SPECTRAL AND TEMPORAL FEATURES AS THE ESTIMATORS OF THE IRRELEVANT SPEECH EFFECT

Toros Senan
Philips Research Laboratories, Eindhoven, The Netherlands.

Mun Hum Park
Philips Research Laboratories, Eindhoven, The Netherlands.

Armin Kohlrausch
Philips Research Laboratories, Eindhoven, The Netherlands.
Human-Technology Interaction, Eindhoven University of Technology, Eindhoven, The Netherlands.

Sam Jelfs
Philips Research Laboratories, Eindhoven, The Netherlands.

Roger Fonolla Navarro
Universitat de Barcelona, Universitat Politècnica de Catalunya, Barcelona, Spain.

Summary

The distractive effects on cognitive processes ascribed to the nature of sound have been examined in the paradigm of "irrelevant sound", where test participants try to perform cognitive tasks in the presence of background sound. By comparing the test scores for different acoustic stimulus conditions in relation to a reference condition from such experiments, the "irrelevant sound (speech) effect" (ISE) can be quantified. The ISE is often explained by the changing state hypothesis: the distinctive segmentation of sound tokens where tokens may be understood as sound segments that can be distinguished from each other in temporal and/or spectral characteristics. A sequence of sounds consisting of different tokens produces significantly more disruption than a steady-state sound. The present work attempts to investigate the relation between the features from both the temporal and spectral domain, and the ISE, by predicting its behaviour separately with two estimators: The average modulation transfer function (AMTF) and the frequency domain correlation coefficient (FDCC). The first parameter is a measure for temporal variations in a sound, the latter one measures spectral variability in the sounds. Background stimuli are synthesized from a pulse train in which modified and unmodified pulses alternate. In order to manipulate the temporal and spectral features in the stimuli, a numerical optimization method was used to generate two sets of background stimuli where one of the two descriptors was kept constant and the other was varied in a systematic way. Therefore, stimulus sets used in this study allow the separate estimation of the role of the two estimators on cognitive performance in tasks involving serial ordering of short-term memory content.

PACS no. 43.50.Qp, 43.55.Hy. 43.66.Lj, 43.66.Ba

1. Introduction

Most of us, at one time or another, have been distracted by external sounds while trying to concentrate on a mental activity. These conditions which cause distraction and a decrease in cognitive performance have been studied under the name of irrelevant sound studies, where test participants attempt to perform cognitive tasks in the presence of background noise. By comparison of the test scores between diverse acoustic conditions from such experiments, the "Irrelevant Sound (speech) Effect" (ISE) can be quantified. This effect has been well documented in the literature [1] most commonly concentrating on open plan offices [2]. When it comes to the task of recalling serially presented test items, a substantial amount of evidence from the literature supports the hypothesis that the ISE occurs from the conflict

between two concurrent serialization processes in our cognitive system, one for the visually presented items and the other for the acoustically presented irrelevant sound. The brain, despite being trained to ignore background sound, processes the visual task as well as the irrelevant sound. The conflict between these parallel processes causes the performance to degrade for the given task. This effect is known as ISE.

The ISE effect is often explained by the changing state hypothesis: The distinctive segmentation between sound tokens [3]. A sequence of sounds consisting of different stimuli produces much more disruption than a steady-state sound. Researchers have investigated the ISE for background noise [4], music [5], or speech [6]. For example, unintelligible foreign speech or reversed speech was found to affect the serial-recall performance to a similar extent as normal speech [7]. These results indicate that the meaning of the irrelevant sound does not affect the degree of disruption, in fact, any sound stimulus containing more than one token that differs in spectral features appears to influence the serial-recall performance. The level of sound within the normal range of hearing (below 80 dBA) does not have any influence on the ISE, as the serial-recall performance was not particularly more disrupted at higher sound levels [8]. Furthermore, a level variation between successive tokens was not found to cause additional disruption [9].

