An attempt to predict ISE by a spectral estimator

Toros Ufuk Senan[1], Armin Kohlrausch[2], Sam Jelfs[1]

[1] Philips Research Laboratories, Eindhoven, The Netherlands (corresponding author)

[2] Human-Technology Interaction, Eindhoven University of Technology, Eindhoven, The Netherlands

Corresponding author's e-mail address: toros.senan@philips.com

ABSTRACT

The distractive effect of background sounds on cognitive performance is investigated using the paradigm of irrelevant sound where the effect is called the irrelevant sound effect (ISE). The effect is quantified by comparing cognitive test scores under different acoustic conditions. Even though the acoustic properties are well established and three predictors have been proposed in literature, a single metric that relates the cognitive distortion to an acoustic feature has not yet been successfully developed. The present work investigates one of these estimators, a spectral parameter proposed to be an ideal metric to predict ISE: frequency-domain correlation coefficient (FDCC). The parameter measures the spectral variation between perceptually distinguishable segments of distracting sounds. In order to evaluate FDCC, alternating noise pulses and noise-vocoded speech stimuli are generated in a way that the spectrum of the adjacent segments of the sound varies systematically. Finally, the stimuli are employed in short-term memory tasks and the parameter is evaluated under the light of the test scores.

INTRODUCTION

Distractive effects of background sounds were investigated under the paradigm of irrelevant sounds and the effect was labelled as the irrelevant speech effect (ISE) [1]. It was soon discovered that non-speech sounds also develop cognitive distortion so the phenomenon was renamed as the irrelevant sound effect while keeping the acronym the same [2, 3]. Laboratory studies showed that the effect is robust: Extraneous sounds impair memory performance. However, the degree of disruption depends on both the properties of the irrelevant sounds [4, 5] as well as those of the cognitive task [6, 7, 8]. And even though the key functions of the ISE has been well established, developing a successful predictor was shown to be a complicated process [9, 10].

Typically, ISE can be quantified in the lab environment where participants perform a certain cognitive task while being exposed to background sounds which are not relevant. The test scores obtained under different acoustic conditions are compared to quantify the effect. One of the most common tasks employed in the ISE literature is a serial-recall task: To-be-remembered items (e.g. letters or digits) are presented in a randomized order on a computer screen together with irrelevant acoustic stimuli to the participants and participants are asked to recall the order of the items presented. Although the brain tries to ignore the acoustic stimuli, it

still processes the visual as well as the acoustic input which eventually diminishes the cognitive performance.

This unavoidable conflict is often explained by the changing state hypothesis [11]. The hypothesis states that an acoustic stimulus containing segments of sound which differ in terms of acoustic properties, produces much more disruption than a steady-state sound. Further studies revealed that while successive sound tokens changing in frequency disrupt the memory performance [12, 13], varying the sound pressure level does not create disruption [14, 15]. The ISE is observed under any acoustic condition which satisfies the changing state hypothesis: background music [16], alternating tones [13, 4], background noise [17] as well as native, foreign and reversed speech [18].

The changing state hypothesis attempts to explain the disruption by structuring a guideline in order to observe the effect. However, the magnitude of the effect is much more complicated to predict since the disruption reaches a maximum using speech [3, 19, 20], suggesting that the proper explanation of the phenomenon should be derived from speech perception properties rather than global acoustic features. Two predictors from the literature follow this reasoning: the speech transmission index (STI) and the fluctuation strength (FS).

The speech transmission index (STI) is a room acoustic measure derived from the results of subjective intelligibility tests typically conducted in enclosed acoustic environments. In order to minimize the costs and the time spent for such intelligibility tests, the concept of the modulation transfer function was applied to predict speech intelligibility [21]. STI is a temporal distinctiveness metric which is defined by the amplitude modulation ratio between the modified signal (e.g. recorded signal) and the original (e.g. source signal) which was applied into a sigmoidal function in order to predict the error rates of the cognitive task within the context of ISE [22]. There are limitations of the model: The model seemed to be suitable for degraded speech stimuli only and it requires a reference signal which is not always available. More recent literature also reports that STI can reveal unrealistic values for non-speech stimuli [9].

A second model, based on the fluctuation strength [23], is a psychoacoustic sensation dating back to the original work of Zwicker and Fastl [24]. FS is perceived when listening to slowly modulated (<20 Hz) sounds and it works in a way that FS value, expressed in vacil, reaches the maximum with fluctuations of approx. 4 Hz, which is close to the value of the average syllable rate in running speech [25]. FS was first used in the study of Schlittmeier *et al.* as a prediction model of ISE [19]. The experimental results and the prediction values shared 50 % of the variance. More specifically, the values generated by the prediction model were within the interquartile range of the error rates for 63 out of 70 cases, including both speech and non-speech sounds. This is quite an impressive outcome. However, most relevant to this paper, the model lacks the ability to distinguish amplitude and frequency modulation. Such a limitation would result in predicting a maximum error rate for amplitude-modulated white noise with the modulation rate of 4 Hz, which in fact does not create an ISE.

