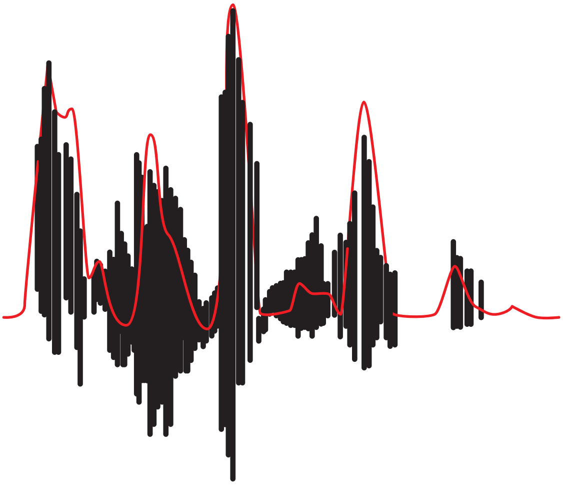
An evaluation of a psychoacoustic model of the changing-state hypothesis

An Evaluation of A Psychoacoustic Model of The Changing-state Hypothesis

Toros Ufuk Senan



An evaluation of a psychoacoustic model of the changing-state hypothesis

Toros Ufuk Senan

The work in this dissertation was financially supported by the European Commission within the ITN Marie Skłodowska-Curie Action project BATWOMAN under the 7th Framework Programme (EC grant agreement Nr. 605867) and Royal Philips.

A catalogue record is available from the Eindhoven University of Technology Library.

ISBN: 978-94-6380-185-0
NUR: 776

Keywords: Irrelevant sound effect, changing-state hypothesis, auditory distraction, irrelevant speech, instrumental measure, short-term memory, spectral variations

Cover design: Nihal Işik Senan
Printed by: ProefschriftMaken —— www.proefschriftmaken.nl.

An evaluation of a psychoacoustic model of the changing-state hypothesis

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit
Eindhoven, op gezag van de rector magnificus
prof.dr.ir. F.P.T. Baaijens, voor een commissie aangewezen door het
College voor Promoties, in het openbaar te verdedigen
op dinsdag 22 januari 2019 om 16.00 uur

door

Toros Ufuk Senan

geboren te Adana, Turkije

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof.dr.ir. A.W.M. Meijers
1e promotor:	prof.dr. A.G. Kohlrausch
Co-promotor:	dr. S. Jelfs (Philips Research Europe)
leden:	prof.dr. S. Schlittmeier (RWTH Aachen)
	prof. D. Jones, PhD (Cardiff University)
	prof.dr.ir. J.H. Eggen
	prof.dr. R. van Ee (Radboud Universiteit Nijmegen)
	dr.ir. M.C.J. Hornikx

Het onderzoek dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

Summary

Title: An evaluation of a psychoacoustic model of the changing-state hypothesis

Short-term memory disruption ascribed to the nature of sound has been studied under the paradigm of “irrelevant sound,” where test participants perform serial-recall tasks in the presence of background sounds. By analyzing the test scores for different acoustic stimuli in relation to a reference condition from such experiments, the “irrelevant sound (speech) effect” (ISE) can be quantified. The ISE is thought to result from the interference of two parallel ordering processes; one for the deliberate processing of visually presented items which the subject needs to memorize in the correct order, and one for the involuntary processing of the acoustically presented irrelevant sound. This interference is observed as a decrease in memory performance and is often explained by the changing state hypothesis: In order to induce an ISE, the background sound must be separable into tokens where successive tokens change in frequency or spectral content.

In this thesis, a spectral predictor, called frequency domain correlation coefficient (FDCC), is evaluated for its ability to predict the ISE. This predictor has been proposed by Park *et al.* (2013) and it is an algorithmic approach to capture the changing-state hypothesis. The metric divides the continuous background sound into temporal tokens and the change between successive states is quantified by computing the correlation coefficient between the power spectra of the successive tokens.

In Chapter 2, noise-vocoded speech (NVS) was employed as the irrelevant sound in two serial-recall experiments. Noise-vocoded speech is a manipulation of speech stimuli where the speech signal is filtered into frequency bands and the intensity envelope of each frequency band is mapped to band-limited white noise covering the corresponding spectral range. The result is a harsh, metallic distorted speech sound. NVS preserves parts of the spectro-temporal shape of the original speech while the number of frequency bands used in synthesis determines its intelligibility. More importantly for this study, by increasing the number of frequency bands, the spectral variation also increases which is reflected in a systematic decrease in the FDCC values. The two experiments consisted of eight NVS conditions, in which the number of frequency bands

varied between 1 to 18, alongside the original speech and the silence condition. The experimental results were compared with the prediction values computed by the FDCC as defined by Park *et al.* (2013) and three other metrics used in the literature to predict ISE: the fluctuation strength (FS), the speech transmission index (STI) and the normalized-covariance metric (NCM). None of the metrics did successfully predict the short-term memory disruption, but by comparing the values of the FDCC, the STI and the NCM we concluded that both temporal and spectral features of the sound should be taken into account in order to adequately predict the ISE.

In Chapter 3, a new variant of the FDCC was developed by integrating a peak detection stage before the correlation, in order to ensure that the token selection process in the metric captures the prominent parts of the sound (e.g., syllables in a word). The new version was evaluated by collecting a large set of stimuli from the literature and comparing the prediction values of the FDCC, both for the old and new versions, and the FS. Results showed that the new FDCC predicts the ISE under distorted / masked speech or speech-like sounds better than the old one and the FS, while for non-speech sounds, the performances of the new FDCC and the FS are similar to each other and better than that of the original FDCC predictor.

In Chapter 4, the newly developed token selection stage of the algorithm was evaluated in two experiments. The first was a serial-recall experiment: A set of irrelevant sound stimuli from the literature was segmented into tokens using the token selection stage of the algorithm by preserving the the original timing of the tokens. In order to avoid on- and offset artifacts at the token boundaries an adaptive low-level noise was added to fill the temporal gaps between the tokens. The expectation was that, if the token selection stage did correctly select all segments which comprise the distractive properties of the sound, the experimental scores for the original and the segmented stimuli should be the same. The expectation was supported in the experiments for continuous speech, for which the segmented version yielded very similar results. In contrast, for other background stimuli, like highly intelligible NVS with six bands, significant differences were observed in the two conditions. The second part of the third study investigated the sensitivity of the segmentation stage by focusing on a set of NVS stimuli from the literature where the original work had tried to evaluate the impact of speech fidelity on the ISE. This had been evaluated by temporally reversing the information

within two-thirds of the frequency bands of the NVS and comparing the serial-recall results with the original NVS. The FDCC values were in line with the experimental results where the reversed NVS was observed to be less disruptive than the original NVS. The outcome indicated that the token selection stage of the algorithm was sensitive enough to capture the temporal differences between the two stimuli, which was reflected in the serial-recall results.

The last set of experiments attempted to independently investigate the impact of temporal and spectral features of irrelevant sounds on serial-recall performance. Two serial-recall experiments were designed by generating white noise pulse sequences where every second pulse was modified in spectral shape and duration. In the first experiment the noise pulses were presented strictly periodically, and no ISE effect was observed for such stimuli. In the second experiment, the information regarding the temporal position, amplitude and duration of each pulse was extracted from continuous speech samples while the pulse modification method was the same as in the first experiment. The result agreed with those from the first experiment: The speech-positioned pulses did not yield any distraction in the serial-recall task. The findings of these experiments suggest that it is not possible to create an ISE by modifying the spectral and temporal features of noise pulses. This observation is an indication that spectral changes in successive parts of a background sound are not a sufficient condition to create an ISE and thus puts some limitations on the changing-state hypothesis and, in consequence, on the predictive power of the FDCC parameter.

The thesis reports studies which attempt to investigate the impact of the spectral variation on serial-recall performance and evaluate the changing-state hypothesis by employing an acoustic metric designed to quantify spectral distinctiveness. The results show that the spectral variation plays an important role and, for speech and masked-speech sounds, the FDCC appears to be a valid predictor of the ISE: The FDCC can be useful for room acoustic applications where noise maskers are used to reduce background-sound-induced distraction. However, the studies also reveal limitations of the metric and, particularly the results of the last study, indicate that the validity of the changing-state hypothesis may be stimulus dependent.

Table of contents

List of acronyms and abbreviations

AFC	Alternative forced choice
AMTF	Average modulation transfer function
BPF	Band-pass filter
CSII	Coherence-based speech intelligibility index
CVC	Consonant-vowel-consonant
FDCC	Frequency domain correlation coefficient
FS	Fluctuation strength
IEC	International Electrotechnical Commision
ISI	Inter stimulus interval
ISE	Irrelevant sound effect
JND	Just-noticeable difference
LPF	Low-pass filter
M	Music
MSC	Magnitude squared coherence
MTF	Modulation transfer function
NCM	Normalized covariance measure
NT	Noise-pulse train
NVS	Noise-vocoded speech
ON	Office noise

PESQ	Perceptual evaluation of speech quality
RMS	Root mean square
SLNC	Silence
SNR	Signal-to-noise ratio
SM	Spectrally modified speech-positioned noise-pulse train
SNT	Speech-positioned noise-pulse train
SPCH	Speech
SS	Segmented speech
STI	Speech transmission index
TM	Temporally modified speech-positioned noise-pulse train

1 | General introduction

1.1 Adverse effects of background sounds on cognitive performance

In contrast to the visual world which might be avoided by deflecting or shutting one's eyes or blocking the view with a physical barrier, the ear comes furnished with no mechanical means of eliminating unexpected or unwanted sounds: The auditory world remains inevitable. The ears always being open not only gives certain advantages when the sounds are relevant, but also the auditory events that are out of one's sight do not go unregistered. This unintentional awareness of one's acoustic environment is very useful for detecting the events which stay out of sight, but are of importance like a car passing by or a fire alarm. However, a by-product of the continuous processing of the surrounding auditory events is that currently irrelevant information intrudes on the organism and may induce mandatory processing, potentially detrimental to any focal task. One demonstration of this concept is staying concentrated on the set of information gathered by reading this text and ignoring the sound of someone chatting nearby. The dilemma is that this necessity to stay focused on the focal task (e.g., comprehending the text) is connected with the need to continue processing the irrelevant auditory information in parallel, such that we can turn our attention to it in the case that our current goal changes or there are significant differences in the environment which require to be acted upon immediately. This open processing system increases the mental workload and therefore creates an unwanted distraction on mental processing (Baldwin, 2016, Ch. 7).

The example mentioned in the first paragraph presents reading comprehension as the focal task and speech stimuli as the distractor. A semantic task, such as reading comprehension, may consist of a semantic component (prior knowledge of the content) and an episodic memory component (familiarity or experience with the examples stated in the text) (Marsh and Jones, 2010). Potential disruption caused by the unattended speech in this complex task may have arisen from either one or both of these two components, depending on the characteristics of the speech (e.g., language, intelligibility, semantic similarity, acoustic properties, number of speakers). Considering the diversity of acoustic environments and types of obstacles we have to overcome in our daily life, the underlying principles of the auditory distraction and the acoustic properties of the background sounds can turn into a multifaceted question and therefore attracts interest from different research fields.

The room acoustics community has a long-standing interest in the effects of noise in spaces such as open-plan offices, classrooms and hospitals considering the work efficiency, health and well-being of the occupants (Biley, 1994; Haka *et al.*, 2009; Jahncke *et al.*, 2013; Reinten *et al.*, 2017). Soundscape designers and urban planners have been working on developing new and sustainable solutions to eliminate negative effects of environmental noise on well-being (Yang, 2005; Galbrun, 2012). Linguists and educational scientists are curious about the impairment of reading and arithmetic capabilities of young children (Ljung *et al.*, 2009) as well as writing skills of adults (van de Poll *et al.*, 2014). Additionally there are numerous high-stress, high-workload environments in which the inadequacy of auditory warnings may have dramatic consequences (Baldwin, 2016, Chp. 2), like medical care (Baldwin, 2016, p. 28 - 29) and aviation (Hermann and Hunt, 2011). The recent development of electrical vehicles also emphasizes the question of safety in traffic settings due to its generation of lower sound pressure levels compared to gasoline or diesel engine cars (Parizet *et al.*, 2014).

Even though the practical settings for these examples are different, the research methods used are similar. Researchers attempt to create an interaction between a focal task and an auditory distractor in a laboratory which is similar to the ones expected in real life in terms of the underlying cognitive principles. For example, in order to asses the impact of background sounds on work efficiency in an open-plan office environment, long experimental sessions (3-4 h) with multiple tasks are constructed (Haka *et al.*, 2009; Jahncke *et al.*, 2013; Haapakangas *et al.*, 2014). The experiment typically takes place in a laboratory shaped as an open-plan office and is accompanied by different sounds which usually resemble the ones in a real life situation (e.g., office noise). The memory tasks are chosen in a way that the task demands are thought to be similar to those of the daily tasks of the office workers and the impact of the background sounds on work efficiency is quantified by comparing the task scores under different acoustic conditions.

As opposed to the applied research methods, a more common approach in fundamental research is to investigate the interaction between a specific type of background sound and a single focal task. One of the most used tasks in the literature is a short-term memory task, the serial-recall task: Typically, participants are asked to recall the order of a list of seven to nine verbal items (e.g., letters, digits) presented one at a time on a screen. This procedure is accompanied by irrelevant background sounds alongside a control condition (e.g., silence), usually delivered via headphones, either during only the presentation and the retention periods, or the whole trial. The serial-recall disruption is quantified by comparing the scores between different acoustic conditions (e.g., speech and silence) and is typically presented as an error rate (%). A large body of evidence shows that the serial-recall task is vulnerable to disruption induced by background sounds (Colle and Welsh, 1976; Salamé and Baddeley, 1982; Jones and Macken, 1993; Banbury and Berry, 1998; Ellermeier and Hellbrück, 1998; Park *et al.*, 2013).

This thesis is particularly concerned with the impact of the background sounds on serial-recall performance, namely, the *irrelevant sound effect* and the hypothesis which attempts to explain the phenomenon, *changing-state hypothesis*. The relation between the acoustical variations within the irrelevant sound and the corresponding serial-recall disruption is the major interest in this thesis, since the studies reported here evaluate a prediction model for serial-recall disruption which was inspired by the changing-state hypothesis. However, it is a complex phenomenon and it has been studied for more than three decades, therefore before moving into the prediction models proposed in the literature and the properties of the changing-state hypothesis, a summary of the theoretical background with a focus on the irrelevant sound effect is presented. This is followed by the description of the acoustic / psychoacoustic metrics proposed as prediction models and a brief literature review regarding the psychoacoustic properties of the changing-state effect. This chapter is concluded by the sections that present the goals and the contributions of this thesis and is finalized by the outline.

1.2 Irrelevant Sound Effect (ISE)

The detrimental effect of background sounds on serial-recall performance has been extensively studied in the literature (for a review, see Banburry *et al.*, 2001) and was first described as a phenomenon called the "acoustic masking in primary memory" in Colle and Welsh (1976), where the recall of visually presented letters was impaired by accompanying continuous speech (spoken text in a foreign language). It had been initially expressed as the *irrelevant speech effect* (ISE) when referring to the serial-recall results observed in speech conditions in Salamé and Baddeley (1982). However, several key findings in the literature revealed that speech is not the only type of sound that disrupts serial-recall performance: Pure tones changing in frequency (Jones and Macken, 1993); instrumental music (Schlittmeier *et al.*, 2008); office noise (Schlittmeier *et al.*, 2012); and bandpass-filtered noise bursts with different center frequencies (Tremblay *et al.*, 2001) were also shown to produce serial-recall disruption. Since it became obvious that the effect was not only observed in speech conditions, it was renamed to the *irrelevant sound effect*. Nevertheless, the speech stimulus is still pointed out as the most disruptive stimulus (Ellermeier and Hellbrück, 1998; Park *et al.*, 2013; Ellermeier *et al.*, 2015; Senan *et al.*, 2018) with an observed mean error rate of 38 - 50 %, while the control conditions typically yield a mean error rate of 26 - 32 %.

Cognitive models

There exist a number of cognitive models proposed by researchers which aim to describe the ISE based on the underlying cognitive processes (Baddeley, 2000; Jones *et al.*, 2000; Neath, 2000). Two short-term memory models follow the idea that the disruption is a consequence of the auditory stimuli and the representations of the verbal items having access to the same space, either being the phonological store (phonological loop model, e.g., Salamé and Baddeley, 1982; Baddeley, 2000) or the primary memory (the feature model, e.g., Nairne, 1990; Neath, 2000). The two accounts, the phonological store and the feature model, although different in terms of the descriptions they provided regarding how the interference occurs, both govern the idea that the recall impairment is due to the similarity in the content between the irrelevant sounds and the items to be recalled, rather than the processes involved.

As mentioned above, several key findings in the literature revealed that speech is not the only type of sound that disrupts serial-recall performance which undermines the assumptions made by the models regarding the role of the content in background sound induced disruption. With respect to the feature model, one of the limitations discussed in the paper by Jones and Tremblay (2000) was that it did not provide an explanation for the non-speech sounds, but instead states that it is a different effect and the model is not extensible to the irrelevant non-speech sounds. In addition to that, the impact of similarity of content on serial-recall performance was investigated using a set of irrelevant speech stimuli (e.g., *ton*, *gnu*, *tee*, etc.) which rhymed or did not rhyme (e.g., *wick*, *tip*, *dub*, etc.) with the visually presented items (digits) and it was shown that the serial-recall results observed in the two conditions were not significantly different (LeCompte and Shaibe, 1997).

As opposed to the phonological loop and the feature models, the *interference-by-process* model proposes that background sounds produce disruption by interfering with the processes involved in the focal task (Jones and Macken, 1993) rather than the similarity of the content. Within the context of the serial-recall disruption, the interference-by-process model proposes that the change within a sound sequence gives rise to an obligatory ordering process, which interferes with the process of seriating the visually presented to-be-remembered items (Jones and Macken, 1993). This has been known as the *changing-state hypothesis*: The irrelevant background sounds should consist of perceptually distinct variations from one distinguishable entity to the next one and the distinctiveness between subsequent items in the sound stream is the primary element of disruption.

The changing-state effect within the concept of serial-recall disruption was supported by several findings in the literature (e.g., Colle and Welsh, 1976; Jones and Macken, 1993; Jones *et al.*, 1999; Schlittmeier *et al.*, 2012; Senan *et al.*, 2018). For example, a steady-state sound stream like “A, A, A, A...” produces a degree of disruption similar to silence while a sound sequence like “A, B, A, B...” or “A, C, E, Z...” induces a significantly higher serial-recall disruption than the steady-state sound stream (Hughes *et al.*, 2005). A sequence of tones changing in frequency disrupts serial-recall performance significantly as opposed to a sequence of repeated tones (Jones and Macken, 1993). In addition to this, music containing many legato passages is shown to produce less disruption than music with many changes in tempo and pitch (Schlittmeier *et al.*, 2008). Further research also showed that the meaning of speech does not play a role in serial-recall performance (Jones *et al.*, 1992; Jones and Macken, 1995a; Buchner, 1996; Marsh *et al.*, 2008), hence supporting the prominent role of the acoustic variation.

Since the interference-by-process model advocates that the similarity of the process, not the content, is the reason behind the disruption, it was proposed to be an efficient explanation for the cases where the focal task is other than serial-recall. One of the findings which support this assumption was demonstrated by a free-recall task where the participants were asked to recall the words they remember from a list of visually presented words. When the background speech consisted of words semantically related with the list to be recalled, the recall performance was lower than what was reported for semantically unrelated background speech. However, when the participants were asked to perform the same task but this time in the order in which the words were presented (serial-recall) then the usual ISE was observed: The serial-recall performance in semantically related or unrelated speech conditions was not significantly different. It has been discussed extensively whether the interference is driven by the retrieval process in the semantic memory (Marsh *et al.*, 2008) or by the similarity of the content as proposed by the phonological loop and the feature models, *interference-by-content*, (Baddeley, 2003). Marsh *et al.* (2008) stated that, if it were the content that was relevant for the disruption, then the semantically related speech would be disruptive in both the free and serial-recall settings (for a review, see Marsh *et al.*, 2009).

An alternative view, the attentional capture, proposes that the impairment in memory induced by irrelevant sounds can be explained by the reorientation of attention triggered by a deviant sound within the auditory stream and is observed regardless of the processing involved in the task (e.g., Hughes *et al.*, 2005; Lange, 2005; Sörqvist, 2010; Vachon *et al.*, 2012). The attention of the listener can be diverted from the focal task if the background sound consists of a prominent, unexpected change in the auditory sequence. For example, if the sound sequence consists of a tone with a frequency ("B") following a succession of tones with a different frequency ("AAAAABA"), it will tend to capture attention as it interferes with the expectation for another "A" (Hughes *et al.*, 2005).

It was investigated whether the serial-recall disruption results from a sequence of deviant sounds, thus whether the attentional capture mechanism can account for any background-sound-induced disruption or the attentional capture and the interference-by-process paradigms are two distinct forms of auditory distraction (Hughes *et al.*, 2005; Hughes *et* *al.*, 2007). Duplex-mechanism account of auditory distraction proposed by Hughes *et al.* (2007, 2013) represents a current framework model for the effect of background sounds on cognitive performance in which also the changing-state hypotheses was integrated as one concretisation of the "interference-by-process" principle. Several findings in the literature favor the duplex-mechanism account (Hughes *et al.*, 2007) and one of the most clear-cut evidences comes from the study of Hughes *et al.* (2007). A serial-recall task was designed in a way that the irrelevant sound sequences contained an additional deviant sound. A letter spoken by a male voice was inserted in an irrelevant sound sequence of female spoken letters. The results showed that the addition of the deviant sound increased the error rate in both the changing-state (10 different letters) and the steady-state (repeated letter) conditions. It was reasoned that, if the attentional capture paradigm was able to account for the serial-recall disruption, then the mean error rates observed in the changing-state and steady-state sequences with added deviant sound would be similar.

The second experiment in the same study revealed another strong evidence using a short-term memory task which does not require a seriation process. The missing item task is another task for measuring verbal short-term memory capacity (Klapp *et al.*, 1983) where one of the visually presented items is missing (e.g., eight of the nine digits presented in the random order from a set of 1-9) and participants are asked to detect the missing item instead of recalling the order of the list. The aforementioned two acoustic conditions were employed in a missing item task and it was shown that adding a deviant sound equally increased the error rate for both sound conditions. It was concluded that the effects of changing-state sequences and deviants are independent and additive.

