Cognitive disruption by noise-vocoded speech stimuli

Citation for published version (APA):

DOI:
10.1121/1.5026619

Document status and date:
Published: 01/03/2018

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow the link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
Cognitive disruption by noise-vocoded speech stimuli: Effects of spectral variation
Toros Ufuk Senan, Sam Jelfs, and Armin Kohlrausch

Citation: The Journal of the Acoustical Society of America 143, 1407 (2018); doi: 10.1121/1.5026619
View online: https://doi.org/10.1121/1.5026619
View Table of Contents: http://asa.scitation.org/toc/jas/143/3
Published by the Acoustical Society of America

Articles you may be interested in
Auditory distraction by speech: Sound masking with speech-shaped stationary noise outperforms -5 dB per octave shaped noise
The Journal of the Acoustical Society of America 143, EL212 (2018); 10.1121/1.5027765

Effects of age on sensitivity to interaural time differences in envelope and fine structure, individually and in combination
The Journal of the Acoustical Society of America 143, 1287 (2018); 10.1121/1.5025845

Speech-material and talker effects in speech band importance
The Journal of the Acoustical Society of America 143, 1417 (2018); 10.1121/1.5026787

Relationship between speech-evoked neural responses and perception of speech in noise in older adults
The Journal of the Acoustical Society of America 143, 1333 (2018); 10.1121/1.5024340

Speech recognition for school-age children and adults tested in multi-tone vs multi-noise-band maskers
The Journal of the Acoustical Society of America 143, 1458 (2018); 10.1121/1.5026795

Effects of reverberation and noise on speech intelligibility in normal-hearing and aided hearing-impaired listeners
The Journal of the Acoustical Society of America 143, 1523 (2018); 10.1121/1.5026788
Cognitive disruption by noise-vocoded speech stimuli: Effects of spectral variation

Toros Ufuk Senan,1,a) Sam Jelfs,1 and Armin Kohlrausch2

1Royal Philips, High Tech Campus 36, Eindhoven, 5656 AE, The Netherlands
2Human-Technology Interaction, Eindhoven University of Technology, Eindhoven, 5600 MB, The Netherlands

(Received 19 August 2017; revised 25 January 2018; accepted 17 February 2018; published online 12 March 2018)

The effect of irrelevant sounds on short-term memory was investigated in two experiments using noise-vocoded speech stimuli (NVSS). Speech samples were systematically modified by a noise-vocoder and a set of stimuli varying from amplitude-modulated white noise to intelligible speech was created. Eight NVSS conditions, composed of 1-, 2-, 4-, 6-, 9-, 12-, 15-, and 18-bands, were used as the distracting stimuli in a digit-recall task next to the speech and silence conditions. The results showed that performance decreased with the number of frequency bands up to the 6-bands condition, but there was no influence of number of bands on performance beyond six bands. The results were analyzed using four acoustic metrics proposed in the literature: the frequency domain correlation coefficient (FDCC), the fluctuation strength, the speech transmission index (STI), and the normalized covariance measure (NCM). None of the metrics successfully predicted the results. However, the parameter values of the FDCC, the STI, and the NCM indicated that a prediction model for irrelevant sound effect should account for both temporal and spectral features of the irrelevant sounds.

© 2018 Acoustical Society of America. https://doi.org/10.1121/1.5026619

I. INTRODUCTION

The detrimental effects on cognitive performance attributed to background sounds have been investigated under the paradigm of irrelevant sound. Typically, subjects perform working memory tasks in the presence of background speech, and the performances under different acoustic conditions are compared to quantify the irrelevant speech effect (ISE; Salamé and Baddeley, 1982). The finding is robust: The presence of background speech heavily impairs the recall performance even though the subjects are instructed to ignore it. It was soon discovered that the effect is not only apparent when the accompanying sound is speech but also occurs under background noise, music, alternating tones, reversed speech, etc. As a result, the recall impairment has been renamed the irrelevant sound effect while keeping the same acronym (ISE) (Banbury et al., 2001; Ellermeier and Zimmer, 2014).

A serial-recall task is a short-term memory task that requires remembering the order of presented items (e.g., a random sequence of the digits 1–9), and is the most widely used memory task in ISE research. Experimental evidence from the literature, within the context of serial-recall tasks, suggests that the ISE is a joint product of the intentional processing on the order of the items and the involuntary processing of the sound. This formulation has been conceptualized as interference by process, which occurs due to two parallel ordering mechanisms: one for the visually presented items and one for the acoustically presented background sound (Hughes et al., 2007). The conflict between these two modalities arises because the brain processes the irrelevant stimuli as well as the visually presented items. In order to create this conflict, the irrelevant stimuli should comprise perceptually distinct features in successive segments of the sound. This observation is manifested in the changing-state hypothesis: The successive tokens of the sound should have different characteristics in terms of acoustic features in order to create the disruption (Jones, 1999). For example, a sequence of identical tones does not degrade the performance while a sequence of tones, changing in frequency, disrupts performance considerably. Several studies modified the temporal (Tremblay and Jones, 1999; Ellermeier and Hellbrück, 1998) and spectral features (Jones et al., 2000; Jones and Macken, 1993) of the irrelevant stimuli in a systematic way in order to examine the relation between the magnitude of disruption and the acoustic properties of the sounds. However, clear-cut evidence, which relates the degree of disruption to a single acoustic descriptor, is yet to be found. Nevertheless, the variation in the spectrum is regarded to be a prominent factor (Jones et al., 1999; Ellermeier and Zimmer, 2014).