A more recent experiment, which compared the aforementioned two models, demonstrated this shortcoming [9]. The experiment employed noise-vocoded speech as the distracting stimulus in the serial-recall task with the 1-band noise-vocoded speech condition, which is a band-limited white noise mapped to the intensity modulation of the original speech, resulting in the highest FS value while showing no difference in the test scores when compared to silence.

The only frequency-domain metric that follows the guidelines of the changing state hypothesis is called the frequency-domain correlation coefficient (FDCC) and was first mentioned in the study of Park *et al* [26]. The study built an adaptive masking scheme to systematically modify the distracting speech stimuli and the experimental results were investigated under the light of the prediction model. Although the results were promising, this is the only study where the behaviour of the model was observed under speech conditions. In another study [27], which

forms the basis of the first experiment in the current paper, white noise pulse trains were generated in order to obtain a set of stimuli where the spectral and temporal features were systematically modified. A temporal metric, the average modulation transfer function, was also presented in the study in order to evaluate the effect of temporal features of the background sounds on memory performance. However, test scores did not yield a meaningful trend, and therefore it was not possible to evaluate the metrics successfully.

Two experiments where the spectral features of the stimuli were manipulated systematically in order to evaluate the estimator are presented in this paper. First, a white noise pulse train was modified such that the temporal and spectral features varied independent from each other. Second, a noise-vocoding technique was used to decrease the spectral parameter value in a controlled manner by increasing the number of frequency bands employed. Spectral and temporal parameters are explained in detail in the next sections, followed by the two experiments. The results are examined in the discussion section and the metric is evaluated in the conclusion section.

FREQUENCY - DOMAIN CORRELATION COEFFICIENT

The frequency domain correlation coefficient (FDCC) is a correlation measure between successive segments of a sound in the frequency domain. It was proposed as a spectral distinctiveness metric [26] and attempts to explain the behaviour of the ISE by determining the spectral similarity between adjacent tokens of the sound.

In order to determine the positions of the sound tokens, the intensity envelope of the signal is obtained by squaring and applying a second order low-pass filter at 10 Hz. The median of the resulting envelope is computed and the signal parts with lower amplitude are eliminated. The median duration of the time intervals of the potential tokens are obtained and determined as a threshold. The intervals shorter than the threshold are discarded. For each of the remaining tokens, octave band-pass filters, with centre frequencies ranging from 125 Hz to 8 kHz, are applied. The power spectrum, *P*, is calculated for each octave band of each token. The FDCC is formulated as follows:

$$\boldsymbol{F} = \frac{\sum_{j=1}^{K=19} \boldsymbol{P}_{i,j}\boldsymbol{P}_{i+1,j}}{\sqrt{\sum_{j=1}^{K=19} \boldsymbol{P}_{i,j}^{2}\boldsymbol{P}_{i+1,j}^{2}}} \qquad (1)$$

where $P_{i,j}$ indicates the 1/3-octave band power spectrum for the token, *i*, in the frequency band, *j*. The FDCC shows variation in the spectrum from one segment of the sound to the next one. A low correlation value is expected to create a large amount of disruption within the context of ISE since it indicates high spectral distinctiveness.

AVERAGE MODULATION TRANSFER FUNCTION

The modulation transfer function (MTF) was defined as the modulation index reduction of the intensity envelope as a function of modulation frequency [22]. The modulation index can be quantified by comparing the temporally modified signal with the reference signal.

First, an octave band analysis is carried out (125 Hz – 8 kHz). Second, the intensity envelope for each octave band is computed by squaring the output. The resulting signal is then low-pass filtered with a cut-off frequency of 30 Hz and analyzed with a 1/3-octave band pass filter for the modulation frequencies ranging from 0.5 Hz to 16 Hz in order to cover the range of modulation frequencies typically found in speech. For the last step, the root-mean-square (RMS) value of the intensity envelope was computed and normalized.

The result was a K x N matrix of modulation index values, where K and N refer to the number of octave bands and number of the modulation frequencies, respectively. The modulation

index for each octave band and for each modulation frequency, $m_{ij,}$ was compared with the reference signal, resulting in a new K x N matrix defining the changes between the modified and reference signal.

$$\boldsymbol{M_{ij}} = \frac{\boldsymbol{m_{ij,x}}}{\boldsymbol{m_{ij,ref}}} \qquad (2)$$

In order to establish a single value for the metric, the resulting MTF matrix was averaged in both dimensions and the parameter, AMTF, was obtained.

EXPERIMENT 1

Stimuli

The first experiment focused on the relation between spectral and temporal features of distracting stimuli and the ISE, by modifying the spectral and temporal parameters independently.

