# Effects of the noise type on listening effort: relationship between subjective ratings and objective measurements

Chiara Visentin[1], Nicola Prodi[1]

[1]University of Ferrara, Department of Engineering, Ferrara, Italy

Corresponding author's e-mail address: chiara.visentin@unife.it

## ABSTRACT

The present study investigates the relationship between the self-rated effort when listening to speech in adverse conditions, and response time, taken as a measure of the cognitive resources deployed for interpreting and responding to the auditory stimulus. Specifically, the peculiar effects of two background noises are assessed: a steady state, speech-shaped noise (SSN) and a fluctuating (ICRA) masker. Matrixed-word listening tests were proposed to a panel of young adults with normal hearing. Twelve realistic acoustic conditions were created by varying speech and noise levels, reverberation and noise type. For each condition intelligibility scores (IS), response time (RT) and self-rating of listening effort were collected. The results were mapped by using the objective short-term metric STIr, whose run-time nature allows the tracking of non-stationary maskers, properly accounting for “listening in the gaps”. Even though the same accuracy was achieved in the two maskers, the conditions with ICRA noise were always rated as more effortful; similarly, RT was significantly higher in fluctuating noise, revealing a greater engagement of cognitive resources.

## INTRODUCTION

Everyday communication in public spaces occurs with the unavoidable presence of reverberation, and in most cases of background noise too. Particular interest is drawn by the fluctuating noise, as for instance may result from unattended voices in the same physical space. A systematic investigation of the joint impact of reverberation and fluctuations on the accuracy in speech recognition tasks has been carried out only in few studies. In particular, relevant results are found in Ref. [1] where a reverberation time as small as 0.25 s is shown to already impact on the fluctuating masker benefit, that is, on the decrease in the speech reception threshold (SRT) that occurs in an anechoic setting because listeners are able to listen in the gaps of the masker. Moreover, depending on the speaker-listener distance, when reverberation time is increased the benefit from noise fluctuations is predicted to be further reduced until probable disappearance.

The outcome in terms of listening effort of the joint effect of reverberation and fluctuations is even less explored and few results are available for their separate effect. As explained in Ref. [2], masker fluctuations are supposed to increase listening effort, due to the involvement of

more complex cognitive resources in the speech reception process. The scoring approach adopted for the estimate of accuracy is not able to describe the listening effort, whereas subjective ratings and measurable quantities are available. In particular, a study of listening effort in stationary noisy, reverberant sound fields, based on subjective ratings [3], highlighted that the Speech Transmission Index (STI) [4] could be a rough predictor for the effect, given the correlation of the two quantities for most of the tested conditions. On the other hand, another subjective rating conceived to describe the same effect, called “listening difficulty” [5], was found to vary considerably even for almost equal intelligibility scores. Thus, despite relevant outcomes in specific studies, subjective ratings are difficult to generalize due to possible individual biases in the scaling adopted [6]. Besides subjective ratings, listening effort can be quantified by a variety of psycho-physiological and behavioral measures [7]. For instance, the usage of pupillometry [8] allowed to outline the increased mental effort when, in an anechoic setting, a single talker interferer and a fluctuating masker were compared under controlled intelligibility conditions. The increased effort for the talker case was primarily associated with a larger informational masking since, apart from minor spectral differences, the meaning of speech was the only characteristic evoking a larger use of cognitive resources. Another viable quantifier of listening effort can be considered the behavioral measure of response time to an auditory stimuli in a single-task paradigm [9, 10],or to a secondary task in dual-task experiments [11,12]. By using the latter measure, the effect of reverberation on listening effort for normal hearing adults in a context of stationary noise was investigated in Ref. [13]. Even when word recognition scores were degraded, the effect of reverberation was minimal; these results were considered inconsistent with the current models of listening effort but the reasons for this inconsistency were unclear. Accuracy results and auditory response time data have been combined into a single indicator termed listening efficiency [14], which allowed outlining different strategies implemented by the children to achieve a given performance in noisy settings [15]. In particular, the study showed that, depending on the specific noise type and on reverberation, older pupils could modulate their resources in order to keep an equivalent listening efficiency.

Then, from the analysis of the available results, it appears that the joint effect of reverberation and fluctuations on listening effort is still unclear, since no studies have examined their concurrent presence. This is the main task of the present work. Normal-hearing adult listeners will be considered and their performance in terms of both accuracy and effort will be evaluated. The acoustical conditions will include variable room reverberation, and a masking effect caused by an incoherent stationary noise or by a fluctuating speech-like signal mixed to the frontal speech signal at various levels. The focus on an incoherent masker is intended to approximate a spatially diffuse disturbance as found for instance inside real-life public spaces, with distributed and unattended speech sources. In this case it is expected that the binaural speech intelligibility is minimized by the least binaural unmasking due to the superposition of incoherent noise sources, and that the listening effort plays a relevant role in the assessment of the performance. As regards the qualification of the listening effort, the auditory response time in a single-task paradigm will be used, and the combined listening efficiency metric will be examined too. Finally, subjective reports of listening effort will be collected and correlated with objective indicators to investigate the subjective/objective consistency.

