ACOUSTICS 2012

Objective and subjective assessment of disturbance by office noise - Relevance of the use of the speech transmission index

A. Ebissou[a], P. Chevret[a] and E. Parizet[b]

[a]INRS, Rue du Morvan, CS 60027, 54519 Vandoeuvre Les Nancy, France
[b]Laboratoire Vibrations Acoustique INSA Lyon, 25 bis avenue Jean Capelle, Villeurbanne, F-69621 Lyon, France
ange.ebissou@inrs.fr

This experiment is part of a study aiming to assess the disturbance experienced by workers in open plan offices. Previous studies have shown that a sound environment rich with speech sounds can be detrimental to one's performance. The magnitude of this Irrelevant Speech Effect (ISE) depends on the intelligibility of the ambient speech. This has led to the use of the STI to model the induced decrement in performance. However, a decrease in performance is only one aspect of the more general concept of disturbance. When attempting to model the ISE in this regard, other components should be explored. In this first experiment, fifty-seven subjects perform a classical seriation task during 10-minutes blocks. They are confronted to sound environments typical of open-plan offices. In each block, a voice emerges from the noise, with a STI value in the 0.3 - 0.7 range. Both their performance and response times are recorded. A silent condition is used as a reference. After each block, they are presented with the NASA-RTLX questionnaire for a subjective assessment of their workload. Comparisons between speech conditions will be made in order to understand the influence of ambient speech intelligibility on objective and subjective disturbance.

1 Introduction

Noise is reported in many studies as the most disturbing physical nuisance in open-plan offices. It has been noted that speech noise is particularly annoying for office workers [1]. A need has thus emerged amongst occupational health specialists for a way to assess the acoustical quality of an office with regards to this problem.

For certain office-related tasks, notably those involving short-term memory, a drop-off in performance can be observed. This phenomenon is usually referred to as Irrelevant Speech Effect. In laboratory settings, it is often explored through a seriation task, where seven to nine elements (digits, letters or words) are presented in rapid succession to participants. After a short retention period, they are required to report back the series in the exact order of presentation. The ISE then seems to be related to speech intelligibility: the more the ambient speech is comprehensible, the harsher the decrease in performance is [2]. This is why an intelligibility index such as the STI has been used to propose target values for an open-plan office of good quality. In the same vein, Hongisto [3] proposed a STI-performance curve to predict the decrement of performance induced by ambient speech of a given intelligibility. Its shape can be seen on Figure 1. According to this model, for unintelligible speech, performance is the same as in silence. A growing impairment occurs for intermediate STI values ranging from 0.3 to 0.5, where the deterioration of performance is already approaching its maximum.

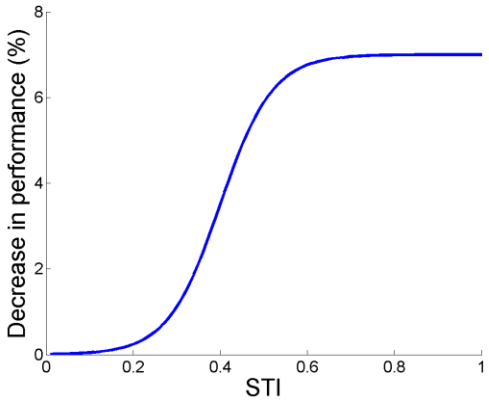

Figure 1: Schematic prediction model, giving the decrease in performance as a function of the STI.

For other common tasks, no ISE can be noticed [4]. As a consequence, when investigating the disturbing nature of ambient speech, one has to find alternative ways to estimate the difficulty of the task as perceived in a particular sound environment. In the ISE-related literature, questionnaires are the most frequently used method to obtain such an assessment. Their use is also relevant to complement performance measurements on short-term memory tasks. In that case, subjective assessments generally agree with objective ones. Participants express a higher level of disturbance when speech intelligibility increases. Nevertheless, both measurements do not seem to provide the same information with regard to the disturbance caused by the sound environment [4]. It should also be noted that the questions asked to participants vary across studies, which can make comparisons difficult. The use of a widespread workload measure could be useful for future reference.