White noise was preferred since its flat spectrum provides an equal gain over all bands which enables modification in a controlled manner. White noise, *G(t),* was shaped by a Hanning window, *W(t)*, with the size of *w*. The pulse, *WG(t)*, was filtered into 1/3-octave bands with center frequencies ranging from 125 Hz to 8 kHz. Out of the nineteen pulses obtained from each 1/3-octave band, seven pulses which correspond to the 7 octave bands were chosen in order to avoid an overlap between different octave bands.

The seven selected bands were summed and a pulse, with the duration of *w,* was synthesized. The pulse (P1) was generated by summing all the seven selected bands, of each new selected pulse x_i, where *i* indicates corresponding octave band.

$$\boldsymbol{P1} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{K=7}} \boldsymbol{x_i(t)} \qquad (3)$$

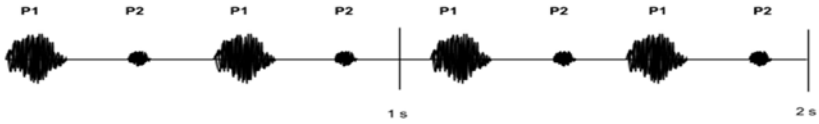

Figure 1: Two seconds segment of the reference signal. P1 and P2, alternate every 250 ms.

First, a half-second basic signal which consisted of two pulses of 50 ms duration was created. The amplitude of the first pulse, P1, was set to 0.9, and the amplitude of the second pulse, P2, was set to 0.3. A 1 min reference signal was generated where the two pulses alternated. The distance between pulses was kept constant at 250 ms for all the stimuli and P2 was modified in the time and frequency domain. A set of stimuli with a wide range of temporal and spectral variations was obtained.

Changes in Time Domain: AMTF Modification

For temporal variation, the width of P2 was systematically modified from 50 ms to 450 ms with a step size of 25 ms. It was observed that enlarging the width of the second pulse enabled AMTF values to drop while spectral parameter values remained constant. A subset of stimuli was created for experiment 1. The AMTF and FDCC values as a function of the pulse width sizes are presented in Figure 2.

Changes in the Frequency Domain: FDCC Modification

A set of different gains to each octave band (125 Hz - 8 kHz) was applied for P2. Since modifying the gain structure of each octave-band would result a change in temporal features of the signal, AMTF values were checked in each octave band after different gains were applied. In order to find the most suitable gain values, which satisfy a variation in FDCC while keeping the AMTF constant, a numerical optimization was applied. The *fmincon* function in MATLAB was used to find the optimal gain, where the maximum number of iterations was set to 100 with the 0.01 tolerance of the AMTF value.

Both of the parameter values for the generated stimuli are presented in Figure 3.

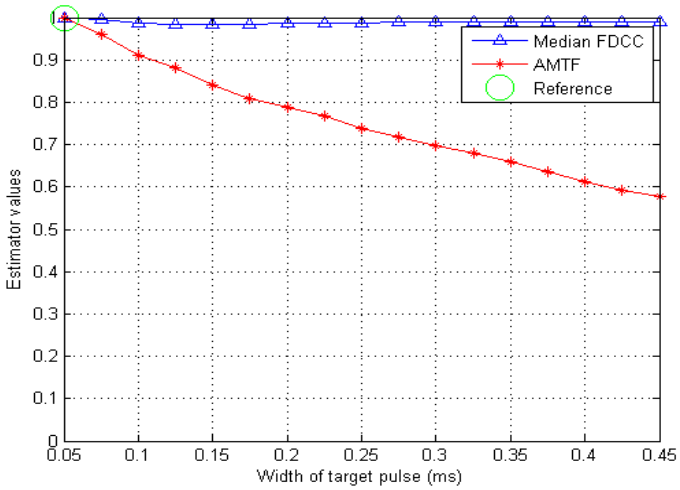

Figure 2: FDCC and AMTF parameter values as a function of the pulse width of the second pulse, P2.

Method and Procedure

The serial-recall task began with three asterisk signs disappearing from the screen one by one indicating that the task will begin in three seconds. Nine digits (1-9) were presented on a computer screen to the participant. Numbers were displayed one by one every second, while each number was shown for 0.7 s followed by a 0.3 s pause. The presentation order was randomized in a way that consecutive numbers were not presented either in ascending or descending order.

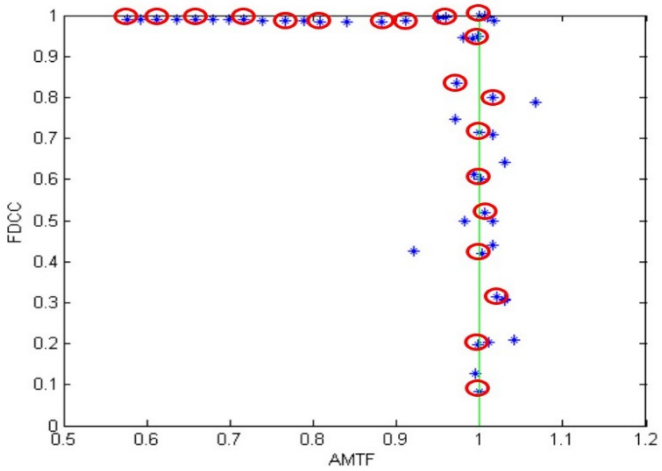

Figure 3: Each blue point represents one stimulus with two parameter values; the y-axis shows the values of the spectral parameter and x-axis shows the temporal parameter. Blue points circled in red represents the parameter values of the stimuli employed in the experiment.