The interference-by-process account gathers its strength from providing an explanation for the auditory distraction which is detached from the identity of the background sounds, but rather related with their process demanding properties, and proposes that the impairment is a result of similarity of the concurrent processes. Particularly important for this study is that, within the context of ISE, the primary ordering process is vulnerable to sounds which posses a variation, a change in state, which intrudes a secondary seriation process.

1.3 Prediction models

The cognitive models summarized so far create a frame of research with respect to qualitative principles (investigating the decrease in the cognitive performance based on the irrelevant sound - focal task interaction and deriving conclusions regarding the interference process) and do not attempt to quantify the disruptive effects of the irrelevant sounds. Most important to this study is that they are not acoustic or psychoacoustic models that can predict the degree of disruption based on acoustic features of the irrelevant sounds. There exist, however, several attempts to explain the phenomenon based on the acoustic features of the irrelevant sounds and the findings derived from the cognitive studies guide these proposals. The following sections explain briefly each proposed model including the motivation behind and the discussions which emerged afterwards.

Speech Transmission Index (STI)

The earliest proposed ISE prediction model (Hongisto, 2005) is based on the speech transmission index (STI), which is a measure used to quantify the quality of a sound transmission path based on the intelligibility of speech (Houtgast and Steeneken, 1985). Listening tests can be conducted to measure speech intelligibility: Typically a narrator is speaking and participants are transcribing what they hear during the test. The speech intelligibility is calculated based on the average percent of correct answers and is usually different for words, sentences or syllables. The direct measurement of speech intelligibility for an actual environment can be very expensive and time consuming therefore physical measures, such as the STI, have been developed (International Electrotechnical Commission, 2003).

The fast fluctuations of the intensity envelope of the speech correspond to the phonemes within words and the slow fluctuations coincide with sentence and word boundaries. These fluctuations, termed modulations, carry the most relevant information contributing to speech intelligibility and can be quantified as a function of modulation frequency (Houtgast and Steeneken, 1973). Any degradation of the modulation spectrum induced by the transmission path is considered to reduce speech intelligibility, as this reduction of the modulation spectrum coincides with the decrease in the modulation depth at one or more modulation frequencies. This change in modulation depth is quantified by the Modulation Transfer Function (MTF). The working hypothesis is that the

MTF of a sound transmission path reflects its performance with respect to speech intelligibility and is widely used in room acoustics (Houtgast and Steeneken, 1985).

Conversion of the MTF values to the STI is a five step procedure and briefly consists of averaging signal-to-noise ratios along the two variables, modulation frequency and the carrier frequency (Houtgast and Steeneken, 1985). The resulting STI values change between 0 and 1, while the former corresponds to silence or stationary noise, the latter refers to a fully comprehensible speech (see Appendix A).

The STI was first proposed as a basis for an ISE prediction model in the study of Hongisto (2005). The author presented findings from a literature review and classified the studies based on different focal tasks. The experimental results were summarized in a table with a focus on the number of subjects, length of the experiment, type of cognitive task, auditory stimuli employed in the experiment and the test performance in the relevant acoustic conditions (Hongisto, 2005, Table 2). The relationship between the performance scores and STI was supposed to follow the results of the subjective speech intelligibility tests of the sentences, obtained from IEC 60268-16. The relationship was non-linear and different for words, sentences and syllables. Subjective speech intelligibility increased rapidly for the STI values above 0.2 and reached 100 % when the STI was above 0.6 (Hongisto, 2005, Fig. 3). For the last stage of the development of the model, the author computed the average of the normalized error rates (performance decrease relative to the control condition) for each type of focal task and employed the smallest value (7 %, proofreading task) in a sigmoid function. The exponential curve reaching 100 % speech intelligibility represented 7 % decrease in performance relative to the silence condition and this point corresponded to an STI value above 0.6. The highest performance, on the other hand, was obtained when no speech is heard, when the STI is equal to 0.

There exist a number of studies which evaluated the STI parameter as an ISE predictor (e.g., Haka *et al.*, 2009; Park *et al.*, 2013; Haapakangas *et al.*, 2014; Ellermeier *et al.*, 2015; Liebl *et al.*, 2016; Senan *et al.*, 2018) and the model was reported to do well in accounting for situations of degraded / masked stimuli, however, it requires prior knowledge of the signal which can be an unknown factor. The choice of the normalization constant (7 %) and the unvarying normalized error rate above an STI value of 0.6 were also discussed as critical factors and considered to be possible reasons behind the limitations of the model (Ellermeier and Zimmer, 2014). The STI parameter is evaluated as an ISE predictor in the experiments reported in Chapter 2 and the computation of the STI is presented in Appendix A.

Fluctuation Strength(FS)

Another attempt to model the ISE was proposed by Schlittmeier *et al.* (2012) which is based on the hearing sensation fluctuation strength (FS) (Fastl and Zwicker, 2007, p. 247 - 256). FS is a psychoacoustic sensation that occurs because of the perception of temporal modulations. Temporal modulations can result in two different percepts: Roughness at higher frequencies of modulation and FS at low frequencies of modulation. Even though the boundary that separates FS from roughness is not strict, amplitude modulations below 20 Hz are attributed to FS and modulation frequencies above 20 Hz are attributed to the sensation of roughness (Fastl and Zwicker, 2007).

The FS reaches its maximum at a 4 Hz modulation frequency and shows a band-pass characteristic (the effect quickly diminishes for higher and lower modulation frequencies), which coincides with the average syllable rate in narrative speech (4 Hz) (Fastl and Zwicker, 2007). The FS value of 1 vacil (the unit of the fluctuation strength) is generated by a stimulus with 100 % amplitude modulation, a carrier frequency of 1 kHz and a level of 60 dB. FS values typically vary between 0 and 2 vacil.

In the study of Schlittmeier *et al.* (2012), a large number of serial-recall measurements (N = 70) were obtained using a variety of irrelevant sounds (e.g., speech, traffic noise, animal sounds, tone sequences, music, etc.) for the development of the prediction model based on the FS. One of the irrelevant sounds (music with staccato passages) which led to the median error rate with smallest interquartile range was selected as a reference sound. The normalized median error rate (7.5 %) obtained in the digit-recall task for the reference sound condition and its FS value (0.68 vacil) were used as normalization constants in the prediction algorithm. The FS values of all of the irrelevant sounds employed in the experiment were computed and used in the prediction model and the resulting normalized error rates were compared with the normalized error rates obtained from experiments.

The model managed to predict the error rates of 63 behavioral measurements out of 70 within the interquartile range of the experimental results. For the irrelevant sounds that particularly caused disruption (e.g., speech, staccato music, office noise), the algorithm produced valid estimates. The model was also successful in predicting the absence of a disruptive effect for continuous, steady-state sound conditions (e.g., legato music, traffic noise, continuous noise).

The FS was analyzed as an ISE prediction model in several studies after it had been proposed (Ellermeier *et al.*, 2015; Liebl *et al.*, 2016; Senan *et al.*, 2018). The results reported in these studies are discussed in detail in Chapter 2 and Chapter 3 where the FS parameter is evaluated. The definition of the parameter is presented in Appendix A.

1.4 Frequency domain correlation coefficient (FDCC)

The last prediction model to be discussed in this section is called the frequency domain correlation coefficient (FDCC) and in contrast to the aforementioned two models, this model was proposed as an ISE prediction model solely in the study by Park *et al.* (2013). In the study, an adaptive noise-masker algorithm was developed and evaluated with respect to its capability to reduce the ISE. The major objective of the study was to investigate if the adaptive masking approach can further reduce the serial-recall disruption when compared to stationary noise-maskers. Introduction of the stationary noise, combined with background speech, has already been stated as a useful option to reduce the ISE (Ellermeier and Hellbrück, 1998; Haapakangas *et al.*, 2014). The reason behind this performance increase is thought to result from the reduction of the spectral distinctiveness between the successive segments of the sound after the masking noise is added. Thus, the changing-state sound became "less" changing in "state" (more "steady-state") after the masker was introduced when compared with its original, unmasked version.

In order to investigate the effect of masking noise on ISE, Park *et al.* (2013) conducted a serial-recall experiment which consisted of two stationary masking noise conditions with two different signal-to-noise ratios; an adaptive masking noise condition; a continuous speech condition and continuous noise as control condition. The results were consistent in terms of the authors' expectations: The adaptive masking stimuli succeeded in improving the serial-recall performance when compared to the stationary masking noise with low noise level while the stationary masking noise condition with high noise level produced a similar degree of disruption as the control condition. The authors computed the STI values for the experimental stimuli and the parameter values were in conflict with the error rate for the adaptive masking conditions. It was concluded that the ISE can not be predicted by only using a temporal distinctiveness measure based on the STI, but it should be approached from a spectro-temporal perspective.

The FDCC was proposed as a spectral similarity metric which attempts to quantify the changing-state effect by following the definition of the hypothesis: The metric segments the stimuli into sound tokens and computes similarity between the power spectra of the successive tokens using correlations. For all audio samples in the study, the FDCC value was computed and the correlation values were found to be in line with the experimental results: The original speech stimuli resulted in the lowest FDCC value (largest spectral distinctiveness between sound segments) and the FDCC values for noise masking conditions followed the trend of the serial-recall results. The authors concluded that the use of the STI and the FDCC are limited as ISE predictors due the former one's focus on speech stimuli and the latter one's statistical approach to the segmentation of the tokens. However, these two models can be improved in a way to cover a wider scope of irrelevant sounds either by introducing a different modulation frequency weighting function for the STI or by arranging a different threshold point for the time intervals between the sound tokens for the FDCC.

Here it should be noted that the term "token", when used within the context of the FDCC, corresponds to the segments of sounds extracted by the token selection stage of the FDCC parameter. The FDCC attempts to relate the magnitude of the spectral variation within the irrelevant sound to the observed serial-recall performance and, as the main focus of this dissertation, it is evaluated as an ISE predictor in every chapter of the thesis. Therefore, the token selection stage of the parameter is also explained, modified and evaluated in this thesis.

On the other hand, the definition of the term "token" varies between different studies in the ISE literature, depending on how the irrelevant sound stimulus was constructed. The term generally refers to the smallest building block of the disruptive sound stimulus. For instance, if the irrelevant sound used is a sequence of words (e.g., *bowls-boy-day-dog-go-than-view*), then each word is considered as a token. Similarly, if the disruptive sound stimulus consists of sine tones then each tone is accepted as a token (e.g., Jones *et al.*, 1999).

The next section provides some examples of the different types of irrelevant sounds used in the literature. It focuses on some of the critical findings from the literature where the relation between the degree of the serial-recall disruption, ISE, and the magnitude of the variation within the irrelevant sounds, causing the changing-state, were investigated.

1.5 Changing-state hypothesis

The changing-state hypothesis states that the degree of the change, the physical mismatch, between successive items within the irrelevant stimulus determines the level of the secondary seriation process. Therefore, the stronger the prominence of the secondary seriation process the stronger the interference should be, hence the magnitude of the changing-state within the irrelevant sound is related to the degree of disruption. A brief literature review regarding the required physical mismatch and its relation to the degree of serial-recall disruption is presented in the following paragraphs.

The domain of the acoustic variation required to create the changing-state effect was investigated in several studies where the spectral and temporal features of the irrelevant sounds were systematically manipulated. The impact of the change in frequency was examined using: low-pass filtered words with different roll-off values (Jones *et al.*, 2000); sequences of sine tones with repeated and changing frequencies (Jones and Macken, 1993); sequences of vowels changing in pitch (Jones *et al.*, 1999); as well as noise-vocoded speech with different numbers of frequency bands (Ellermeier *et al.*, 2015; Senan *et al.*, 2018). These studies reported that changes in frequency within the irrelevant sound produced significant serial-recall disruption when compared to the control conditions.

It was shown that within the normal hearing range the serial-recall disruption is not dependent on sound intensity: The recall performance is approximately the same whether the sound pressure level is 80 dB(A) (e.g., someone talking loudly) or 45 dB(A) (e.g., a quiet library) (Colle, 1980; Ellermeier and Hellbrück, 1998). This attribute is also observed when the sound level is modified within or between the trials (Tremblay and Jones, 1999). In addition to this, modulating the irrelevant sound with a fixed or varying envelope (Jones *et al.*, 1992) or changing the duration between the concurrent tokens of the irrelevant sounds (Tremblay and Jones, 1999) were shown to be ineffective in producing an ISE.

A number of studies investigated the relation between the degree of disruption and the acoustic features of the irrelevant sounds. Due to the aforementioned findings, an assumption about the degree of disruption was that the magnitude of the ISE should be reduced when the spectral variation within the irrelevant sound was reduced.

In the first experiment reported in the study of Jones *et al.* (2000), a set of 14 male-spoken words was systematically degraded by applying a low-pass filter with roll-off values changing between 0 to 24 dB/octave, with a step size of 6 dB and the recall performance increased monotonically as a function of the filter roll-off value. 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 the degradation was increased and the effect was reported to be significantly linear.

In Jones *et al.* (1999), the first experiment also supported this assumption: Three irrelevant sound conditions were created by employing sine-tones and were labeled on the basis of frequency differences between the members of each set; small, medium and large. There were four acoustic conditions, including the silence condition. The results were in line with the expectation of observing a monotonic increase in the mean error rate as a function of frequency difference between the tones.

In the second experiment (Jones *et al.*, 1999, Exp. 2), the authors increased the variation in the stimuli by introducing timbral differences to the tone sequences using square (energy only at odd harmonics) and sawtooth (energy at all harmonics) waves. Four sound conditions were created: (1) repeated timbre, repeated frequency (only one of the tones repeating per sequence); (2) changing timbre, repeated frequency (a sequence of three tones with a frequency of 440 Hz changing in timbre only); (3) repeated timbre, changing frequency (one of the tones changing in frequency only); and (4) changing timbre, changing frequency (both attributes changing randomly in a sequence). The results showed that simultaneous variations in two parameters (timbre and frequency) produced lower mean error rate when compared to variation in only one parameter: Repeated timbre-repeated frequency (1) and changing timbre-changing frequency (4) generated lower mean error rates ($\cong$ 40 %) than the repeated timbre-changing frequency and changing timbre-repeated frequency conditions ($\cong$ 45 %).

It was argued that the non-monotonic relation between the magnitude of the changing-state within the sound sequences and the observed recall performance in the aforementioned experiment results from the introduction of a large variation in the sound sequence. If the differences within the sequence are within a perceptual limit that preserves the characteristics of the sequence, then the sequential information is retained; when the difference is too large, the sequential information is diminished and therefore the secondary ordering process is weakened.

Jones *et al.* (1999, Exp. 3) investigated the role of perceptual limits on the ISE by introducing pitch changes to the semitones from relatively modest (two and five semitones) to the extreme (10 semitones) in the third experiment with the expectation that if the pitch changes are above the “limit” (five semitones) then the sequence would be perceived as two distinct streams with repeated frequency instead of one with changing frequency and the impairment would drop. The disruption increased as a function of pitch difference and reached the maximum at the five-semitones condition, while the error rates observed for the two- and the ten-semitones conditions were similar. The results in the study clearly demonstrated that the function relating the pitch difference and degree of disruption is not monotonic and the perceptual organization of the auditory stimulus, in this case the auditory stream segregation based on pitch perception (Bregman, 1990), plays a key role.

The role of perceptual organization on the ISE was investigated further by Jones and Macken (1995a) focusing on the spatial location of the irrelevant sounds. In the first experiment two sound sequences were created by recording three letters spoken by a female voice: “U”, “C” and “O”. The first material was recorded monaurally and presented diotically. The second sequence was created for stereophonic presentation in which “U” was played on the left channel, “O” was played on the right channel and “C” was played on both channels. When the stereophonic recording was played over headphones it led to the perception of three repeating streams based on their location instead of one varying sequence. The serial-recall results also supported this distinction of the streams: The mean error rate observed in the stereo sound condition was lower than what was observed in the mono condition.

A similar experiment was conducted by Jones and Macken (1995b) using a speech stimulus as the irrelevant sound: Six overlapping voices were presented from either one spatial location (babble speech) or from six loudspeakers located in six different positions. It was observed that when the voices were played back from different locations there was a significantly higher degree of disruption when compared to the babble speech condition. This was explained by the role of stream segmentation on the ISE: When the voices overlap the cues for segmentation, such as amplitude variations, are reduced due to energetic masking, and speech gradually develops into a relatively steady-state sound with occasional words and syllables noticeable; when the individual voices are presented from different spatial locations, the speech is perceptually reorganized and the segmentation cues are recovered.

Another example of the role of segmentation was demonstrated using non-speech sounds: A continuous tone randomly varying in pitch did not produce an ISE, but when the same sound was interrupted by short segments of silence, the ISE was observed (Jones *et al.*, 1993). The result showed that the magnitude of the physical change itself, the degree of the variation in pitch, is not sufficient to produce an ISE: In order to generate the changing-state effect, the change has to occur from one segment to another.

The key findings related to the features of the changing-state hypothesis are not limited to those reported above, however, this overview should be sufficient to give an idea about the acoustic and perceptual aspects involved in the serial-recall disruption. The impact of the spectral variations within the irrelevant sounds (Ch. 2-5), the role of segmentation (Ch. 4), the role of temporal variations within the irrelevant sounds and the role of auditory stream segregation (Ch. 5) on the ISE are the features of the changing-state hypothesis which are discussed in detail in the corresponding chapters.

1.6 Goals of the thesis

The major goal of this dissertation is to evaluate and improve an ISE predictor, the FDCC, with respect to its ability to quantify the magnitude of the changing-state effect observed for various types of auditory stimuli, from degraded, masked or continuous speech to sequences of noise-bursts. This predictor is the only predictor designed for such a purpose in the literature and thus follows the definition of the changing-state hypothesis. Therefore, the FDCC was employed not only for evaluating and detecting its limitations, but also used as a medium to observe the relationship between the magnitude of the spectral variation that the irrelevant sounds comprise and the ISE observed in such acoustic conditions because this is typically not a monotonic relation.

The use of a spectral metric gave further insight into the role of speech intelligibility on the ISE. The systematically manipulated speech-like stimuli were evaluated by both the STI and the FDCC parameters. It was observed that the variation applied in the frequency domain was also reflected in changes in the time domain. When both variations could be quantified, these could be independently evaluated and their independent roles, with respect to the ISE, could be analyzed. This was demonstrated in the studies reported in Chapter 2. In the same chapter, the normalized covariance measure (NCM) was also evaluated as a potential ISE predictor.

The role of the changing-state syllables, vowels and consonants was also investigated in a way that the token selection stage of the FDCC was evaluated based on a small set of short speech samples in Chapter 3. It was found that when the token selectivity of the segmentation procedure was revised with a focus on syllables, the prediction accuracy was increased. More interestingly, when the metric was improved, the prediction accuracy was increased not only for speech sounds, but also for non-speech sounds as well. Furthermore, the developed token selection stage was used as a basis to investigate the role of segmentation on the ISE in Chapter 4.

Moreover, the magnitude of the spectral variation was shown to yield different results in ISE experiments depending on the type of irrelevant sounds, such as pure tones varying in frequency vs pitch shifted vowels (Jones *et al.*, 1999), which eventually had led to a discussion about the role of speech and non-speech irrelevant sounds on the ISE (LeCompte *et al.*, 1997). In Chapter 5, we generated a set of non-speech stimuli which possessed the same magnitude of spectral variation and similar temporal characteristics as speech. The role of spectral variation on the ISE was evaluated using both speech and non-speech stimuli in a serial-recall experiment and the use of the FDCC allowed us to base our conclusions on a physical measure.

The FDCC was evaluated regarding its ability to predict the ISE and its applicability as a physical measure in the ISE studies. We aimed to reveal the promising aspects as well as the limitations of the metric and, when possible, we attempted to improve the prediction accuracy of the FDCC.

1.7 Outline

The prediction models explained above, the STI, the FS, the FDCC and the NCM are evaluated using a distorted speech stimulus, noise-vocoded speech (NVS), in two serial-recall experiments in Chapter 2 where the number of frequency bands used in generating the NVS stimuli is varied. The noise-vocoding technique used to generate the NVS stimulus allowed us to investigate the role of spectral features on ISE as well as the speech intelligibility.

In Chapter 3, the two ISE predictors, the FDCC and the FS are evaluated using a large dataset (N = 91) obtained from four studies in the literature. The Pearson correlation values between the two metric values and the corresponding normalized error rates are computed and compared. In the same study, the token selection stage of the FDCC is also modified and evaluated by comparing it with the previous version.

Two studies are presented in Chapter 4: In the first study, the recently developed token selection stage of the FDCC is analyzed by segmenting continuous speech into tokens and employing both the continuous and segmented versions in a serial-recall task. The experiment also consisted of segmented non-speech stimuli obtained from the literature and a between-subject analysis is conducted for the continuous and segmented non-speech sounds. In the second study, a set of specially crafted NVS stimuli is obtained from the literature and the roles of speech fidelity and the spectral variation on ISE are investigated.

The spectral and temporal features of the irrelevant sounds are investigated in Chapter 5 using two metrics, the average modulation transfer function (AMTF) and the FDCC. Two sets of noise-pulse trains are generated and used in two experiments, in which the spectral and temporal features of the noise stimuli are modified independently. The first set contains periodic noise-pulse sequences while in the second set, the position, the amplitude and the duration of each pulse is derived from speech samples which are extracted using the token selection stage of the FDCC algorithm. In both sets, the second pulse of each noise-pulse train is independently modified in time or frequency content and the resulting modified noise-pulses are used in serial-recall experiments.

In Chapter 6 the results and conclusions derived from each chapter are summarized. The revealed advantages and limitations of the FDCC are presented. Further improvements that can be applied to the FDCC are also discussed.

2 | Cognitive disruption by noise-vocoded speech stimuli: Effects of spectral variation[1]

Abstract

The effect of irrelevant sounds on short-term memory was investigated in two experiments using noise-vocoded speech (NVS) stimuli. 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 NVS conditions, composed of 1, 2, 4, 6, 9, 12, 15 and 18-bands, were used as the irrelevant sound 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-band condition but there was no influence of number of bands on performance beyond 6 bands. The results were analyzed using four acoustic metrics proposed in the literature: the frequency domain correlation coefficient, the fluctuation strength, the speech transmission index and the normalized covariance measure. None of the metrics successfully predicted the results. However, the parameter values of the frequency domain correlation coefficient, the speech transmission index and the normalized covariance measure indicated that a prediction model for irrelevant sound effect should account for both temporal and spectral features of the irrelevant sounds.

[1] This chapter is a modified version of:
Senan, T. U., Jelfs, S., and Kohlrausch, A. (2018). "Cognitive disruption by noise-vocoded speech stimuli: Effects of spectral variation," J. Acoust. Soc. Am. 143(3), 1407-1416.

