8 shows the results across participants detailed for each target word. A significant interaction was found between listening condition and word position ($\chi^2(4)=24.83$, $p<0.001$). In condition A, it was found that participants were significantly slower to respond on the first and the third word of the sequence with respect to the other words; no statistically significant difference was found between the RT of the first and the third word. Worsening the listening conditions significantly increased the RT of these two words alone (A vs B – first word: z=4.72, p<0.001; third word: z=2.57, p=0.046), whereas no statistically significant difference was found between the two conditions for the remaining words.

Discussion

The aim of Experiment 2 was to compare the WST-W4 to existing speech-in-noise tests of similar characteristics, as to understand the benefits of the new test when describing speech reception in favorable listening conditions.

Firstly, concerning IS, it is noteworthy the correct recall of all the five words composing the MST in condition A, which does not happen with the WST-W5. Indeed, in ideal conditions, the syntactic structure linking the target words of the MST supports the recall of the whole sentence: grouping the target items in a meaningful way facilitates the memory performance. As hypothesized, for the listening conditions examined in the experiment no differences were observed in the IS results of MST. Indeed, this speech-in-noise test was conceived and optimized for the measure of the SRT, as required in clinical practice. When operating at STI close or higher than 0.47, ceiling effects prevent to get more insight on the effects of listening conditions on speech recognition accuracy. As concerns the DRT, the IS results are fully comparable with the reference STI-IS curve reported by Steeneken (2014). It is interesting to notice the IS reduction between the listening conditions, stemming from the consonant confusion. In fact, the addition of a small amount of stationary noise is likely to produce an effective energetic masking, especially on the low-energy consonants, increasing the complexity of the task. The highest IS decrease between the listening conditions was found for the WST-W4. When the semantic link between the words of the sentence is removed a more cognitive-demanding task can be obtained and, compared to the other speech-in-noise tests, WST-W4 ensures a greater variation of accuracy in the speech recognition task. On the basis on the present results and supposing a continuously decreasing course of its psychometric curve over the STI interval, one could argue that the WST-W4 would allow detecting variations of favorable listening conditions when the other two tests would not. To better detail this behavior, further experiments will be carried out at intermediate listening conditions, with a STI comprised between 0.47 and 1. Moreover, it would be of interest to achieve the full WST-W4 psychometric function and compare it with those of MST and DRT. This task was outside the scope of the present investigation, which was focused on better listening conditions only.

Secondly, concerning the onset RT, it was found that it was differently affected by stimulus type across the listening conditions. The highest difference between conditions was found for WST-W4, suggesting that longer processing time is required for this type of stimulus when listening conditions worsen. This is a further advantage of WST-W4. The lowest difference was found for the DRT. The result is probably yielded by the task being relatively easy and thus requiring a limited engagement of cognitive resources (at least for adult, normal-hearing participants). As concerns the MST, the analysis of the RT pattern across the words of the sentence points out that onset RT only partly represent the processing time associated to this kind of test. Already in ideal listening conditions, RT of the third word of the sentence was comparable with onset RT. The effect is kept in worsened acoustic conditions. Then, as already argued in the Discussion of Experiment 1, a total RT or a weighted RT might be better suited for the purpose, even in favorable listening conditions and this points to a limit in using onset RT for MST. On the other hand the main disadvantage of WST-W4 is that it does not closely taper the higher-level processes that aid sentence understanding in noise which are on the contrary partly included in the MST evaluations.

Overall conclusions

(1) The Words Sequence Test (WST) in the Italian language was developed, based on a phonetically balanced corpus of disyllabic words specifically organized in a matrixed form. The speech test material consists of sequences of meaningful words without syntactic links that must be recognized and recalled in serial order. A closed-set format implemented in a touchscreen application was selected for the test presentation.

(2) Several sequence lengths (from two up to six words) have been tested in quiet (STI=1) and in presence of reverberant, steady-state noise (STI=0.47) in order to determine the optimal sequence length for young normal-hearing adults. The results indicate that a four words sequence (WST-W4) is short enough to be correctly remembered in quiet conditions, where an accuracy undistinguishable from the 100% score was obtained. In the reverberant noisy condition a significant decrease of the IS results was found, indicating that down to STI=0.47, and thus, in the interval of acoustic conditions targeted for the acoustical design of communication spaces, the WST-W4 allows discriminating between different conditions.

(3) Due to the increased relevance that listening effort has gained in the last years, it was decided to include a feasible measure carrying information related to aspects of this construct. Based on previous studies the response time to the auditory stimulus in a single-task experiment was selected as the cognitively-informed outcome variable that was suitable for the scope. The quantity was retrieved both as onset RT and subsequent RTs and differences were outlined. The statistical model revealed that RT was very sensitive to both listening conditions and sequence type. This valuable characteristic was traced back to the processing occurring both before and during items selections. Thus, RT was effectively used to complement speech intelligibility with information on cognitive load during the completion of the task.