After the presentation stage, a 10s retention period was inserted where the 10 asterisk signs disappeared one by one on the computer screen. For the recall stage, the participant was asked to click on the corresponding key buttons of the number pad on the screen in the correct order. The lay out of the key buttons was arranged in a randomized manner for every trial in order to eliminate the visual cue. In addition to that, the number keys disappeared from the screen after being pressed so the same number could not be selected more than once, and there was no option to correct the key input.

The experimental design consisted of five blocks where every session began with the training block which consisted of four trials without any sound (silence). After the training block, the responsible researcher asked the participant if there were any problems regarding the test conditions before advancing in the session.

There were 22 trials in each of the remaining four blocks. Ten trials were accompanied by a white noise pulse train with different FDCC values and 10 trials with different AMTF values. One trial in each block was carried in silence. The first trial of every block was the dummy trial, and the rest of the stimuli were presented in the randomized order.

There were 5 min breaks between the blocks and the pilot tests showed that the whole session took 60-65 mins to be complete, including the breaks.

Participants

Ten volunteers from the Philips Research Laboratories in Eindhoven participated in the experiment (4 = females, 6 = males). They all reported normal hearing and vision by signing the corresponding bullets in the informed consent form and this was double checked at the end of the first block. The age range of the participants was 18-50 years.

Material and Apparatus

The experiment was designed using MATLAB (R2014a) and run on a Hewlett-Packard computer. Auditory stimuli were presented diotically in MATLAB via a PC soundcard (RME Hammerfall DSP Multiface). The sessions took place in a double-walled IAC soundproof booth (Industrial Acoustics Company GmbH) at Philips Research Eindhoven and Beyer-Dynamic DT 990 headphones were used for playback. The average sound level of the stimuli was calibrated to 60 $dB_{LAeq1min}$.

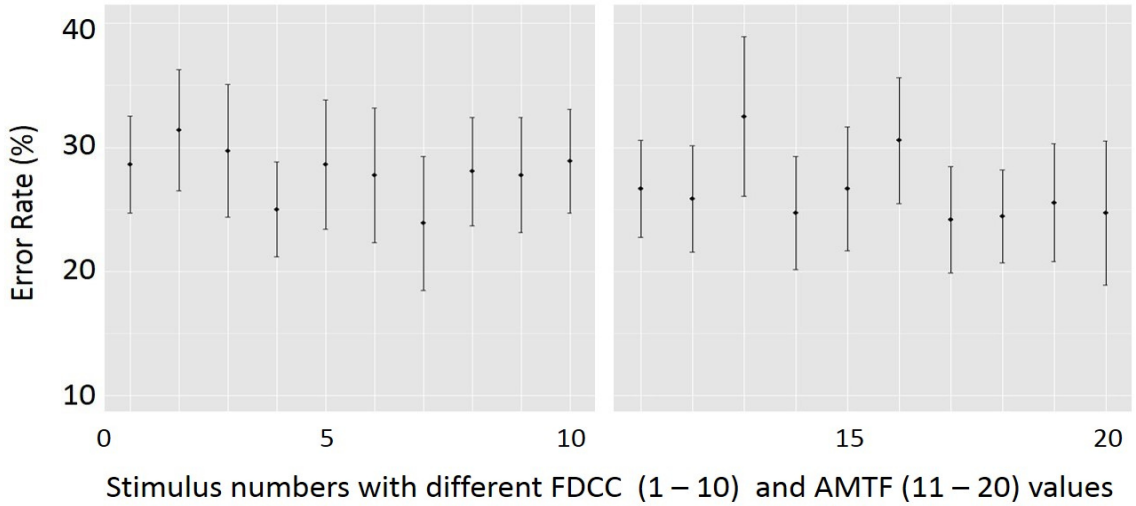

Figure 4: Mean error rate percentages for 10 participants using the noise stimuli with different parameter values. The x-axis represents the stimulus number with 10 different values of FDCC (1-10), and 10 different values of AMTF (11-20). Error bars represent the standard error of the mean (SEM).

Results

Only the digits recalled in their previously-presented serial position were marked as a correct response. The first trial, being the dummy trial, of each block was discarded from the analysis and therefore error rates represent the percentage of the incorrect answers for 21 trials in each test block. The mean error rates as a function of parameter values, for both estimators, are presented in Figure 4 for 10 participants.