2.1 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 in 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 using 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) (Banburry *et al.*, 2001; Ellermeier and Zimmer, 2014).

A serial-recall task is a short-term memory task which requires remembering the order of visually 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 of the order of the items and of 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 (Ellermeier and Hellbrück, 1998; Tremblay and Jones, 1999) and spectral features (Jones and Macken, 1993; Jones *et al.*, 2000) 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).

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 the ISE, by following the definition of the 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 the 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 the ISE is not a straightforward process since the degree of disruption using 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 which 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 the 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 by dynamic envelope compression introduced by hearing aids, the 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). The NCM determines the transmission index values from the band pass 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 is therefore evaluated in the current study as an ISE predictor.

The present study describes a stimulus synthesis procedure which 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 (NVS) stimulus 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 NVS 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). In the study, a letter-recall task with seven letters is reported for one control (steady-state white noise) and four NVS conditions: 3-bands, 6-bands, 9-bands 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-bands 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 the STI performed 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 for 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 NVS stimuli 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 studies by employing a finer grain on the number of frequency bands and investigating the behaviour of the spectral parameter, the frequency domain correlation coefficient. 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. The following section explains the spectral descriptor and relates the parameter values to the auditory stimuli. Section 2.3 and 2.4 describe the experimental procedures for the two experiments. The results of the aforementioned prediction models are compared in Section 2.5 in the light of the experimental results, where each parameter is explained in Appendix A including two new metrics which are not analyzed in this chapter. The chapter ends with the discussion (Sec. 2.6) and conclusion (Sec. 2.7) sections.

2.2 Spectral estimator and stimuli

2.2.1 Spectral estimator

The spectral estimator used in this study is called frequency domain correlation coefficient (FDCC). The 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).

The 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 quantifying 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 which 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 P is calculated for each band of each token.

The FDCC for two successive tokens is defined as follows:

$$FDCC_i = \frac{\sum_{j=1}^{19} P_{i,j} P_{i+1,j}}{\sqrt{\left(\sum_{j=1}^{19} P_{i,j}^2\right)\left(\sum_{j=1}^{19} P_{i+1,j}^2\right)}} \tag{2.1}$$

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

2.2.2 Noise-vocoded speech (NVS)

As mentioned in the first section, 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, NVS stimuli were generated by dividing the speech signal between 50 and 8000 Hz into 1, 2, 4, 6, 9, 12, 15 and 18 Hanning-shaped band-pass 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, software is available at www.praat.org). The band pass filters have a roll off of 6 dB/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 to create a set of stimuli which changes from broadband amplitude-modulated white noise to highly intelligible NVS when increasing the number of frequency bands. The cut-off frequencies and the number of frequency bands of the stimuli are presented in Table 2.1.

Table 2.1: Noise-vocoded speech stimuli with the number of frequency bands and upper frequency boundaries.

Nr. of bands	Exp. used in	Cut-off frequencies (Hz)
18 Bands	(Exp. 1 and 2)	98, 157, 229, 317, 425, 558, 720, 918, 1160, 1457, 1820, 2265, 2809, 3474, 4289, 5286, 6506
15 Bands	(Exp. 1)	110, 184, 280, 402, 560, 756, 1010, 1332, 1742, 2265, 2931, 3782, 4863, 6242
12 Bands	(Exp. 1)	126, 229, 369, 558, 814,1160, 1630, 2265, 3125, 4289, 5865
9 Bands	(Exp. 1)	157, 317, 558, 918, 1457, 2265, 3474, 5286
6 Bands	(Exp. 1 and 2)	229, 558, 1160, 2265, 4289
4 Bands	(Exp. 2)	370, 1160, 3125
2 Bands	(Exp. 2)	1160
1 Band	(Exp. 2)	...

In the study by Ellermeier *et al.* (2015), the experimental results showed that when four or more bands were used in the native language condition, the recall performance got very close to the level of original speech. The explanation of this may be linked to the almost-speech-like intelligibility of the NVS conditions when more than four bands are employed (Davis *et al.*, 2005). However, when the spectral parameter values are examined (see Fig. 2.1), the spectral similarity continues to decrease beyond the 4-bands condition. In fact, NVS stimuli 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 for intelligible and unintelligible NVS conditions. The NVS conditions with 6 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 NVS stimuli for the two experiments are explained in detail in the Sections 2.3 and 2.4.

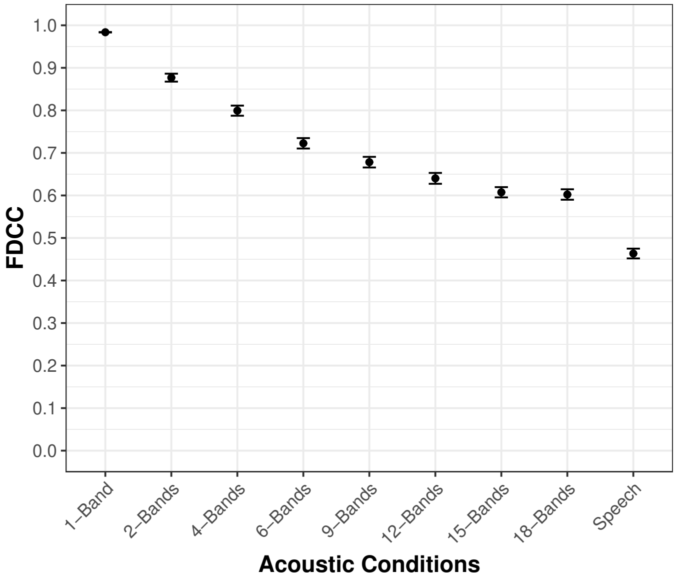


Figure 2.1: Parameter values of the frequency domain correlation coefficient (FDCC) for all acoustic conditions used in the two experiments. A value of 1 indicates no spectral difference between successive sound tokens of the stimulus. Error bars represent the standard error of the mean (SEM).

2.3 Experiment 1

For the first experiment, five different NVS stimuli 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.

2.3.1 Method

Participants

Fifteen native speakers of Dutch (eight females and seven males, age range between 18-50 years) 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.

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 10 long speech exemplars were used in the original speech condition and were transformed into noise-vocoded speech stimuli for every condition (for the procedure, see Sec. 2.2.2). In addition to the silent training condition, there were seven acoustic conditions; 6-band NVS, 9-band NVS, 12-band NVS, 15-band NVS, 18-band NVS, silence (SLNC) and original speech (Speech).

Apparatus

The experiment was run on a Hewlett-Packard computer using MATLAB (R2014b). 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 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 $\text{dB}_{\text{LAeq1min}}$. A computer screen was positioned outside of the soundproof booth and was visible through the double-glass window.

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 1-9) 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, to correct the previously pressed key input or to select a number key more than once. The auditory stimuli were played 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 these 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 10 trials and took approximately 7.5 min to complete. There were 2 min breaks after each block and one experimental session took approx. 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.

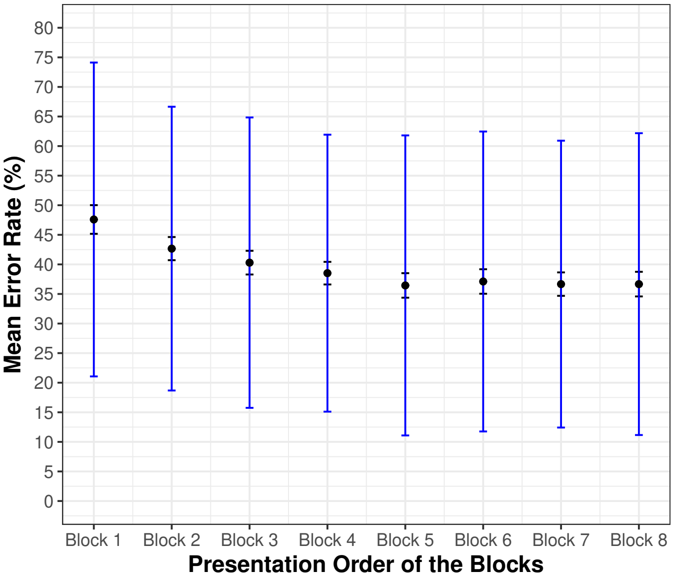


Figure 2.2: Mean error rate (%) as a function of the order of the presentation blocks. Dark error bars represent the SEM and blue error bars represent the standard deviation (SD).

2.3.2 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.

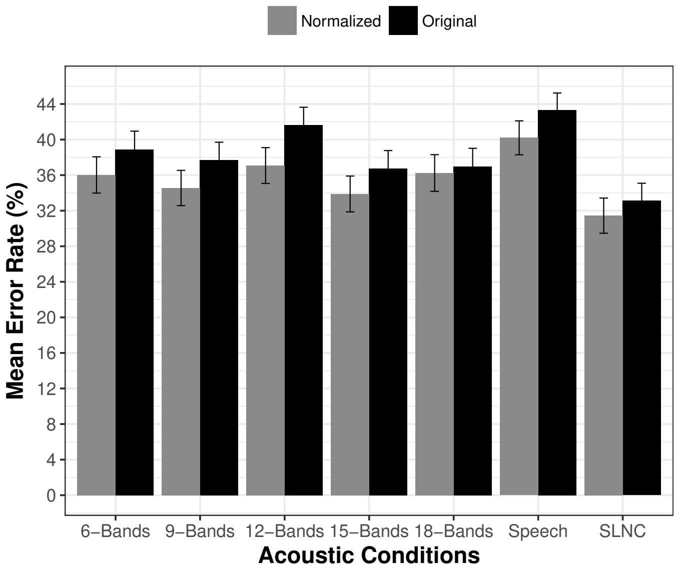


Figure 2.3: Recall performance of 15 participants as a function of the acoustic conditions, represented as mean error rates (%). Grey and dark bars represent normalized and original scores, respectively. Error bars represent the SEM.

In Fig. 2.2, 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 ANOVA, $F(7, 98) = 3.205$; $p < .005$. 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 2.3 presents the mean error rates as a function of the acoustic conditions calculated for both original and the normalized scores. The difference between the mean error rates for SLNC (31.5 %) and the speech condition (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 (Jones and Macken, 1993; Tremblay and Jones, 1999; Schlittmeier *et al.*, 2012). 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 silence and the speech conditions. As it appears, the recall performances varied in a rather random manner for the various NVS 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, $F(6, 84) = 2.198$, $p < .05$, $\eta^2 = 0.01$. Tukey HSD tests were conducted on all possible pairs of sound conditions. Only one pair of groups was found to be significantly different (p <.05): SLNC (M = 31.4, SD = 24.2) and Speech (M = 40.2, SD = 23.4); p = .015.

2.3.3 Discussion

The data did not show a significant increase in mean error rates with an increase in the number of frequency bands within the NVS 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.2.2, Fig. 2.1) 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 the experimental results may be due to the small sample size or the choice of the number of frequency bands and will be discussed in Section 2.5. Nevertheless, the continuous decrease of FDCC values as a function of number of the frequency bands clearly shows that the interpretation of the parameter needs to be adapted, in order to be a valuable predictor of the ISE.

2.4 Experiment 2

For the second experiment, the NVS 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 NVS conditions used in the first experiment were also preserved in the experimental design.

2.4.1 Method

Participants

Twenty-five participants (15 females and 10 males, age range between 18-50 years) participated who were recruited via the JF Schouten subject database of the Eindhoven University of Technology, Eindhoven, The Netherlands. Twenty of twenty-five participants were university students and all 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 Internal Committee Biomedical Experiments (ICBE) of Philips Research and by the Human Technology and Interaction department, Eindhoven University of Technology.

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 three NVS conditions: The 9-bands, 12-bands, 15-bands were replaced with 1-band, 2-bands and 4-bands. The 6-bands, 18-bands, Speech and SLNC conditions were preserved.

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). 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 $\text{dB}_{\text{LAeq1min}}$. 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.

Procedure

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

2.4.2 Results

A similar learning effect, as shown in Fig. 2.2 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, $F(7,\ 168) = 11.59$; $p < .001$. The mean error rate percentage difference between the first (TRNG) and the last block was ≈ 20 % and half of the 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 2.4 shows the error rate percentages in different acoustic conditions. The mean error rate in the SLNC and speech conditions were 27.5 % and 35.5 %, respectively. The lowest mean error rate was observed in the 1-band NVS condition, 27 %, and the highest was observed in 6-band NVS, 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. 2.1), except for the 18-band NVS condition: The decrease in the mean error rate in this condition is not reflected in the parameter values.

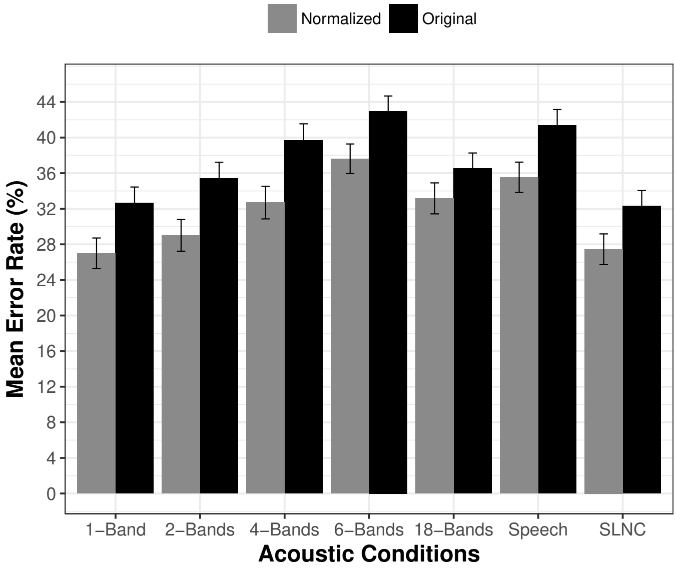


Figure 2.4: Recall performance of 25 participants as a function of the experimental conditions, presented as mean error rates (%). Grey and dark bars represent normalized and original scores, respectively. Error bars represent the SEM.

The effect of acoustic conditions on recall performance was highly significant and confirmed by a one way repeated measures ANOVA, $F(6,\ 144) = 7.939$, $p < .001$, $\eta^2 = 0.02$. 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 ($p < .05$). The statistical results are summarized in Table 2.2.

2.4.3 Discussion

The normalized mean error rate for SLNC (27.5 %) and the speech conditions (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 (Jones and Macken, 1993; Tremblay and Jones, 1999; Schlittmeier *et al.*, 2012). The increase in the mean error rates as a function of the number of frequency bands between the 1-band and 4-band 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-band vocoder condition was unexpected when compared to the 20-band condition in the same study.

The results are similar to those from the first experiment in terms of SLNC and speech conditions with an exception that the speech stimulus did not lead to the highest error rate. Although the speech and 6-band 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-band Japanese NVS condition: The mean error rate in the 4-band NVS condition was slightly higher than in the original speech condition 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 NVS conditions, 6-bands 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-band and 18-band pairs differed significantly (original score; $p = .035$, normalized; $p = .314$). 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 1 to 6 bands supports the effect of spectral variation within the ISE, the FDCC parameter, in its current form, is not adequate to address the complexity

Table 2.2: Pairwise comparison of test performances for all possible pairs using the Tukey HSD test. Only the pairs with statistical significance are reported.

Statistically significant pairs	Mean error rate (%)	p values
SLNC - Speech	27.5 - 35.5	$p < .002$
SLNC - 6-Bands	27.5 - 37.6	$p < .001$
1-Band - 6-Bands	27 - 37.6	$p < .001$
1-Band - 18-Bands	27 - 33.2	$p < .05$
1-Band - Speech	27 - 35.5	$p < .001$
2-Bands - 6-Bands	29 - 37.6	$p < .001$
2-Bands - Speech	29 - 35.5	$p < .05$

of the phenomenon. In fact, the experimental results indicate that the cognitive response towards the spectral variation is not straightforward and a perceptual approach to the spectral features may be needed to predict the ISE.

2.5 Prediction models from the literature

The experimental results have been further analyzed by evaluating three objective metrics as prediction models for ISE: the fluctuation strength (Fastl and Zwicker, 2007), the speech transmission index (Steeneken and Houtgast, 1980) and the normalized covariance measure (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 to 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 in the silence condition was subtracted from this value. The resulting value was accepted as a relative error rate which showed the magnitude of disruption in 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 sound conditions in the two experiments are presented in Fig. 2.5.

The first metric, proposed in the study of Schlittmeier *et al.* (2012), is derived from the hearing sensation fluctuation strength (FS) (Fastl and Zwicker, 2007) and aims to predict the ISE resulting from both speech

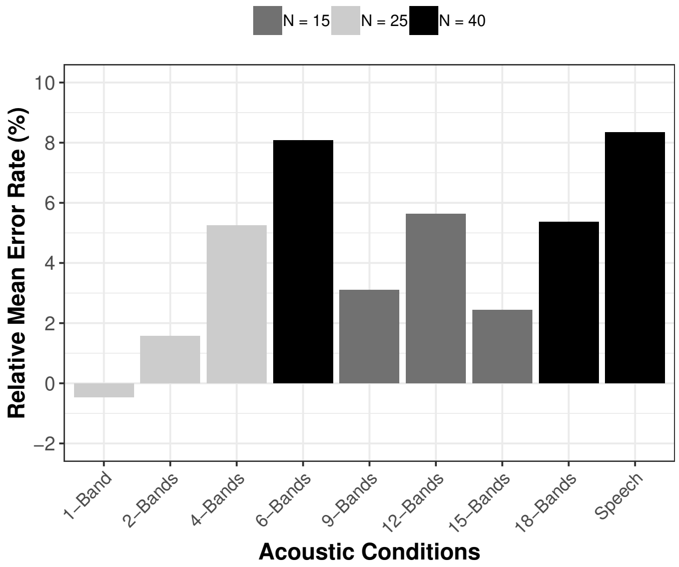


Figure 2.5: Relative mean error rate with respect to silence for the irrelevant sound conditions of the two experiments. Dark gray, light gray and black bars represent the different sample sizes of 15, 25 and 40, respectively.

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, known to be the syllable rate of speech. In the study where the metric was proposed 70 behavioural measurements, obtained for 44 sound conditions (e.g., native, foreign and babble speech, tones changing in frequency, music, office noise, etc.), were presented. The FS values of the experimental stimuli and the corresponding experimental results were used in an algorithm (Schlittmeier *et al.*, 2012, Eq. 2) which estimated the disruption observed in a particular sound condition. The algorithm was successful in predicting the performance drop for 63 out of the 70 measurements and the FS parameter values shared 55 % of the overall variance with the experimental results (Schlittmeier *et al.*, 2012).

All 90 stimuli (10 long sentences x 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. 2.6.

First, it can be seen that the FS model can not make accurate predic-

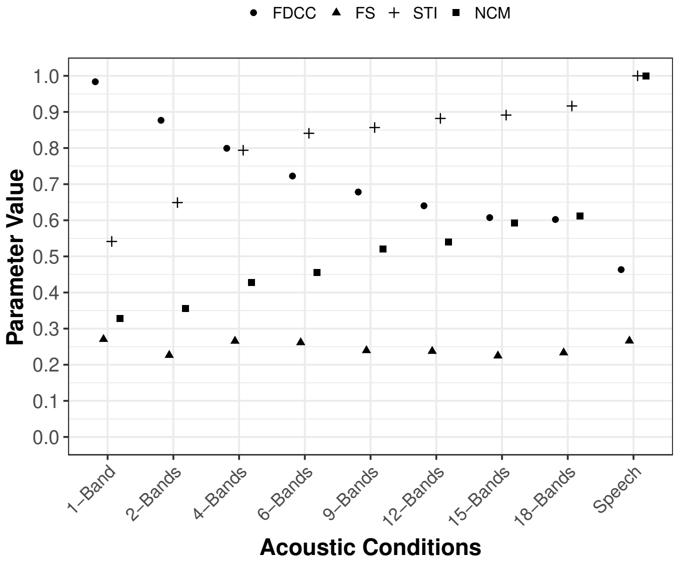


Figure 2.6: Mean parameter values of the frequency domain correlation coefficient, the fluctuation strength, the speech transmission index and the normalized covariance measure for the noise-vocoded speech and the original speech stimuli.

tions 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, fluctuation strength tended to decrease (from a maximum of 0.69 to 0.49) with increasing number of vocoded channels for both German and Japanese NVS stimuli. The 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.

The second metric analyzed for the NVS conditions is based on the speech transmission index (STI), an estimator for speech intelligibility, which was proposed as an ISE prediction model in the study of Hongisto (2005). In the study, the author presented a table which summarized the experimental results obtained from studies that investigated the impact of speech on task performance (Hongisto, 2005, Table 2). The results were grouped based on the types of the focal tasks employed in the experiments and the relative mean error rates for each group were computed. The obtained minimum relative mean error rate (7 %, proofreading task) was used in a sigmoid function which was built using the curve of subjective speech intelligibility results derived from IEC 60268-16 (International Electrotechnical Commission, 2003) with the generalization that all task performances were disturbed at least 7 % by intelligible speech. The model proposed that task performance starts to decrease at an STI value of 0.2 and reaches its minimum when the parameter value reaches 0.7. In the present study, only the STI values of acoustic conditions were computed and the sigmoid function proposed in the study of Hongisto (2005) was not used.

The computation of the STI requires a reference signal and for the analysis of NVS stimuli, the original sentences were used as the reference and STI values were computed for every vocoding condition (10 test sentences x 8 noise vocoding conditions). The mean values of the parameter are presented in Fig. 2.6. When the values of the STI parameter are investigated, it can be seen that the parameter value increases with the number of the frequency bands of the NVS stimuli. However, relative mean error rates for the NVS conditions with more than 6 frequency bands do not follow the trend of the parameter values, while the relative error rates for 1 to 6-band NVS shows resemblance.

The last metric, the normalized covariance measure, 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. 2.6. 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 the original speech condition (NCM: 1) and 18-bands (NCM: 0.61), 0.39, is larger than the overall range of the parameter values for the NVS 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 the shape of the curve and the large difference between 18-bands and original 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.

2.6 General discussion

The experiments explored serial-recall performance, a standard short-term memory measure, using 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 to examine the extent to which the detrimental effects of speech and speech-like stimuli on recall performance can be predicted by a spectral parameter.