In Jones et al. (2000), a set of male-spoken words was systematically degraded by applying a low-pass filter with roll-off values changing between 0 and 24 dB, with a step size of six, and the recall performance increased monotonically as a function of the filter roll-off values. In the second experiment, the degradation of the speech stimuli was increased to an extent that the speech was transformed into amplitude-modulated noise. As a result, the interference by the irrelevant speech stimuli decreased when compared to the first experiment, supporting the concept of changing state.

The changing-state hypothesis provides a framework for the perceptual conditions required to observe the effect, but does not try to quantify the magnitude of the ISE. An attempt to quantify ISE, by following the definition of changing-state
hypothesis, has been introduced in the study of Park et al. (2013), who defined the frequency domain correlation coefficient (FDCC). In the study, the effect of adaptive masking on ISE was investigated and the spectral distinctiveness of the sound tokens was quantified by a spectral model. The results were promising, but this is the only study where speech stimuli were used in combination with a spectral estimator. The FDCC was later used in another study where spectral and temporal features of white noise pulse trains were modified systematically; however, no clear trend in the experimental results was observed (Senan et al., 2015).

Quantifying the magnitude of ISE is not a straightforward process since the degree of disruption under speech stimuli is larger than observed for any other irrelevant sound and the acoustic features responsible for such a distinction are not clear. This dilemma is taken into account in the literature by focusing on speech-specific acoustic properties of the irrelevant sounds while developing a prediction model.

The psychoacoustic hearing sensation of fluctuation strength (FS; Fastl and Zwicker, 2007), a measure that quantifies the perception of slow (<20 Hz) amplitude modulation was used as the basis of a prediction model (Schlittmeier et al., 2012). The numerical predictions for the degree of disruption stayed within the interquartile range of the experimental results for 63 out of 70 types of background stimuli. However, the model lacked the ability to identify whether two successive sound tokens are spectrally similar or not, which is critical in the ISE research (Ellermeier and Zimmer, 2014).

In Hongisto (2005), speech intelligibility was used as a temporal parameter to predict ISE in an open-plan office environment. The speech transmission index (STI) is a physical measurement method to estimate speech intelligibility. The STI can be derived from the reduction in the modulation index of the intensity envelope of a signal after the signal has traveled through a sound transmission system (Steeneken and Houtgast, 1980). The STI has been shown to be a promising ISE model for speech and speech-like stimuli, and has been shown to predict intelligibility accurately when the speech degradation is induced by additive noise and/or reverberation. However, when the speech is processed non-linearly, such as dynamic envelope compression introduced by hearing aids, STI fails to predict the intelligibility. This shortcoming has been addressed and several modifications have been proposed in the literature (Steeneken and Houtgast, 1980; Goldsworthy and Greenberg, 2004). Among those modifications the normalized covariance measure (NCM) correlated moderately well with the subjective intelligibility scores of noise-suppressed sentences (Ma et al., 2009) and vocoded sentences where both sine wave and white noise were employed as the carrier signals (Chen and Loizou, 2011b). NCM determines the transmission index values from the bandpass filtered intensity envelopes of the clear and degraded signals, just like the STI, but quantifies the intelligibility by computing the covariance between the intensity envelopes rather than using the modulation transfer function for determining the change in the modulation depth. Unlike the STI, the NCM has not been used as an ISE predictor in the literature. Although both measures are in need of a reference signal to be computed, which can be a limitation, the NCM gives better results in terms of intelligibility for vocoded speech (Chen and Loizou, 2011a) and therefore is evaluated in the current study as an ISE predictor.

The present study describes a stimulus synthesis procedure that allows for the creation of a continuum from the most disruptive acoustic condition (speech) to a non-disruptive one (amplitude-modulated broadband noise) by systematically modifying the spectral features of the stimuli. In order to accomplish this, speech is processed by a noise vocoder, which is a technique used in cochlear implant and speech perception studies (Shannon et al., 1995; Roberts et al., 2011). Noise-vocoded speech is a manipulation of natural speech that is generated by filtering speech into frequency bands, extracting the amplitude envelope of each band, and using it to modulate band-limited white noise in the corresponding frequency band. In the final step, all amplitude-modulated noise bands are combined to create the noise-vocoded speech. The result is a harsh, metallic, distorted speech sound. Despite lacking many qualities of natural speech, noise-vocoded speech stimuli (NVSS) can still be as intelligible as speech, depending on the number of employed frequency bands. Due to the way it is synthesized, noise-vocoded speech is intelligible primarily as a result of intensity variations (Shannon et al., 1995). Modifying the number of frequency bands in NVSS allows to manipulate the spectral variation of the stimuli and, important to this study, it creates an ideal case to investigate the effect of spectral variation on cognitive performance.

Noise-vocoded speech was first used in the irrelevant speech paradigm in the Master’s thesis of Dorsi (2013) at State University of New York (SUNY) New Paltz. In the study, a letter-recall task with seven letters is reported under one control (steady-state white noise) and four NVSS conditions: 3-, 6-, 9-, and 12-bands. The cut-off frequencies were obtained by dividing the spectrum into logarithmically spaced frequency bands. The results showed that the recall performance decreased when the number of frequency bands was increased from 3- to 9- and 12-bands.