The first observation was that the experimental results did not yield a meaningful trend as a function of parameter values. In fact, in contrast with the literature, the highest score was not obtained under the silence condition while the error rate (28 %) was as high as expected [26, 19]. Second, the spectral modification (FDCC) did not disrupt the performance any more than the temporal changes (AMTF), as opposed to the hypothesis. The mismatch between performance scores and parameter values, AMTF and FDCC, showed that the estimators are currently inadequate to predict ISE.

EXPERIMENT 2

For the second experiment, a noise-vocoding technique was used to create the distractive stimuli. There were two main reasons behind the choice: First, the lack of ISE in the first experiment might have occurred because of the non-speech like temporal structure of the stimuli and second, noise-vocoded speech (NVS) stimuli can be systematically modified in the frequency domain by changing the number of frequency bands it consists of.

Stimuli

Noise-vocoded speech is a manipulation of running speech where speech is filtered into frequency bands and the fluctuations of each frequency band are mapped to band-limited white noise.

NVS was generated by dividing the speech signal between 50 and 8000 Hz into Hanning-shaped bandpass filtered frequency bands. The width of the transition region between passband and stopband, the – 6dB cut-off frequency point, was determined by dividing the upper cut-off frequency of each band by 10. The speech signals were processed by employing the scripts used by a speech comprehension study [28] in Praat software (available at www.praat.org).

The cut-off frequencies were determined using the Greenwood function which forms the mathematical basis of the cochlear implant array placement [29]. The number of frequency bands with the cut-off frequencies of the final stimuli are shown in Table 1.

NVS was synthesized by replacing information in each frequency band with amplitude-modulated band-limited noise and combining the resulting modulated noise bands. This technique enables the acquisition of a set of stimuli where the non-disruptive amplitude modulated white noise, 1-band NVS, was systematically transformed into a highly disruptive intelligible NVS by increasing the number of frequency bands. More important to the current study, such transformation allows an organized modification of the spectrum (see Fig. 5).

The original sentences were taken from a speech reception study [30]. The samples consist of 10 lists with 13 short Dutch sentences (6-8 s) per list. Short sentences were concatenated in order to create 10 long sentences (42-55s) and these sentences form the basis of the NVS stimuli. There were 7 acoustic conditions employed in the experiment: 1- band NVS, 2-bands NVS, 4-bands NVS, 6-bands NVS, 18-bands NVS, original speech and silence (SLNC).

Method and Procedure

The serial-recall task used in the first experiment was also employed for the second experiment, with differences in the experimental design as well as the stimuli. There were 8 blocks in each session where the first block was the training block, which consisted of 8 trials in the silence condition, and the remaining 7 blocks consisted of different acoustic conditions which were mentioned above. The presentation order of the block was randomized in a way that the silent block (control condition) was always the fourth. Each block consisted of 10 trials which were derived from 10 different original speech samples. One session took 60-65 min to complete including 2 min breaks after each block.

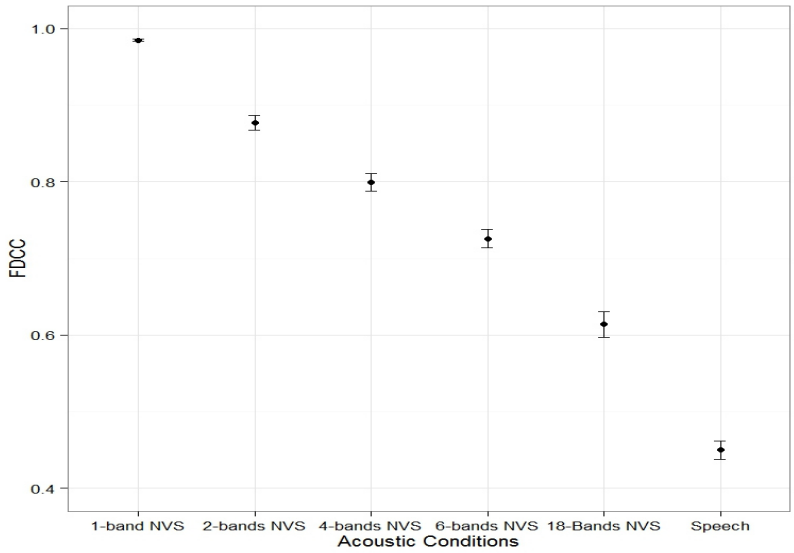

Figure 5: FDCC values as a function of the number of frequency bands of NVS and original speech stimuli. Error bars represent SEM.

Table 1: Noise-vocoded speech stimuli with the number of frequency bands and lower frequency boundaries.

Number of frequency bands	Cut-off frequencies
1-band NVS	50
2-bands NVS	50, 1160
4-bands NVS	50, 370, 1160, 3125
6-bands NVS	50, 229, 558, 1160, 2265, 4289
18-bands NVS	50, 98, 157, 229, 317, 425, 558, 720, 918, 1160, 1457, 1820, 2265, 2809, 3474, 4289, 5286, 6506

Participants

Twenty-five native Dutch-speaking subjects (15 = females, 10 = males, between 18-50 years old) participated in the second experiment. They all reported normal hearing and vision which was a pre-condition of the recruitment procedure. They signed the informed consent forms where the criteria were cross-checked. Participants were paid a modest compensation fee for their contribution.