The first goal of this study is to improve our understanding of the variations of performance with speech intelligibility, as measured by the STI. By exploring the [0.3-0.5] range, it aims to offer some material to further the work undertaken by Hongisto. By using an objective measure of performance and a subjective report of difficulty, the following experiment also seeks to compare these two types of assessments. Their respective evolutions with regard to the ambient speech intelligibility will be observed in order to compare their ability to discriminate between ambient speeches of different intelligibility levels, as measured by the STI.

2 Experiment

The experimental group consisted of 57 participants, one half being students and the other half recruited through an outsourcing company specialised in clinical trials. There were 32 women and 25 men, aged from 22 to 73 years, 36 years on average. All participants reported normal hearing and were paid for their participation.

Lists of sentences in French (299 sentences in total), designed for audiometry purposes, were used as the speech signal. The recordings were provided by the Collège National d'Audiométrie along with background noise rich with multi-talker babble [5]. The total duration exceeded 12 minutes. Four values of the STI were implemented: 0.3, 0.4, 0.5 and 0.7. A model proposed by Hongisto [6] predicts the STI between two nearby workstations. For a given office setting, the model provides octave-band attenuations which are applied to the spectrums of both speech and noise signals in order to obtain the corresponding STI value. The resulting spectrums are displayed in Figure 1. In this experiment, four settings were chosen, each leading to one of those four STI values.

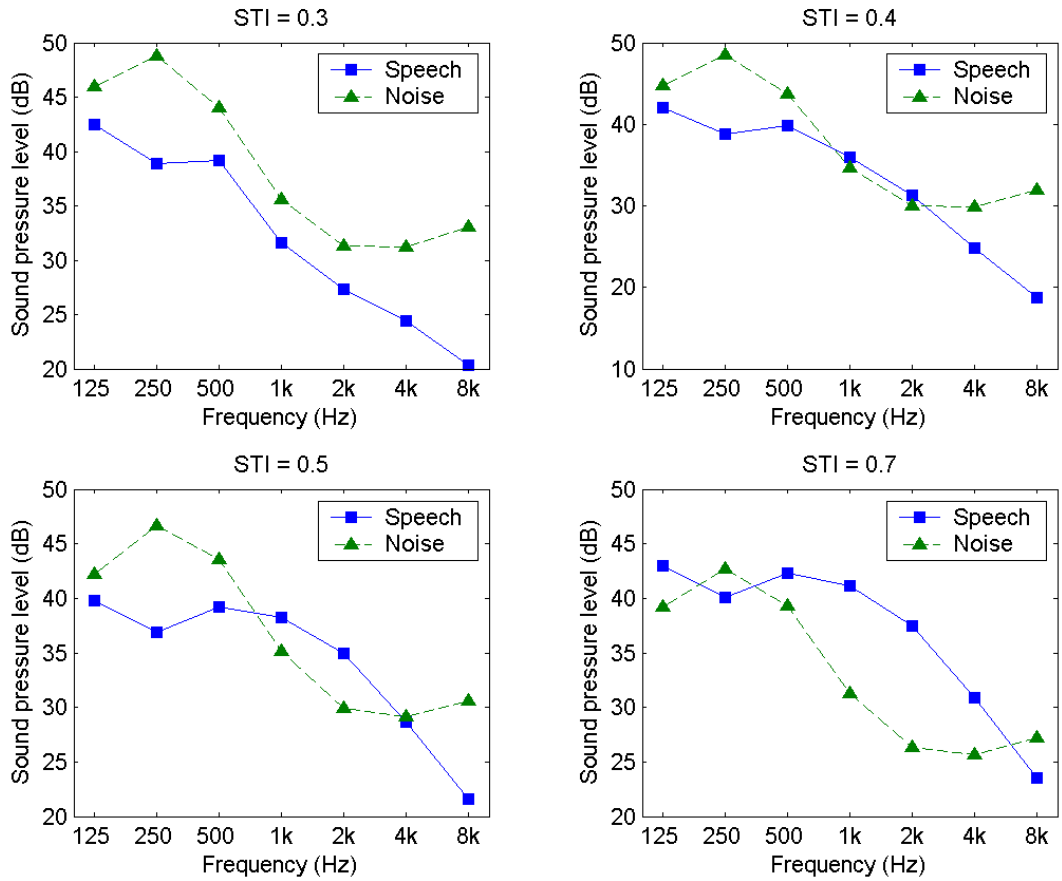

Figure 1: Time-averaged sound pressure levels of the speech and noise signals for each of the four noisy conditions. In each case, the total A-weighted SPL was 46 dB(A).