(4) For the WST-W4 the onset RT increased with the worsening of the listening conditions and in both conditions it proved to be appropriate to represent the main processing time associated to the sequence. More generally, it was found that depending on the sequence length or on the listening condition, the RT pattern across the sequence items could change. In those cases, the usage of a more comprehensive RT measure, for instance the total RT or a weighted combination of the single words RT, could be explored. A preliminary investigation indicated a statistically significant increase of total RT in condition B with respect to condition A for all WST sequences form W2 to W6 and, consistently with the results for onset RT, the greatest effect was found for WST-W4.

(5) Then the WST-W4 was compared with current speech-in-noise tests (Diagnostic Rhyme Test and Matrix Sentence Test), showing that in the [0.47; 1] STI interval the new test exhibits a larger sensitivity in the IS results. Similarly, with the WST-W4 the greatest RT variation was found between the two listening conditions compared to both DRT and MST, corroborating the test potentials in tracing the increase of cognitive load and of the required processing effort via longer processing times.

Acknowledgments

The authors would like to thank Professor Alessandra Salvan and Professor Nicola Sartori (Department of Statistical Sciences, University of Padova, Italy) for their advices on the statistical analyses. The authors acknowledge Hörtech GmbH, Oldenburg (D) for providing the speech recordings of the matrix sentence test in the Italian language.

References

American National Standards Institute. (2009). *Method for measuring the intelligibility of speech over communication systems ANSI/ASA S3.2:2009*. New York.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research, 3,* 12-28.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in cognitive sciences*, *4*(11), 417-423. [27]

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of learning and motivation*, *8*, 47-89.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of verbal learning and verbal behavior*, *14*(6), 575-589.

Baer, T., Moore, B. C. J., Gatehouse, S. (1993). Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: effects on intelligibility, quality, and response times. *Journal of Rehabilitation Research and Development, 30*(1), 49-72.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed effects mod els using lme4. *Journal of Statistical Software, 67*(1), 1-48.

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

Choi, S., Lotto, A., Lewis, D., Hoover, B., Stelmachowicz, P. (2008). Attentional modulation of word recognition by children in a dual-task paradigm. *Journal of Speech, Language, and Hearing Research, 51*(4), 1042-1054.

Cowan, N., Towse, J. N., Hamilton, Z., Saults, J. S., Elliott, E. M., Lacey, J. F., ... & Hitch, G. J. (2003). Children's working-memory processes: a response-timing analysis. *Journal of Experimental Psychology: General*, *132*(1), 113.

Cowan, N. (2010). The magical mystery four: how is working memory capacity limited, and why? *Current Directions in Psychological Science, 19*(1), 51-57.

Everitt, B. S., Hothorn T. (2010). A handbook of statistical analysis using R. Second Edition. Chapman and Hall/CRC, New York.

Gagnè, J-P., Besser, J., Lemke, U. (2017). Behavioral assessment of listening effort using a dual-task paradigm: a review. *Trends in hearing*, *21*, 1–25.

Gatehouse, S., & Gordon, J. (1990). Response times to speech stimuli as measures of benefit from amplification. *British Journal of Audiology, 24*(1), 63-68.

Gathercole, S., & Alloway, T. P. (2008). *Working memory and learning: A practical guide for teachers*. Sage Publications, London (Chap. 1).

Giordano, R. (2006). Note sulla fonetica del ritmo dell'italiano (Notes on the phonetic of the rhythm in the Italian language). *Analisi prosodica. Teorie, modelli e sistemi di annotazione, Atti del II Convegno Nazionale AISV*.

Gustafson, S., McCreery, R., Hoover, B., Kopun, J. G., & Stelmachowicz, P. (2014). Listening effort and perceived clarity for normal hearing children with the use of digital noise reduction. *Ear and hearing, 35(*2), 183-194.

Hagerman, B. (1982). Sentences for testing speech intelligibility in noise. *Scandinavian Audiology*, *13*(1), 57-63.

Hecker, M. H. L, Stevens, K. N., Williams, C. E. (1966). Measurements of reaction time in intelligibility tests. *Journal of the Acoustical Society of America, 39*, 1188 – 1189.

Hällgren, M., Larsby, B., Lyxell, B., & Arlinger, S. (2001). Evaluation of a cognitive test battery in young and elderly normal-hearing and hearing-impaired persons. *Journal of the American Academy of Audiology*, *12*(7), 357-370.

Houben, R., van Doorn-Bierman, M., & Dreschler, W. A. (2013). Using response time to speech as a measure for listening effort. *International Journal of Audiology, 52*(11), 753-761.

International Electrotechnical Commission. (2011). *Sound system equipment – Part 16: Objective rating of speech intelligibility by speech transmission index: IEC 60286-16*. Geneva, Switzerland.

International Organization of Standardization. (2003). *Ergonomics-Assessment of Speech Communication ISO9921:2003*. Geneva, Switzerland.

Kahnemann, D. (1973). Attention and effort. Englewood Cliffs, NJ: Prentice Hall.

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, *61*(5), 1337-1351.

Kjellberg, A., Ljung, R., & Hallman, D. (2008). Recall of words heard in noise. *Applied Cognitive Psychology*, *22*(8), 1088-1098.