Material and Apparatus

The experiment was run on a Hewlett-Packard computer using MATLAB (R2014b). All acoustic conditions were delivered diotically in MATLAB via a PC soundcard (M-Audio Transit). The participants were positioned in a double-walled IAC soundproof booth in the auditory lab of the Human Technology Interaction department at the Eindhoven University of Technology, and Sennheiser HD Linear 265 headphones were used for playback. The average sound level of the stimuli was calibrated to 60 $dB_{LAeq1min}$.

Results

The analysis began by looking for a learning effect. The test scores were analyzed based on the presentation order, regardless of the sounds delivered. Mean error rate differences between test blocks and the training block were calculated and the results were summed with the mean error rate of the corresponding test block. The curve of the learning effect showed that the overall performance had a sudden increase until the end of the third block and then it was stabilized. The effect was highly significant and confirmed by repeated measures ANOVA, F (7, 168) = 11.59; p < .001. Statistical analysis presented in the following refers to the corrected scores.

Figure 6 shows the test scores, represented as error rate percentages per acoustic conditions. The difference between the original mean error rate in the speech (41 %) and silence (32.3 %) conditions was slightly lower than what is reported in the literature, while silence condition yielded similar performance when compared to the literature [13, 19, 15]. The decrease in the performance as a function of the number of frequency bands between the 1-band and 4-bands conditions was similar to the study of Ellermeier *et al*. [9] while the decrease in mean error rate in the 18-band NVS condition was unexpected.

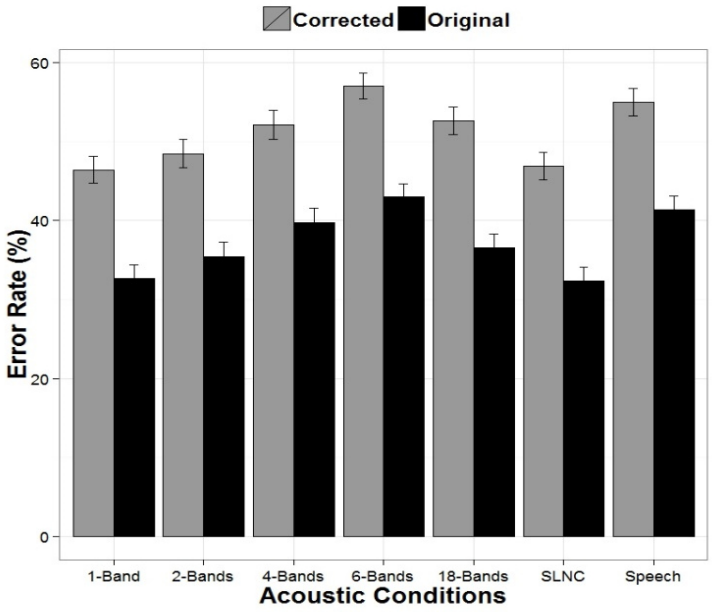

Figure 6: Mean error rate percentages of 25 participants as a function of the experimental conditions. Dark and grey bars represent original and corrected scores, respectively. Error bars represent SEM.

The effect of sound on memory performance was significant and confirmed by a one way repeated measures ANOVA, F (6, 144) = 5.378; p < .001. Post hoc analyses were conducted

given the statistically significant ANOVA result. All possible pairs were compared by Tukey HSD tests. Seven pairs of statistically significant acoustic conditions ($p < .05$) are presented in the Table 2.

DISCUSSION

The experiments were designed in order to evaluate the FDCC parameter, a spectral distinctiveness metric, which follows the guidelines of the changing state hypothesis by focusing on the spectral properties of the successive sound segments. The stimuli were built in a way that the behavior of the parameter could be observed in relation to the ISE by manipulating the estimator values systematically.

Two types of stimuli were generated: white noise pulse trains and noise-vocoded speech. The motivation behind the choice of the stimuli was based on two criteria: It should be possible to modify the temporal and spectral features in a controlled way and the stimuli would be expected to cause ISE.

Table 2: Pairwise comparison of test performances for the all possible pairs by Tukey HSD test. Only the pairs with statistical significance are reported.