Participants worked over two sessions on consecutive days. Each session included three 10-minute blocks, separated by 5-minute pauses. One of them was carried out in silence, which served as a control condition for which STI = 0. For each of the other two, one of the four sound environments was used. At the end of the two sessions, each participant had been confronted to the four STI values. The order of the sentences was randomised before each block. The sound condition sequence was also balanced between subjects. The experiment took place in a sound-treated booth. Participants were alone in the booth, behind a desk. They faced a computer screen and were equipped with a mouse. Behind the screen was a loudspeaker, which created the sound environment. The A-weighted SPL was 46 dB(A) for all four noisy conditions.

Participants had to acheive a seriation task. A permutation of integers from 1 to 9 was presented, one number at a time (700 ms on, 300 ms off; MS Sans Serif font, 3 cm high). After the last number had disappeared, a five-second pause was observed. The individual numbers then came into view in a 3x3 response array. Participants were asked to reproduce the series in the exact order of presentation by clicking on the buttons. After clicking, the number disappeared and could not be selected again. The next series began 3 seconds after the ninth number was proposed. The response phase was self paced, which means that participants carried out as many series as they could in the 10-minute block. Performance was measured by first reporting the number of figures incorrectly placed in any series. The average number of errors per series in a block was then calculated.

After each block, participants were presented with a computer-based French version of the NASA-TLX questionnaire [7]. This workload assessment method comprises six questions, each of them addressing a particular component of the mental workload as detailed in Table 1.

Table 1: Overview of the NASA-TLX questionnaire

Question 1	Mental demands of the task
Question 2	Physical demands of the task
Question 3	Temporal demands of the task
Question 4	Self-rated performance
Question 5	Effort level
Question 6	Frustration

For each question, participants gave out a score comprised between 0 and 100 using a 21-point scale. The lower the score, the more comfortable the situation was reported to be. A Raw Task Load Index (RTLX) was eventually calculated by averaging the six scores [8].

At the beginning of the first session, subjects benefited from a 4-minute trial session, in silence, followed by a discovery of the questionnaire.

3 Results

Statistical analyses were led using Stata v12.0. A high degree of heteroscedasticity for both performance and RTLX excluded the use of a classical ANOVA. Moreover, the data presented a hierarchical structure: blocks were carried out on a given session by a given participant. Therefore, within-subject and within-session correlations had to be taken into account. These observations led to the use of a three-level mixed-effects regression model, with a dummy variable for each condition. After modelling, it was then possible to assess separately between-subject variability, the learning effect between sessions and the residual variance.

3.1 Performance

In the silent conditions, participants averaged 2.2 errors per series. A significant effect of the sound condition could be found ($\chi^2 = 22.84$, d.f. = 4, $p < 0.05$). For STI = 0.7, the decrease in performance (DP) was at its highest at 5%. Figure 3 shows the decrease in performance for each of the four noisy conditions along with the corresponding standard errors.

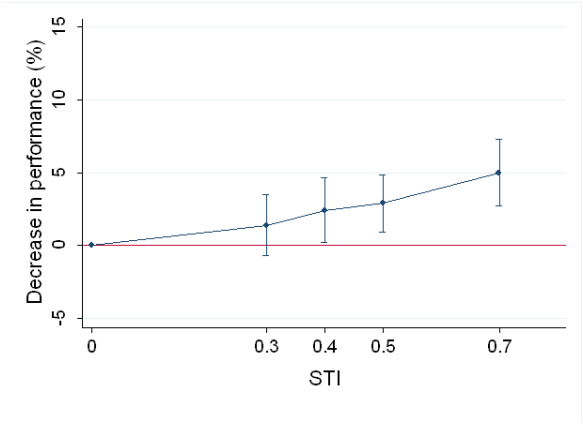

Figure 3: Decrease in performance for the seriation task in sound environments of various STI values.