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. 2.5). Varying the number of frequency bands between 6 and 18 (represented by dark gray bars in Fig. 2.5) 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. 2.5) 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 to 6-band conditions followed the trend of parameter values, but reached a ceiling in the 6-band NVS. 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 rate was 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. 2.5, where the relative error rate reached their maximum in the 6-band NVS condition. The nature of such a limit is hard to explain but it was observed that this maximum appeared when the NVS 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-band NVS are readily intelligible. The intelligibility thresholds as a function of the number of bands, were, however reported differently in the two ISE studies. In Ellermeier *et al.* (2015), the 4-band NVS 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 NVS and the original speech conditions might have created an increase of the perceived clarity of speech content in NVS stimuli: When the original speech and/or NVS with higher number of frequency bands (e.g. 18-bands) appear early in an experimental session, perceived intelligibility of NVS 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 recognized than the subjects who did not hear the clear speech. The same study also reported that hearing NVS stimuli 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 NVS stimuli 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). The role of intelligibility on the ISE is only negligible 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-band conditions and that the 6-band NVS stimuli are highly intelligible. It can be argued that when the stimulus reached a certain level of preserving the temporal and the spectral features of the original speech, its detrimental impact reaches a maximum. One can think that if the 6-band condition already preserves the acoustic cues needed to create maximum disruption, NVS conditions with more than 4 bands should be equally distractive as the speech. From this perspective, 6-bands, 9-bands, 12-bands, 15-bands, 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-band 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-band NVS condition.

When analyzing the STI and the NCM values 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-band, 2-band and 4-band NVS conditions are 0.75, 0.81 and 0.89, respectively. These parameter values are elevated by a constant amount with respect to our measurements depicted in Fig. 2.6. Note that the 1-band condition in their study already reaches a parameter value above 0.7, which is the STI beyond which performance is not expected to deteriorate further (Hongisto, 2005). Second, the STI value of 0.54 attributed to 1-band NVS condition in the current study indicates that the STI overestimates the intelligibility levels of NVS stimuli regardless of this inconsistency.

The subjective intelligibility scores (Dorman, *et al.*, 1997) and NCM values have been reported in the study of Chen (2011) where 2-band, 4-band and 6-band NVS stimuli 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. 2.6 are slightly lower than what is reported in that 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 performance and the intelligibility predictions of these two metrics, it should be noted that the STI and the NCM are affected by the increase in the number of frequency bands of NVS stimuli 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.

2.7 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 literature for the second experiment: The performance decreases up to 6-band noise-vocoded speech stimuli. 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 irrelevant sound effect can not be predicted by the current structure of the spectral parameter: The FDCC values indicate a change in frequency spectrum beyond the 6-bands which does 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, speech transmission index and fluctuation strength, also failed to make accurate estimates for the magnitude of the disruption for noise-vocoded speech stimuli. It appears that the complexity of the paradigm exceeds the capabilities of a single acoustic metric, therefore a more sophisticated model which accounts for this complexity should be developed.

3 | Psychoacoustic modelling of the changing-state hypothesis in short-term memory experiments involving serial recall

Abstract

Previous research has shown that the FDCC was able to predict the serial-recall results to a certain extent but also demonstrated limitations. The present study investigated those limitations and attempted to improve the parameter's prediction accuracy by integrating a peak detection phase into the token segmentation stage of the algorithm. The improved metric was evaluated by employing a large set of irrelevant sound stimuli from the ISE literature and comparing the experimental results with prediction values of the two versions of the FDCC and the fluctuation strength parameters.

3.1 Introduction

The irrelevant sound effect (ISE), the detrimental impact of background sounds on short-term memory, has been extensively studied in the literature and it was shown that in order to observe the effect, certain combinations of the focal task and the irrelevant sound stimuli should be formed: While some focal tasks are affected by certain irrelevant sounds, other tasks stay unaffected by the presence of the same disruptive stimuli. In addition to this complexity, the magnitude of disruption also depends on both the properties of the focal task (Hughes *et al.*, 2007) and the irrelevant stimuli (Jones, 1999; Jones *et al.*, 1999). Although the critical aspects of the ISE have been well established and a framework has been created, predicting the magnitude of disruption was shown to be a complicated challenge (Ellermeier *et al.*, 2015; Liebl *et al.*, 2016).

As demonstrated in the previous chapter, the ISE can be quantified by behavioral experiments where participants try to perform cognitive tasks in the presence of irrelevant background sounds. Typically, disruptive sounds consist of a reference condition (e.g., silence or steady-state noise) and the ISE is quantified by averaging the error rates over all trials in each acoustic condition and subtracting the average error rate of the reference condition from these values. The results serve as a measure of relative error rate per participant for each sound condition. Moreover, the individual error rates are averaged in order to obtain the normalized error rate for each sound condition. Further on in this study, the term *normalized error rate* is used to indicate the performance drop when compared to the silence condition while the term *error rate* refers to the absolute error rates.

One of the most commonly employed cognitive tasks in the ISE literature is the serial-recall task where participants try to recall the order of to-be-remembered items (e.g., letters or digits) presented in a randomized order on a computer screen while being exposed to the irrelevant acoustic stimuli. Even though the participants are instructed to ignore the background sounds, the brain still processes the auditory input as well as the visual, which may degrade cognitive performance.

This inevitable conflict is thought to result from the interference of two parallel ordering processes; one for the conscious processing of the visually presented items, and one for the involuntary processing of the acoustic input. In order to reach a degree of interference, the acoustic properties of the irrelevant sound should vary in time. This is explained by the changing state hypothesis (Jones *et al.*, 1996): An acoustic stimulus should be distinguishable into perceptually discrete segments, and each segment of the stimulus must differ, in terms of spectral features, from the one that precedes it.

The properties of the changing-state hypothesis have been investigated by studies which revealed that while spectral variation impairs the short-term memory performance (Jones and Macken, 1993; Jones *et al.*, 2000), changes in the sound pressure level between successive sound segments do not create disruption (Ellermeier and Hellbrück, 1998; Tremblay and Jones, 1999). In fact, the short-term memory disruption is observed in any acoustic condition which satisfies the changing state hypothesis, regardless of the information it consists of: background music (Perham and Vizard, 2011), alternating tones (Jones and Macken, 1993; Jones *et al.*, 1999), alternating band-pass filtered noise bursts with different center frequencies (Tremblay *et al.*, 2001) as well as native, foreign and reversed speech (Jones *et al.*, 1990).

The changing state hypothesis aims to explain the short-term memory disruption by providing a framework and defining the requirements in order to observe the effect. However, the relationship between the magnitude of the changing state and short-term memory disruption seems to be a complex problem since the performance drop reaches a maximum using speech stimuli as the irrelevant sound (Tremblay *et al.*, 2000; Schlittmeier *et al.*, 2012; Ellermeier and Zimmer, 2014). This finding is in conflict with the changing-state hypothesis as it ascribes a special role to speech stimuli and suggests that the phenomenon can be explained by focusing on speech specific properties instead of global acoustic features. One of the proposed ISE predictors, the speech transmission index (STI), follows this reasoning.

The STI is a speech intelligibility metric which quantifies the temporal change of the original speech after it is transmitted through a medium (e.g., phone line, a room) (Steeneken and Houtgast, 1980): It is defined by the amplitude modulation ratio between the modified signal (e.g., recorded signal) and the original (e.g., source signal). The STI was proposed as a basis for predicting the cognitive disruption induced by background sounds, by processing the parameter values with a sigmoid function (Hongisto, 2005). A couple of limitations were observed: The computation of the STI requires a reference signal which may not always be available and the model seems to be efficient for speech and masked speech stimuli only. It was also reported that the metric can overestimate the intelligibility of 1-band noise-vocoded speech, which is actually amplitude-modulated white noise (Ellermeier *et al.*, 2015).

Other than the STI, two metrics have been proposed as ISE predictors in the literature which do not ascribe any special role to the speech stimuli: the fluctuation strength (FS) (Schlittmeier *et al.*, 2012) and the frequency domain correlation coefficient (FDCC) (Park *et al.*, 2013).

The FS is based on a psychoacoustic sensation (Fastl, 1982; Fastl and Zwicker, 2007) which is perceived when listening to slowly modulated ($<$ 20 Hz) sounds. The unit of FS is called vacil and it reaches a maximum with fluctuations of approx. 4 Hz, which is close to the average syllable rate in continuous speech (Jones *et al.*, 1992). The FS was first employed as an ISE prediction metric in the study of Schlittmeier *et al.* (2012). The model was evaluated with a diverse and large set of stimuli and 70 behavioral measurements were conducted using a serial-recall task. The predicted error rates were within the interquartile range of the experimental results for 63 out of 70 measurements.

Currently, the FS model is the most successful prediction model within the context of ISE, although some limitations were reported: The lack of ability to discriminate between amplitude and frequency modulation is a limitation which was observed in another study where noise-vocoded speech stimuli were used as the irrelevant sound (Ellermeier *et al.* 2015; Chapter 2). The model was also shown to be inadequate in another ISE study where various degrees of masked speech were employed as the irrelevant sound stimuli (Liebl *et al.*, 2016). These three studies indicate that the FS model has shortcomings when used with degraded speech stimuli.

The third metric proposed for predicting ISE, the FDCC, follows a different approach than the aforementioned ones and attempts to quantify the changing-state hypothesis: The metric divides the sound into segments and computes a correlation between the power spectra of the successive segments of the sound. The FDCC, by definition, does not distinguish between speech and non-speech sounds and only focuses on the change in the spectrum from one segment to another. The metric was first proposed in the study of Park *et al.* (2013), where the authors systematically modified the speech stimuli by employing an adaptive masking scheme and serial-recall results were evaluated using, among others, the FDCC metric. The results were promising, however, the experimental results reported in Chapter 2, where noise-vocoded speech stimuli with 1 to 18 frequency bands were used as irrelevant sounds, showed that the metric predicted the serial-recall performance successfully to a certain extent but demonstrated limitations (Chapter 2).

The aforementioned two prediction models, the FS and the FDCC, are particularly important for the current study since the present study attempts to evaluate the spectral descriptor, the FDCC, by employing four sets of irrelevant sound stimuli from the literature for which the FS values are also available. The relationship between the experimental data and the parameter values of the FDCC is investigated and the results are used to compare the prediction ability of the FDCC with that of the FS. In addition to this, the token selection stage of the FDCC is also modified and the parameter values of the two versions, FDCC_{old} (as defined in Chapter 2) and FDCC_{new}, are evaluated. The FDCC_{new} is defined in Sec. 3.2, where a detailed explanation of the modification is included. The experimental data, collected irrelevant sound stimuli and the descriptions of the experimental procedures are presented in Sec. 3.3. The results are presented in Sec. 3.4 and the study is finalized by the conclusion section, Sec. 3.5.

3.2 FDCC_{new}

The FDCC is a correlation measure between successive segments of a sound in the frequency domain. It was proposed as a spectral similarity metric (Park *et al.*, 2013) and attempts to explain the behavior of the ISE by following the definition of the changing-state hypothesis. It should be noted that the token segmentation stage of the FDCC in the present study is different from the one reported in Chapter 2 and in the previous studies (Park *et al.*, 2013; Senan *et al.*, 2015). Throughout the rest of this thesis, the term FDCC_{old} will be used to refer to the computation of the FDCC using the token selection stage explained in Chapter 2 and FDCC_{new} will be used to denote the recently developed version which is explained below.

The FDCC_{new} mainly consists of two stages: segmentation of the sound into tokens and computation of the correlation between the power spectra of the extracted successive tokens. For the token segmentation, the intensity envelope of the sound is obtained by squaring and applying a second order Butterworth low-pass filter at 10 Hz which is followed by a peak detection stage where the built-in MATLAB (The MathWorks Inc., Natick, MA) function, *findpeaks* (R2007b), is applied on the extracted intensity envelope. The width of each peak is determined by establishing a reference horizontal line at half of the peak amplitude and measuring the distance between the two points where the descending signal intercepts the reference horizontal line. If the two consecutive peaks are too close to each other that the signal starts ascending before it drops down to the half peak amplitude, the sample point where the ascending begins is chosen as the interception point. As a result, the tokens are selected in a way that the peak locations correspond to the mid-point of the tokens and the durations of the tokens are determined by the width of the corresponding peaks. The sound levels of the extracted tokens are computed and the ones which are 15 $\mathrm{dB_{LAeq}}$ lower than the average sound level are discarded. The choice of 15 dB is derived from the definition of the speech transmission index (STI, see Appendix A for details), as it is the lower limit of the signal-to-noise ratio defined by this metric (Steeneken and Houtgast, 1980; International Electrotechnical Commission, 2003).

The previous version, $\mathrm{FDCC_{old}}$, (see Sec. 2.2.1) differs in the stage after extracting the intensity envelope of the signal: First the median of the envelope was computed and the segments with envelope values above the median were accepted as feasible tokens. Second, the durations of the feasible tokens were computed and the median interval duration was obtained. Tokens which were shorter than the median duration were discarded.

The motivation behind the change was to increase the sensitivity of the token selection procedure of the algorithm in order to capture a more detailed image of the stimulus regarding fluctuations in the time domain, with the expectation that a finer degree of spectral variation can be quantified. The token selection stage of the algorithm was evaluated manually by monitoring its ability to detect at least the syllables in 13 short speech samples (3-5 s each) in Dutch (Plomp and Mimpen, 1979, list 1) since sequences made up of one syllable words and vowels were shown to produce an ISE (e.g., Jones *et al.*, 1999; Schlittmeier *et al.*, 2012; Dorsi *et al.*, 2018). The role of vowels and consonants in a one-word syllable (consonant-vowel-consonant (CVC)) on the ISE has been investigated in the literature and it was shown that changing the vowel was most effective in terms of revoking the ISE when compared to changing the initial and the last consonants of a CVC syllable (Hughes *et al.*, 2005). However, in the present study we deliberately refrained from further fine-tuning the algorithm for detecting only the vowels because the metric is expected to account for both speech and non-speech sounds and it might lead to an overfitting of the algorithm to the set of stimulus used for evaluation. The short speech samples used for the evaluation were the same samples that were concatenated to form one long speech sample of 50 s, and had been used as one of the 10 irrelevant speech stimuli in Chapter 2.

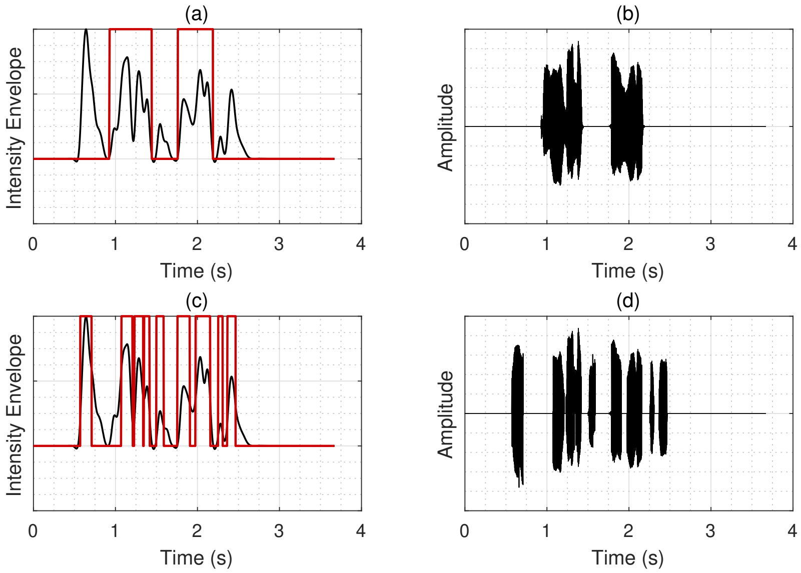


Figure 3.1: Plots on the left side of the figure show the intensity envelopes of the same speech sample where the red lines represent the boundaries of the tokens determined by the token selection stages of the FDCC_{old} (a) and the FDCC_{new} (c). Plots on the right side present the tokens extracted using the FDCC_{old} (b) and the FDCC_{new} (d).

Three of those short speech samples, taken from Plomp and Mimpen (1979), are used to demonstrate the difference of the two segmentation procedures. The first short speech sample consists of nine syllables; *mor* - *gen* - *wil* - *ik* - *maar* - *een* - *li* - *ter* - *melk*. The token extraction stage of FDCC_{old}, used in section 2.2.1, results in two tokens, *gen* - *wil* - *ik* and *een* - *li*, where several syllables are grouped into one token. When the same speech sample is segmented into tokens by the new token selection procedure, it can be observed that there are nine short tokens which correspond to the position of the nine syllables. The token boundaries and the extracted tokens using both the FDCC_{old} and the FDCC_{new} are presented in Fig. 3.1.

The second example consists of eight syllables: *de* - *ze* - *kerk* - *moet*

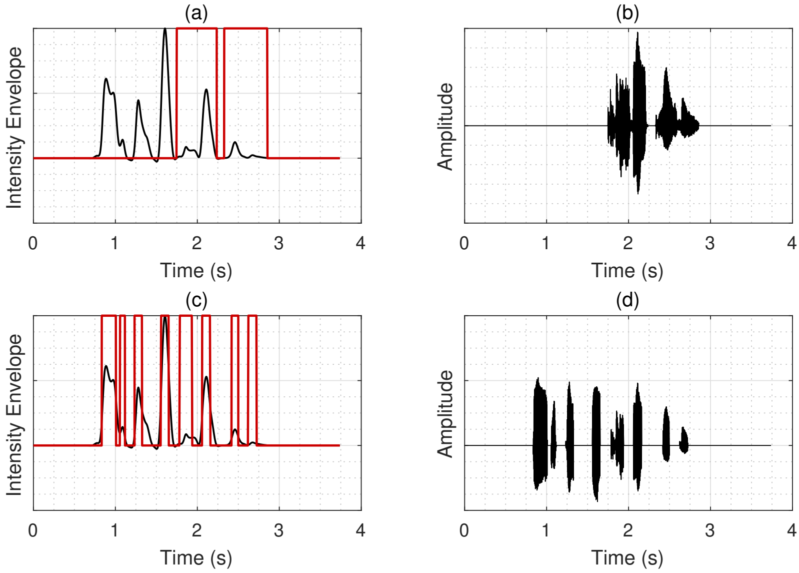


Figure 3.2: The intensity envelopes of the same speech sample are presented in the plots on the left side of the figure where the red lines represent the boundaries of the tokens determined by the token selection stages of the FDCC$_{\text{old}}$ (a) and the FDCC$_{\text{new}}$ (c). Plots positioned on the right side of the figure show the tokens extracted using the FDCC$_{\text{old}}$ (b) and the FDCC$_{\text{new}}$ (d).

- *ge* - *sloopt* - *wor* - *den*. The FDCC$_{\text{old}}$ results in two tokens, *ge* - *sloopt* and *wor* - *den*, while the FDCC$_{\text{new}}$ results in eight tokens which correspond to the positions of the syllables. The token boundaries and the selected tokens using the two versions are presented in Fig. 3.2.

The last example demonstrates that the tokens extracted using the FDCC$_{\text{new}}$ can contain a single consonant instead of a syllable as well. The third speech sample consists of eight syllables: *de* - *nieu* - *we* - *fiets* - *is* - *ge* - *sto* - *len*. The token boundaries and the selected tokens determined using FDCC$_{\text{old}}$ and FDCC$_{\text{new}}$ are presented in Fig. 3.3. The FDCC$_{\text{old}}$ results in two tokens, *de* - *nieu* - *we* and *is* - *ge*, while the FDCC$_{\text{new}}$ produces nine tokens: *de* - *nieu* - *we* - *fiet* - *s* - *is* - *ge* - *sto* - *len*. It can be seen that the FDCC$_{\text{new}}$ divides the speech sample into nine tokens which is more than the syllables it consists of, by picking up the consonant *s*. On the other hand, the FDCC$_{\text{old}}$ can not detect the word *fiets* and produces two tokens which contain several syllables.

When the two methods are compared, it can be seen that the peak detection stage yields a more detailed representation of the short speech

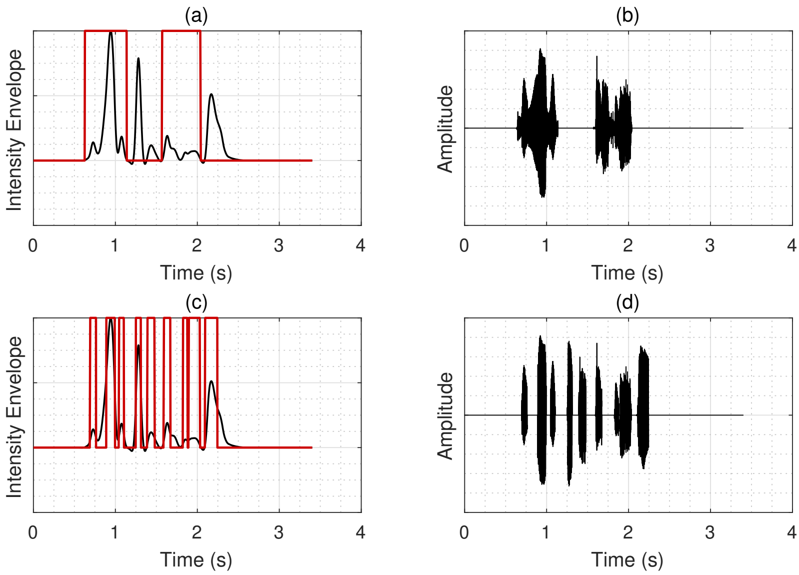


Figure 3.3: The intensity envelopes of the same speech sample are shown in the plots positioned on the left side of the figure where the red lines denote the boundaries of the tokens determined by the token selection stages of the $\mathrm{FDCC_{old}}$ (a) and the $\mathrm{FDCC_{new}}$ (c). Plots on the right side of the figure present the tokens extracted using the $\mathrm{FDCC_{old}}$ (b) and the $\mathrm{FDCC_{new}}$ (d).

stimulus used in the examples above: The 13 short speech samples (50 s) used in developing the new token selection stage are analyzed and it was found that the $\mathrm{FDCC_{old}}$ results in 34 tokens (0.68 tokens / second) while the $\mathrm{FDCC_{new}}$ results in 130 (2.6 tokens / second). The impact of the two approaches on the predicted FDCC values will be examined in this study.

The rest of the computation is the same for both versions: Each extracted token is filtered through one-third octave band filters with center frequencies between 125 Hz and 8000 Hz and the power, P, is calculated for each band of each token. The FDCC is formulated as follows:

$$FDCC_i = \frac{\sum\limits_{j=1}^{19} P_{i,j}P_{i+1,j}}{\sqrt{\left(\sum\limits_{j=1}^{19} P_{i,j}^2\right)\left(\sum\limits_{j=1}^{19} P_{i+1,j}^2\right)}} \tag{3.1}$$

where $P_{i,j}$ indicates the power in token i and frequency band j. Finally, the FDCC values are averaged across the extracted pairs of tokens.