Klink, K. B., Schulte, M., & Meis, M. (2012a). Measuring listening effort in the field of audiology—a literature review of methods (part 1). *Zeitschrift für Audiol*, *51*(2), 60-67.

Klink, K. B., Schulte, M., & Meis, M. (2012b). Measuring listening effort in the field of audiology—a literature review of methods (part 2). *Zeitschrift für Audiol*, *51*(3), 96-105.

Koch, X., & Janse, E. (2016). Speech rate effects on the processing of conversational speech across the adult life span a. *The Journal of the Acoustical Society of America*, *139*(4), 1618-1636.

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Uslar, V., Brand, T., & Wagener, K. C. (2015). The multilingual matrix test: Principles, applications, and comparison across languages: A review. *International Journal of Audiology*, *54*(sup2), 3-16.

Lemke, U., & Besser, J. (2016). Cognitive load and listening effort: concepts and age-related considerations. *Ear and hearing*, *37*, 77S-84S.

Lenth, R.V. (2016). Least-squares means: the R package. *Journal of Statistical Software, 69*(1), 1-33.

Lewis, D., Schmid, K., O'Leary, S., Spalding, J., Heinrichs-Graham, E., & High, R. (2016). Effects of noise on speech recognition and listening effort in children with normal hearing and children with mild bilateral or unilateral hearing loss. *Journal of Speech, Language, and Hearing Research, 59,* 1218-1232.

McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). 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*.

Mealings, K. T., Demuth, K., Buchholz, J., & Dillon, H. (2015). The development of the Mealings, Demuth, Dillon, and Buchholz Classroom Speech Perception Test. *Journal of Speech, Language, and Hearing Research*, *58*(4), 1350-1362.

Murdock, B.B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64*(5), 482-488.

Pals, C., Sarampalis, A., van Rijn, H., & Başkent, D. (2015). Validation of a simple response-time measure of listening effort. *The Journal of the Acoustical Society of America, 138*(3), EL187-EL192.

Pellegrino, F., Coupé, C., & Marsico, E. (2011). Across-language perspective on speech information rate. *Language*, *87*(3), 539-558.

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., ... & Naylor, G. (2016). Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear and hearing*, *37*, 5S-27S.

Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America*, *97*(1), 593-608.

Plomp, R. (1986). A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired. *Journal of Speech and Hearing Research, 29,* 146-154.

Prodi, N., Visentin, C., & Farnetani, A. (2010). Intelligibility, listening difficulty and listening efficiency in auralized classrooms. *The Journal of the Acoustical Society of America, 128*(1), 172-181.

Prodi, N., Visentin, C., & Feletti, A. (2013). On the perception of speech in primary school classrooms: Ranking of noise interference and of age influence. *The Journal of the Acoustical Society of America*, *133*(1), 255-268.

Puglisi, G. E., Warzybok, A., Hochmuth, S., Visentin, C., Astolfi, A., Prodi, N., & Kollmeier, B. (2015). An Italian matrix sentence test for the evaluation of speech intelligibility in noise. *International Journal of Audiology, 54*(sup2), 44-50.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rennies, J., Schepker, H., Holube, I., & Kollmeier, B. (2014). Listening effort and speech intelligibility in listening situations affected by noise and reverberation. *The Journal of the Acoustical Society of America*, *136*(5), 2642-2653.

Rönnberg, J., Rudner, M., Foo, C., & Lunner, T. (2008). Cognition counts: a working memory system for ease of language understanding (ELU). *International Journal of Audiology, 47*(sup2), S99-S105.

Sarampalis, A., Kalluri, S., Edwards, B., & Hafter, E. (2009). Objective measures of listening effort: Effects of background noise and noise reduction. *Journal of Speech, Language, and Hearing Research*, *52*(5), 1230-1240.

Steeneken, H.J.M. (2014). Forty years of speech in intelligibility assessment (and some history). *Proceedings of IOA 40th Anniversary Conference.* Birmingham, UK (Keynote lecture).

Surprenant, A. M. (1999). The effect of noise on memory for spoken syllables. *International Journal of Psychology*, *34*(5-6), 328-333.

Tillan-Haverkate van den, M., de Ronde-Brons, I., Dreschler, W. A., Houben, R. (2017). The influence of noise reduction on speech intelligibility, response times to speech, and perceived listening effort in normal-hearing listeners, *Trends in hearing*, *21*, 1-13.

Tonelli L., Panzeri M., & Fabbro F. (1998). Un'analisi statistica della lingua italiana parlata (A statistical analysis of the spoken Italian language). *Studi Italiani di Linguistica Teorica e Applicata 3*, 501–514.

Towse, J. N., Cowan, N., Hitch, G. J., Horton, N. J. (2008). The recall of information from working memory: insights from behavioral and chronometric perspectives. *Experimental Psychology*, *55*(6), 371-383.

Uslar, V. N., Carroll, R., Hanke, M., Hamann, C., Ruigendijk, E., Brand, T., & Kollmeier, B. (2013). Development and evaluation of a linguistically and audiologically controlled sentence intelligibility test. *The Journal of the Acoustical Society of America, 134*(4), 3039-3056.