The present work attempts to build a relation between the features from both the temporal and spectral domain and the ISE, by predicting its behavior with two estimators: The Average Modulation Transfer Function (AMTF) and the Frequency Domain Correlation Coefficient (FDCC). A set of stimuli is created where each one assures an independent relation with one of the two proposed metrics. For the first of stimuli the AMTF is modified while the FDCC is kept constant, for the second set of stimuli the FDCC is varied while the AMTF is kept constant.

The details of the metrics used are described in Secs. 2.1 and 2.2. Temporal and spectral modifications are discussed in Secs. 3.1 and 3.2, the results and conclusions are presented in Secs. 4 and 5, respectively.

2. Estimators of the ISE

Previous studies proposed various features related to the ISE. The concept of the Modulation Transfer Function (MTF) has been applied to predict the intelligibility of speech in a variety of room conditions and used to evaluate temporal distinctiveness [10]. The Speech Transmission Index (STI) indicates the ratio of the modulation indices between a modified signal and the original, which is weighted in the frequency domain. Other estimators were investigated in the literature: the STI was modified into a sigmoidal function which was shown to model the error rates of the various tasks [11]. The STI was used as an estimator of the ISE and compared with the test results of the serial-recall task. The conclusion was that the STI itself cannot be used to predict the results. In the same study, the FDCC was suggested as a useful parameter to rate the behavior of ISE, but it was argued that neither the STI nor the FDCC may be used alone to describe the performance of the ISE. The fluctuation strength, a psychoacoustic effect perceived when listening to slowly modulated sounds was also used to predict the behavior of the ISE [12].

In the following sections MTF and FDCC are used to represent the spectral and temporal properties which influence the irrelevant speech effect.

2.1. Average Modulation Transfer Function (AMTF)

The MTF describes the reduction of the modulation index of the intensity envelope as a function of modulation frequency. If a signal is modified in the temporal domain and then compared to the reference, the changes in the modulation index can be quantified using the MTF.

To obtain the MTF, an octave band analysis was carried out in order to cover the range of frequencies between 125 Hz and 8 kHz. Jones *et al.* [3] examined that the human voice in its range of modulation frequencies has a strong peak at around 4 Hz. For speech intelligibility investigation, and the relation with the MTF, a range of modulation frequencies between of 0.5 Hz and 16 Hz was chosen. The intensity envelope of the input signal x is obtained for each octave band. The input signal is filtered with Butterworth octave band pass filters (BPF) with center frequencies ranging from 125 Hz to 8 kHz. The output is squared and then low-pass filtered (LPF) with a cutoff frequency of 30 Hz. The resulting intensity envelope is analyzed for each modulation frequency with a 1/3-octave BPF with center

frequencies ranging from 0.5 Hz to 16 Hz. The root-mean-square (RMS) of the filtered intensity envelope, y_{ij}, (where i indicates the i-th octave band, and j the j-th modulation frequency) is computed and normalized by the mean of y_{ij}. For the elements of the resulting *K-by-N* matrix, m_{ij}, K is the number of octave bands and N is the number of modulation frequencies. Given the modulation index for each octave band, and for each modulation frequency, m_{ij} is compared with the corresponding values for the reference signal, obtaining a new K x N matrix describing the changes between the modified signal and the reference.

$$M_{ij} = \frac{m_{ij,x}}{m_{ij,ref}}, \quad (1)$$

The estimator used in this study to describe the performance of each stimulus is obtained by averaging the MTF matrix in both dimensions (i,j) resulting in a single value, AMTF ($\overline{M}$).

2.2. Frequency Domain Correlation Coefficient (FDCC)

The frequency domain correlation coefficient (FDCC) was proposed as a spectral distinctiveness metric [13] and defined as a correlation measure between nearby segments or tokens of a sound in the frequency domain. It was suggested to be a meaningful spectral estimator for the behavior of the ISE. The following describes the procedure to obtain the FDCC.