In a more recent study by Ellermeier et al. (2015), a similar approach was followed while introducing slight changes to the number and the cut-off frequencies of the spectral bands. The digit-recall experiment consisted of one control condition and five acoustic conditions where the number of frequency bands was increased. The cut-off frequencies were determined using the Bark scale (Zwicker, 1961). Japanese and German speech was presented to both Japanese and German participants. Participants were randomly assigned to one of the language conditions, including their non-native language. In addition to the language component, this study also applied STI and FS prediction models and compared experimental results with parameter values. The conclusion was that STI fared much better, but that there needs to be a spectral model in order to accurately predict the ISE. Another study where STI and FS models were compared, this time in an open-plan office context, also showed that the STI model was generally more successful at predicting recall performance under the masked speech conditions while the FS model failed except for the (unmasked) speech condition (Liebl et al., 2016).
In the present study, we aim to investigate the effect of NVSS on serial-recall following a similar approach and evaluating three metrics proposed in the ISE literature and one metric that has not been used as an ISE predictor before in the light of experimental results. We intend to broaden the scope of previous work by employing a finer grain on the number of frequency bands and investigating the behaviour of the spectral parameter, the FDCC. Two experiments were conducted where different ranges of the number of frequency bands were used in order to cover the spectral parameter values varying between non-speech to speech stimuli. Section II explains the spectral descriptor and relates the parameter values to the auditory stimuli. Sections III and IV give the explanations of the experimental procedures for the two experiments. The results of the aforementioned prediction models are compared in Sec. V in the light of experimental results, and the paper is finalized by the discussion (Sec. VI) and Conclusion (Sec. VII).

II. SPECTRAL ESTIMATOR AND STIMULI

A. Spectral estimator

The spectral estimator used in this study is called the FDCC. FDCC was proposed as an ISE estimator in the study of Park et al. (2013) where adaptive and stationary masking noises were developed and applied to the speech stimuli in order to observe the effect of masking on the ISE.

Masking of the background speech by stationary noise is known to be effective in reducing the irrelevant-sound-related cognitive disruption (Ellermeier and Hellbrück, 1998; Haapakangas et al., 2014). This finding was attributed to the reduction of distinctiveness in the power spectrum of the successive sound tokens since the amount of spectral variation, therefore the magnitude of the changing state, decreased with the introduction of background noise (Park et al., 2013).

FDCC attempts to follow the definition of the changing-state hypothesis. The computation of the FDCC begins with dividing the sound into tokens, which is followed by magnifying the spectral difference between the successive ones. The intensity envelope of the signal is obtained by squaring and applying a second-order Butterworth low-pass filter at 10 Hz. In order to determine the borders of the tokens, first the median of the envelope is computed and the segments with envelope values above the median are accepted as feasible tokens. Second, the time intervals of the feasible tokens are computed and the median interval duration is obtained. Tokens that are shorter than the median interval are discarded. Each of the remaining sound tokens is filtered through 19 one-third octave band filters with center frequencies ranging from 125 Hz to 8 kHz, and the power spectrum \( P \) is calculated for each band of each token. The FDCC for two successive tokens is defined as follows:

\[
\text{FDCC}_i = \frac{\sum_{j=1}^{19} P_{i,j} P_{i+1,j}}{\sqrt{\sum_{j=1}^{19} P_{i,j}^2 P_{i+1,j}^2}} ,
\]

where \( P_{i,j} \) indicates the power spectrum for the token \( i \) and the frequency band \( j \). Finally, the FDCC values are averaged across the number of tokens extracted. The estimator can highlight changes in the frequency domain where a high correlation value indicates less distinctiveness, therefore, more spectral similarity between nearby tokens.

B. NVSS

As mentioned in Sec. I, noise-vocoded speech is a modification of natural speech where speech is filtered into frequency bands and the intensity variation of each frequency band is mapped to band-limited white noise. Finally, the resulting amplitude-modulated noises of various frequency bands are combined to generate the stimuli. For the current study, NVSS were generated by dividing the speech signal between 50 and 8000 Hz into 1, 2, 4, 6, 9, 12, 15, and 18 Hanning-shaped bandpass filtered frequency bands by modifying the scripts used in a speech comprehension study (Davis et al., 2005) in Praat software (Institute of Phonetic Sciences, University of Amsterdam, Amsterdam, The Netherlands). The bandpass filters have a rolloff of 6 CdB/\( w \), where \( w \), the width of the regions between pass and stop bands on both sides of the frequency band, was determined by dividing the upper cut-off frequency of each band by 10.

The cut-off frequencies were determined by an exponential function developed by Greenwood (1961). The Greenwood function relates the location of the inner ear hair cells with the frequencies at which they are activated; hence, it is considered to be the mathematical basis of the cochlear implant array placement Greenwood (1990).

Sentences were re-synthesized by replacing information in each frequency band with amplitude-modulated bandpass noise. This procedure allows the creation of a set of stimuli that changes from broadband amplitude-modulated white noise to highly intelligible noise-vocoded speech when increasing the number of frequency bands. The cut-off frequencies and the number of frequency bands of the stimuli are presented in Table I. In the study by Ellermeier et al. (2015), the experimental results showed that when four or more bands were used