The estimator can highlight changes in the frequency domain where a high correlation value indicates more similarity and therefore higher serial-recall performance. The FDCC is inversely related to the changing state hypothesis and therefore the parameter values are presented as spectral distinctiveness values, 1-FDCC$_{\mathrm{new}}$ and 1-FDCC$_{\mathrm{old}}$, when indicated.

3.3 Experimental data and stimuli

Four studies from the ISE literature were chosen based on four criteria: The cognitive task used in the experiments should be a serial-recall task, the FS values should be (made) available, the final set of stimuli collection should include a rich diversity in terms of types of sounds and the irrelevant sound stimuli should be provided in order to compute FDCC values.

The first set of stimuli is taken from the study of Park *et al.* (2013). The study investigated the impact of masked speech sounds on serial-recall performance where an adaptive masking scheme was proposed. There were five sound conditions alongside silence; stationary white noise; continuous unmasked speech; speech with a low level stationary noise masker; speech with a high level stationary noise masker; and speech with an adaptive noise masker. The experiment was conducted in two parts and there were 20 subjects in the first part of the experiment while only 11 of them participated in the second part. The first part of the experiment was divided into three sessions and the sessions were completed in three consecutive days. The second part of the experiment was conducted one month after the last day of the first part of the experiment and was completed in one day. The authors of the study have taken the time schedule of the experiment into account while computing the normalized error rates for each sound condition: The last experimental block of each day was the silence block and the reported normalized error rates were computed by subtracting, for each participant, the error rate obtained in the last block of each day (silence condition) from those in the other test conditions on the same day (for a detailed explanation of the potential bias regarding the normalization, please see Appendix A in Park *et al.* (2013)).

The background sounds were played back via headphones (diotically) at a sound pressure level of 55-65 dB(A) and there were 20 trials for each sound condition. The parameter values of the STI and the FDCC for each sound condition were reported in the discussion section of the study.

The largest set of sounds is obtained from the study of Schlittmeier *et al.* (2012). It included 44 sound conditions which consist of native, foreign and babble speech as well as office noise, traffic noise, music, tone sequences, and animal sounds. These sounds were used in 70 behavioral measurements where 18 to 36 subjects had to perform digit-recall tasks for three to seven of the different sound conditions. All background sounds were presented via headphones (diotically) or loudspeakers at a sound pressure level of 35-60 dB(A) and there were 15-20 trials for each sound condition.

The experiments were conducted in different settings and some of the auditory stimuli were used in more than one of the 70 experimental measurements. For more detailed information the reader is referred to Table 1 in Schlittmeier *et al.* (2012). The authors stated that they computed median error rates instead of mean error rates because of the asymmetric distribution of the data. The study reported the predicted error rates based on the FS values which were computed using the software PAK (Müller-BBM VibroAkustik Systeme GmbH).

The third set of stimuli is obtained from the study of Liebl *et al.* (2016) who had seven sound conditions in the serial-recall task next to the silence condition: continuous speech-like noise; masked speech sounds with different signal-to-noise ratios, 0, -3 and -6 dB; unmasked native speech; variable speech-like noise; and pink noise. The continuous speech-like noise had spectral characteristics similar to the spectral shape of male speech without the temporal structure. The masked speech sounds were generated by using the continuous speech-like noise as a masker at different sound pressure levels. The unmasked native speech was unmodified running speech in German and variable speech-like noise is similar to the continuous speech-like noise as they both preserve the spectral characteristics of natural speech, while the variable speech-like noise also preserves the temporal structure. In addition to these, pink noise was also included in the experiment as it has been commonly employed in research focusing on the impact of sound conditions on performance (e.g., Ellermeier and Hellbrück, 1998).

There were, in total, 24 participants who participated in the digit-recall experiment where each had to perform 12 trials in each of the eight sound conditions. The irrelevant sound conditions were all presented binaurally via headphones at a sound pressure level of 55 dB(A). The FS values were computed using the software ArtemiS (HEAD acoustics GmbH, Herzogenrath, Germany).

The last set of stimuli analyzed in this study are the noise-vocoded speech (NVS) stimuli employed in Chapter 2 of this thesis. NVS is a manipulation of continuous speech where the speech stimulus is filtered into frequency bands and the intensity envelope of each frequency band is mapped to band-limited white noise. NVS stimuli were generated by dividing the speech signal between 50 and 8000 Hz into a different number of Hanning-shaped bandpass filtered frequency bands. This technique enabled the creation of a set of stimuli where the non-disruptive amplitude modulated white noise was transformed into a disruptive intelligible NVS stimulus by increasing the number of frequency bands. The organized modification of the spectrum allowed the authors to evaluate the FDCC, which was the major objective of the study.

Nine sound conditions, 1-, 2-, 4-, 6-, 9-, 12-, 15-, 18-band NVS and speech were employed in two digit-recall experiments alongside silence. There were 15 subjects in the first experiment and 25 participants were enrolled in the second experiment. Each sound condition was played back 10 times in one experimental session for each participant. The acoustic conditions were presented via headphones (diotically) at a sound pressure level of 60 dB(A). The ArtemiS software (HEAD acoustics GmbH, Herzogenrath, Germany) was used to compute the FS values of the NVS stimuli.

3.4 Results

The FDCC$_{\text{new}}$ and the FDCC$_{\text{old}}$ are computed for every sound condition in the four sets of stimuli. The FS values are used as reported in the three original studies and calculated using Artemis for the stimuli obtained from the study of Park *et al.* (2013), since they were not reported in the paper. The Pearson correlation between the normalized error rates and values of the both versions of the spectral distinctiveness metric, 1-FDCC$_{\text{new}}$ and 1-FDCC$_{\text{old}}$ as well as the FS are computed. The significance of the difference between the two correlation coefficients, for 1-FDCC$_{\text{new}}$ and FS, obtained for the 91 data points is computed based on Fisher's r to z transformation (Fisher, 1921). The spectral distinctive metric values presented in the figures are the 1-FDCC$_{\text{new}}$ values while the 1-FDCC$_{\mathrm{old}}$ values are only reported in the text.

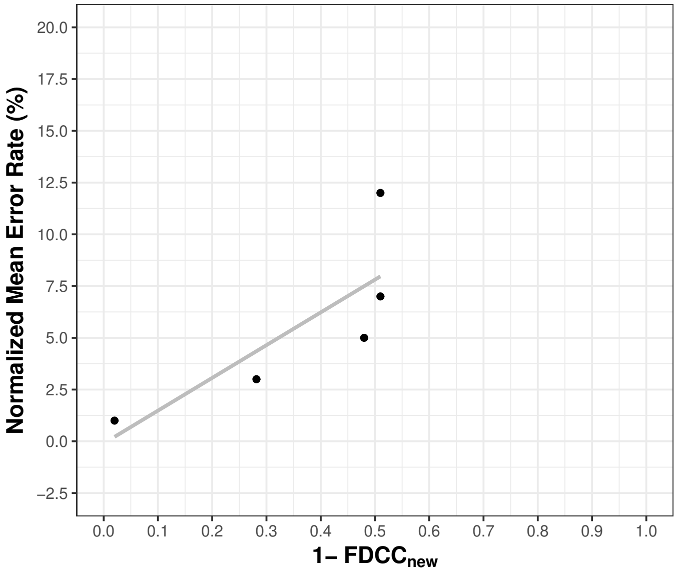


Figure 3.4: The normalized mean error rates as a function of the spectral distinctiveness values, 1-FDCC$_{\mathrm{new}}$, derived from the study of Park *et al.* (2013). Each point represents a serial-recall measurement for a certain sound condition.

The normalized mean error rates of the five experimental measurements obtained from the study of Park *et al.* (2013) as a function of the spectral distinctiveness metric are presented in Fig. 3.4. A Pearson correlation coefficient is computed in order to investigate the relationship between the normalized error rates derived from the experiments and the values of the spectral distinctiveness metric. A statistically insignificant correlation is observed between the two variables, r = 0.80 ($p > 0.05$), and for the 1-FDCC$_{\mathrm{old}}$, the Pearson's r is 0.87 ($p > 0.05$). The resulting correlation values are lower than the Pearson's r between the normalized error rates and the FS values for the same set of stimuli, r = 0.92 ($p < 0.05$).

The normalized median error rates of the 70 experimental measurements obtained from the study of Schlittmeier *et al.* (2012) as a function of the spectral distinctiveness metric are presented in Fig. 3.5. A Pearson correlation coefficient is computed in order to investigate the relationship between the normalized error rates derived from the experiments and the values of the spectral distinctiveness metric. A highly significant positive correlation is observed between the two variables, r = 0.76

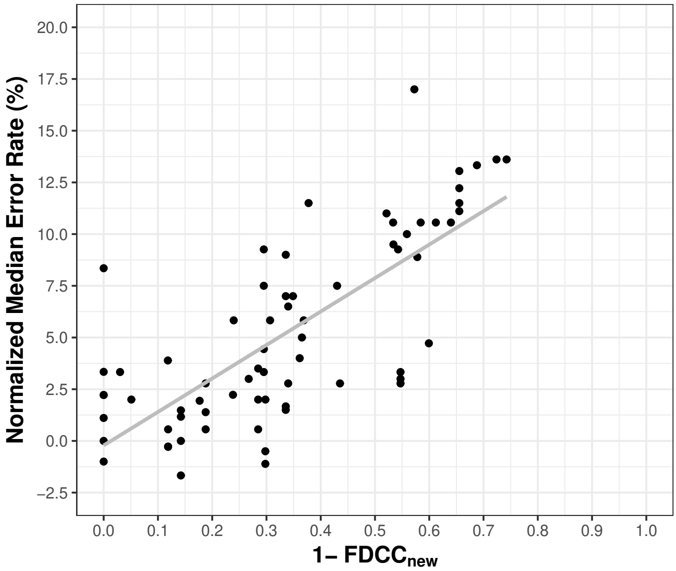


Figure 3.5: The normalized median error rates as a function of the spectral distinctiveness values, 1-FDCC$_{new}$, derived from the study of Schlittmeier *et al.* (2012). Each point represents a serial-recall measurement for a certain sound condition.

($p < 0.01$). The correlation value computed for 1-FDCC$_{old}$ is 0.58 ($p < 0.01$). The correlation value obtained from the spectral distinctiveness metric, 1-FDCC$_{new}$, is slightly higher than the Pearson's r computed for the normalized error rates and the FS values reported for the same sound conditions, r = 0.67 ($p < 0.01$).

The normalized mean error rates corresponding to the seven sound conditions in the study of Liebl *et al.* (2016), and the spectral distinctiveness values of each stimulus are presented in Fig. 3.6. The normalized error rates and the spectral distinctiveness values are highly correlated yielding a Pearson correlation value of r = 0.93 ($p < 0.01$), which is similar to the one computed for 1-FDCC$_{old}$, 0.95. The computed correlation values are higher than the correlation values obtained from the normalized error rates and the FS values, r = 0.31 ($p > 0.05$).

The normalized error rates corresponding to the nine sound conditions obtained from Chapter 2, as a function of the spectral distinctiveness metric are presented in Fig. 3.7. The normalized error rates and the spectral distinctiveness values are moderately correlated producing a Pearson correlation coefficient of, r = 0.68 ($p < 0.05$). The same computation for the 1-FDCC$_{old}$ yields a correlation value of r = 0.70 ($p < 0.05$).

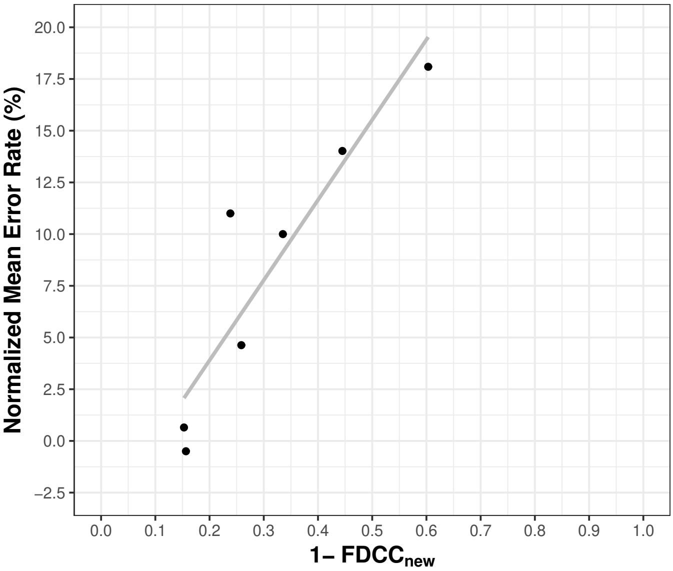


Figure 3.6: The normalized mean error rates as a function of the spectral distinctiveness values, 1-FDCC$_{\mathrm{new}}$, derived from the study of Liebl *et al.* (2016). Each point represents a serial-recall measurement for a certain sound condition.

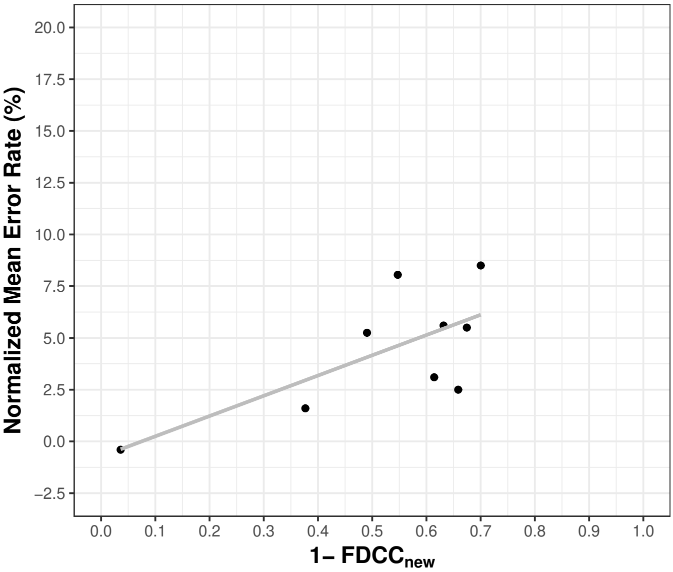


Figure 3.7: The normalized mean error rates as a function of the spectral distinctiveness metric, 1-FDCC$_{\mathrm{new}}$, derived from Chapter 2. Each point represents a serial-recall measurement for a certain sound condition.

The normalized error rates and the FS values have a lower correlation value, r = 0.27 ($p > 0.05$).

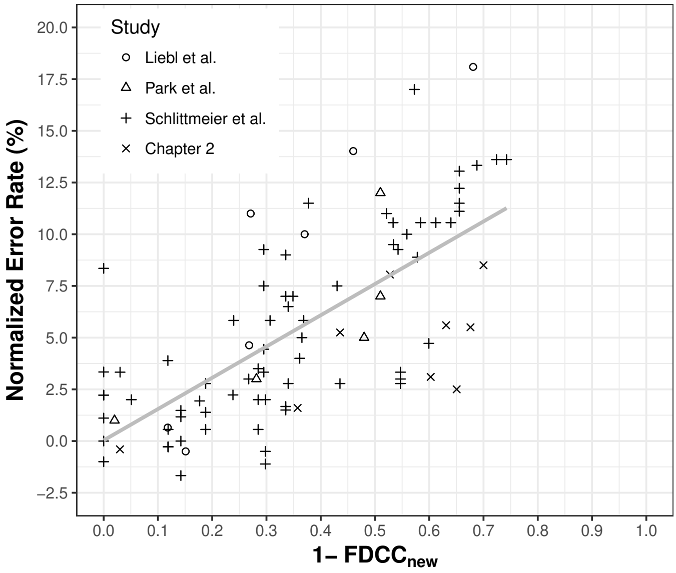


Figure 3.8: The normalized error rates as a function of the spectral distinctiveness values, 1-FDCC$_{\mathrm{new}}$, for 91 behavioral measurements. The four data sets are represented by four different symbols.

The experimental results and the parameter values of the two metrics for the aforementioned four studies are combined to form 91 data points. The normalized error rates corresponding to the 91 sound conditions as a function of the spectral distinctiveness metric are presented in Fig. 3.8 and as a function of the FS model in Fig. 3.9.

The normalized error rates and the spectral distinctiveness metric values are significantly correlated yielding a Pearson correlation value of r = 0.70 ($p < 0.01$). In comparison, Pearson's r computed for 1-FDCC$_{\mathrm{old}}$ has a value of r = 0.59 ($p < 0.01$). The same computation using the FS values result in a lower correlation value of r = 0.50 ($p < 0.01$). The two Pearson correlation values computed for the 1-FDCC$_{\mathrm{new}}$ and FS are significantly different ($p < 0.05$). On the other hand, the r value for the 1-FDCC$_{\mathrm{old}}$ is not significantly different from the correlation values computed for the other two metrics.

An additional analysis was conducted in order to asses the influence of the outliers within the 91 data points for the two metrics and the

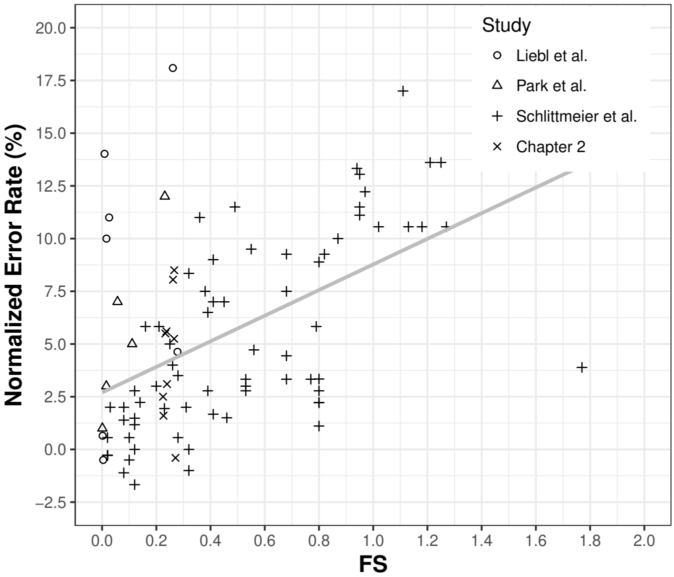


Figure 3.9: The normalized error rates as a function of the fluctuation strength metric for the total of 91 behavioral measurements. The four data sets are represented by four different symbols.

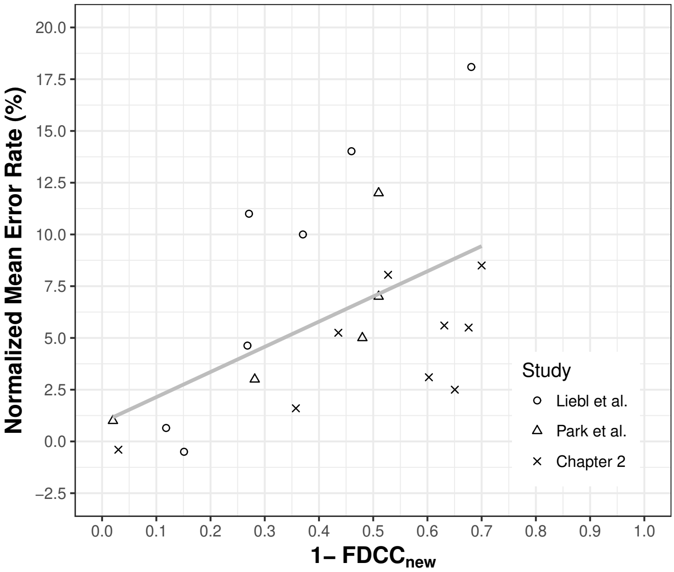


Figure 3.10: The normalized error rates as a function of the spectral distinctiveness values, 1-FDCC$_{\text{new}}$, are presented for the 21 behavioral measurements obtained from the three studies. The three data sets are represented by three different symbols.

normalized error rates by computing the Cook's distance. One of the sounds, the synthesized duck's quacking sequence, which yields the maximum FS value among the 91 sounds (the data point positioned close to the lower-right corner in the Fig. 3.9, FS = 1.77), obtained from the study of Schlittmeier *et al.* (2012), is detected as an influential outlier. The original study reported that there was no significant behavioral effect observed for the sound condition (normalized median error rate = 3.89 %) and the FS value overestimated the serial-recall performance.

When the duck sound and the associated recall score are removed from the data set, the Pearson's r between the normalized error rates and the FS values increased to 0.55 ($p < 0.01$). The two correlation values, computed for the 1-FDCC$_{\text{new}}$ and FS, are not significantly different after the exclusion.

The final comparison for the Pearson correlation values of 1-FDCC$_{\text{new}}$ and the FS is realized by excluding the data obtained from the study of Schlittmeier *et al.* (2012) from the total set of data points, in order to observe the behavior of the two metrics for speech and modified / degraded speech-like sounds such as adaptive / stationary masked speech, variable speech-like noise and NVS.

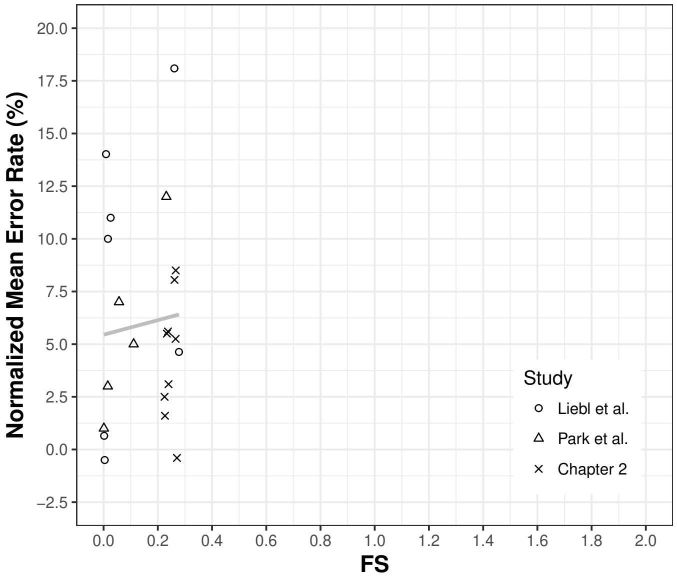


Figure 3.11: The normalized error rates as a function of the fluctuation strength values are displayed for the 21 behavioral measurements. Each symbol represents a different dataset.

The behavioral results and the parameter values of the two metrics derived from the stimuli of the three studies are combined to form 21 data points. The normalized error rates corresponding to the 21 sound conditions as a function of the spectral distinctiveness metric are presented in Fig. 3.10 and as a function of the FS in Fig. 3.11.

The normalized error rates and the values of the spectral distinctiveness metric are significantly correlated producing a Pearson correlation value of r = 0.53 ($p < 0.05$), while the previous version of the metric, 1-FDCC$_{\text{old}}$, yields a higher value or, r = 0.72 ($p < 0.01$). The Pearson correlation between the FS values and normalized error rates reveals a very low, insignificant correlation value of r = 0.08 ($p > 0.05$).