Ward, G., Tan, L., & Grenfell-Essam, R. (2010). Examining the relationship between free recall and immediate serial recall: the effects of list length and output order. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(5), 1207-1241.

Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record, 58*(3), 475-482.

Zekveld, A. A., Heslenfeld, D. J., Festen, J. M., & Schoonhoven, R. (2006). Top–down and bottom–up processes in speech comprehension. *NeuroImage*, *32*(4), 1826-1836.

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and hearing*, *31*(4), 480-490.

Figure captions

Figure 1. Temporal pattern of a typical six words sequence, preceded by the carrier phrase. The starting and ending points of the carrier phrase and each sequence item are also highlighted.

Figure 2. Speech intelligibility scores (IS) for listening condition A (anechoic speech signal, no noise) and listening condition B (speech in reverberated noisy conditions). The results refer to the speech-in-noise test with sequences of different length (two to six words – W2 to W6). The bottom and the top of the boxes are the first and the third quartiles of the IS distributions, the central, bold line is the median value and the circle is the mean value; 99% of the IS lay within the whiskers. The outliers are shown as filled points outside the whiskers.

Figure 3. Mean values across participants of speech intelligibility scores (IS) as a function of the word position, for each sequence length (W2-W6) of the Words Sequence Test (WST). Results refer to the listening conditions A (anechoic speech signal, no noise) and B (speech in reverberated noisy conditions). The corresponding standard deviations are reported in Table 2.

Figure 4. Response time (RT) of the first target word, for listening condition A (anechoic speech signal, no noise), and listening condition B (speech in reverberated noisy conditions). The results refer to the speech-in-noise test with sequences of different length (two to six words – W2 to W6). The bottom and the top of the boxes are the first and the third quartiles of the RT distributions, the central, bold line is the median value, and the circle is the mean value; 99% of the RT data lay within the whiskers. The outliers are shown as points outside the whiskers.

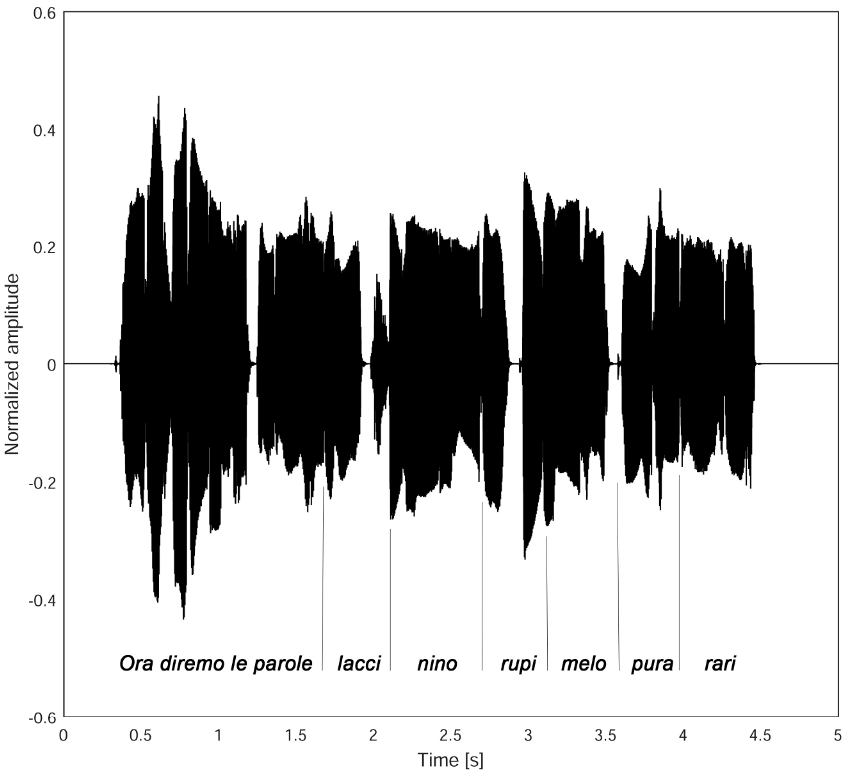
Figure 5. Mean values across participants of response time (RT) as a function of the word position, for each sequence length (W2-W6) of the Words Sequence Test (WST). Results refer to the listening conditions A (anechoic speech signal, no noise) and B (speech in reverberated noisy conditions). The corresponding standard deviations are reported in Table 3.

Figure 6. Boxplots of the speech intelligibility scores (IS) for listening condition A (anechoic speech signal, no noise) and listening condition B (speech in reverberated noisy conditions). The results refer to the Diagnostic Rhyme Test (DRT), the Matrix Sentence Test (MST), and the Words Sequence Test with a sequence of four words (WST-W4). The bottom and the top of the boxes are the first and the third quartiles of the IS distributions, the central, bold line is the median value, and the circle is the mean value; 99% of the IS lay within the whiskers. The outliers are shown as points outside the whiskers.