The envelope of the signal is obtained by squaring and applying a second order low-pass filter at 10 Hz. In order to segment the signal in to tokens the median of the envelope is computed and used as an amplitude threshold. The segmented signal parts with an envelope lower than this threshold are discarded. For the potential tokens the time intervals are obtained, and the median duration is computed. Tokens which are shorter than the median duration are also eliminated. For each token, octave band pass filters are applied with center frequencies ranging from 125 Hz to 8 kHz. For each octave band of each token the power spectrum P is calculated. The FDCC is defined as follows:

$$F = \frac{\sum_{j=1}^{K=7} P_{i,j} P_{i+1,j}}{\sqrt{\sum_{j=1}^{K=7} P_{i,j}^2 P_{i+1,j}^2}}, \quad (2)$$

where $P_{i,j}$ indicates the 1/3-octave band power spectrum for the i-th token in the j-th frequency band and F the FDCC. The FDCC can underline changes in the frequency domain that the MTF cannot express, where a high correlation value indicates less distinctiveness in the frequency domain, therefore more similarity between nearby tokens.

3. Stimuli

The present work studies the relation between temporal and spectral features of the irrelevant sounds and the ISE, by guaranteeing the effects of FDCC and AMTF on the ISE, separately.

White noise was chosen in order to control the features of each octave band, because a flat spectrum in the frequency domain ensures an equal gain over all bands. White noise, $G(t)$, and a Hanning window $W(t)$, of size w, were used to define the pulse shape. A 1/3-octave band filter with center frequencies ranging from 125 Hz to 8 kHz was used to perform the decomposition of $WG(t)$ into all bands. A total of twenty-one pulses were obtained from each 1/3-octave band. Seven 1/3-octave bands whose center frequencies are the same as those of the 7 full octave bands were selected.

The pulse of size w was generated by summing all the seven selected bands, of each new selected pulse x_i, where i indicates the i-th octave band.

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

In order to cover the full range of modulation frequencies involved in speech intelligibility a 1 min basic signal is created where every half second two pulses (P1) of 50 ms alternate. From the basic signal the reference signal is defined as the 1 min signal where the pulse P1 and P2 alternate as shown in Figure 1. The second pulse, P2 is derived from P1 after applying temporal and spectral changes.

The pulses are separated by a distance of 250 ms which is kept constant for all the stimuli created in this study. The amplitude of the pulse P1 is set to 0.9 and P2 is set to 0.3.

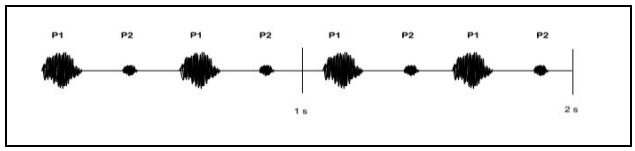


Figure 1: An illustration of two seconds of the reference signal. P1 and P2 alternate every half second.

3.1. Modifying the Modulation Transfer Function: Time domain changes

In order to modify the AMTF without altering the FDCC, the pulse width of P2 is modified from 50 ms to 450 ms and tested with the AMTF estimator. When the width of P2 is increased, the AMTF value decreases, while the FDCC remains constant. This feature is used in the current study in order to create a subset of the audio stimuli. The results are shown in Figure 2.

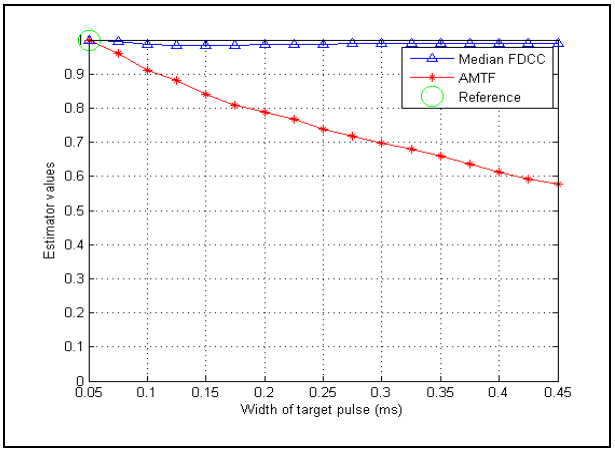


Figure 2: Relation between the width of the target pulse and value of the estimators (AMTF and FDCC).