## METHODS

### Participants

The experiment was proposed to 47 participants, all of them native Italian speakers. They were recruited among the students and the academic staff of the local University, and paid a small allowance for their participation. All the participants self-reported normal hearing. Due to

the extended design of the experiment and the large number of conditions investigated, the participants were divided in two homogeneous groups, each of them presented with a different subset of conditions. A panel of 21 listeners (12 men, 9 women; mean age 27.3 yr, σ: 6.1 yr) evaluated a first set of four listening conditions. The remaining subset of eight conditions was presented to a group of 26 listeners (17 male, 9 female); they were between 19 and 41 years of age (average: 26.7 yr, σ: 7.3 yr). The two groups can be considered similar in the gender distribution; no significant difference was found between the age distributions when a Wilcoxon rank sum test was performed ($p$=0.47).

### Speech material

The speech material used for the experiment was the recently developed matrixed word test (MWT) in the Italian language [16]. The test bases on sequences of four disyllabic (CVCV structure), meaningful words. The items were selected among the corpus of the already available Diagnostic Rhyme Test in the Italian language [17], thus respecting the language-specific phonetic distribution. Twenty-eight words were organized in a (7x4) base word matrix, and test sequences were created by randomly selecting one word from each matrix column in succession. The combination of low-contest material and absence of a syntactic structure linking the target words, allows for a sensitive discrimination of listening conditions already on the perceptual level (i.e., intelligibility scores). Furthermore, the number of items within a sequence of the MWT was optimized as to maximize the difference in response times between listening conditions, highlighting the different impact on the cognitive resources. The test was conceived to be presented in a closed-set format, allowing for a reliable retrieval of RT data.

The word sequences, embedded in a carrier phrase ("Ora diremo le parole…" / "*Now let's say the words…"*), were recorded by a native Italian speaker. She was instructed to speak at a conversational rate, maintaining a constant vocal effort and the same intensity across all the words composing a sequence (i.e., avoiding the intensity decrease on the last sequence item). Care was taken that, consistently with literature results [18], a speech rate of 4-5 syll/sec., corresponding to about 110BPM, was ensured for the target part of the sequence. The recordings took place in a silent room, at a sampling frequency of 44.1 kHz. Each sequence was then filtered as to match the long-term spectrum of a female talker suggested in the IEC 60268-16 standard [4], and set to the same root-mean-square. Eight test lists, composed by 12 randomly selected sequences, were used for the experiment; within a list, all of the words of the base matrix were evenly represented.

### Listening conditions

A steady-state noise (SSN), spectrally shaped to match the long-term spectrum of a female talker [4], and a single speaker continuous fluctuating noise (Italian ICRA noise) were selected to produce an energetic masking of the target words. The latter noise was obtained by processing, according to the established procedure [19], Italian phrases spoken by a native female speaker at a normal vocal effort. The resulting signal maintains the envelope of the speech signal but does not carry any informative content, being the processed speech completely unintelligible. Then, the two background noises had the same spectral properties but differed in their temporal envelope.

Twelve listening conditions were created by combining the two types of noise, 3 reverberation conditions and 2 signal-to-noise ratios (SNRs). The reverberant conditions were obtained by convolving the anechoic signals with binaural impulse responses (BRIRs) simulated in a rectangular room of (4x8x12) m. The speech signal was convolved with the impulse response

of a source placed at a distance of 2.5 m from the receiver, having the directivity pattern of a talker. The noise was instead convolved with the sum of the impulse responses from four omnidirectional sources located at the room lower corners. In order to obtain a diffuse noise condition, and lose the directional characteristics of the noise, a broadband mixing of the phases was performed by convolving the sum of the impulse responses with a short sample of white noise. The absorption properties of the room boundaries in the simulations were evenly varied as to obtain different listening conditions. The resulting reverberation times ($T_{30}$, averaged across 500-2000 Hz) were 0.30 s, 0.65 s and 1.03 s. During the experiment, the reverberated speech was reproduced at a fixed level of 63 dB(A), measured at the listener position; the reverberated noise level was varied as to achieve the desired SNRs of -3 and -6 dB.