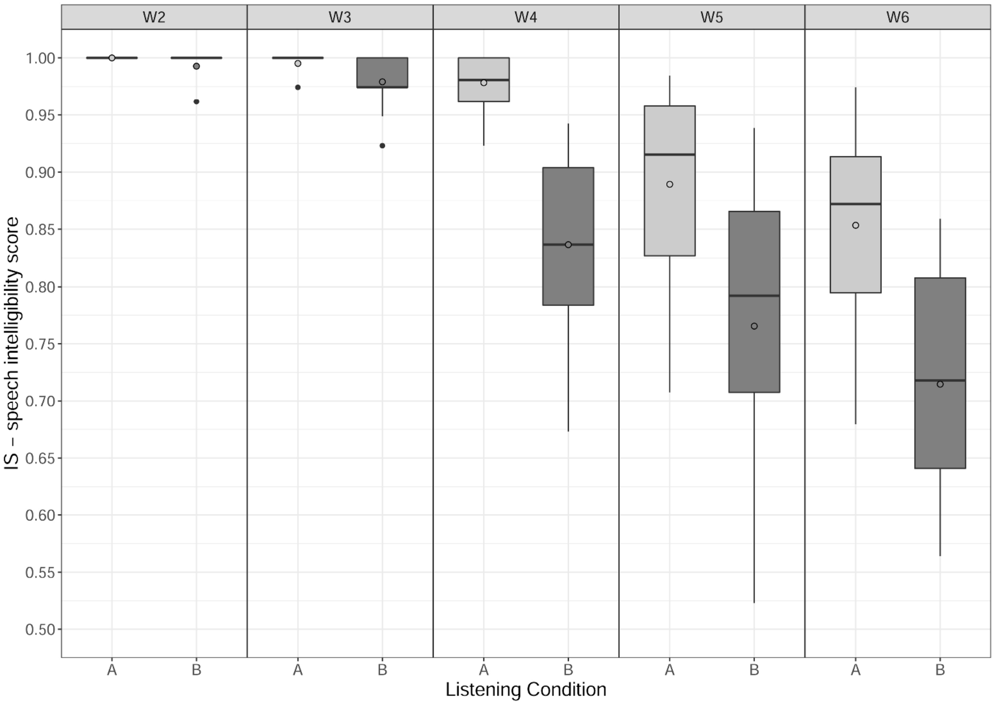
Figure 7. Boxplots of the onset Response Time (RT) for listening condition A (anechoic speech signal, no noise) and listening condition B (speech in reverberated noisy conditions). The results refer to the Diagnostic Rhyme Test (DRT), the Matrix Sentence Test (MST), and the Words Sequence Test with a sequence of four words (WST-W4). The bottom and the top of the boxes are the first and the third quartiles of the RT distributions, the central, bold line is the median value, and the circle is the mean value; 99% of the RT data lay within the whiskers. The outliers are shown as points outside the whiskers.

Figure 8. Mean values and standard deviations across participants of the response time (RT) as a function of the word position, for the Matrix Sentence Test (MST). Results refer to the

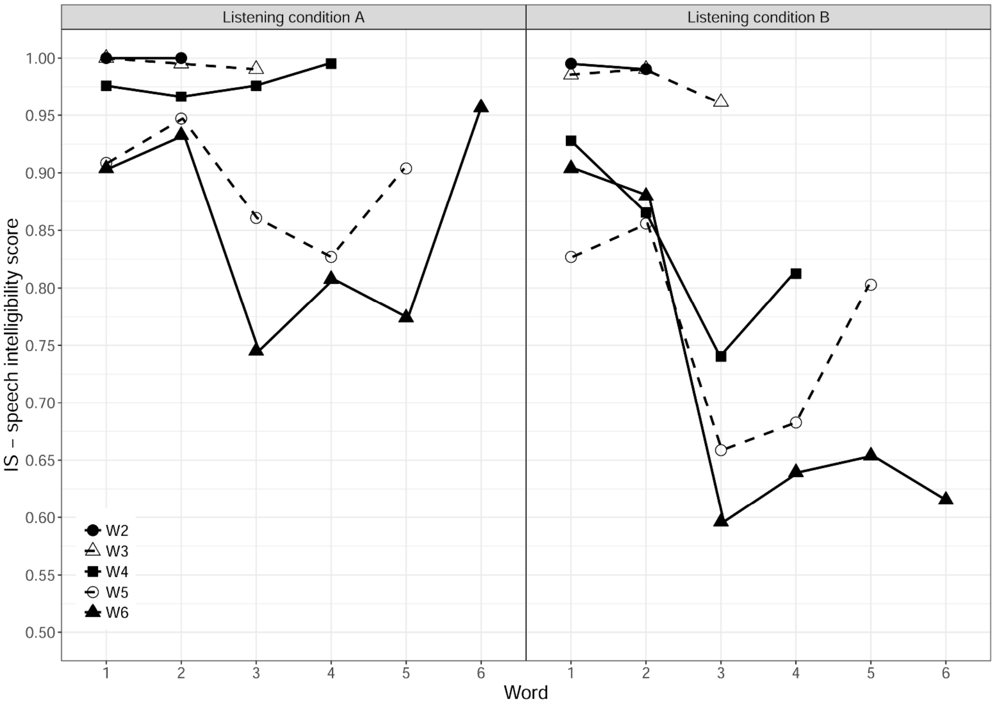
listening conditions A (anechoic speech signal, no noise) and B (speech in reverberated noisy conditions).