3.2. Modifying the FDCC: Frequency domain changes

The FDCC is a correlation measure that describes the similarity of power spectra between neighboring tokens. In order to change this relation the power spectrum of each band is modified. The general idea is to apply a set of different gains to each octave band (125 Hz - 8 kHz) for P2. Modifying the gain of each octave band will also have an effect on the modulation index. Therefore, the AMTF in each octave band is investigated after different gains were applied.

A total of eleven gains were applied in each octave band of P2, ranging from θ = 0 to θ = 1 in steps of Δθ = 0.1. The width of P2 is kept at 50 ms.

When the gain of each octave band is increased the MTF decreases in modulation frequencies of 0.5 Hz, 1 Hz, 2 Hz and 16 Hz, while in the case of modulation frequencies of 4 Hz and 8 Hz, the MTF index remains constant. Knowing the behavior of the MTF for a given θ_i value, the AMTF is calculated from the expression (4).

$$M_{ij}(\theta_i) = \begin{cases} 1 & if\ j = 4{,}8\ Hz \\ \propto_j \theta_i^2 + \beta_j \theta_i + \gamma_j & if\ j = 0.5{,}1{,}2{,}16\ Hz \end{cases} \quad (4)$$

where α_j, β_j, γ_j are the coefficients of the quadratic function, j indicates j-th modulation frequency, θ_i the gain applied to the i-th octave band. FDCC and AMTF are represented in two equations (5), (6) as the functions of octave band gains. Given these equations, a combination of θ_i gains will be found to satisfy a variation of FDCC while keeping the AMTF constant.

$$\bar{M} = \frac{1}{NK} \sum_{i=1}^{K=7} \sum_{j=1}^{N=6} M_{ij}(\theta_i), \quad (5)$$

$$F = \frac{\sum_{i=1}^{K=7} \theta_i}{\sqrt{K} \sqrt{\sum_{i=1}^{K=7} \theta_i^2}} \quad (6)$$

Numerical optimization is applied to all seven θ_i gains given the desired values of AMTF and FDCC. The cost function, $Q(\Theta)$, is presented as a constrained minimization problem where Θ is a vector containing all θ_i values. The *fmincon* function in MATLAB was used to find the optimal Θ, with maximum of 100 iterations and a tolerance of 0.01. The gradient of the cost function is also required to implement the optimization. For every θ_i, the gradient was calculated, Θ was found given F values ranging from 0 to 1 in steps of 0.1 with $\bar{M}$ fixed to 1.

This optimization method was tested with the following settings: FDCC ranges from 0 to 1, AMTF was fixed to 1. Several trials revealed that Θ gains associated with FDCC ≤ 0.2 were the most suitable values to be used as initial values. Three trials were run with three different initial values Θ_0.

The results show that the intended purpose is achieved and the AMTF is kept constant while a wide range of FDCC values is obtained. The parameters of the audio stimuli generated by the techniques presented in Sec. 3 are shown in Fig. 3.

4. Experiment

4.1. Method and procedure

A single trial of the serial-recall test began with nine digits (1-9) presented to a participant on a computer screen. Numbers were flashed one by one every second, while each number was shown for 0.7 s followed by a 0.3 s pause. The order of presentation was randomized, and no two or more consecutive numbers were presented either in ascending or descending order.