The objective characterization of the acoustic conditions presented in the experiment was achieved by using the short-term STIr method [20, 21]. The metric relies on a frame-based application of the IEC60268 standard [4] indirect method for the calculation of the STI and allows for a meaningful mapping of the listening conditions even in presence of a time-varying masker. The approach was deemed suitable given the absence of non-linear processing (e.g., frequency shifts, jitter…), that could impair the MTF calculation form the impulse response underlying the indirect STI method. Firstly the MTF was evaluated from the noiseless simulated impulse response, then speech and noise signals were framed in time segments and the SNR in each analysis window was calculated. The respective frame values of STI were later averaged and the STIr of the entire recording was obtained. The optimal length of the analysis window to be used in the short-term metric calculation is still a topic under discussion in the literature (see Ref. [20] for a review). In this work, a duration of 186 ms was selected, corresponding to $2^{13}$ points of the FFT at a sampling rate of 44.1 kHz. The window length closely reflects the typical duration of a CV syllable in the Italian language, where an average syllable duration between 160 and 190 ms is expected when speaking at a normal speaking rate [18]. As the objective metric was calculated starting from binaural signals and noises, a better-ear criterion was applied to predict binaural STIr values [22].The calculated STIr values ranged between 0.18 and 0.37 for the conditions with the SSN masker and between 0.22 and 0.40 for the corresponding listening conditions with the ICRA masker. For reference purposes, the STI values calculated with the SSN masker were found to vary between 0.22 and 0.48, thus varying between a “bad” and a “fair” speech intelligibility [4].

**Procedure**

The experiments took place in a sound treated room. A three-dimensional audio rendering system based on seven pairs of loudspeaker processed for trans-aural rendering surrounded the listener who was seated in the center of the room. In front of her/him, a touch-screen was located to be used for the items selection. The stimulus material was presented using an in-house LabView-based script, dialoguing via MIDI with the software engine of the audio system, an Audiomulch application with VST plug-ins for real-time auralization.

Prior to the experiment, a training session of 12 test sequences, presented at a fixed SNR of +10 dB in stationary noise and anechoic conditions, was proposed to the participants. The aim was to familiarize the listeners with the test procedure, and to reduce the influence of training effects (expected due to the limited vocabulary of the test) during actual measurements. After the training, participants were presented with the listening tests. In order to minimize the influence of sequential and learning effects, acoustic conditions and test sequences were counterbalanced across the participants using a balanced Latin-Greek square design. Furthermore, to avoid listeners' fatigue, a small break was proposed after the conclusion of the first half of the experiment. During the test, participants were presented with a sequence at a time; the background noise started almost 1000 ms before the carrier phrase and ended

simultaneously with the final item. After the last word had been presented, the base word matrix was displayed on the touch screen. Participants had to mark the identified words in serial order, and it was not possible to change a response once it had been selected. As the word in the last column was selected, the next sequence was automatically presented. After responding to 12 test sequences for each listening condition, the participants were asked to subjectively rate the effort. Two different scales were employed. The former subgroup of participants rated the conditions on a 7-categories Likert scale, with the categories labelled from “minimum effort” to “maximum effort”. The latter subgroup, rated the listening conditions on a visual analog scale (VAS). The anchors of the scale were ``minimum effort'' and ``maximum possible effort''.

The percentage of correctly recognized words within a sequence was used as a measure for speech intelligibility (IS). Furthermore, for each sequence, the response time (RT) was automatically collected during the test. RT is defined as the time elapsed between the end of the audio playback of the test sequence and the selection of the first word on the touch screen. Consistently with literature [10, 23] the RT of the first target word was considered as representative of the whole sequence. For each participant, the IS and RT results in a listening condition were obtained as the averages across the 12 sequences composing the test (i.e., average of 4x12 intelligibility scores and of 12 response times). The listening efficiency (DE) was calculated for each listening condition as the ratio between the average intelligibility score, and the average response time [14]. As concerns the self-reported measurements of effort, the scoring on the VAS were calculated as the distance from the left-hand extreme of the scale to the point select by the participant, normalized over a scale of 10 points. All scoring from Likert scale were instead converted on a 0-10 scale. The data were aggregated and analyzed together, as literature results [24] suggest that VAS and single-item Likert questions measuring the same construct are highly correlated and can thus be made consistent.

## RESULTS

### Statistical analysis

A direct, point-to-point comparison of the two noises by means of the IS or RT data could only be effective if they were exactly mapped by the same STIr values. On the contrary, the same values of $T_{30}$ and SNR did not output the same STIr for the two noises due to the fluctuating nature of the ICRA one, which was tracked by the short-time analysis. Then, the effect of SSN and ICRA noises was tested by a detailed statistical comparison of the respective regression curves fitted to the results of each metric under evaluation (IS, RT,DE and subjective rating of effort).