<table>
<thead>
<tr>
<th>Number of bands</th>
<th>Experiment used</th>
<th>Cut-off frequencies (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>1 and 2</td>
<td>98, 157, 229, 317, 425, 558, 720, 918, 1160, 1457, 1820, 2265, 2809, 3474, 4289, 5286, 6506</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>110, 184, 280, 402, 560, 756, 1010, 1332, 1742, 2265, 2931, 3782, 4863, 6242</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>126, 229, 369, 558, 814,1160, 1630, 2265, 3125, 4289, 5865</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>157, 317, 558, 918, 1457, 2265, 3474, 5286</td>
</tr>
<tr>
<td>6</td>
<td>1 and 2</td>
<td>229, 558, 1160, 2265, 4289</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>370, 1160, 3125</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1160</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>( \ldots )</td>
</tr>
</tbody>
</table>
in the native language condition, the recall performance got very close to the level of original speech. The explanation of this may be linked with the almost-speech-like intelligibility of the NVSS conditions when more than four bands are employed (Davis et al., 2005). However, when the spectral parameter values are examined (see Fig. 1), the spectral similarity continues to decrease beyond the 4-bands condition. In fact, NVSS with more than four frequency bands cover almost half of the variation in the parameter values between broadband noise and speech. Therefore, in the current study, we distribute the whole range of parameter values over two experiments to investigate the impact of spectral variation under intelligible and unintelligible NVSS conditions. The NVSS conditions with six frequency bands and beyond are investigated in the first experiment, and the second experiment focuses on the lower number of frequency bands where the test performance is known to change drastically between conditions (Ellermeier et al., 2015). The choices of the NVSS for the two experiments are explained in detail in Secs. III and IV.

III. EXPERIMENT 1

For the first experiment, five different NVSS conditions were chosen in a way that the corresponding FDCC values cover the lower half of the parameter range between 1-band and the original speech stimuli.

A. Method

1. Participants

Fifteen native speakers of Dutch (eight females, seven males, age range between 18 and 50 yr) participated in the experiment, which was performed at the Philips Research Laboratories in Eindhoven, The Netherlands. Participants reported normal hearing and vision during the intake as part of the informed consent. All participants were volunteering Philips employees. Safety of participants, data privacy, ethical compliance, and framework of the experimental design were documented, controlled, and approved by the Internal Committee Biomedical Experiments (ICBE) of Philips Research.

2. Stimuli

The original speech sentences were taken from a speech reception study by Plomp and Mimpen (1979). There were 10 lists of 13 sentences in Dutch, spoken by a female speaker. Each trial was synthesized by concatenating 13 sentences (6–8 s each) from the same list and forming one 42–55 s long speech sample. The resulting ten long speech exemplars were used in the original speech condition and were transformed into NVSS for every condition (for the procedure, see Sec. II B). In addition to the silent training condition, there were 7 acoustic conditions: 6-bands NVSS, 9-bands NVSS, 12-bands NVSS, 15-bands NVSS, 18-bands NVSS, silence (SLNC), and original speech (Speech).

3. Apparatus

The experiment was run on a Hewlett-Packard computer (Hewlett-Packard Enterprise Company, Palo Alto, CA) using MATLAB (R2014b, Natick, MA). All background sounds were generated in MATLAB at a 44.1-kHz sampling rate to a resolution of 16-bits and played out diotically using a personal computer (PC) soundcard (RME Hammerfall DSP Multiface, RME Audio, Haimhausen, Germany). The participants were placed in a double-walled IAC soundproof booth (Industrial Acoustics Company GmbH, Niederkrüchten, Germany) at Philips Research Eindhoven and Beyer-Dynamic DT 990 headphones (Beyerdynamic GmbH & Co., Heilbronn, Germany) were used for playback. The average sound level of the stimuli was calibrated to 60 dBAeq1min. A computer screen was positioned outside of the soundproof booth and was visible through the double-glass window.

4. Procedure

A single trial began with three asterisks disappearing one by one, indicating that the presentation stage is going to begin in three seconds. In the presentation stage, nine digits (sampled from 1 to 9) were presented to the subjects on a computer screen, flashing one by one every second. Each number was shown for 0.7 s followed by a 0.3 s pause. The presentation was randomized in a way that two consecutive numbers were not shown in descending or ascending order. A 10 s retention period was inserted before the recall stage and then participants were asked to recall the correct order of the numbers by clicking on the corresponding box from the number pad that appeared on the screen. The layout of the number pad was randomized in every trial so that the visual cue of the key positions was eliminated. There was no possibility to skip a number, correct the previously pressed key input, or select a number key more than once. The auditory stimuli were played

FIG. 1. Parameter values of the FDCC for all acoustic conditions used in the two experiments. A value of 1 indicates no spectral difference between successive sound tokens of the stimulus. Error bars represent the standard error of the mean (SEM).
back continuously throughout the trial (e.g., during presentation of the digits, retention, and recall).

The digit-recall task consisted of eight blocks. The first block of each session was the training block and consisted of eight trials in the silence condition. The rest of the blocks corresponded to different acoustic conditions and were randomized in a controlled manner such that the control condition (SLNC) always appeared after the second and before the sixth block. Each block consisted of ten trials and took \( \sim 7.5 \) min to complete. There were 2 min breaks after each block, and one experimental session took \( \sim 60 \) min to complete. A detailed explanation of the procedure was given to each participant both in written form and orally before the session began. They were told that repeating the digits out loud and/or using their fingers to remember the earlier digits was not allowed, there was no time constraint to complete the recall stage so that they could use as much time as they needed, and the whole session was being monitored from an additional screen in duplicate mode by the researcher.

B. Results

The recall of the correct digit in the correct order was evaluated as a correct response, and the performance was measured as error rate (%) out of nine digits. Before the analysis of the error rates per condition, the data were checked for a possible learning effect.