Statistically significant pairs	**Mean error rate (%)**	***p* values**
SLNC – Speech	46.9 - 54.9	$p < .002$
SLNC – 6-Bands	46.9 – 57.0	$p < .001$
1-Band – 6-Bands	46.4 – 57.0	$p < .001$
1-Band – 18-Bands	46.4 – 52.6	$p = .042$
1-Band - Speech	46.4 – 54.9	$p < .001$
2-Bands – 6-Bands	48.4 – 57.0	$p < .001$
2-Bands - Speech	48.4 – 54.9	$p = .025$

The first experiment was structured with the hypothesis that the stimuli with a wide range of spectral parameter values should disrupt memory performance much more than the temporally modified, spectrally static stimuli. However, the data clearly discarded the hypothesis by presenting a lack of correlation between cognitive performance and the parameter values. In fact the memory performance, on average, showed no difference with the control condition. A possible explanation for the lack of distractive properties of the stimuli could lie in the periodic structure of the noise pulses. Even though the syllable rate (4 Hz) was taken as the noise pulse occurrence rate, the background stimuli of the first experiment were very different from speech stimuli. As a result, the lack of disruption under the background sounds in this experiment obstructed the clarity of the parameter evaluation. It would not be incorrect to say that the spectral parameter needs to be improved in any case since it failed to predict the lack of cognitive distortion in this particular experiment.

This shortcoming was addressed in the second experiment by generating noise-vocoded speech stimuli which can systematically be modified in the frequency-domain and have sufficient properties to create an ISE. NVS stimuli allowed us to increase spectral variation by increasing the number of frequency bands which also led to the transformation of a non-

disruptive, amplitude modulated band limited noise to highly disruptive and intelligible speech-like stimuli.

The results for the second experiment were in line with the literature in terms of silence and speech conditions except that 6-bands NVS had the largest distractive effect on memory performance instead of the original speech. However, similar results were reported in the literature where NVS was employed as distractive stimuli. Both in the current study and the study of Ellermeier *et al* [9], there were no significant differences between the most disruptive NVS condition and the speech condition.

The increase in the performance as a function of the number of frequency bands was in line with the ISE studies employing NVS as distractive stimuli, except the 18-bands NVS condition [31, 9]. The increase in the performance under 18-bands NVS condition is puzzling, since it was expected to yield a magnitude of disruption at least similar to those of 6-bands and speech.

For the second experiment, the FDCC parameter was partly successful: The performance drop between 1 to 6 bands NVS condition was reflected in the parameter values. However, the spectral parameter generated distinct FDCC values for 6-bands and speech stimuli, which was not the case according to the experimental results. In addition, the continuation of the parameter value drop beyond 6-bands turned out to be untrue, since the 18-bands NVS condition yielded a better performance than 6-bands NVS. Nevertheless, the systematic increase in the magnitude of disruption from 1 to 6 bands indicates that there is a strong effect of spectral components within ISE.

CONCLUSION

- Periodic noise pulses with temporal and spectral variation failed to create ISE. Although neither one of the estimators managed to predict the outcome, the lack of performance loss blurred the evaluation of the two estimators.
- There were no significant differences in the experimental results between temporally and spectrally modified stimuli. It was expected that spectral variation would yield larger disruption than the temporal variation, which turned out to be incorrect, at least for the current study.
- Memory performance under methodically manipulated speech stimuli did not result in a systematic decrease as the number of frequency bands increased, which was in contradiction to spectral parameter's prediction.
- The ISE cannot be estimated by the spectral parameter, in its current form, as the nearly monotonic decrease in the FDCC values contradict with the ceiling effect, observed under 6-bands NVS condition. The parameter should be adapted by investigating the reason behind this discrepancy.
- Although the parameter failed to make accurate predictions within the context of ISE, it anticipated the trend in cognitive performance until a certain extent which, supports the role of spectral variations in the phenomenon.

Acknowledgements

The work has received funding from the European Union Seventh Framework Programme under grant agreement No 605867, FP7-PEOPLE-2013-605867.

REFERENCES

[1] Salamé, P., & Baddeley, A. (1982). Disruption of short-term memory by unattended speech: Implications for the structure of working memory. *Journal of Verbal Learning and Verbal Behavior*, *21*(2), 150–164.

[2] Banbury, S. P., Macken, W. J., Tremblay, S., & Jones, D. M. (2001). Auditory distraction and short-term memory: phenomena and practical implications. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 43*(1), 12-29.

[3] Ellermeier, W., & Zimmer, K. (2014). The psychoacoustics of the irrelevant sound effect. *Acoustical Science and Technology*, *35*(1), 10–16.

[4] Jones, D. M., Alford, D., Bridges, A., Tremblay, S., & Macken, W. J. (1999). Organizational factors in selective attention: The interplay of acoustic distinctiveness and auditory streaming in the irrelevant sound effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 466–473.

[5] Jones, D. M. (1999). The cognitive psychology of auditory distraction: The 1997 BPS Broadbent Lecture. *British Journal of Psychology*, *90*(2), 167–187.

[6] Hughes, R. W., Vachon, F., & Jones, D. M. (2005). Auditory attentional capture during serial recall: violations at encoding of an algorithm-based neural model? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*(4), 736–749.