Participants were significantly impaired by the noise when the STI exceeded 0.5 ($\chi^2 = 8.18$, d.f. = 1, $p < 0.05$, Bonferroni-adjusted). The protocol was not powerful enough to give rise to significant differences between the intermediate STI values.

An agglomerative hierarchical clustering showed that the group of participants could be divided into two separate groups with distinctive sensitivities to the sound environment, as shown in Figure 4. Neither age nor sex had a significant influence on the classification.

The first group is composed with high-performing subjects (32 members, 1.8 errors per series on average in the silent condition). Their behaviour depended very little on the sound condition, as no significant effect of STI value could be found.

The average performance in the second group was lower (2.8 errors per series on average in the silent condition, $\chi^2 = 7.89$, d.f. = 1, $p < 0.05$). Most noticeably, its members displayed a higher sensitivity with regards to the sound environment. The decrease in performance reached 11% for STI = 0.7. The effect of STI value on DP was significant ($\chi^2 = 93.36$, d.f. = 4, $p < 0.05$). Noise was significantly detrimental to performance when compared to the silence condition ($\chi^2 = 15.43$, d.f. = 1, $p < 0.05$). More errors are made for the two highest levels of STI (0.5 and 0.7) than for the two lowest ones (0.3 and 0.4; $\chi^2 = 11.46$, d.f. = 1, $p < 0.05$). The difference in DP was higher between the 0.4- and 0.7-levels (5.6 percentage points) than between the 0.3- and 0.5-levels (4.5 percentage points). This result does not comply with the shape of the curve in the prediction model.

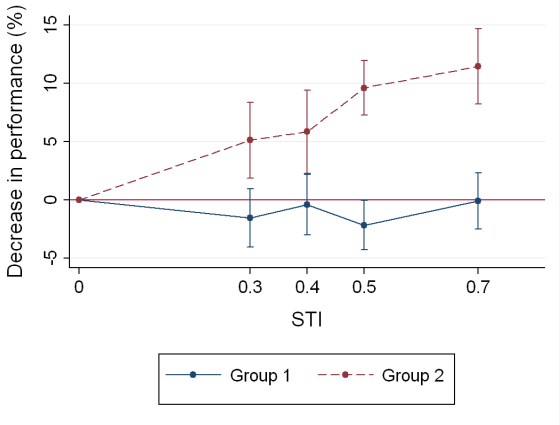

Figure 4: Decrease in performance for the two groups of participants.

Adding speech noise affected negatively the good execution of a seriation task. This phenomenon grew stronger as the ambient speech noise intelligibility increased. Nevertheless, any analysis should allow for the high level of between-subject variability. A sizable portion of the panel exhibited no sensitivity to noise. The relationship between the decrease in performance and the value of the STI did not conform to the curve proposed by Hongisto. This sigmoid function was chosen because it mimics the shape of the curve describing the dependence of subjective intelligibility of sentences on the STI. This assumption suggests that the level of performance in the task at hand depends on the meaning of the ambient speech. However, it does not appear to be the case for many tasks, such as simple seriation tasks [9]. A shape approaching the curve relative to CVC-syllables [5] may be more adequate, as these elements do not carry any meaning.

3.2 Mental workload

There was a significant effect of the sound condition ($\chi^2 = 114.63$, d.f. = 4, $p < 0.05$). The reported RTLX was much higher in a noisy environment than in silence, as shown in Figure 5. This difference should not, however, be attributed solely to a change in STI. It may be a known fact that performance in a seriation task is not sensitive to the noise level [9], but it is not necessarily the case for mental workload assessments.