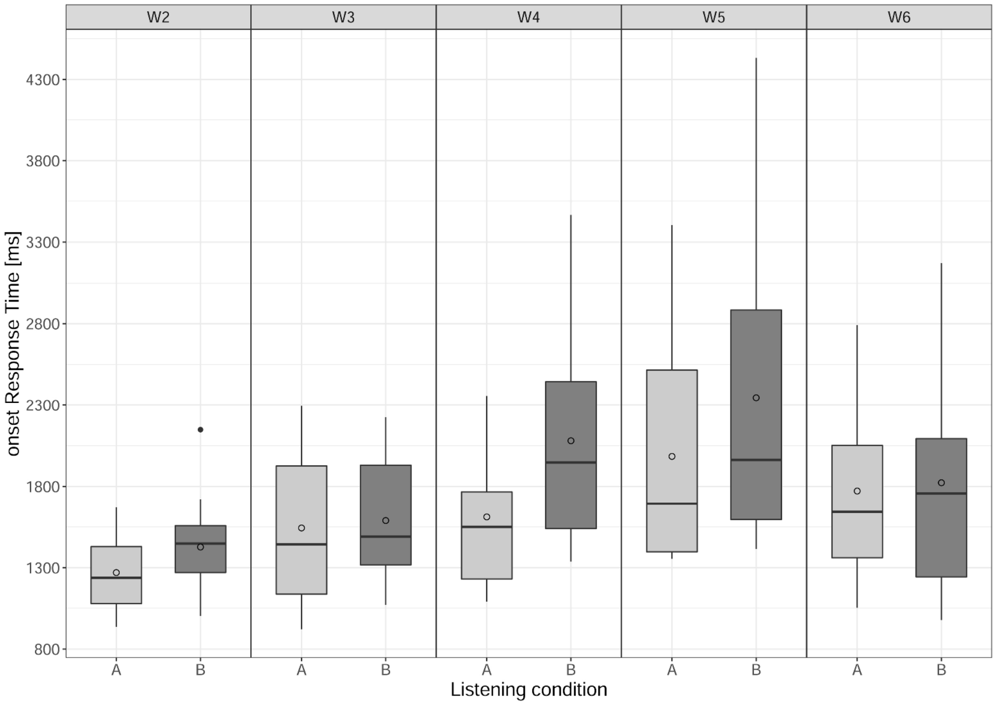
209x181mm (300 x 300 DPI)



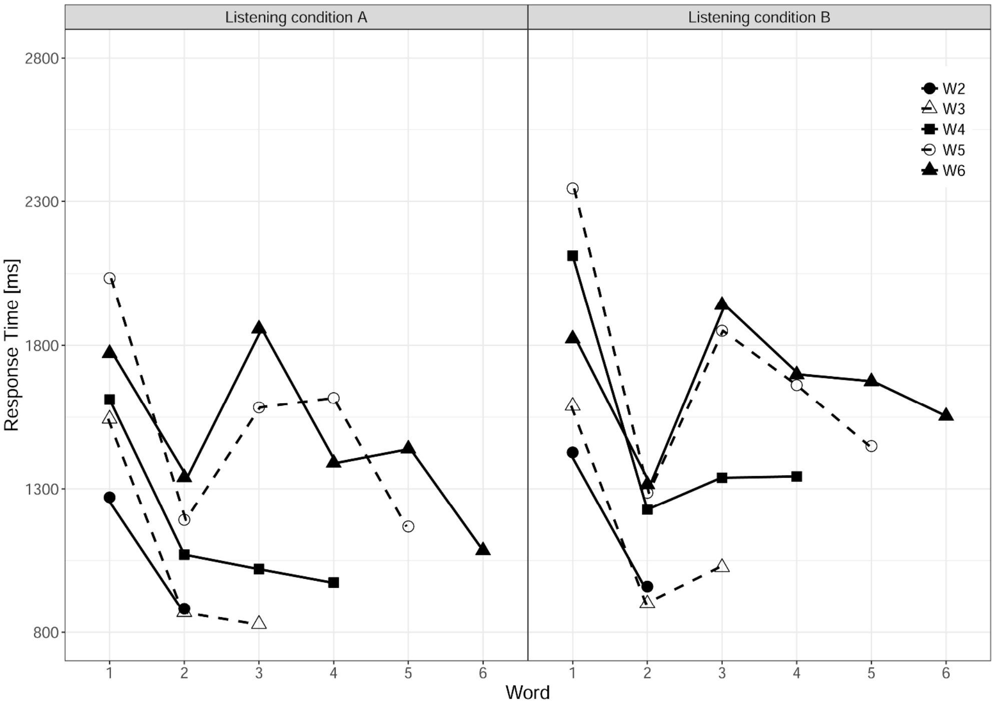
296x209mm (300 x 300 DPI)