3.5 Conclusion

For the current study, a large set of sounds was collected and analyzed in order to evaluate two parameters, the FDCC and the FS. All sound stimuli were analyzed using the two versions of the FDCC and the FS. The resulting values were investigated in relation to the behavioral measurements. The analysis was conducted for each set of stimuli individually and finally for all the experimental data.

A significant difference between correlation coefficients was observed when the metric values, the FDCC$_{\text{new}}$ and the FS, were compared for the 91 data points. Here it should be noted that the correlation values of the FDCC$_{\text{new}}$ and the FS were not significantly different when the outlier was removed from the data set. Regarding that the FS parameter attempts to predict the ISE for any sound condition, as a novel metric, the outlier can only be an indication of a limitation which has to be analyzed closely rather than removing the data point. The authors verbalized the limitation by stating that the FS parameter failed to predict ISE for artificial sounds in the original study (Schlittmeier *et al.*, 2012).

A further analysis was conducted using the degraded / masked speech and speech-like stimuli from the three studies (Park *et al.* 2013; Liebl *et al.* 2016; Chapter 2). It was observed that the correlations between both versions of the FDCC and the data collected from these three studies were higher than that for the FS. Both versions of the FDCC parameter tend to be more efficient than the FS parameter for such a set of stimuli, which is in line with what was reported in the associated literature (Ellermeier *et al.* 2015; Liebl *et al.* 2016; Chapter 2). However, the change in the token segmentation stage resulted in a decrease in the correlation value between the FDCC parameter and behavioural measurements collected under degraded / masked speech stimuli.

It should be pointed out that the sound set collected from the study of Schlittmeier *et al.* (2012) contains sounds which are more likely to be heard in daily life, such as music, office and traffic noise when compared with noise-vocoded and masked speech employed in the other three studies. Despite the overestimation of the recall performance for the duck sound, which is unlikely to be heard in daily life in its synthesized version, the FS parameter reached a similar level of performance as the $\mathrm{FDCC}_{\mathrm{new}}$ parameter for this set of stimuli.

The results reported for the total set of stimuli show that the two investigated metrics are similar in terms of prediction accuracy with the $\mathrm{FDCC}_{\mathrm{new}}$ generating a slightly higher correlation value, which favors the role of spectral variation within the context of serial-recall disruption. It can also be observed from the results that the integration of the peak detection stage to the spectral parameter increased the prediction accuracy of the algorithm. It should be noted that there is an obvious limitation coming from the small number of data points which does not allow us to derive a concrete conclusion regarding the performance of the descriptors. Nevertheless, the $\mathrm{FDCC}_{\mathrm{new}}$ parameter, in its current shape, performs at least equally well as an ISE predictor when compared to the FS parameter.

4 | Evaluation of the token selection stage of the $\mathrm{FDCC_{new}}$

Abstract

The token selection stage of the $\mathrm{FDCC_{new}}$ introduced in Chapter 3 was evaluated in two studies. The first study employed a set of stimuli which was generated by segmenting the continuous versions of the stimuli into tokens using the token selection stage of the $\mathrm{FDCC_{new}}$ and adding low-level adaptive noise in order to prevent onset-offset artifacts. The set of segmented irrelevant stimuli was used in a serial-recall task and the results were compared with the serial-recall results observed using the continuous versions of the same sounds. It was observed that the $\mathrm{FDCC_{new}}$ was sensitive enough to capture disruptive properties of the irrelevant speech, but not of 6-band noise-vocoded speech (NVS). In the second study, a set of modified NVS stimuli from the literature was obtained, and the $\mathrm{FDCC_{new}}$ values for the stimuli were computed. The results showed that $\mathrm{FDCC_{new}}$ was able to predict the difference in serial-recall results between the unmodified and modified NVS stimuli.

4.1 Introduction

The impact of spectral variation and the use of spectral features as a basis for an ISE prediction model were investigated in the previous two chapters. In Chapter 2, a set of special stimuli, noise-vocoded speech (NVS), was generated and employed in a serial-recall task. The noise-vocoding technique allowed us to create a set of stimuli where the spectral features of the irrelevant sounds were systematically modified by increasing the number of frequency bands employed in the NVS stimuli. This was particularly important for the experiment since the obtained spectral variation was reflected in a frequency domain metric, the frequency domain correlation coefficient (FDCC$_{\text{old}}$), and the relation between the serial-recall results and the FDCC$_{\text{old}}$ values was the major objective of the study. The serial-recall performance showed almost a linear decrease as a function of the increase in the number of frequency bands up to a critical point, but not beyond that. The experiments demonstrated that spectral features of the irrelevant sounds are relevant in terms of generating an ISE, however the spectral variation and the test performance were not linearly related.

In Chapter 3, the role of spectral variation on the ISE and the performance of the FDCC$_{\text{new}}$ as a prediction model were further investigated by employing a large set of stimuli (N = 91) from the literature and comparing the metric values with the serial-recall results reported in the studies. Ninety-one sounds, obtained from four different studies, were analyzed by the FDCC algorithm, using both the initial (FDCC$_{\text{old}}$) and the improved versions (FDCC$_{\text{new}}$), and the resulting metric values were compared with the experimental scores. It was shown that the recent change in token selection stage increased the Pearson correlation value between the spectral metric and the serial-recall results by 0.11, from 0.59 (FDCC$_{\text{old}}$) to 0.70 (FDCC$_{\text{new}}$) for the 91 data points.

The change in the token selection stage was realized by applying a peak detection stage and the token selection accuracy was analyzed by manually monitoring the syllable extraction ability of the metric using 13 short Dutch speech samples (3-5 s) taken from the study of Plomp and Mimpen (1979). The result was demonstrated in Sec. 3.2 by presenting the selected tokens of three of the 13 short speech samples.

The present chapter further investigates the newly developed token selection stage of the algorithm in two experiments. The first experiment employed a serial-recall task where the irrelevant sounds were segmented into tokens by the token selection stage of the FDCC$_{\text{new}}$ and used in the serial-recall experiment by their segmented versions only. The expectation is that, if the token selection stage of the algorithm is efficient in preserving the disruptive properties of the stimuli, then the serial-recall results should be similar to those of the unsegmented, original stimuli. The sound stimuli were chosen from sounds used in Chapter 2 and in the ISE literature (Schlittmeier *et al.*, 2012) and the choice was motivated by the observed correlation between the reported recall-performance and the computed FDCC$_{\text{new}}$ values for the continuous versions.

The second experiment evaluated if the token selection stage of the FDCC$_{\text{new}}$ is accurate enough to detect the disruptive parts of the irrelevant stimuli by focusing on a set of specially crafted NVS stimuli from the literature (Dorsi *et al.*, 2018): The NVS stimuli were modified by temporally reversing the information within two-thirds of the frequency bands with the aim of altering speech fidelity of the NVS. The original work had attempted to investigate the impact of speech fidelity on the ISE by comparing the serial-recall results of the selectively reversed NVS with the unmodified NVS. The FDCC$_{\text{new}}$ values of the two sets of NVS stimuli were computed and the results were compared with the serial-recall results reported in the original study.

The computation of the FDCC$_{\text{new}}$ in the present chapter is identical with what is reported in Sec. 3.2. The first experiment, including the method, the stimuli, results and the discussion, is presented in Sec. 4.2. The second experiment is described in Sec. 4.3 and the study is finalized by the general discussion and the conclusion, in sections 4.4 and 4.5, respectively.

4.2 Experiment 1

In the previous chapter, the token selection stage of the spectral metric had been modified with a focus on speech stimuli and it was observed that the integration of the peak detection stage increased the ISE prediction performance, when compared with the initial version. The token selection stage was modified by manually monitoring its ability in terms of extracting syllables, and the resulting FDCC$_{\text{new}}$ values correlated well with serial-recall scores obtained using speech and speech-like stimuli.

It had been demonstrated in the study by Jones *et al.* (1993) that segmentation plays a crucial role in serial-recall disruption. Speech stimuli were regarded as to be inevitably perceived as segmented because of its natural acoustic variations, while for a non-speech stimulus this was not the case: When a pure tone slowly varying in pitch was interrupted by short segments of silence (100 ms), a significant serial-recall disruption was observed but the uninterrupted version of the same stimulus failed to produce an ISE (Jones *et al.*, 1993, Exp. 1). In the second experiment, short segments of silence (100 ms) were used to replace the first 100 ms of every vowel resulting in an alternating sequence of silence and vowel. The serial-recall disruption observed for the interrupted vowels was not significantly different from the uninterrupted vowel condition. The authors stated that speech and words possess clear cues for segmentation, while for some non-speech sounds, those cues may not be prominent and this might give an explanation regarding the inconsistent effects observed for different kinds of music, such as music with a slow tempo or pitch variations (Schlittmeier *et al.*, 2008).

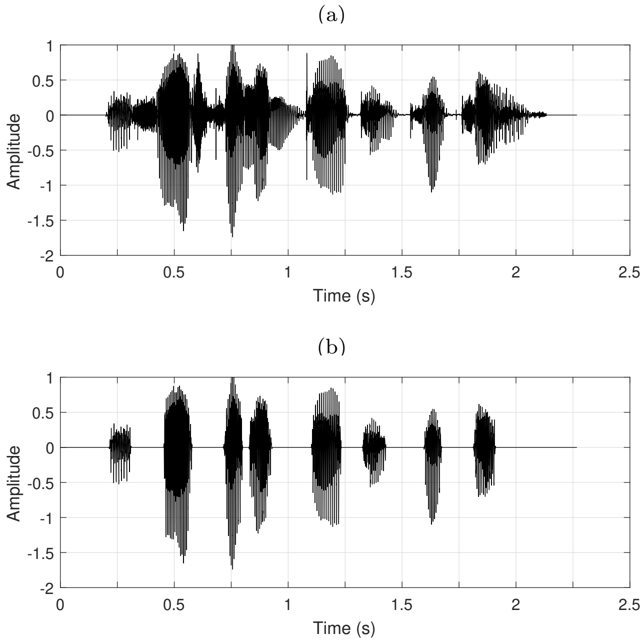
In the second experiment mentioned above, the interruption of the vowel sequence by segments of silence did not change the ISE. If the vowels in the sequences would be successfully detected by the token selection stage of the FDCC$_{\text{new}}$ then the FDCC$_{\text{new}}$ values of the two conditions, interrupted and uninterrupted vowel sequences, would be the same. In the same experiment, the locations of the interruptions were periodic (first 100 ms of each 500 ms vowel) and the vowel-to-vowel structure of the original sequence was preserved. For the present study, we replace the periodically positioned vowels with continuous speech and the locations of the interruptions are determined by the token selection stage of the FDCC$_{\text{new}}$: If the token selection stage of the algorithm is capable of preserving the disruptive parts of the temporally complex speech stimuli then the segmented and the continuous speech should produce a similar degree of disruption.

On the other hand, the results of the first experiment reported in the study of Jones *et al.* (1993) showed that the segmented and continuous versions of the same non-speech stimulus might result in different levels of ISE depending on the segmentation cues it provides. Therefore the behaviour of the token selection stage of the algorithm for non-speech and distorted speech stimuli is also investigated. Music with legato passages is chosen as the irrelevant non-speech stimulus in order to observe the influence of the token segmentation process on a sound with slow amplitude and pitch variations. The degraded speech stimulus is the 6-band NVS which was shown to be almost as intelligible as speech, (Davis *et al.*, 2005) and the 1-band NVS stimulus serves as steady-state speech that can be segmented. The last stimulus is office noise which consists of various types of speech and non-speech sounds. Irrelevant stimuli used in the literature were regenerated or obtained from the authors (Schlittmeier *et al.* 2012; Chapter 2), and the serial-recall results of the experiment using segmented stimuli were compared with the results reported in the literature for the corresponding unsegmented stimuli.

The stimulus segmentation process is explained in detail in the following section. The acoustic conditions employed in the serial-recall task with the FDCC$_{\text{new}}$ values of the segmented and the continuous versions of each condition, as well as the mean error rates of the continuous stimuli reported in the literature are presented in Sec. 4.2.2.

4.2.1 Segmentation process

A segmented irrelevant stimulus was generated by dividing the continuous irrelevant stimulus into tokens based on the information derived from the FDCC$_{\text{new}}$ algorithm. The individual tokens were concatenated, while preserving their original order and position in time, in order to form a segmented stimulus. A short speech sample and the segmented version of the same sample are presented in Fig. 4.1a and 4.1b, respectively.


(a)

(b)

Figure 4.1: Continuous (a) and segmented (b) versions of a short speech sample in Dutch.

A set of irrelevant continuous and segmented speech stimuli was employed in a pilot serial-recall task and it was found that the segmented stimuli produced significantly higher mean error rates than the continuous speech. This was due to the onset-offset artifacts (e.g., clicks) introduced by the segmentation process, which are not part of the continuous speech stimuli.

In order to avoid audible artifacts, an adaptive low-level noise was introduced. First, the long term average frequency spectrum of each continuous sample was extracted and applied to white noise with the same duration of the corresponding continuous stimulus. The continuous stimulus was segmented into tokens based on the information derived from the FDCC$_{\text{new}}$ and the root-mean-square (RMS) of each segmented token was computed. The level of the parts of the noise which correspond to the positions of the segmented tokens in the continuous stimulus was adjusted to be 15 dB lower than that of the corresponding segmented tokens. The change in the amplitude between the adjacent noise tokens were shaped by linear ramps where the slope of each ramp was determined by the time gap and the amplitude differences between the adjacent tokens. Finally, 10 ms Hanning onset and offset ramps were introduced to each token of the segmented stimulus and the segmented stimulus was summed with the low-level adaptive noise. An example of the segmented speech and the resulting noise-added segmented speech is presented in Fig. 4.2a and 4.2b, respectively.

A similar approach had been followed in the study of Jones *et al.* (1993, Exp. 3): The non-disruptive tone used in the first experiment in the same study, a pure tone gradually changing in frequency, was interrupted with white noise low-pass filtered at 4 kHz (spectrally similar noise) and high-pass filtered at 4 kHz (spectrally dissimilar noise) instead of silence. The tone interrupted with spectrally similar noise and the continuous, uninterrupted tone condition produced a similar degree of disruption, while the tone interrupted with spectrally dissimilar noise created a significantly higher mean error rate than the other two conditions.

The authors stated that filling the silences with the segments of white noise which covered the frequency range of the varying pitch of the original signal preserved the perceived continuity while the bursts of high-pass filtered noise enhanced the segmentation process and increased the error rate. Therefore, the aforementioned method used to prevent the onset-offset artifacts is expected not to affect the serial-recall scores in

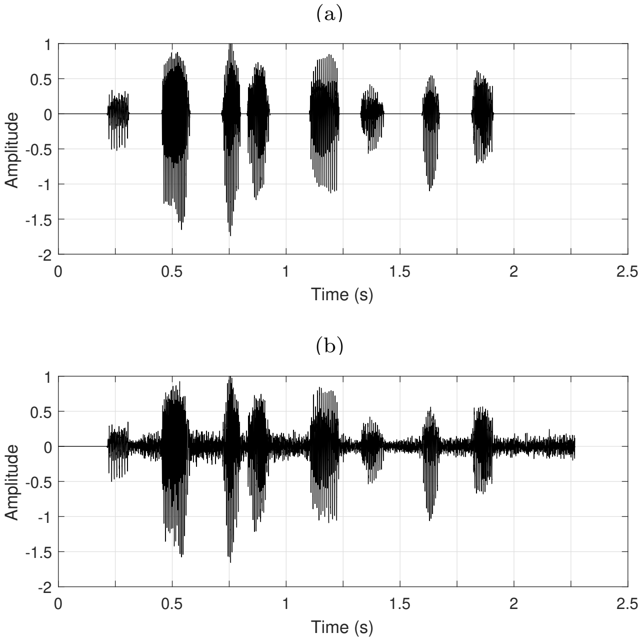


Figure 4.2: The top panel (a) shows a segmented version of the short speech sample before the low-level noise was added and the bottom panel (b) presents the same segmented short speech sample after the addition of the low-level adaptive noise.

the present experiment.

4.2.2 Method

Participants

Twenty-three participants (10 females and 13 males, age range between 18-50 years) participated in the serial-recall task who were recruited via the JF Schouten subject database of the Eindhoven University of Technology, Eindhoven, The Netherlands. All participants were university students and native Dutch speaker. As part of the recruitment procedure, subjects were chosen by specifying the necessary criteria of healthy vision and hearing as well as no history of memory related disorder. The eligibility criteria were double checked by the experimenter and participants who satisfied the criteria signed the informed consent forms before the experiment began. They were paid a small compensation fee determined by the university department board. The experimental procedure was examined and approved by the Human Technology Interaction department, Eindhoven University of Technology and the Internal Committee Biomedical Experiments (ICBE) of Philips Research.

Stimuli

The segmentation procedure described in Sec. 4.2.1 was applied to five sets of stimuli: 1-band NVS, music, 6-band NVS, office noise and speech. The four segmented conditions, 1-band NVS, music, 6-band NVS, office noise, are between-subject variables and the serial-recall scores of the segmented versions obtained in the present experiment are compared with the continuous versions reported in the original studies.

The speech condition is a within-subject variable: Both the segmented and the continuous speech were employed in the serial-recall task in order to compare the normalized error rates of the two conditions directly. Alongside the six sound conditions, a control condition (SLNC) was also employed in the experiment and the average sound level of each stimulus was calibrated to 65 dB$_{\text{LAeq1min}}$.

NVS

The 42 to 55 second long speech samples were generated by concatenating short sentences in Dutch which were obtained from a speech reception study (Plomp and Mimpen, 1979) and from the Dutch matrix test (Houben *et al.*, 2014). The two sets of NVS stimuli were generated by bandpass filtering the speech samples into 1 (50 - 8000 Hz) and 6 bands (50 - 229 Hz, 229 - 558 Hz, 558 - 1160 Hz, 1160 - 2265 Hz, 2265 - 4290 Hz, 4290 - 8000 Hz) using sixth-order Butterworth filters. For the 6-band NVS, the cut-off frequencies were determined by an exponential function developed by Greenwood (1961) as explained in Sec. 2.2.2. The envelopes of each frequency band of the speech signals were extracted by half-wave rectification and low-pass filtering (second-order Butterworth filter) with a 50-Hz cutoff frequency. The resulting envelopes were applied to modulate band-limited white noise, which was band-pass filtered with the same Butterworth filters and the noise-modulated envelopes of each band were finally combined.

Ten long (42-55 s) 1-band NVS and 10 long (42-55 s) 6-band NVS stimuli were generated using 20 different long speech samples. The 20 NVS stimuli were segmented into tokens and summed with adaptive low-level noise as explained in the previous section. The segmented 1-band NVS yielded an FDCC$_{\text{new}}$ value of 0.99 and the FDCC$_{\text{new}}$ revealed 1257 tokens for 10 1-band NVS stimuli (2.6 tokens per second). The FDCC$_{\text{new}}$ value for the segmented 6-band NVS is 0.55 and for the 6-band NVS, there were 1827 tokens detected (3.8 tokens per second) in total.

***Music* (M)**

A 16 min long legato music passage, one of the irrelevant sound conditions used in the study of Schlittmeier *et al.* (2012, stimulus Nr. 41 in Fig. 1 in the original study), was divided into 1 min long music samples. Ten 1 min long music samples were chosen randomly out of the 16 and were segmented into tokens in order to use in the serial-recall experiment. The segmented M yielded an FDCC$_{\text{new}}$ value of 0.78 and the total number of tokens for 10 long music samples was 1839 (3 tokens per second).

***Office noise* (ON)**

A 15 min long office noise sound was provided by the authors of the same study as the music sample (Schlittmeier *et al.*, 2012, stimulus Nr. 25 in Fig. 1 in the original study). The office noise sample contained various types of sounds, such as phone ringing, conversations by different speakers, classical music, as well as low-level machinery sounds, sounds of keyboard typing and footsteps. The long office noise sample was divided into 1 min long versions and 10 of the 1 min long office noise samples were segmented into tokens as explained in the previous section. The segmented ON yielded an FDCC$_{\text{new}}$ value of 0.61 and the total number of the tokens detected using the FDCC$_{\text{new}}$ for 10 long ON stimuli were 999 (1.6 tokens per second).

***Segmented speech* (SS)**

The segmented speech was generated by segmenting the 42 to 55 second long speech samples which were generated by concatenating 2 to 4 second long, female and male spoken sentences in Dutch (Houben *et al.*, 2014). Ten long SS stimuli, five female and five male spoken, formed the SS condition in the serial-recall task. The FDCC$_{\text{new}}$ value of the SS condition is 0.40 and there were in total 1840 tokens (3.8 tokens per second) detected for 10 SS stimuli.

***Continuous speech* (SPCH)**

Female and male spoken short sentences in Dutch (2-4 s) were concatenated to form ten long (42-55 s) speech samples. Ten long speech samples yielded an FDCC$_{\text{new}}$ value of 0.38 and were used in their original continuous form in the serial-recall experiment. The SS and SPCH conditions were generated by using different speech samples and the conditions are expected to yield similar mean error rates in the serial-recall experiment. The number of tokens detected using the FDCC$_{\text{new}}$ for 10 long SPCH stimuli were 1759 (3.5 tokens per second).

The $FDCC_{new}$ values of the experimental sound conditions

The FDCC_{new} values of the six sets of stimuli with their continuous and segmented versions are presented in Fig. 4.3.

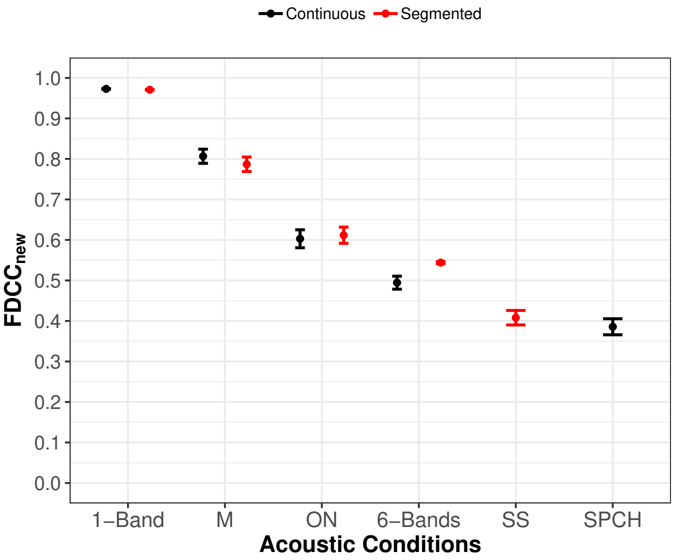


Figure 4.3: The FDCC_{new} values of the continuous and segmented versions of the 1-band noise-vocoded speech (1-band), music (M), office noise (ON) and 6-band noise-vocoded speech (6-bands) stimuli. The parameter values of the segmented (SS) and continuous speech (SPCH) stimuli are also presented. The error bars represent the standard error of the mean (SEM).