Due to the limited number of data points that build up the regressions to be compared ($n$=6), classical tests, such as the $t$-test, which are used for the comparison of slopes and intercepts of linear regressions, may provide unreliable results. To tackle this problem, the regressions were compared based on the distribution of the residuals. At first a best-fitting regression curve was calculated for a group of data, together with its corresponding residuals (e.g., regression curve and residuals for SSN data, named $SSN_{SSN}$). These residuals were statistically distributed in a symmetrical pattern around zero. Secondly, keeping fixed the previous regression curve, the residuals of the other data set referred to the curve were calculated (e.g., the residuals of the ICRA noise data with respect to the SSN regression curve, named $ICRA_{SSN}$). These residuals, depending on the relative position of the two set of data (and thus, of the regression curves) could either be distributed around a zero value, or be wholly positive/negative. Then, the two sets of residuals (e.g. $SSN_{SSN}$ and $ICRA_{SSN}$) were

statistically compared by using a stochastic ordering procedure applied to their distributions. This process is based on a permutation approach [25] under the null hypothesis of equality in distribution of the populations. In case of rejection of the null hypothesis, the presence of a significant difference is established between the residuals, meaning that the observed differences in the residuals distributions were due not to random variations but to the best fit of the regression line to just one of the set of data. Based on the sign of the residuals, on the direction of the significant comparisons ($SSN_{SSN}$>$ICRA_{SSN}$ or $SSN_{SSN}$<$ICRA_{SSN}$), and confirmed by the reversed testing (comparison of the residuals of SSN and ICRA data with respect to the ICRA regression curve) conclusions about the presence of differences between the two regression curves could be drawn. In particular, when significant and consistent differences were found for both direct and reverse testing, it could be assumed that the two set of data could be better represented by two distinct regression curves, at least in the STIr interval under analysis.

### Effects of noise type

Figure 1 depicts IS for each listening condition, calculated as the average across all participants; the data are plotted as a function of the STIr values and divided according to the background noise. Specifically, a psychometric function was chosen to describe the listeners' IS as a function of the objective metric STIr. The function is defined by the $STIr_{50}$ parameter (corresponding to the STIr required for a 50% intelligibility score) and by its slope in [% / STIr] at the same point. The logistic curves showed in Fig. 1 are the best-fitting regressions, found using a non-linear least squares method. For the SSN data, a significant *p*-value is associated to the estimate of each parameter of the function ($p$<0.001 for both $STIr_{50}$ and slope). The $R^2$ value, taken as an indicator of how well the logistic function fitted the data has a value of 0.99. Concerning the IS results under ICRA noise, the fit of the data on the logistic curve is good ($R^2$=0.96) and the parameters estimates are found to be both significant ($STIr_{50}$: $p$=0.003, *m:* $p$=0.004). The parameter of the two logistic curves are reported in Tab.1, together with the calculated $STIr_{80}$ values (STIr required to obtain an 80% intelligibility score).

**Table 1:** $STIr_{50}$ values and slopes of the best fitting regression logistic functions for the IS results, depending on the type of background noise. The $STIr_{80}$ values (STIr required to obtain an 80% score) calculated from the regression curves are also reported.

| Background noise | $STIr_{50}$ [-] | Slope at $STIr_{50}$ [% / *STIr*] | $STIr_{80}$ [-] |
|---|---|---|---|
| SSN | 0.16 | -10.87 | 0.29 |
| ICRA | 0.15 | -9.48 | 0.30 |

The direct and reverse statistical comparisons of the residuals always returned *p* values greater than the significance level of 0.05, implying that the following equalities were verified between the distributions of residuals: $SSN_{SSN}$=$ICRA_{SSN}$ and $SSN_{ICRA}$=$ICRA_{ICRA}$. The deviations among the residuals could be then considered as due to random variations, implying that the IS results under SSN and under ICRA noise could be fitted by a common regression curve without significant change in the residuals. Consequently, it could be said that, when using the STIr for objectively mapping the data, the type of background noise has no effect on IS results for the MWT test, at least in the considered STIr interval [0.18; 0.40].

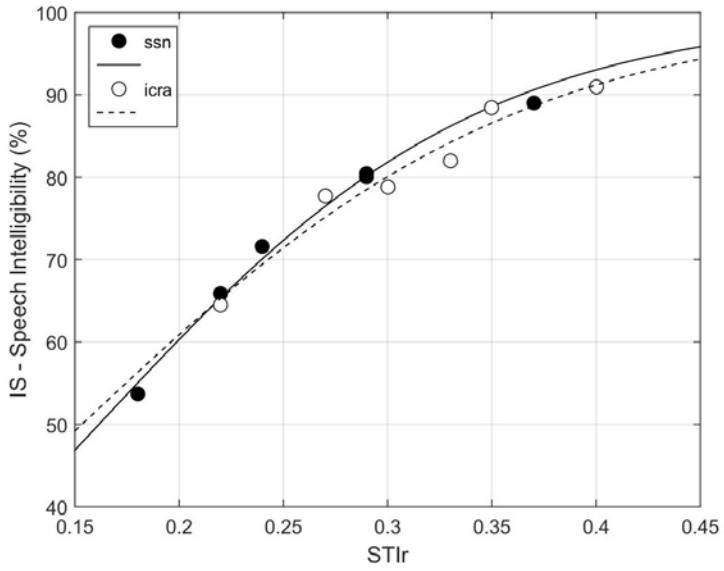


**Figure 1**: Relationship between the speech intelligibility scores IS (%) and the short-term speech transmission index (STIr) values, for the two masking noises: steady-state (SSN) and fluctuating (ICRA). The best-fitting regression curves with a logistic shape are also included.