Once again, no significant difference could be made between the intermediate STI values. In order to show whether both objective and subjective assessments of disturbance vary in the same way with the STI, separate analysis were led with the same groups that were presented in §3.1. The results are shown in Figure 6. Subjects drawn from the first group reported lower RTLX scores than their

counterparts from the second group ($\chi^2 = 16.96$, d.f. = 1, $p < 0.05$). Nevertheless, the clustering failed to distinguish different response patterns. For both groups, the three intermediate conditions were not significantly different from one another.

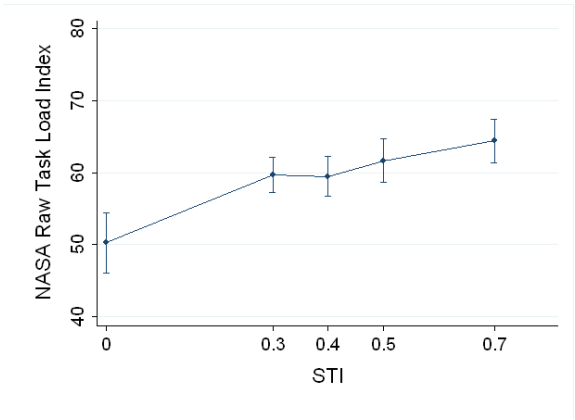

Figure 5: NASA-RTLX scores for the seriation task in sound environments of various STI values.

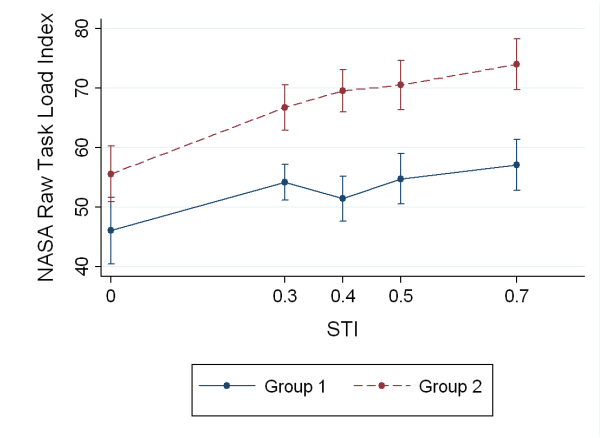

Figure 6: NASA-RTLX scores for the two groups of participants.

The NASA-TLX was conceived and validated as a global object, taking in consideration all questions. This makes any study of responses to an individual question theoretically unsound. Nevertheless, it can be a useful diagnostic tool to understand which questions bear the most importance in the variations of the global rating. The range of predicted scores for all 6 questions in noisy conditions, expressed as a relative increase from the lowest predicted score for each question, is presented in figure 7.

It appears that questions 2 (related to the physical demand of the task) was the one for which the range of predicted scores was the widest. This is quite surprising because the task does not seem to be physically strenuous, which should make the question irrelevant to our experiment. The study of variance partition coefficients (VPC) shines a light on that fact: 70% of the total variance of the scores for question 2 could be explained by between-subject variability and only 2% by the sound condition. These numbers tend to put the previous observation regarding the high relative increase into perspective. As a matter of fact, despite the seemingly great differentiation, the residual error was too important for the effect of STI level to be significant. No significant differences between noisy conditions could either be noticed for questions 1 and 3. These results indicate an increase in STI did not change the way participants perceive the difficulty of the task.

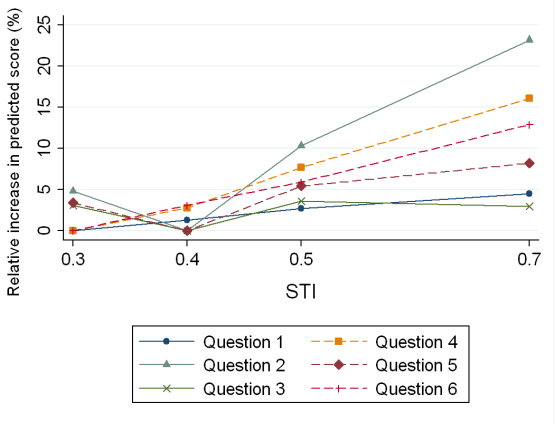

Figure 7: Increase in predicted score relative to the minimum for noisy conditions for the 6 questions of the NASA-TLX questionnaire.