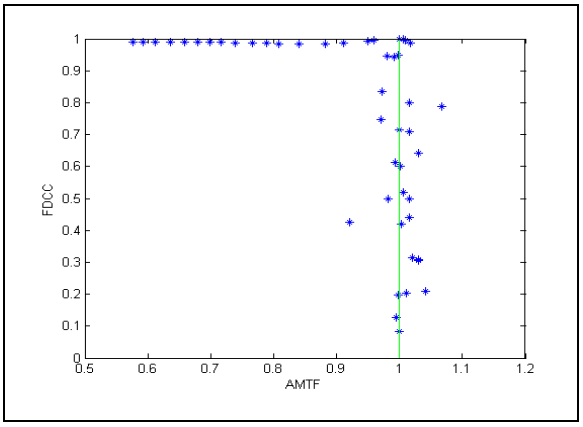


Figure 3: FDCC and AMTF of the final stimuli, where each point belongs to one audio file. The lines show the ideal positions of the parameter value combinations.

Following the presentation of nine digits, a 10s retention period was given, and then the participant was asked to recall the numbers in the correct order and indicate them via the graphical user interface. The layout of the keys was randomized in every trial so that the participant could not utilize the visual cue of the key positions to memorize the order. Also, when pressed by the participant, the key disappeared from the screen so that no number key could be selected more than once, and there was no facility available for the participant to correct the key input.

The serial-recall test design consisted of five blocks. The first block was the training block, enabling the participant to learn the test procedure by running four trials without any background sounds (silence). The instructor then checked if the participant had any questions or problems about the test procedure or the environment before continuing the experiment. The remaining four blocks consisted of 22 trials each. In each trial, a different background stimulus was presented 3 seconds before the first digit's appearance until the participant presses the last button at the end of the recall section. The order of the trials was randomized for every block except the dummy trial being the first one of each block.

Four blocks were presented with 5 min breaks between the blocks. During the pilot test, it was found that five test blocks can be completed in about 60-65 minutes, including the breaks.

4.2. Participants

A total of 10 participants from the Philips Research Laboratories in Eindhoven took part in the experiment (four female, six male). They were between 23-31 years old (median = 27.5 years) and all reported normal hearing and vision.

4.3. Material and apparatus

The experiment was run on a Hewlett-Packard computer using MATLAB (R2014a). All background sounds were presented diotically in MATLAB via a PC soundcard (RME Hammerfall DSP Multiface). The participants were placed 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 $dB_{LAeq1min}$.

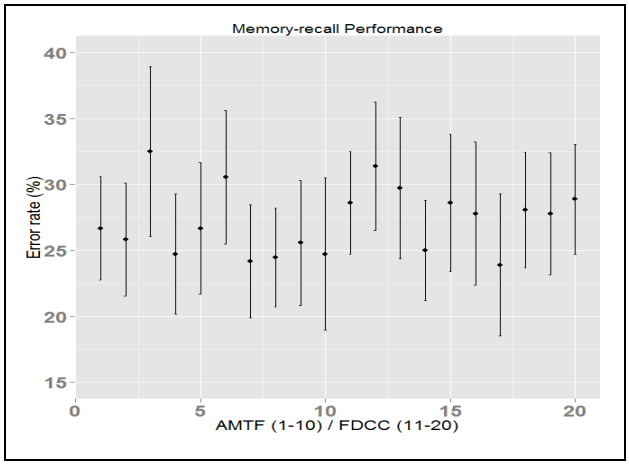
4.4. Results

Each digit not recalled in its previously presented serial position was scored as an error, and the score of the very first trial of each block was discarded, resulting in a total of 21 scores available for the data analysis in each test block. The overview of the test score with different parameters of the estimators is given in Figure 4 for a group of 10 subjects. The participants recalled the numbers presented earlier better than those later with the exception of the last one or two digits. Results indicate that the stimuli did not create a significant distraction effect. The lowest error rate (highest score) is expected to be achieved under silence condition, which in fact is not the result. However, the error rate under silence condition (28%) is in agreement with the literature [8]. The data also does not show any systematic change in the error rate as a function of the two parameters. Spectral modifications (FDCC) were expected to result in higher error rates than temporal changes (AMTF). The data, however, does not show any significant performance difference for the parameter range realized by the 21 background stimuli. That is, the definition of the estimators needs to be adapted, because the observed effect of the stimuli with wide variations in the FDCC parameter is not in line with the literature.