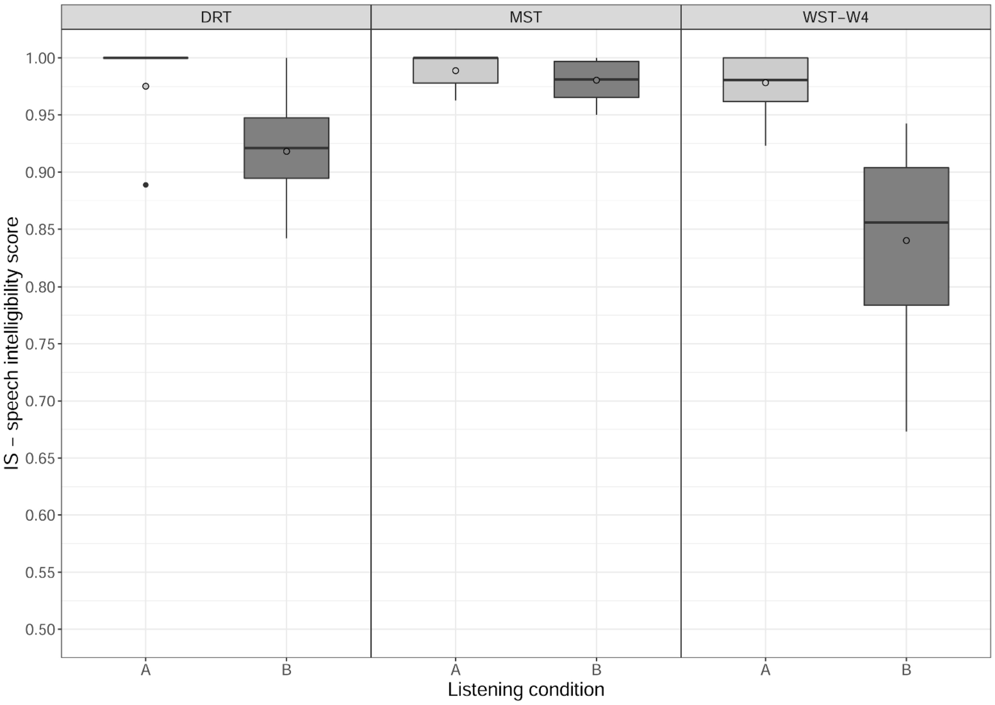
296x209mm (300 x 300 DPI)



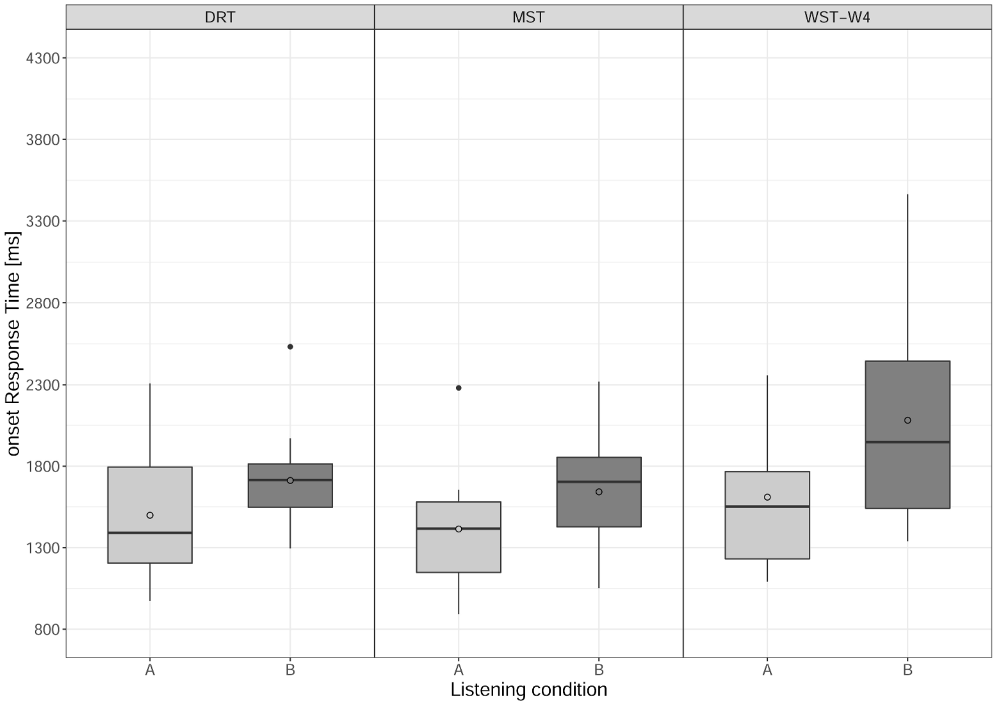
296x209mm (300 x 300 DPI)