It should be noted that question 3 was treated very differently from subject to subject. While some considered primarily the time pressure they felt during the presentation of the numbers, others expressed that, as the response phase was self-paced, there was no particular temporal demand to report. This could account for the limited variations of the reported scores for this item. On a related issue, response time was also measured. No significant effect of the sound condition could be found.

When considering the noisy conditions only, the effect of the STI value on reported scores reached significance for questions 4 ($\chi^2 = 9.61$, d.f. = 3, $p < 0.05$) and 5 ($\chi^2 = 9.30$, d.f. = 3, $p < 0.05$). These questions refer respectively to self-rated performance and effort level. Participants reported that an increase in STI forced them to work harder but still altered the quality of their output.

Adding speech noise increased the mental workload of a seriation task as reported by participants. The workload was stronger for ambient speech of high intelligibility. Noise affected less the perception of difficulty by the subjects than the appreciation of their work during the execution of the task. The information given by these subjective reports was not redundant with what the objective measurement of performance could bring by itself.

4 Conclusion

According to both objective and subjective measurements of disturbance, adding speech noise to a sound environment is detrimental to one's work. The level of disturbance grows as the intelligibility of the ambient speech increases, albeit not in the same way for both types of assessments. One should nevertheless be aware of the high level of between-subject variability when analyzing the results of such an experiment. This problem may be mitigated by enlisting a greater number of participants. Further experiments are planned and a few remarks pertaining to the making up of an adequate protocol can be made.

Regarding the performance measure, the results show that the shape of the general STI-performance curve may

not be adequate for every type of tasks. Future experiments should be centered on other office-related tasks. Moreover, it could be interesting to explore the intelligibility spectrum in a novel way, such as modifying the global signal-to-noise ratio without altering the STI.

The mental workload assessment failed to distinguish the intermediate conditions. In order to compare correctly the scores for STI = 0 to the other levels, future protocols should include a sound condition consisting of unintelligible background noise, at the same level than other noisy conditions. It would also be profitable to implement a time constraint on the response phase. This change could possibly modify the way participants perceive the difficulty of the task as a function of ambient speech intelligibility.

References

[1] S.P. Banbury, D.C. Berry, "Office noise and employee concentration: Identifying causes of disruption and potential improvements", *Ergonomics* 48, 25-37 (2005)

[2] S.P. Banbury, W.J. Macken, S. Tremblay, D.M. Jones, "Auditory distraction and short-term memory: Phenomena and practical implications", *Human Factors: The Journal of the Human Factors and Ergonomics Society* 43, 12-28 (2001)

[3] V. Hongisto, "A model predicting the effect of speech of varying intelligibility on work performance", *Indoor Air* 15, 458-468 (2005)

[4] S.J. Schlittmeier, J. Hellbrück, R. Thaden, M. Vorländer, "The impact of background speech varying in intelligibility: Effects on cognitive performance and perceived disturbance", *Ergonomics* 51, 719-736 (2008)

[5] http://www.college-nat-audio.fr/accueil.html (as seen on February 29th, 2012)

[6] V. Hongisto, J. Keränen, P. Larm, "Simple model for the acoustical design of open-plan offices", *Acta Acustica united with Acustica* 90, 481-495 (2004)

[7] S.G. Hart, L.E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research". In P.A. Hancock, N. Meshkati (Eds.), *Human Mental Workload*, North Holland Press, Amsterdam (1988)

[8] J.C. Byers, A.C. Bittner, S.G. Hill, "Traditional and raw task load index (TLX) correlations: Are paired comparisons necessary?" In A. Mital (Ed.), *Advances in Industrial Ergonomics and Safety I*, Taylor and Francis (1989)

[9] S. Tremblay, D.M. Jones, "Change of intensity fails to produce an irrelevant sound effect: Implications for the representation of unattended sound", *Journal of Experimental Psychology: Human Perception and Performance* 25, 1005-1015 (1999)