The RT results are presented in Fig. 2, each point being the average value pooled across all subjects. Differently to IS, RT data could not be described by a psychometric function (being the upper and lower asymptotes unknown) and then a logarithmic regression was assumed as the best fit of the data set, described by the function: RT=A+B log(STIr). The response time data for the SSN masker ranged from 2.38 to 1.62 s, following the expected decreasing trend with the improvement of the acoustical conditions. The estimates of both coefficients of the curve were significant (A: $p$=0.046; B: $p$=0.002) and $R^2$=0.93. The regression curve for ICRA data covered RT values varying between 2.36 and 1.71 s. Significant $p$ values were found for both coefficients ($p$=0.006 for A, and $p$<0.001 for B) and $R^2$=0.95. The statistical comparison of the residuals showed the presence of a significant difference between the distributions ($p$<0.001), expressed by the inequality $SSN_{SSN}$ <$ICRA_{SSN}$. The finding was confirmed by the reversed analysis whose result was $SSN_{ICRA}$ < $ICRA_{ICRA}$ with $p$<0.001. Then, it can be said that, at least in the present STIr interval, the RT data are best fitted by two separate regression curves, which follow an almost parallel course. That of ICRA regression lays always above SSN and the bias has a weak dependence on STIr, which is comprised in the range from 160 ms to 200 ms.

The combined analysis of IS and RT data, through the joint metric of listening efficiency DE, led to the results showed in Fig. 3. Again, each experimental point represent the average efficiency pooled across the participants. Similarly to RT data, a logarithmic curve was chosen to fit the DE results. The best-fitting curve for SSN data ranged from 0.26 to 0.57 $s^{-1}$, with a $R^2$=0.97; both parameters estimates were statistically significant ($p$<0.001). The regression curve for ICRA data covered DE values varying between 0.30 and 0.57 $s^{-}$1 with the $R^2$ = 0.99 and both parameters statistically significant ($p$<0.001). When the presence of differences between the two regression curves was tested through the comparison of the residuals, significant results were always found ($SSN_{SSN}$>$ICRA_{SSN}$: $p$<0.001; $SSN_{ICRA}$>$ICRA_{ICRA}$: $p$<0.001). Then, the DE results could be described by two different regression curves, the differences between them not due to random variations but to the effect of background noise. With both maskers, the listening efficiency increased with the improvement of acoustical conditions; in the STIr interval under analysis, the differences in the impact of the two noises

on the speech reception process were constant and equal to 0.05 $s^{-1}$. Due to the absence of differences between the IS of the two noises in the considered STIr interval, DE results were entirely driven by the RT values; then, the significant increase in the response latencies observed under the ICRA noise, was reflected in a lower listening efficiency.

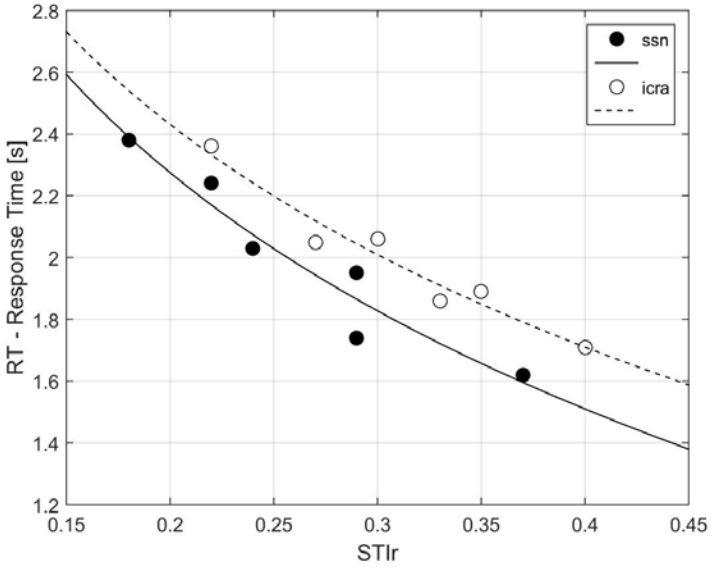


**Figure 2**: Relationship between the response time RT [s] and the short-term speech transmission index (STIr) values, for the two masking noises: steady-state (SSN) and fluctuating (ICRA). The best-fitting regression curves with a logarithmic shape are also included.