It can be observed that the segmentation process did not change the FDCC_{new} values of the stimuli by more than 0.05, which is the difference observed between the continuous and segmented versions of the 6-band NVS stimuli. In addition to this, the FDCC_{new} values show a systematic change as a function of the acoustic conditions and this is expected to be reflected in the serial-recall results.

Apparatus

The apparatus used for the current experiment was the same as that reported in Sec. 2.4.1.

Procedure

The GUI used for the serial-recall and the experimental procedure regarding the visual part of the experiment was the same as reported in previous experiments. The serial-recall task design consisted of six blocks.

The first block was the training block and consisted of 14 trials without any irrelevant sound (silence). After the training block, the experimenter controlled if the participant comprehended the focal task and if he/she had any questions or problems about the test procedure or the environment. Each of the remaining five block also consisted of 14 trials, where two trials of each condition were presented in randomized order in each block. Six blocks were completed in approximately 65-70 min, including 2 min breaks between each block.

4.2.3 Results

The experimental results in seven acoustic conditions, represented as error rates (%), are shown in Fig. 4.4.

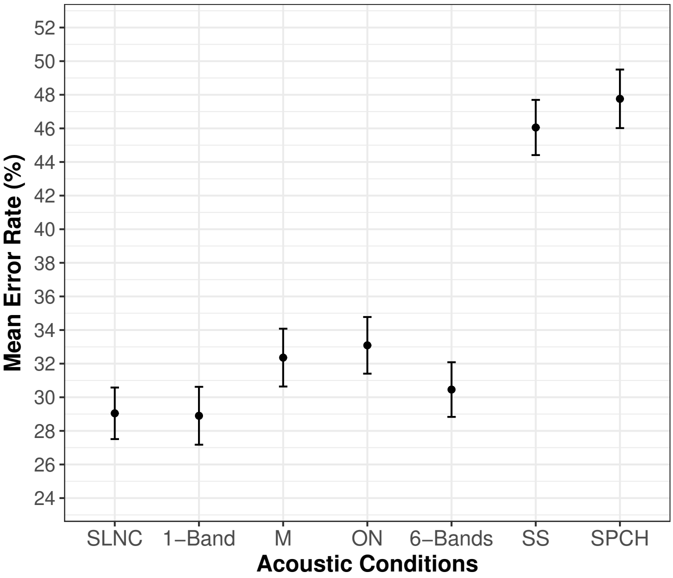


Figure 4.4: Recall performance of 23 participants are represented as mean error rates (%) for seven acoustic conditions: silence (SLNC), 1-band noise-vocoded speech (1-band), music (M), office noise (ON), 6-band noise-vocoded speech (6-bands), segmented (SS) and continuous speech (SPCH). Error bars represent the SEM.

The mean error rate observed in the SLNC condition (29.05 %) is in line with what is reported in the literature. The mean error rate observed in the SPCH condition (47.75 %), resulting in a normalized error rate of 18.5 %, is at the upper end of the range reported in the literature for speech stimuli (8 - 20 %).

A one way repeated measures ANOVA determined that there was a highly significant impact of background sound on serial-recall, $F(6, 132)$

= 15.89, $p < .001$, $\eta^2 = 0.08$. Post hoc analyses were conducted due to the ANOVA result: Each sound condition was compared with the SLNC condition using the Bonferroni correction ($p = 0.008$ for six pairs). The analysis revealed that the mean error rates for SPCH (M = 47.75, SD = 26.29) and SS (M = 46.05, SD = 24.79) were significantly higher than that of the SLNC ($p < 0.001$). The SS and SPCH were also compared by a pairwise t-test and there was no significant difference observed ($p > 0.05$).

The data were further investigated by comparing the results with the literature. Before doing so, the test scores were normalized: For each participant, the mean of the error rates was computed over all trials in each condition. The mean SLNC condition performance of each participant was subtracted from the mean scores of each condition: A normalized error rate per condition for each participant was obtained. The mean and the median of these participant-based normalized error rates were computed. Here it should be noted that the representation of the serial-recall results as normalized mean error rates was applied throughout the thesis, which is also the common practice in ISE studies. However, the continuous versions of the stimuli M and ON were obtained from the study of Schlittmeier *et al.* (2012), where the test scores were reported as the normalized median error rates instead of mean error rates. For those two conditions, the normalized median error rates are reported in this study.

The normalized mean / median error rates as a function of the acoustic conditions used in the present study and those of the continuous versions reported in the literature are presented in Fig. 4.5: The normalized mean error rates for 1- and 6-band NVS are those reported in Sec. 2.4.2, and the normalized median error rates are those reported for the continuous M and ON stimuli in the study of Schlittmeier *et al.* (2012).

Independent t-tests revealed no significant difference between the data for the continuous 1-band NVS condition (normalized mean error rate = -0.4 %, N = 25) obtained from Chapter 2 and the data for the segmented 1-band NVS (normalized mean error rate = -0.04 %) collected in the present study ($p > 0.05$). The continuous 6-band NVS (normalized mean error rate = 8.05 %, N = 40) in the same study yielded significantly higher normalized mean error rate than the segmented 6-band NVS (normalized mean error rate = 1.8 %, $p < 0.01$).

The normalized median error rate reported for the continuous music

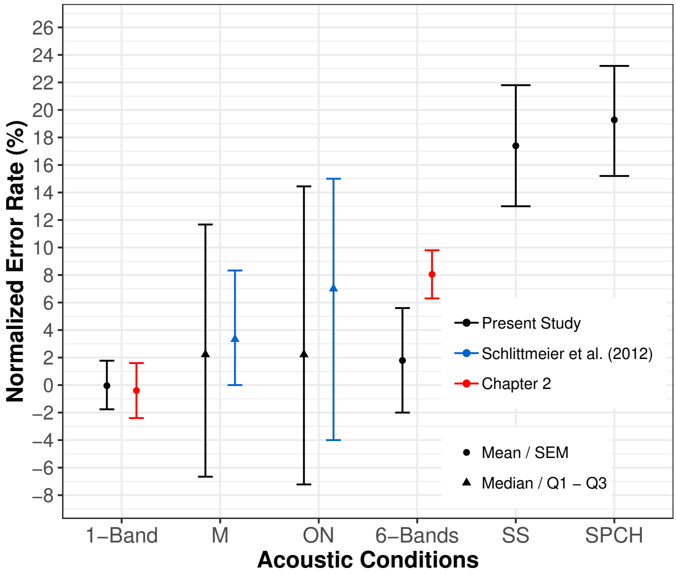


Figure 4.5: Recall performance of 23 participants are represented as normalized mean and median error rates (%) for six acoustic conditions: 1-band noise-vocoded speech (1-band), music (M), office noise (ON), 6-band noise-vocoded speech (6-bands), segmented (SS) and continuous speech (SPCH). Serial-recall results observed in the continuous versions of the stimuli obtained from literature are presented in red and blue. Error bars represent the SEM and interquartile ranges between the first (Q1) and the third (Q3) quartiles.

condition in the study of Schlittmeier *et al.* (2012) and the normalized median error rate observed for the segmented M in the present study are very similar, 3 % and 2.25 %, respectively. The continuous office noise stimuli in the same study yielded normalized median error rate of 7 % while the normalized median error rate for segmented ON condition in the present study is 2.2 %. Paired comparisons for these two conditions were not conducted, since the data are not available to us and the sample size is unknown. It was reported in the study of Schlittmeier *et al.* (2012) that the data were not symmetrically distributed in all behavioral experiments, so regenerating a dataset from random numbers which satisfies the statistics reported would not be reliable. However, the normalized error rates reported for the two studies can be used as a basis for comparison since the difference between the maximum and minimum normalized mean error rates observed in the present study (19.3 % for SPCH – 1-band NVS) is very similar to the maximum difference between the normalized median error rates (18.67 % for speech in Japanese – legato music) reported in the study of Schlittmeier *et al.* (2012). It can be observed that the segmented ON condition was slightly less disruptive than the continuous office noise.

4.2.4 Discussion

It was shown that the token selection stage of the $\mathrm{FDCC}_{\mathrm{new}}$ successfully captured the disruptive parts of the speech stimuli: The mean error rates observed in the SS and the SPCH conditions were very similar. This was expected since the token selection stage of the $\mathrm{FDCC}_{\mathrm{new}}$ algorithm was crafted based on speech stimuli with a focus on syllables. The extracted tokens from a speech stimulus created an almost equal degree of disruption as the continuous speech.

The general trend of the $\mathrm{FDCC}_{\mathrm{new}}$ values for the acoustic conditions was also reflected in the serial-recall results, except for the ON and the 6-band NVS conditions: The token selection stage of the algorithm failed to detect the disruptive tokens of the ON and the 6-band NVS stimuli.

When the ON stimuli were analyzed, the first noticeable difference between the ON and the rest of the acoustic conditions was the low number of tokens extracted from the stimuli. The ON stimuli yielded an average of 1.6 tokens per second, while the 1-band NVS stimuli were made up of approx. 2.6 tokens per second and the other conditions consisted of three or more tokens per second. In addition to this, the token frequency was not normally distributed along the trials: Half of the ON trials consisted of 40 tokens on average, which is close to 0.5 tokens per second. This means that there were long segments of silence in some trials, steady-state noise in this case, which might have increased the serial-recall performance. Nevertheless, this observation reveals a limitation in the definition of the $\mathrm{FDCC}_{\mathrm{new}}$: The $\mathrm{FDCC}_{\mathrm{new}}$ does not take into account the temporal distance between the successive tokens. While the lack of a token distance criterion is not critical for speech stimuli, it might lead to problems for non-speech stimuli with very low token rates.

The more apparent discrepancy was observed in the normalized mean error rates for continuous and segmented 6-band NVS conditions. The 6-band noise-vocoder was shown to produce different serial-recall results in the literature: In the study of Dorsi *et al.* (2018), the reported normalized mean error rate was 3.5 %, while in Chapter 2 the 6-band NVS yielded a normalized error rate of 8.05 %.

One of the differences between the two studies is that in Chapter 2 the same speech samples were used for generating all NVS conditions, as well as the continuous speech condition. This is particularly important because NVS is known to become more intelligible if NVS with higher number of frequency bands is heard before the NVS with lower number of frequency bands, which is known as pop-out effect (Davis *et al.*, 2005). The presentation order of the blocks in the same experiment was not counter-balanced to avoid such an effect, hence the 6-band NVS condition might have appeared after 9-, 12-, 18-band NVS and speech in an experimental session. This problem in the experimental procedure was discussed in Sec. 2.5 in detail and it was concluded that the objective intelligibility metrics used in the same study can only be used as temporal distinctiveness metrics because the actual perceived intelligibility of the NVS stimuli may be very different. It was also stated that the intelligibility of speech should not play a role in ISE since reversed and foreign speech disrupted the serial-recall performance similar to native speech stimuli (Jones *et al.*, 1990). Furthermore, the 4-band NVS stimuli ($\approx$ 8.5 %) used in the study of Ellermeier *et al.* (2015) yielded a similar mean error rate as the original speech condition ($\approx$ 10.5 %) in the same experiment, where the two conditions were generated by using different speech samples.

When the mean error rates and the FDCC$_{\text{new}}$ values of the continuous 6-band NVS and the segmented 6-band NVS are compared, it can be clearly seen that the change in the power spectra between the sound tokens, as quantified by the FDCC$_{\text{new}}$, can not be responsible for the serial-recall disruption observed for continuous or segmented 6-band NVS condition. In addition to this, an important characteristic of the 6-band NVS was mentioned to be its high level of intelligibility in Chapter 2 and speech perception scores for 4-band NVS were the same as those of speech in the study of Ellermeier *et al.* (2015), which eventually implies a role for the intelligibility.

For the present study, these interpretations are rather speculative since there is not enough evidence to ascribe a special role to the intelligibility of speech within the context of ISE based on the scores of the segmented 6-band NVS. There is however another study which investigated the relation between the speech-specific properties of the NVS within the context of ISE from a slightly different perspective: In the study of Dorsi *et al.* (2018), the impact of NVS on ISE, with a focus on speech specific properties of the NVS, was investigated by manipulating the speech fidelity of the NVS stimuli while preserving the overall changing-state complexity by keeping the number of frequency bands constant. The stimuli were later used in a serial-recall task. The stimuli used in the study of Dorsi *et al.* (2018) were provided to us by the authors and we have investigated the experimental results of this study by computing the $\mathrm{FDCC}_{\mathrm{new}}$ values of the NVS stimuli.

4.3 Experiment 2

The second experiment focuses on a study where Dorsi *et al.* (2018) have generated specific NVS stimuli and used them in a serial-recall experiment. In the study of Dorsi *et al.* (2018), the NVS stimuli were employed in three experiments with the aim of investigating the effect of speech-fidelity on auditory distraction. The first two experiments were serial-recall tasks: The first experiment (N = 81, between subjects) employed 3-, 6-, 9-, and 12-band NVS stimuli and the second experiment (N = 77, within subjects) employed 6-, 12- and 18-band of both unmodified and selectively reversed NVS. The first experiment was designed in order to investigate the impact of the number of frequency bands on the ISE and similar results were observed as those of the studies in the literature (Ellermeier *et al.* 2015; Chapter 2).

The second experiment aimed at evaluating the impact of speech fidelity on ISE with the argument that the temporally reversed NVS stimuli, such as 6-band reversed NVS, lacked speech fidelity which the original 6-bands provided while simultaneously preserving the changing-state complexity. The results showed that the temporally reversed NVS disrupted the serial-recall significantly less than the original NVS: The authors argued that speech fidelity was an important factor in the ISE.

The third experiment was a missing item task (N = 77, within subjects) which was designed to dissociate whether the effect of the speech fidelity observed in the second experiment was due to interference-by-process or due to attentional capture. The selectively reversed NVS stimuli did not alter the missing-item performance significantly and the authors concluded that the effect of speech fidelity could not be attributed to attentional capture.

For the present study, we focus on the second experiment reported in the study of Dorsi *et al.* (2018): The significant difference between the serial-recall results obtained using the selectively reversed and the typical NVS observed in the second experiment is particularly interesting since the frequency spectra of the two types of NVS, for instance 6-band reversed NVS and typical NVS, are identical. The set of NVS

Table 4.1: Noise-vocoded speech stimuli with the number of frequency bands and upper frequency boundaries.

Nr. of bands	Cut-off frequencies (Hz)
18 Bands	66, 88, 116, 154, 205, 271, 360, 477, 632, 838, 1112, 1474, 1954, 2590, 3434, 4552, 6034
12 Bands	76, 116, 178, 271, 414, 632, 965, 1474, 2249, 3434, 5241
6 Bands	229, 558, 1160, 2265, 4290

stimuli employed in the second experiment was analyzed in terms of the FDCC$_{\text{new}}$ values of these stimuli.

4.3.1 NVS

As it was explained in detail in Sec. 2.2.2 of Chapter 2, NVS is a manipulated speech stimulus which is generated by filtering the original speech into frequency bands and mapping the intensity envelope of each frequency band to band-limited white noise. The resulting speech-enveloped noise bands are summed to create a harsh, metallic distorted speech.

The noise-vocoding technique used in the study of Dorsi *et al.* (2018) is very similar to that reported in the second chapter: The NVS stimuli were generated by dividing the speech signal between 50 and 8000 Hz into 6, 12, and 18 Hanning-shaped bandpass filtered frequency bands by modifying the same Praat scripts (Praat software, Institute of Phonetic Sciences, University of Amsterdam, Amsterdam, The Netherlands) that were used in Chapter 2, which were initially employed in a speech comprehension study (Davis *et al.*, 2005). The chosen cut-off frequencies are presented in Table 4.1.

Seven words, *bowls*, *boy*, *day*, *dog*, *go*, *than* and *view*, were used to generate the NVS stimuli. The NVS conditions were created by concatenating the seven noise-vocoded words in a randomized order for each NVS condition and each NVS stimulus was repeated twice during each trial to reach the duration of the presentation stage.

4.3.2 Selectively reversed NVS

The selectively reversed NVS conditions were generated by temporally reversing the lower two-thirds of the vocoded channels of each noise-vocoded word. The method applied in the study of Dorsi *et al.* (2018) was similar to that of the selectively reversed sine wave-speech used in the study of Viswanathan *et al.* (2014). The cross-over frequency of the reversal corresponds to 1474 Hz in the 12- and 18-band NVS conditions while in the 6-band NVS condition it is 2265 Hz (see Table 4.1). The authors concluded that the reversal applied to the NVS distorted the speech information, and was therefore used as a technique to manipulate the speech fidelity in the experiment.

The reversal method was applied to the seven noise-vocoded words individually and the selectively reversed noise-vocoded words were concatenated in randomized order to generate a selectively reversed NVS stimulus for each condition. Each stimulus was repeated two times, until the end of the presentation stage, in each trial.

4.3.3 Summary of the experimental procedure reported in the study of Dorsi *et al.* (2018)

Participants

The serial-recall experiment consisted of 77 participants from the University of California, Riverside. All participants were reported to be native English speakers with normal hearing and normal or corrected to normal vision.

Stimuli

There were seven acoustic conditions presented in the experiment: Three NVS, 6-, 12-, 18-band, three selectively reversed versions of the same NVS conditions and the silence condition (SLNC). The playback level of the sound conditions were reported to be 70 dB.

Apparatus and procedure

The serial-recall tasks followed the general protocol followed in ISE studies. The to-be-remembered items were seven different consonants and were presented to the subjects via a computer screen in randomized order. The subjects were instructed to write down the order of the presented item 1 second after the last letter was shown. The paper did not report any information about the retention period or where the experiment took place.

Each sound condition was repeated six times in one experimental session and each trial consisted of one randomly selected irrelevant stimulus. The irrelevant sounds were played back through headphones.

Results

The authors first analyzed the effect of sound on serial-recall results and it was reported that both unmodified (M = 35 %, SD = 20.15 %, $p <$ .01) and reversed NVS (M = 32.5 %, SD = 19.5 %, $p <$ .01) conditions created significantly higher serial-recall disruption than SLNC (M = 29 %). Here it should be noted that the SD values were not reported in the original study, and they were regenerated based on the SEM values presented in the figure showing the normalized error rates in the paper (Dorsi *et al.*, 2018, Fig. 2). The SLNC condition was not presented in that figure and the mean error rates for the SLNC condition was only shown in a table (Dorsi *et al.*, 2018, Table 1), without reporting the SD and SEM values.

The mean error rates reported in the study were normalized by subtracting the mean error rate observed in the SLNC condition, and the normalized mean error rates (%) as a function of the acoustic conditions are regenerated and presented in Fig. 4.6.

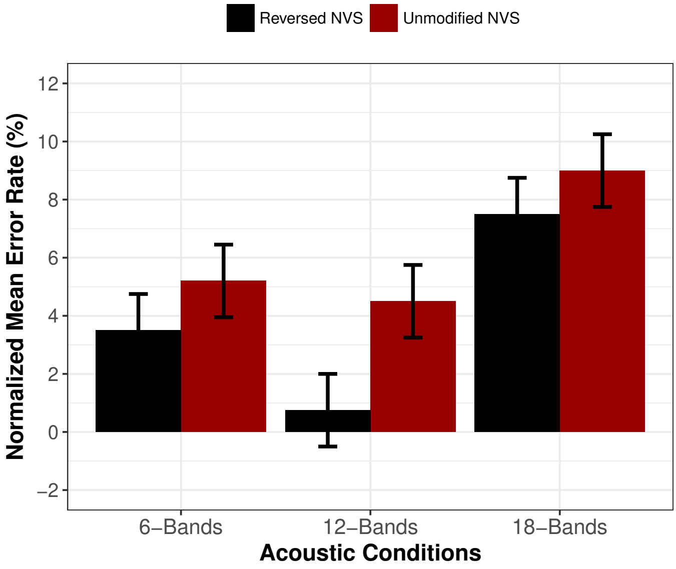


Figure 4.6: Normalized mean error rates (%) as a function of the reversed and unmodified noise-vocoded speech conditions for 77 participants. Error bars represent the SEM.

It was reported in the study (Dorsi *et al.*, 2018) that the unmodified NVS conditions, regardless of the number of the frequency bands employed, produced significantly higher normalized mean error rates than those of the reversed NVS conditions ($p < .05$). There were no further statistical analysis reported in the original study.

4.3.4 Analysis of the experimental results with respect to the FDCC_{new}

The FDCC_{new} values for each acoustic condition were calculated from the long NVS stimuli which were actually used in the experiment. The parameter values, as a function of the acoustic conditions, are presented in Fig. 4.7.

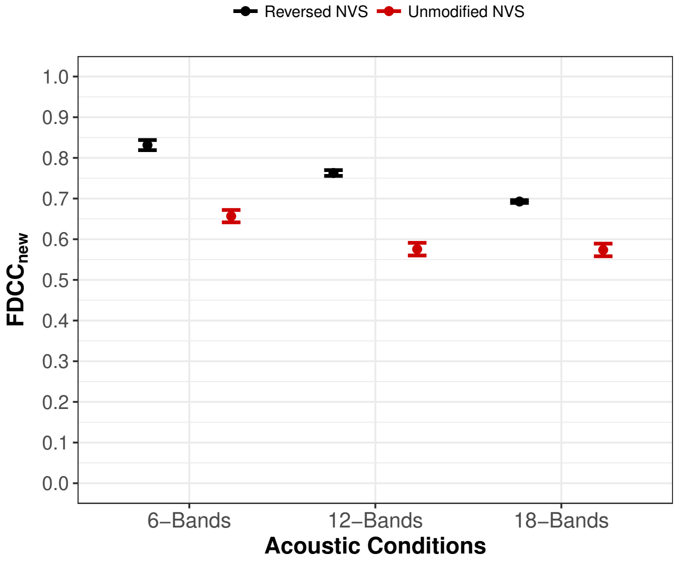


Figure 4.7: FDCC_{new} values as a function of the reversed and unmodified noise-vocoded speech conditions. Error bars represent the SEM.

For both types of the unmodified and the reversed NVS stimuli, there is a general trend of a decrease as a function of the number of frequency bands, but it should be noted that the FDCC_{new} values of unmodified NVS with 12 (0.58) and 18 bands (0.57) are very close.