In Fig. 2, the test scores are plotted based on their presentation position, regardless of the acoustic conditions presented. The exponential decay in the mean error rate is clearly visible, which indicates that the participants either developed a new strategy or got better in their own by experience. The difference in mean error rate percentage between the first (TRNG) and the last block amounted to 11.4%, and nearly half of the mean error rate reduction was completed at the end of the second block. The effect was confirmed to be highly significant by a repeated measures analysis of variance (ANOVA), \( F(7,98) = 3.205; p < 0.005 \). In order to account for this learning effect, performance scores were normalized: The mean error rate difference between the last block and each presentation position was calculated, and the resulting value was subtracted from the original mean error rate of the corresponding acoustic conditions.

Figure 3 presents the mean error rates as a function of the acoustic conditions calculated for both the original and normalized scores. The difference between the mean error rates for SLNC (31.5%) and Speech (40.2%) is slightly smaller than what is reported in the literature (\( \approx 12\% \)), but the original mean error rate for the silent condition is in line with the literature (Schlittmeier et al., 2012; Jones and Macken, 1993; Tremblay and Jones, 1999). However, it is evident that there is no systematic variation in the error rates as a function of the number of frequency bands; the scores in the vocoder conditions are spread between SLNC and Speech. As it appears, the recall performances varied in a rather random manner for the various NVSS stimuli. The statistical analyses were conducted for normalized data.

The effect of sound on recall performance was confirmed by a one-way repeated measures ANOVA, \( F(6,84) = 2.198; p < 0.05 \). Tukey honestly significant difference (HSD) tests were conducted on all possible pairs of sound conditions. Only one pair of groups was found to be significantly different \( (p < 0.05): \) SLNC \( [M = 31.4, \text{ standard deviation (SD)} = 24.2] \), and Speech \( (M = 40.2, \text{ SD} = 23.4); p = 0.015 \).

C. Discussion

The data did not show a significant increase in mean error rates with an increase in the number of frequency bands.

FIG. 2. (Color online) Mean error rate (%) as a function on the order of the presentation blocks. Darker error bars represent the SEM and lighter error bars represent the standard deviation (SD).

FIG. 3. Recall performance of 15 participants as a function of the acoustic conditions, represented as mean error rates (%). Gray and dark bars represent normalized and original scores, respectively. Error bars represent the SEM.
within the NVSS conditions. In fact, there was no significant difference between conditions except for the comparison of SLNC and Speech. On the other hand, the FDCC values for these conditions (see Sec. II, Fig. 1) showed a systematic decrease as the number of bands increased.

These results indicate that the spectral variation in the current stimulus set is not crucial for recall performance. The lack of any meaningful trend in experimental results may be due to the small sample size or the choice of the number of frequency bands, and will be discussed in Sec. V. Nevertheless, the continuous decrease of FDCC values as a function of the number of frequency bands clearly shows that the interpretation of the parameter needs to be adapted in order to be a valuable predictor of the ISE.

IV. EXPERIMENT 2

For the second experiment, the NVSS evolving from non-speech to speech-like stimuli were chosen where both spectral variation and speech intelligibility increase as a function of the number of frequency bands. In order to allow further comparison, a couple of NVSS conditions used in the first experiment were also preserved in the experimental design.

A. Method

1. Participants

Twenty-five participants (15 females and 10 males, age range between 18 and 50 yr) recruited via the JF Schouten subject database of the Eindhoven University of Technology, Eindhoven, The Netherlands took part. Twenty of 25 participants were university students, and all of the participants were native-Dutch speakers. As part of the recruitment procedure, subjects were chosen by specifying the necessary criteria: healthy vision and hearing, no history of memory related disorder, and speaking Dutch as native language. They provided their informed consent before starting the experimental session where all the criteria were cross-checked. They were paid a small compensation fee for their participation. The experimental procedure was approved by the ICBE of Philips Research and the Human Technology and Interaction Department, Eindhoven University of Technology.

2. Stimuli

The original speech stimuli and the noise-vocoding technique were identical with the first experiment. The only difference was the choice of the number of frequency bands in 3 NVSS conditions: the 9-, 12-, and 15-bands were replaced with 1-, 2-, and 4-bands, respectively. The 6-bands, 18-bands, Speech, and SLNC conditions were preserved.

3. Apparatus

The experiment was run on a Hewlett-Packard computer using MATLAB (R2014b). All acoustic conditions were generated at a sample rate of 44.1 kHz with 16-bits resolution and delivered diotically in MATLAB via a PC soundcard (M-Audio Transit, Cumberland, RI). 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 (Sennheiser electronic GmbH & Co., Wedemark, Germany) were used for playback. The average sound level of the stimuli was calibrated to 60 dB\text{L}\text{Aeq}1\text{min}. One Philips computer screen was positioned inside the sound booth and one outside to enable a real-time monitoring of the experiment for the responsible researcher.

4. Procedure

The experimental procedure was exactly the same as in the first experiment.

B. Results

A similar learning effect as that shown in Fig. 2 was observed when the error rates were computed as a function of the block order. The effect was highly significant and confirmed by a repeated measures ANOVA, $F(7,168) = 11.59; p < 0.001$. The mean error rate percentage difference between the first (TRNG) and the last block was $\approx 20\%$, and half of mean error rate decrease was completed at the end of the second block. The error rates were normalized as in the first experiment, and the statistical analysis reported in this section refers to the normalized test scores. Figure 4 shows the error rate percentages under different acoustic conditions. The mean error rate in the SLNC and Speech was $27.5\%$ and $35.5\%$, respectively. The lowest mean error rate was observed in the 1-band NVSS condition, $27\%$, and the highest was observed in 6-bands NVSS, $37.6\%$. The increase in the mean error rate as a function of the number of frequency bands follows the trend of the parameter values (see Fig. 1), except for the 18-bands NVSS condition: The decrease in the mean error rate in this condition is not reflected in the parameter values.