5. Conclusions

The experiment was designed with the hypothesis that the stimuli which have very different values in temporal and spectral estimators would lead to a

significant reduction in memory-recall performance.



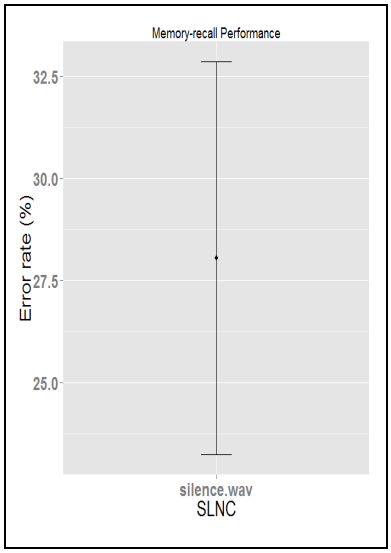


Figure 4 : (Up) Mean error rates based on 10 participants under different acoustic conditions. X-axis represents the stimulus number with 10 different values of AMTF (1-10), 10 different values of FDCC(11-20). (Left) Mean error rate under silence condition (28%).

The data clearly showed that there is no correlation between memory-recall performance and the variations of parameters of the stimuli. In fact, all modifications, on average, had the same score with the silence condition. Previous research [13] showed that spectral changes in speech stimuli created clear distraction effects. We can conclude that these observations indicate the inadequacy of the current definition of FDCC to predict the relations between background stimulus properties and memory-recall performance. A possible explanation for this lack of efficiency of our stimuli might lie in their regular structure, which overall makes the background stimuli very different from the speech stimuli. This shortcoming will be addressed in a new experiment by defining a stimulus which has properties to create an ISE while enabling to modify temporal and spectral features independently.

Acknowledgement

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

References

[1] K. Stokes, K. M. Arnell: New considerations for the cognitive locus of impairment in the irrelevant-sound effect. Memory & Cognition 40.6 (2012): 918-931.

[2] S. P. Banbury, D. C. Berry: Office noise and employee concentration: Identifying causes of disruption and potential improvements. Ergonomics 48.1 (2005): 25-37.

[3] D. Jones, C. Madden, C. Miles: Privileged access by irrelevant speech to short-term memory: The role of changing state. The Quarterly Journal of Experimental Psychology 44.4 (1992): 645-669.

[4] F. Chen, L. Wong, Y. Hu: A Hilbert-fine-structure-derived physical metric for predicting the intelligibility of noise-distorted and noise-suppressed speech. Speech Communication 55.10 (2013): 1011-1020.

[5] N . Perham, J. Vizard: Can preference for background music mediate the irrelevant sound effect? Applied Cognitive Psychology 25.4 (2011): 625-631.

[6] J. D. Larsen, A. Baddeley, J. Andrade: Phonological similarity and the irrelevant speech effect: Implications for models of short-term verbal memory. Memory 8.3 (2000): 145-157.

[7] D. M. Jones, C. Miles, J. Page: Disruption of proofreading by irrelevant speech: Effects of attention, arousal or memory? Applied Cognitive Psychology 4.2 (1990): 89-108.

[8] W. Ellermeier, J. Hellbrück: 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 (1998): 1406.

[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 25.4 (1999): 1005.

[10] T. Houtgast, H. Steeneken: A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. The Journal of the Acoustical Society of America 77.3 (1985): 1069-1077.

[11] V. Hongisto: A model predicting the effect of speech of varying intelligibility on work performance. Indoor Air 15.6 (2005): 458-468.

[12] S. Schlittmeier,et al: Algorithmic modeling of the irrelevant sound effect (ISE) by the hearing sensation fluctuation strength. Attention, Perception, & Psychophysics 74.1 (2012): 194-203.

[13] M. H. Park, A. Kohlrausch, & A. van Leest: Irrelevant speech effect under stationary and adaptive masking conditions. The Journal of the Acoustical Society of America 134(3) (2013): 1970-1981.