[7] Hughes, R. W., Vachon, F., & Jones, D. M. (2007). Disruption of short-term memory by changing and deviant sounds: support for a duplex-mechanism account of auditory distraction. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *33*(6), 1050–1061.

[8] Marsh, J. E., Hughes, R. W., & Jones, D. M. (2008). Auditory distraction in semantic memory: A process-based approach. *Journal of Memory and Language*, *58*(3), 682–700.

[9] Ellermeier, W., Kattner, F., Ueda, K., Doumoto, K., & Nakajima, Y. (2015). Memory disruption by irrelevant noise-vocoded speech: Effects of native language and the number of frequency bands. *The Journal of the Acoustical Society of America*, *138*(3), 1561–1569.

[10] Liebl, A., Assfalg, A., & Schlittmeier, S. J. (2016). The effects of speech intelligibility and temporal-spectral variability on performance and annoyance ratings. *Applied Acoustics*, *110*, 170–175.

[11] Jones, D. M., Beaman, P., & Macken, W. J. (1996). The object-oriented episodic record model. In (S. Gathercole, ed.) Models of Short-term Memory (pp. 209-238). London, UK: Erlbaum.

[12] Jones, D. M., Alford, D., Macken, W. J., Banbury, S. P., & Tremblay, S. (2000). Interference from degraded auditory stimuli: linear effects of changing-state in the irrelevant sequence. *The Journal of the Acoustical Society of America*, *108*(3 Pt 1), 1082–1088.

[13] Jones, D. M., & Macken, W. J. (1993). Irrelevant tones produce an irrelevant speech effect: Implications for phonological coding in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 369–381.

[14] Ellermeier, W., & Hellbruck, J. (1998). Is level irrelevant in "irrelevant speech"? Effects of loudness, signal-to-noise ratio, and binaural unmasking. *Journal of Experimental Psychology. Human Perception and Performance*, *24*(5), 1406–1614.

[15] Tremblay, S., & Jones, D. M. (1999). 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(4), 1005.

[16] Perham, N., & Vizard, J. (2011). Can preference for background music mediate the irrelevant sound effect? *Applied Cognitive Psychology*, *25*(4), 625–631.

[17] Chen, F., Wong, L. L. N., & Hu, Y. (2013). A Hilbert-fine-structure-derived physical metric for predicting the intelligibility of noise-distorted and noise-suppressed speech. *Speech Communication*, *55*(10), 1011–1020.

[18] Jones, D. M., Miles, C., & Page, J. (1991). Disruption of proof-reading by irrelevant speech: Effects of attention, arousal or memory? *Applied Cognitive Psychology, 4(2), 89-108.*

[19] Schlittmeier, S. J., Weissgerber, T., Kerber, S., Fastl, H., & Hellbrück, J. (2012). Algorithmic modeling of the irrelevant sound effect (ISE) by the hearing sensation fluctuation strength. *Attention, Perception & Psychophysics*, *74*(1), 194–203.

[20] Tremblay, S., Nicholls, a P., Alford, D., & Jones, D. M. (2000). The irrelevant sound effect: does speech play a special role? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *26*(6), 1750–1754.

[21] Houtgast, T., & Steeneken, H. (1973). The modulation transfer function in room acoustics as a predictor of speech intelligibility. *Acta Acustica united with Acustica*, *28*(1), 66-73.

[22] Hongisto, V. (2005). A model predicting the effect of speech of varying intelligibility on work performance. *Indoor Air*, *15*(6), 458–468.

[23] Fastl, H. (1982). Fluctuation strength and temporal masking patterns of amplitude-modulated broadband noise. *Hearing Research*, *8*(1), 59–69.

[24] Fastl, H., & Zwicker, E. (2006). *Psychoacoustics: facts and models* (Vol. 22). Springer Science & Business Media.

[25] Jones, D. M., Madden, C., & Miles, C. (1992). Privileged access by irrelevant speech to short-term memory: the role of changing state. *The Quarterly Journal of Experimental Psychology*, *44*(4), 645–669.

[26] Park, M., Kohlrausch, A., & van Leest, A. (2013). Irrelevant speech effect under stationary and adaptive masking conditions. *The Journal of the Acoustical Society of America*, *134*(3), 1970–81.

[27] Senan, T., Park, M. H., Kohlrausch, A., Jelfs, S., & Navarro, R. F. (2015). SPECTRAL AND TEMPORAL FEATURES AS THE ESTIMATORS OF THE IRRELEVANT SPEECH EFFECT.

[28] Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology. General*, *134*(2), 222–241.

[29] Greenwood, D. D. (1990). A cochlear frequency-position function for several species—29 years later. *The Journal of the Acoustical Society of America*, *87*(6), 2592–2605.

[30] Plomp, R., & Mimpen, A. M. (1979). Improving the reliability of testing the speech reception threshold for sentences. *Audiology*, *18*(1), 43–52.

[31] Dorsi, J. (2013). Recall disruption produced by noise-vocoded speech: A study of the irrelevant sound effect.