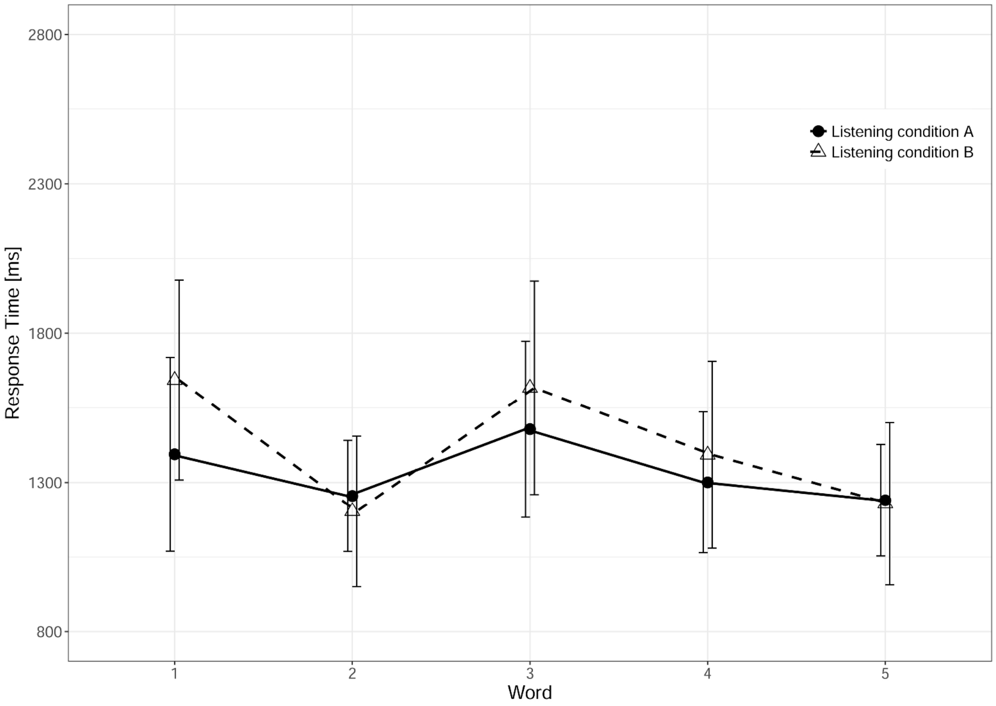
296x209mm (300 x 300 DPI)



296x209mm (300 x 300 DPI)



296x209mm (300 x 300 DPI)



296x209mm (300 x 300 DPI)

Table 1

Base word matrix of the words sequence test (WST). The first row of the matrix reports the sequence of the vowel contexts; words within each column differ for the initial consonantal feature, but have the same vowel context. In bold characters the words composing a randomly built up sequence of six words.

/a/	/i/, /e/	/o/, /u/	/i/, /e/	/o/, /u/	/a/
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

Table 2

Standard deviations (%) of the IS results averaged across participants for the Words Sequence Test (WST); the corresponding mean values are depicted in Figure 3. The data are detailed for each listening condition (A *vs* B) and word serial position (word 1 to word 6) within the sequence (W2 to W6).

		word 1	**word 2**	**word 3**	**word 4**	**word 5**	**word 6**
W2	**A**	0.0	0.0	-	-	-	-
	B	1.9	2.6	-	-	-	-
W3	**A**	0.0	1.9	2.6	-	-	-
	B	3.1	2.6	4.9	-	-	-
W4	**A**	4.6	5.6	4.6	1.9	-	-
	B	8.6	8.2	16.6	10.5	-	-
W5	**A**	9.0	6.7	12.9	15.3	11.4	-
	B	13.9	14.3	20.6	17.7	12.2	-
W6	**A**	8.6	12.5	17.3	16.3	13.6	6.3
	B	9.9	12.2	16.2	18.1	17.5	13.2

Table 3

Standard deviations [ms] of the RT results averaged across participants for the Words Sequence Test (WST); the corresponding mean values are depicted in Figure 5. The data are detailed for each listening condition (A *vs* B) and word serial position (word 1 to word 6) within the sequence (W2 to W6).

		word 1	**word 2**	**word 3**	**word 4**	**word 5**	**word 6**
W2	**A**	230	152	-	-	-	-
	B	297	250	-	-	-	-
W3	**A**	501	129	175	-	-	-
	B	383	140	255	-	-	-
W4	**A**	411	305	262	226	-	-
	B	624	335	411	481	-	-
W5	**A**	776	346	535	515	364	-
	B	949	462	684	578	447	-
W6	**A**	526	507	553	388	444	265
	B	704	311	525	454	287	398

Table 4

Number and age of the listeners participating in the experiments, divided according to the type of listening test performed: words sequence test (WST), diagnostic rhyme test (DRT), matrix sentence test (MST). In parenthesis, the mean and the standard deviation of the participants' ages are indicated.

	sample size	*M*	*F*	*age*
WST	16	7	9	18-35 (m: 27.0, σ: 5.0)
DRT	18	13	5	19-29 (m: 23.9, σ: 2.9)
MST	18	12	6	21-40 (m: 26.2, σ: 5.3)
all	52	32	20	19-40 (m: 25.7, σ: 4.6)