The effect of acoustic conditions on recall performance was highly significant and confirmed by a one-way repeated
measures ANOVA, $F(6,144) = 7.939; p < 0.001$. Post hoc analyses were conducted given the statistically significant ANOVA result. All possible pairs were compared by Tukey HSD tests. Seven pairs of acoustic conditions were found to be significantly different ($p < 0.05$). The statistical results are summarized in Table II.

C. Discussion

The normalized mean error rate for the SLNC (27.5%) and Speech (35.5%) differ by 8%, which is slightly lower than what is reported in the literature (≈12%). Nevertheless, the error rate in the original silent condition is as high as reported in the literature (Schlittmeier et al., 2012; Jones and Macken, 1993; Tremblay and Jones, 1999). The increase in the mean error rates as a function of the number of frequency bands between the 1-band and 4-bands conditions was similar to the finding in the study of Ellermeier et al. (2015). However, the (non-significant) decrease in mean error rate in the 18-bands vocoder condition was unexpected when compared to the 20-bands condition in the same study.

The results are similar to those from the first experiment in terms of SLNC and Speech with an exception that the speech stimulus did not lead to the highest error rate. Although the Speech and 6-bands conditions yielded similar scores, speech is known to be the most distractive sound in the ISE literature. Similar results were reported in Ellermeier et al. (2015) for native Japanese speaking subjects in the 4-bands Japanese NVSS condition: The mean error rate in the 4-bands NVSS condition was slightly higher than in the Speech, but there was no significant difference for the pair and the mean scores were very close.

Further analyses of the data can be performed by looking closer at the common NVSS conditions, 6- and 18-bands. Both original and normalized mean error rates for the two conditions were similar in both experiments. When the post hoc tests were applied to the original data of the second experiment, the 6- and 18-bands pairs differed significantly (original score; $p = 0.035$, normalized; $p = 0.314$). The change occurred against the predictions of the FDCC parameter. There were no other differences in terms of pairwise comparison between the original and normalized data.

While the systematic increase in the error rates from one to six bands supports the effect of spectral variation within the ISE, the FDCC parameter, in its current form, is not adequate to address the complexity of the phenomenon. In fact, the experimental results indicate that the cognitive response toward the spectral variation is not straightforward and a perceptual approach to the spectral features may be needed to predict the ISE.

V. PREDICTION MODELS FROM THE LITERATURE

The experimental results have been further analyzed by evaluating three objective metrics as prediction models for ISE: the FS (Fastl and Zwicker, 2007), STI (Steeneken and Houtgast, 1980), and NCM (Goldsworthy and Greenberg, 2004).

The two experiments have different sample sizes and sample population characteristics. However, in order to investigate the impact of the full range of the number of frequency bands and observe the behaviour of the aforementioned metrics in a single axis, the data obtained from the two experiments were combined by a second normalization step. First, normalized error rates were averaged over all trials in a sound condition for each participant and the performance under SLNC was subtracted from this value. The resulting value was accepted as a relative error rate, which showed the magnitude of disruption under the specific sound condition for each participant. Second, the obtained relative error rates were averaged for every sound condition. The relative mean error rates for all the sound conditions in the two experiments are presented in Fig. 5.

The first metric, proposed in the study of Schlittmeier et al. (2012), is derived from the hearing sensation FS (Fastl and Zwicker, 2007) and aims to predict ISE resulting from both speech and non-speech stimuli. The model focuses on the slow (<20 Hz) amplitude and frequency modulations of the signal, and the FS reaches the maximum around the modulation rate of 4 Hz, which is also known to be the syllable rate in speech. In the study where the metric was proposed as a prediction model for ISE, the FS parameter values and

<table>
<thead>
<tr>
<th>Statistically significant pairs</th>
<th>Mean error rate (%)</th>
<th>$p$ values</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLNC–Speech</td>
<td>27.5–35.5</td>
<td>$p &lt; 0.002$</td>
</tr>
<tr>
<td>SLNC–6-bands</td>
<td>27.5–37.6</td>
<td>$p &lt; 0.001$</td>
</tr>
<tr>
<td>1-band–6-bands</td>
<td>27–37.6</td>
<td>$p &lt; 0.001$</td>
</tr>
<tr>
<td>1-band–18-bands</td>
<td>27–33.2</td>
<td>$p &lt; 0.05$</td>
</tr>
<tr>
<td>1-band–Speech</td>
<td>27–35.5</td>
<td>$p &lt; 0.001$</td>
</tr>
<tr>
<td>2-bands–6-bands</td>
<td>29–37.6</td>
<td>$p &lt; 0.001$</td>
</tr>
<tr>
<td>2-bands–Speech</td>
<td>29–35.5</td>
<td>$p &lt; 0.05$</td>
</tr>
</tbody>
</table>

FIG. 5. Relative mean error rate with respect to silence for the irrelevant sound conditions employed in two experiments. Dark gray, light gray, and black bars represent the different sample sizes of 15, 25, and 40, respectively.
All 90 stimuli (10 long sentences × 9 acoustic conditions) used in the present study were analyzed with the FS model (using ArtemiS 12.01 Sound Quality software, HEAD acoustics, Herzogenrath, Germany), and the parameter values for the irrelevant sound conditions are presented in Fig. 6. First, it can be seen that the FS model cannot make accurate predictions about the test scores: The parameter values stay almost constant for all the acoustic conditions. Second, the absolute parameter values (FS) in the current study are quite different from those reported for the corresponding conditions in the study of Ellermeier et al. (2015). In their study, FS tended to decrease (from a maximum of 0.69 to 0.49) with increasing number of vocoded channels for both German and Japanese NVSS. This issue has been discussed with the authors of the study and it is thought to result from the different choice of parameters in the software for the two studies.