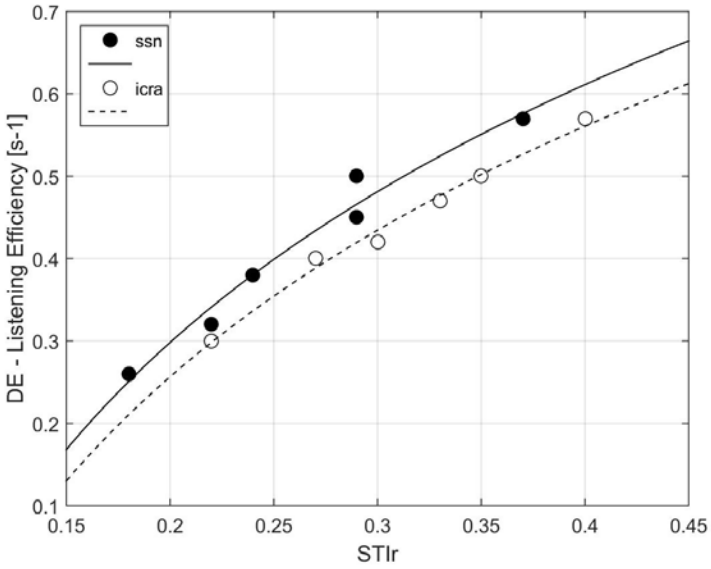


**Figure 3**: Relationship between the listening efficiency DE [$s^{-1}$] and the short-term speech transmission index (STIr) values, for the two masking noises: steady-state (SSN) and fluctuating (ICRA). The best-fitting regression curves with a logarithmic shape are also included.

Finally, the data concerning the subjective effort ratings were analyzed; the average results across participants are presented in Fig. 4. As expected, the effort rating decreased with the improvements of listening conditions, ranging between the values of 8 and 2 (on a 0-10 scale). For both noise types, a good data fit was obtained with the logarithmic curves, obtaining $R^2$=0.98 (SSN) and $R^2$=0.92 (ICRA noise). In both cases the estimates of the curve parameters were significant (*p* always lower than 0.001). The statistical analysis revealed the presence of significant differences among the residuals ($SSN_{SSN}$<$ICRA_{SSN}$: *p*=0.009; $SSN_{ICRA}$<$ICRA_{ICRA}$: *p*=0.047). Then, the two set of date could be described with different regression curves. As for the other metric under analysis, the regression curves for the two noises follow a parallel course in the considered STIr interval, with the ICRA noise always showing higher ratings. The difference between the effort ratings of the two noises had a constant value of 0.8.

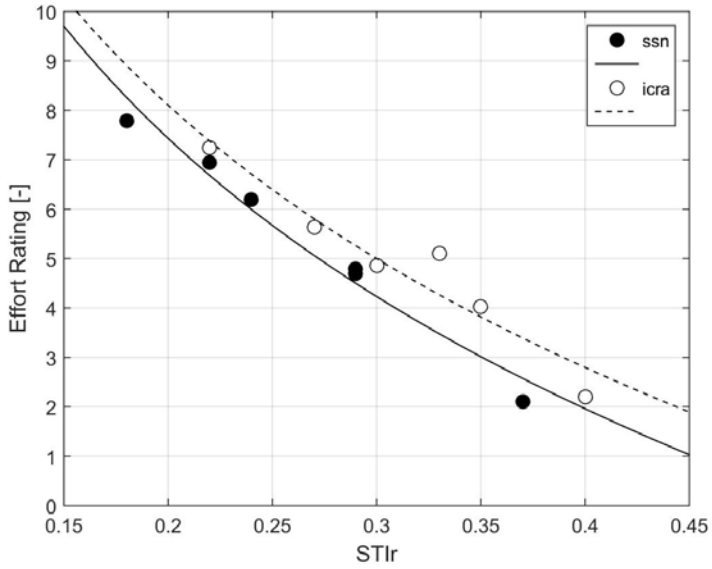


**Figure 4**: Relationship between the subjective effort rating and the short-term speech transmission index (STIr) values, for the two masking noises: steady-state (SSN) and fluctuating (ICRA). The best-fitting regression curves with a logarithmic shape are also included.

## DISCUSSION

The statistical analysis of the results previously discussed returns information about the research questions of this work. Indeed, the statistical comparison of the residuals allowed drawing indications about the goodness of fit of each curve with respect to its data set, thus inferring statements about the statistical effect of the background noise type. A more sophisticated statistical analysis is under development, based on a permutation approach, effectively and directly dealing the statistical comparison of the regression curves, through the comparison of their estimated parameters. The permutation approach was chosen as being capable of dealing with small sample sizes of the distributions to be compared and non-linear regression curves.

Despite the limitations of the statistical method, some interesting aspects are pointed out by the results of the experiment. First, the use of the STIr metric allowed a meaningful comparison of speech reception data under stationary and fluctuating noises. Indeed, the objective metric is able to follow the time course of speech and noise, effectively tracking down the fluctuations of both signals. Thus, given the same reverberation time and the same long-term SNR, the STIr would be systematically larger for the fluctuating noise, which in the

short-term approach shows higher SNRs in correspondence of the dips of the masking noise. Then, the advantage experienced when listening in fluctuating noises, due to the “listening in the gaps” phenomenon, is already accounted for by the objective metric. This is reflected by the non-significant differences in the psychometric curves describing the intelligibility under the two different noises. It is noteworthy that the frame length of 186ms is consistent with the syllable duration in the Italian language of testing. On the other hand, also shorter time frames are suggested in the literature [20, 26] and the choice of one specific value is still a debated issue. By applying the STIr metric, the IS data show that, at least for the peculiar speech material used in the present experiment, no differences are detected in the accuracy results under SSN or ICRA noise. This finding is consistent with results of Ref. [1], stating that FMB would be no longer effective for comparable reverberation times.

Second, a slowing down of the participants’ RTs was systematically observed under the fluctuating noise. For both masking noises, the expected decreasing trend of RT results with the improvement of listening conditions was observed. Indeed, several studies dealing with the measure of RT in single-task paradigms pointed out that participants tend to respond slower in the more challenging acoustic conditions [9, 27], the difference reflecting an increase in the listening effort (i.e., the attentional and cognitive resources requested for speech reception [7]). An average RT bias of 180 ms under the ICRA noise versus the SSN was measured over the STIr interval under evaluation, and this finding shall be interpreted as a greater impairment put by the fluctuating noise on the listeners’ speech reception performance. Then, even though both noises produced an energetic masking of the speech signal, without any informative content, the presence of fluctuations in ICRA noise called for a greater amount of explicit information processing to accomplish the task. Since the bias is almost constant across the STIr interval, it can be hypothesized that the dependence of the effect on the acoustical conditions is quite weak in the analyzed range. This finding is consistent with the framework pointed out in Ref. [2], stating that when fluctuating noise was present during a speech perception task, listeners increasingly capitalized on cognitive resources (e.g., working memory) without any specific reference for the sound field. Then, present results generalize the findings to reverberated conditions for cases of rather unfavorable speech reception.

Third, differences between the two noises were also detected by the self-reported effort ratings of the participants. Interestingly, the pattern of the results was in agreement with the RT results, with the ICRA noise always evaluated as more effortful than the SSN. In fact, even though a similar accuracy performance was achieved under both noises, the presence of fluctuations is perceived as making listening more effortful, mirroring the longer response time required to accomplish the task. The results of the present experiment suggest the presence of a correlation between the behavioral measure and the subjective evaluation of listening effort, at least in the challenging listening conditions presented during the task. The presence of the correlation should be extensively checked, especially in the best acoustic conditions, where the inter correlations between subjective ratings and other evaluation methods for the listening effort may become unreliable [28] due to the intrinsic subjective nature of the self-reported measures (e.g., effects of age, individual thresholds, individual cognitive capacity).

It is to be remarked that the present acoustically unfavorable conditions are close to what is experienced in realistic settings inside reverberated public spaces, when spatially distributed unattended voices mask a target frontal voice placed at a critical communication distance. Despite the accurate reproduction of the acoustic communication channel (disregarding visual cues), the message exchanged is not fixed, and can be modeled in various ways having increasing complexity. From this point of view, the choice of the MWT can be regarded as close to a worst case. In fact, this test was also compared to other conventional tests [16] (in particular to the DRT in the Italian language and to the Matrix Sentence Test in Italian [29])

and it was found that longer latencies characterized MWT both in quiet and in noisy settings. When considering cognitively less demanding material the scenario would be modified accordingly, with very probable involvement of RT. Thus, the present choice helps in setting a trend for the behavior of effort that pertains to challenging situations as anecdotally reported by users, in particular when the benefit of context and of syntax is negligible. Further studies will include the study of less acoustically demanding conditions and a broader panel of listeners according for instance to age and to language proficiency.

## CONCLUDING REMARKS

This study explored the joint impact of reverberation and modulated background noise on both accuracy and listening effort in a speech reception task. The results pointed out that, when a short-term objective metric is used to describe properly the temporal envelope of the masker, the accuracy results are similar under stationary or fluctuating noise, the FMB being both reduced due to the presence of reverberation and accounted for by the STIr metric. Conversely, a longer response time is required for speech reception when fluctuations interfere with the speech signal, even without carrying any informative content. The finding could be related to an increased amount of explicit cognitive processing required to accomplish the speech reception task. A similar result appeared also from the analysis of the self-rating of listening efforts: despite individual differences and similar accuracy results, listeners perceived the presence of fluctuations as making listening more effortful.

## REFERENCES


[1] George, E. L., Festen, J. M., & Houtgast, T. (2008). The combined effects of reverberation and nonstationary noise on sentence intelligibility. *J. Acoust. Soc. Am., 124(2),* 1269-1277.

[2] Rönnberg, J., Rudner, M., Lunner, T., & Zekveld, A. A. (2010). When cognition kicks in: Working memory and speech understanding in noise. *Noise and Health*, *12(49),* 263-269.