Hongisto (2005) used the curve of subjective speech intelligibility results derived from IEC 60268-16 (International Electrotechnical Commission, 2003) to build the sigmoidal function where the performance drop starts to increase at the STI value of 0.2 and reaches its maximum level when the parameter value comes to 0.6. The computation of the STI requires a reference signal, and for the analysis of NVSS, the original sentences were used as the reference and STI values were computed for every vocoding condition (8 noise vocoding conditions × 10 test sentences). The mean values of the parameter are presented in Fig. 6.

When the values of the STI parameter are investigated, it can be seen that the parameter value increases monotonically with the number of frequency bands of the NVSS. However, relative mean error rates for the NVSS conditions with more than six frequency bands do not follow the trend of the parameter values, while the relative error rates for 1- to 6-bands NVSS shows resemblance.

The last metric, the NCM, also requires a reference signal to be computed, and the original speech sentences were used as the reference signals for every irrelevant sound condition. The mean parameter values are presented in Fig. 6. The absolute parameter values increase from 0.32 to 0.59 almost linearly between 1- and 15-bands, and the value for the 18-bands condition reaches 0.61. The difference between Speech (NCM: 1) and 18-bands (NCM: 0.61), 0.39, is larger than the overall range of the parameter values for the NVSS conditions.

The increase in the parameter values of the NCM, as a function of the number of frequency bands, yields a similar trend as the STI except for the shape of the curve and the large difference between 18-bands and Speech. Nevertheless, the NCM, as well as the other objective metrics observed above, is not able to predict the magnitude of the disruption, at least for the acoustic conditions employed in the current study.

VI. GENERAL DISCUSSION

The experiments explored serial-recall performance, a standard short-term memory measure, under speech and systematically distorted speech-like stimuli. The FDCC varied systematically (FDCC: 1 - 0.6) when the number of frequency bands increased and by the addition of speech stimuli (FDCC: 0.48), a continuum from non-speech to speech stimuli was obtained. Such a stimulus set allowed the extent to which the detrimental effects of speech and speech-like stimuli on recall performance can be predicted by a spectral parameter to be examined.

The results of the experiment indicated that while the increase in spectral variation follows the increase in the number of frequency bands, the memory performance does not follow the trend over the full range (see Fig. 5). Varying the number of frequency bands between 6 and 18 (represented by dark gray bars in Fig. 5) did not produce any clear trend in recall performance, it seemed like the increase in spectral variation did not have a substantial effect. However, using smaller numbers of frequency bands (represented by light gray bars in Fig. 5) demonstrated that, to a certain extent, the spectral parameter can be a promising predictor of the recall performance.

The mean error rate increase between 1- and 6-bands conditions followed the trend of parameter values, but reached a ceiling in the 6-bands NVSS. The indication of a critical value in the number of frequency bands was not completely unexpected: In Ellermeier et al. (2015), the largest difference in error rates were observed between 1-band and 4-bands, and in the study of Dorsi (2013) the significant differences between error rates only occurred when the number of bands increased from 3- to 9- and 12-bands, not from 6- to 12-, or from 9- to 12-bands. The non-linear relation between the magnitude of disruption and spectral variation can be observed in Fig. 5, where the relative error rates reached their maximum in the 6-bands NVSS condition. The nature of such a limit is hard to explain, but it was observed that this threshold appeared when the NVSS became intelligible.

Loizou et al. (1999) stated that intelligibility of noise-vocoded speech reaches a ceiling at 9-bands, and it is possible that there are no gains in speech perception beyond this ceiling. Davis et al. (2005) showed that 10-bands NVSS are
readily intelligible. The intelligibility thresholds as a function of the number of bands, however, were reported differently in the two ISE studies. In Ellermeier et al. (2015), the 4-bands NVSS condition yielded a 75%–80% syllable identification. In the study of Dorsi (2013), although the intelligibility was proven to increase from 3- to 12-bands, the overall rate of speech comprehension was very low, below 0.5%, for the 3-bands and reached 5% for the 12-bands conditions. While these two studies differ in some aspects and a direct comparison would not be possible, the extreme low intelligibility found in the study of Dorsi (2013) is at odds with other work.

The subjective intelligibility scores reported in the literature should only be taken as indicative values for the current study since the usage of the same sentences for both NVSS and the original speech conditions might have created an increase of the perceived clarity of speech content in NVSS: When the original speech and/or NVSS with higher number of frequency bands (e.g., 18-bands) appear early in an experimental session, perceived intelligibility of NVSS with lower number of bands (e.g., 4-bands, 2-bands) dramatically increases. Such change in perception is referred as pop-out effect and has been demonstrated in the study of Davis et al. (2005): Subjects who listened to the clear speech between two identical vocoded versions reported a significantly higher percentage of words correctly than the subjects who did not hear the clear speech. The same study also reported that hearing NVSS repetitively improved subject’s comprehension performance (15%) even for repetitions of the same vocoded sentences (without pop-out possibility). The authors have discussed that information presented by the clear speech eased learning to understand vocoded sentences, which indicates an influence of top-down processing. On the other hand, the performance increase without pop-out was attributed to the impact of repetitions of low level acoustic features on learning to understand vocoded speech.