[3] Rennies, J., Schepker, H., Holube, I., & Kollmeier, B. (2014). Listening effort and speech intelligibility in listening situations affected by noise and reverberation. *J. Acoust. Soc. Am., 136(5),* 2642-2653.

[4] IEC 60268-16 (2011). Sound system equipment – Part 16: Objective rating of speech intelligibility by speech transmission index (International Electrotechnical Commission, Geneva, Switzerland).

[5] Sato, H., Bradley, J. S., & Morimoto, M. (2005). Using listening difficulty ratings of conditions for speech communication in rooms. *J. Acoust. Soc. Am., 117(3),* 1157-1167.

[6] Jones, F. N., & Marcus, M. J. (1961). The subject effect in judgments of subjective magnitude. *Journal of Experimental Psychology*, *61(1),* 40-44.

[7] 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, 53(7),* 433-440.

[8] Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing*, *33(2),* 291-300.

[9] 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.

[10] Pals, C., Sarampalis, A., van Rijn, H., & Başkent, D. (2015). Validation of a simple response-time measure of listening effort. *J. Acoust. Soc. Am., 138(3),* EL187-EL192.

[11] Hicks, C. B., & Tharpe, A. M. (2002). Listening effort and fatigue in school-age children with and without hearing loss. *Journal of Speech, Language, and Hearing Research*, *45(3),* 573-584.

[12] Gosselin, P. A., & Gagne, J. P. (2011). Older adults expend more listening effort than young adults recognizing speech in noise. *Journal of Speech, Language, and Hearing Research*, *54(3),* 944-958.

[13] Picou, E., & Ricketts, T. A. (2015). Using dual-task paradigms to assess listening effort in children and adults. *J. Acoust. Soc. Am., 137(4),* 2236-2236.

[14] Prodi, N., Visentin, C., & Farnetani, A. (2010). Intelligibility, listening difficulty and listening efficiency in auralized classrooms. *J. Acoust. Soc. Am., 128(1),* 172-181.

[15] Prodi, N., Visentin, C., & Feletti, A. (2013). On the perception of speech in primary school classrooms: Ranking of noise interference and of age influence. *J. Acoust. Soc. Am., 133(1),* 255-268.

[16] Visentin, C., & Prodi, N. (2016). *Potentials of a matrixed word test for assessing speech reception in rooms with reverberation and noise*. Paper presented at EuroRegio2016, Porto, Portugal.

[17] 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.

[18] Giordano, R. (2005). *Note sulla fonetica del ritmo dell'italiano*. Paper presented at the 2nd Convegno Nazionale dell'Associazione Italiana di Scienze della Voce (AISV), Salerno, Italy.

[19] Dreschler, W. A., Verschuure, H., Ludvigsen, C., & Westermann, S. (2001). ICRA noises: artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment. *Audiology, 40(3),* 148-157.

[20] Payton, K.L., & Shrestha, M. (2013) Comparison of a short-time speech-based intelligibility metric to the speech transmission index and intelligibility data. *J. Acoust. Soc. Am., 134(5),* 3818-3827.

[21] van Schoonhoven, J., Rhebergen, K. S., & Dreschler, W. A. (2017). Towards measuring the Speech Transmission Index in fluctuating noise: Accuracy and limitations. *J. Acoust. Soc. Am., 141(2),* 818-827.

[22] Van Wijngarden, S., & Drullman, R. (2008). Binaural intelligibility prediction based on the speech transmission index, *J. Acoust. Soc. Am. 123(6),* 4514–4523.

[23] 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. *J. Acoust. Soc. Am., 134(4),* 3039-3056.

[24] Hasson, D., & Arnetz, B. B. (2005). Validation and Findings Comparing VAS vs. Likert Scales for Psychosocial Measurements. *International Electronic Journal of Health Education*, *8*, 178-192.

[25] Bonnini, S., Prodi, N., Salmaso, L., & Visentin, C. (2014). Permutation approaches for stochastic ordering. *Communications in Statistics-Theory and Methods, 43(10-12),* 2227-2235.

[26] Rhebergen, K. S., Versfeld, N. J., & Dreschler, W. A. (2006). Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. *J. Acoust. Soc. Am.,120(6),* 3988-3997.

[27] McGarrigle, R., Dawes, P., Stewart, A. J., Kuchinsky, S. E., & Munro, K. J. (2016). Pupillometry reveals changes in physiological arousal during a sustained listening task. *Psychophysiology*.

[28] Rudner, M., Lunner, T., Behrens, T., Thorén, E. S., & Rönnberg, J. (2012). Working memory capacity may influence perceived effort during aided speech recognition in noise. Journal of the American Academy of Audiology, 23(8), 577-589.

[29] G.E. Puglisi, A. Warzybok, S. Hochmuth, C. Visentin, A. Astolfi, N. Prodi, and B. Kollmeier (2015). An Italian matrix sentence test for the evaluation of speech intelligibility in noise. International journal of audiology, 54(Sup2), 44-50