The aforementioned limitation, induced by the experimental design, rules out the possibility of predicting the actual perceived intelligibility of NVSS used for the current study and therefore drawing conclusions based on subjective intelligibility is not possible. However, it has been shown that intelligibility of speech should not particularly determine such a threshold within a serial-recall context: Reversed speech and foreign speech diminished the performance similar to native speech stimuli (Jones et al., 1990). This is only true if intelligibility is interpreted as the ability to understand the semantic information delivered by speech stimuli, it may also be interpreted as “preserving the acoustic cues needed to be intelligible” since the reversed and foreign speech create a similar degree of disruption on recall as the original speech: They preserve the temporal and the spectral features of the unaltered speech stimuli.

When data of the experiment are analyzed using this interpretation, a first thing to notice is that the test scores are similar for the Speech and the 6-bands conditions and the 6-bands NVSS are highly intelligible. It can be argued that when the stimulus reached a certain level of preserving the temporal and spectral features of the original speech, its detrimental impact reaches a maximum. One can think that if the 6-bands condition already preserves the acoustic cues needed to create maximum disruption, NVSS conditions with more than four bands should be equally distractive as the speech. From this perspective, 6-, 9-, 12-, 15-, 18-bands and speech stimuli can be counted as one acoustic condition, and the average relative error rate for these six conditions is 5.5%, which is similar with what is observed for the 4-bands condition, 5.25%. In such a case, the trend of the mean error rate increase as a function of the number of frequency bands would stabilize beyond 4-bands, but the spectral parameter, FDCC, would still be an inadequate metric since the predicted parameter values continue to decrease beyond 4-bands NVSS.

The measured parameter values of the STI and NCM are compared by the values reported in the literature. The first thing to notice is the inconsistency between the absolute parameter values of the STI and those reported in the literature. In the study of Ellermeier et al. (2015), the reported STI values of 1-, 2-, and 4-bands NVSS conditions are 0.75, 0.81, and 0.89, respectively. These parameter values are elevated by a constant amount with respect to our present measurements depicted in Fig. 6. Note that the 1-band condition in their study already reaches a parameter value above 0.6, which is the STI beyond which performance is not expected to further deteriorate (Hongisto, 2005). Second, the STI value of 0.54 attributed to 1-band NVSS condition in the current study, which shows that the STI overestimates the intelligibility levels of NVSS regardless of this inconsistency.

The subjective intelligibility scores (Dorman et al., 1997) and NCM values are reported in the study of Chen (2011) where 2-, 4-, and 6-bands NVSS correspond to 50%, 90%, and 100% of words correctly recognized with the approximate NCM values of 0.44, 0.52, and 0.55, respectively. The values presented in Fig. 6 are slightly lower than what is reported in the same study: 0.35, 0.42, and 0.45. The small differences may be due to the different parameter choices made in the generation of the vocoded sentences. Nevertheless, when compared to the STI, the NCM parameter predicts more realistic values in terms of intelligibility, as reported in the literature.

Regardless of the role of intelligibility on serial-recall and the intelligibility predictions of these two metrics, it should be noted that the STI and NCM are affected by the increase in the number of frequency bands of NVSS in a systematic way, just like the FDCC. This may be an indication that both temporal and spectral information may be needed to successfully predict the phenomenon.

VII. CONCLUSION

(1) Cognitive disruption observed in the presence of systematically modified speech stimuli yielded a decrease in the serial-recall performance. The outcome is in agreement with what is reported in the literature: Serial-recall performance is vulnerable to noise-vocoded speech and speech stimuli.

(2) The mean error rate increase as a function of the number of frequency bands is in line with the literature for the second experiment: The performance decreases up to 6-bands NVSS. The ceiling effect observed in the second
experiment and the data of the first experiment indicate that there may be no influence of spectral variation beyond 6-bands on cognitive performance.

(3) The ISE cannot be predicted by the current structure of the spectral parameter. The FDCC values show changing-state activity beyond 6-bands NVSS, which do not create an interference in seriation process. There may be some perceptual cues that the FDCC is currently unable to recognize. The FDCC needs to be adapted by investigating the reason behind this mismatch.

(4) The prediction models proposed within the ISE literature, STI, and FS also failed to make accurate estimates for the magnitude of the disruption for NVSS. It appears that the complexity of the paradigm exceeds the capabilities of a single acoustic metric, therefore, a more sophisticated model that accounts for this complexity should be developed.

ACKNOWLEDGMENTS

The work has received funding from the European Union Seventh Framework Programme under Grant No. 605867, FP7-PEOPLE-2013-605867. The authors would like to thank Drasko Masovic, who made the FS values available, Fei Chen, for providing the MATLAB script for the NCM, and to thank Drasko Masovic, who made the FS values available, for the insightful comments that undoubtedly improved the quality of the paper. Portions of the data were presented at the 12th Congress of the International Commission on Biological Effects of Noise (ICBEN 2017, Zurich, Switzerland).

1Software is available at www.praat.org (Last viewed January 25, 2018).