More relevant to the present study, the normalized error rates observed for the pairs of the unmodified and reversed NVS stimuli, based on the number of channels employed, were reflected in the FDCC_{new} values: The FDCC_{new} values of the unmodified NVS stimuli are lower than those of the reversed NVS stimuli with the same number of frequency bands.

The reason for finding higher FDCC$_{\text{new}}$ values for the selectively-reversed NVS was investigated by choosing one of the 12-band NVS as an exemplar to analyze, since the difference between the normalized mean error rates of the unmodified and the reversed 12-band NVS is the largest (difference in normalized error rate = 3.8 %) when compared with the two other conditions.

First the cross-over frequency of the reversal was examined by plotting the spectrogram of the first two words of the unmodified and the reversed 12-band NVS stimuli and it was observed that the cross-over frequency is around 1474 Hz as reported in the study. The spectrograms of the first two words of the unmodified 12-band NVS and the selectively-reversed 12-band NVS are presented in Fig. 4.8a and 4.8b, respectively.

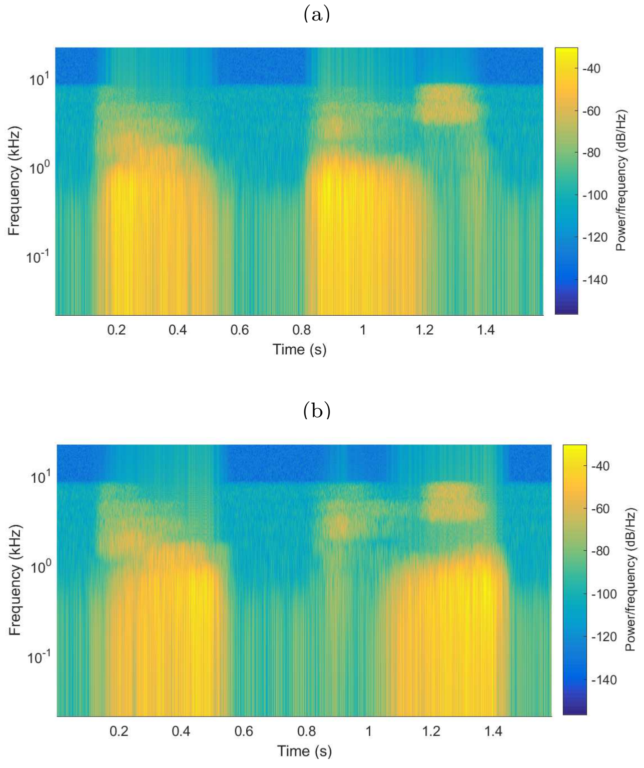
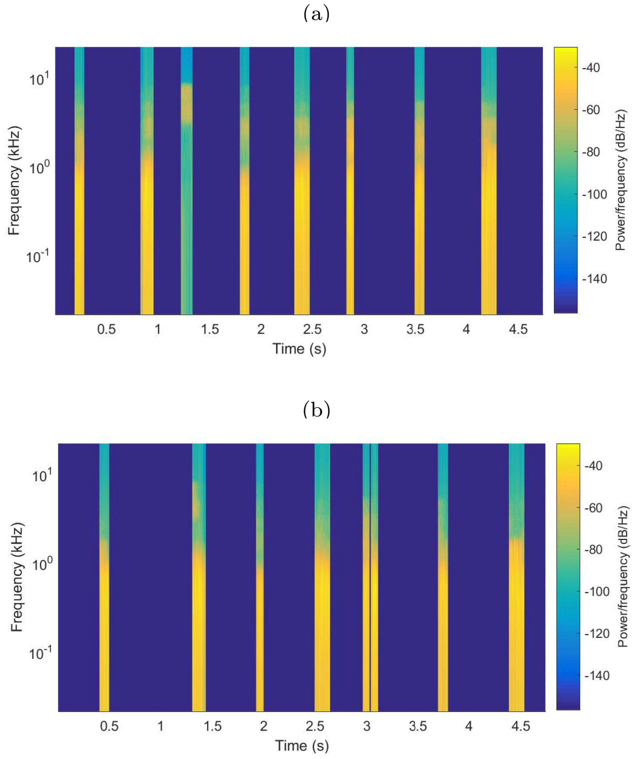


Figure 4.8: The top panel (a) shows the spectrogram of the first two words, *go* - *bowls*, of the unmodified 12-band noise-vocoded speech and the spectrogram of the same two words of the selectively-reversed 12-band noise-vocoded speech are presented in the bottom panel (b).

Second, the original and the selectively-reversed NVS stimuli were segmented into tokens by the token selection stage of the FDCC$_{\text{new}}$ and the spectrograms of the segmented tokens were examined. The two spectrograms are presented in Fig. 4.9a and 4.9b. It can be clearly observed that


(a)

(b)

Figure 4.9: The spectrogram of the selected tokens of the original 12-band noise-vocoded speech are presented in the top panel (a) and the bottom panel (b) shows the spectrogram of the selected tokens of the selectively-reversed 12-band noise-vocoded speech.

the spectra of the majority of the tokens, selected using the FDCC$_{\mathrm{new}}$, are rich in energy between 50 to 1000 Hz in the two spectrograms. The exception is the third token of the unmodified 12-band NVS (Fig. 4.9a): The third token, which represents the end of the letter, *s*, at the end of the second token, *bowls*, only comprises energy above the cross-over frequency.

When the spectrogram of the selectively-reversed 12-band NVS (Fig. 4.9b) is examined, it can be observed there is no such token like the third one of the original 12-band NVS: When the lower two-thirds of the bands were reversed, the process created an overlap between the regions below and above the cross-over frequency, and the singular high-frequency rich token, *s*, became part of a token with broader frequency spectrum.

The closer look at the correlation values between the adjacent tokens revealed that the FDCC_{new} value obtained between the second and the third tokens (0.0010) and the third to fourth tokens (0.0018) of the original 12-band NVS stimuli were the lowest correlation values obtained from the two stimuli. All unmodified and reversed NVS conditions were analyzed, and the same outcome was observed for all sounds: After selectively reversing the NVS stimuli in the time domain, the resulting overlap increased the similarity between the adjacent tokens, hence the low FDCC_{new} values associated with the letter *s* were not present anymore.

The unmodified and the selectively-reversed 12-band NVS stimuli yielded different serial-recall results as well as different FDCC_{new} values and the trend of the FDCC_{new} values was reflected in the normalized mean error rates. The authors have stated that this is a strong indication that speech may have a special role, however, the reduction in the speech fidelity was achieved by altering the magnitude of the spectral variation in the NVS stimuli using the method of selectively-reversing the lower two-thirds of the frequency bands: The frequency spectra of the long unmodified and the reversed NVS stimuli are identical, but the changing-state complexity they comprise is different since the definition of the changing-state hypothesis takes the time-domain information into account. The variation in the power spectra, from one token to another, has changed when the information within the lower two-thirds of the frequency region was temporally reversed. The reversal technique applied here simply aligned the “outlying” high-frequency energy with low-frequency energy, and therefore the spectral variation between the tokens was reduced. The FDCC_{new} was sensitive enough to detect the reduction in the magnitude of change produced by the reversal technique for this particular set of stimuli, which was also reflected in the serial-recall results.

4.4 General discussion

Two experiments were conducted in order to asses the accuracy of the token selection stage of the FDCC_{new} algorithm, with respect to its ability to detect the parts of the sound which comprise the states where between-changes would be disruptive in serial-recall.

The first experiment showed that the token selection stage was accurate enough to detect the disruptive properties of the speech stimuli, but it also revealed that the lack of a token distance criterion in the FDCC_{new}

is a limitation, since it does not take into account the duration between two adjacent tokens in a sound. The second limitation was observed in the segmented 6-band NVS condition: The mean error rate observed in the first experiment for the 6-band NVS condition was significantly lower than what is reported in Chapter 2. Both conditions yielded similar FDCC_{new} values, 0.50 for the continuous 6-band NVS and 0.55 for the segmented 6-band NVS, which showed that the FDCC_{new} was not able to predict the differences in disruptive performance for the two conditions and the token selection stage of the FDCC_{new} was not accurate enough to detect the disruptive properties of the continuous 6-band NVS stimuli. The change in the power spectra of the two adjacent tokens in the 6-band NVS was preserved for both the segmented and continuous versions, but the magnitude of the disruption observed for two conditions was significantly different.

This is particularly interesting since the segmented speech did not create a significantly different disruption when compared to continuous speech, but for the distorted speech condition, a significant difference was observed: When the speech-specific acoustic properties were reduced compared to the original speech, such as in 6-band NVS, the segmentation process further reduced its disruptive capabilities. The segmentation process discarded the non-token parts, but did not change the frequency domain changes which were thought to be ISE relevant. Clearly, there is more than spectral features that play a role in serial-recall performance for the 6-band NVS condition in the present study.

The aforementioned observation can be rephrased: When a degraded speech stimulus, 6-band NVS, was segmented, the speech-specific properties (natural acoustic variations) were reduced further to a level which is below the ISE threshold. On the other hand, when the continuous speech were segmented with the same technique, the speech specific properties were preserved to a certain degree, which was enough to produce an ISE. This interpretation ascribes a special role to speech stimuli, which violates the changing-state hypothesis and subsequently the FDCC_{new}. Nevertheless, increasing the number of frequency bands in NVS increases the speech intelligibility (Davis *et al.*, 2005; Ellermeier *et al.*, 2015) while increasing the magnitude of spectral variation as reflected in the FDCC_{new} values in Chapter 2. This was supported by the systematic increase in the values of the speech transmission index and the normalized covariance measure reported in the same study. As a result, the reason behind the serial-recall disruption observed for NVS stimuli in Chapter 2 is not known for certain. In order to clarify this, the spectral features of the irrelevant sound stimuli should be modified in a way that the temporal features of the same stimuli should not be affected.

The second experiment investigated a set of stimuli, which was used to evaluate the impact of speech fidelity on ISE from a different perspective: Distorting the information in speech was accomplished by temporally reversing the information in the lower two-thirds of the frequency bands in the NVS stimuli. The reversal was successfully detected by the FDCC_{new} since it changed how the frequency spectrum varied in time. There was no subjective intelligibility test conducted in the study of Dorsi *et al.* (2018), so the argument regarding the reduction in the speech fidelity based on the process of selective reversal is simply the author's interpretation. The interpretation sounds likely, however, this was achieved by modifying the spectral features of the stimuli so the observed increase in the serial-recall performance after the reversal supports the changing-state hypothesis and does not ascribe any speech specific property to the ISE in that experiment.

4.5 Conclusion

1. The token selection stage of the FDCC_{new} successfully detected the disruptive tokens in continuous speech stimuli, but failed to extract the disruptive tokens in 6-band noise-vocoded speech.

2. The FDCC_{new} values for the segmented stimuli followed the trend of the serial-recall results, except for the office noise and 6-band noise-vocoded speech conditions: Even though the continuous 6-band noise-vocoded speech and the segmented 6-band noise-vocoded speech yielded similar FDCC_{new} values, the between-subject analysis showed that the serial-recall results of the two conditions were significantly different.

3. The FDCC_{new} was sensitive enough to capture frequency domain changes introduced by the selective reversal technique applied to the noise-vocoded speech stimuli. The serial-recall results observed for the unmodified and the revered noise-vocoded speech conditions were reflected in the FDCC_{new} values.

5 | Spectral and temporal features as the estimators of the irrelevant sound effect[1]

Abstract

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 $\mathrm{FDCC}_{\mathrm{new}}$. The first parameter is a measure for temporal variations in a sound, the latter one measures the spectral variability within a sound stream. For the first experiment, background stimuli were synthesized from a noise-pulse train in which modified and unmodified pulses alternate. In order to modify 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. In the second experiment, alternating noise-pulses were generated by a similar approach but with a difference: The information regarding duration, amplitude and the positions of the pulses was derived from speech samples, hence resulting in an irregular inter noise-pulse interval. Both sets of stimuli were used as irrelevant sounds in a serial-recall experiment and the impact of temporal and spectral features was investigated.

[1] Part of this work was published in:
Senan, T. U., Park, M., Kohlrausch, A., Jelfs, S., and Navarro, R. F. **(2015)**. "Spectral and temporal features as the estimators of the irrelevant speech effect," in *Proceedings of Euronoise 2015*, edited by C. Glorieux (Maastricht, The Netherlands), pp. 1925–1930

5.1 Introduction

The short-term memory performance decrease induced by background task-irrelevant sounds has been investigated thoroughly in the literature (for a review, see Banburry *et al.*, 2001). Researchers have looked into the interaction between the background sounds and the magnitude of the disruption through digit or letter recall tasks. The phenomenon was first observed in close relation with background speech (Colle and Welsh, 1976) and there were debates about speech stimuli having a special role (see, e.g., Baddeley, 1997). Soon, it was shown that both foreign and reversed speech (Jones *et al.*, 1990) disrupted serial-recall to a similar degree as native, intelligible speech.

The claim that speech has a special role in the ISE was also disputed by studies showing that a repeated speech token is not more disruptive than silence (Jones *et al.*, 1992), and also by experiments where the acoustic features of the speech or non-speech background sounds were manipulated in a systematic manner and the degree of manipulations was reflected in the serial-recall scores (Jones and Macken, 1993). Speech tokens in alternating order, such as C-H-U-J, produced a significantly larger magnitude of disruption when compared to a repeated set of tokens (e.g., J, J, J, J) (Jones *et al.*, 1992). Similar results were found when the same principle of within sequence variation was applied to non-speech sounds, such as pure tones changing in frequency (Jones and Macken, 1993).

In several studies, the magnitude of acoustic variation, whether the change is between one discrete token to another or within a continuous sound, was shown to be directly related to the serial-recall performance (Jones and Macken, 1993; Jones *et al.*, 1993). An acoustic variation can be understood as the magnitude of acoustic change from one sound segment to another, hence was defined as the changing-state hypothesis: The irrelevant sound must be segmentable into perceptually discrete tokens and each token should be acoustically different from the one that preceded it (Jones *et al.*, 1993; Hughes *et al.*, 2007). The acoustic determinant of the changing-state was investigated in literature by taking a closer look into the temporal and spectral features of the irrelevant sounds.

For instance, it was shown that modifying the intensity of the tokens of the sounds within an irrelevant sound stream did not create an ISE (Tremblay and Jones, 1999). The serial-recall performance was also shown to not be affected by changing the time gap between successive segments of the background sounds (Tremblay and Jones, 1999) or modulating a sound source with random and fixed envelope (Jones *et al.*, 1992), hence undermining the role of temporal features on the ISE.

The impact of spectral variation on the ISE, which is the major source of interest in this thesis, has also been investigated thoroughly in the literature (e.g., Jones *et al.*, 1999, 2000; Ellermeier and Zimmer, 2014). The relation between the change in frequency and the ISE was examined by means of various types of auditory stimuli, such as pitch shifted vowels (Jones *et al.*, 1999); sequencing tones with repeated and changing frequencies (Jones and Macken, 1993); alternating band-pass filtered noise bursts with different center frequencies (Tremblay *et al.*, 2001); low-pass filtered stimuli with different roll-off values (Jones *et al.*, 2000); as well as noise-vocoded speech (NVS) stimuli (Ellermeier *et al.* 2015; Chapter 2). All these studies support the notion that the spectral variation in a background sound disrupts serial-recall performance remarkably.

The attributed high prominence of frequency change within the context of the ISE, however, did not evolve into a successful prediction model, due to the complex nature of the problem. In fact, throughout this thesis we have studied the only descriptor which attempts to predict the ISE based on the magnitude of the changing-state in the frequency domain. The results, so far, supported the findings in literature: The changes in the frequency domain are relevant, if not necessary, for creating an ISE. However, limitations were also revealed, which indicated that the magnitude of the spectral variation within the irrelevant sounds may not be enough by itself to predict the size of the ISE.

For instance, the results of the experiments in Chapter 2 showed that the STI and the NCM resembled the trend of serial-recall scores using NVS stimuli to a similar extent as the FDCC$_{\text{new}}$ did: The parameter values followed the performance scores up to 6-band NVS successfully, and failed beyond that. The increase in the number of frequency bands employed in NVS not only increased the spectral variation but also modified the temporal features of the stimuli in a systematic way and this systematic change in temporal features of the NVS was quantified by the STI and the NCM (for the definitions of the two metrics, see Appendix A). The relation between the trend of the two parameter values, actual test results, and the impact of speech intelligibility on the ISE were discussed in detail in Sec. 2.6. Nevertheless, the fact that the performance of the spectral descriptor was similar to those of the two temporal descriptors suggested that both the spectral and the temporal features of irrelevant sounds must be examined, independently from each other, in order to comprehend the sole impact of spectral features of background sounds on the ISE.

The role of spectral variations on the ISE was investigated by calculating the FDCC_{new} for a large set of stimuli from the literature in Chapter 3. The FDCC_{new} values correlated well with the serial-recall results obtained for 91 data points (Pearson's r = 0.70, $p < 0.01$). Although the value strongly links the change in frequency with the ISE, it also supports that the magnitude of spectral variation within an irrelevant sound stream may not be the only reason behind the ISE.

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 FDCC_{new}. A set of white noise-pulse train stimuli was created for which the values of the two metrics can be modified independently. For the first experiment, the noise stimuli were organized in a way that there were 10 noise-pulse train conditions where only the AMTF was modified, and 10 noise-pulse train conditions where the FDCC_{new} values were varied while the AMTF was kept constant. The results of the first experiment revealed some limitations regarding the experimental design and the FDCC_{new}. The reasons behind the limitations were further investigated by designing another set of noise-pulse train stimuli, where the rhythmic structure of the stimuli was inspired by continuous speech. The speech-positioned noise-pulse train stimuli were employed in a second serial-recall experiment.

The details of the two metrics are described in section 5.2. Temporal and spectral modifications applied to the stimuli used in the first experiment are explained in sections 5.3.1 and 5.3.2, the experimental procedure for the first experiment is presented in sections 5.4. The second experiment, including the motivation, procedure, stimuli, results and discussion is described in section 5.5 through section 5.5.4. The study is finalized by the general discussion and the conclusion sections, 5.6 and 5.7, respectively.

5.2 Estimators of the ISE and stimuli

In the following sections the AMTF and the FDCC$_{\text{new}}$ are used to quantify the spectral and temporal properties of the irrelevant sound stimuli in order to observe the influence of the acoustic properties of background sounds on the ISE. The AMTF is explained in the following section and the computation of the FDCC$_{\text{new}}$ is identical with what is reported in section 3.2.

5.2.1 Average Modulation Transfer Function

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 the temporal distinctiveness (Houtgast and Steeneken, 1985). 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 is carried out in order to cover the range of frequencies between 125 Hz and 8 kHz. For speech intelligibility investigations, a range of modulation frequencies between 0.5 Hz and 16 Hz is chosen. The intensity envelope of the input signal, x, is obtained for each octave band by filtering the input signal with a second-order Butterworth band-pass filters (BPF), squaring the output, and then applying a second-order Butterworth low-pass filter (LPF) to the signal with a cutoff frequency of 30 Hz. The resulting intensity envelope is analyzed for each modulation frequency with an octave-band filter 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 the envelope, $\overline{y_{ij}}$. 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. 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:

$$M_{ij} = \frac{m_{ij,x}}{m_{ij\,ref}} \tag{5.1}$$

Chapter 5

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 $(\overline{M})$.

5.3 Noise-pulse train (NT)

A white noise pulse (P1) was generated as the basis of the reference signal, for which the FDCC$_{\text{new}}$ and the AMTF can be manipulated independently. White noise, G(t), and a Hanning window W(t), of size w, were used to define the pulse shape. A one-third octave-band filter with center frequencies ranging from 125 Hz to 8 kHz was used to perform the decomposition of $WG(t)$ into 21 bands. Seven out of the 21 bands, whose center frequencies are 125 Hz, 250 Hz, 500 Hz, 1000 Hz, 2000 Hz, 4000 Hz and 8000 Hz, were selected and the P1 of size w was generated by summing the selected seven bands:

$$\text{P1} = \sum_{i=1}^{K=7} x_i(t) \tag{5.2}$$

where i indicates the i-th octave band.

A half-second sample was generated where two pulses, P1 and P2, of 50 ms separated by 250 ms alternate. The amplitude of P2 was lowered to be one-third of P1. A 1 min basic signal was formed by concatenating a sequence of half-second samples containing P1 and P2, and was defined as the reference signal. An illustration of the reference signal is presented in Fig. 5.1.

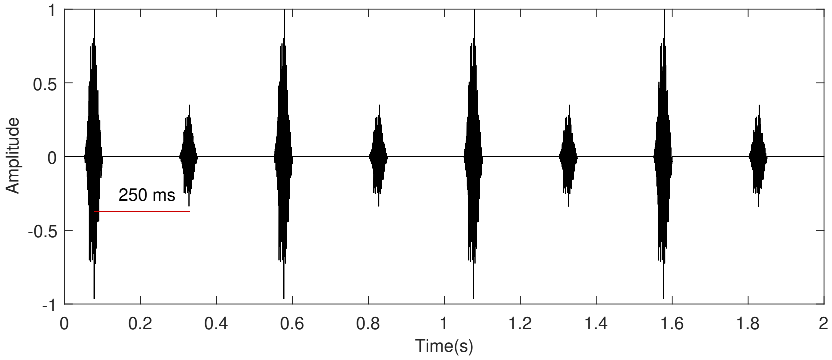


Figure 5.1: A two second segment of the reference noise-pulse train (NT) stimulus is presented above, where the two pulses, P1 and P2, alternate within every half second segment.

The reference signal served as a basis, from which the independent temporal and spectral modifications were achieved by processing P2 only, by the methods described below.

5.3.1 Modifying the temporal features

In order to modify the AMTF without changing the $\mathrm{FDCC_{new}}$ values, the pulse width of P2 was modified from 50 ms to 450 ms in steps of 25 ms. It was observed that, as the width of P2 increased, the AMTF decreased, while the $\mathrm{FDCC_{new}}$ remained constant. This technique was used to create a set of periodic 1 min long noise-sequences with AMTF values ranging from 0.58 - 1 (see Fig. 5.2), which was later employed in the experiments.

5.3.2 Modifying the spectral features

The spectral features of the reference noise-pulse train stimulus were modified by applying gains between 0 and 1 to seven octave bands (125 Hz - 8 kHz) of P2. The applicable gains were selected using a gain optimization procedure because applying gains to octave bands does change the AMTF value as well. The gain optimization procedure was designed to find the optimal gain values which would keep the AMTF value of the spectrally modified noise-pulse train constant. The procedure was used to generate a set of periodic 1 min long spectrally modified noise-pulse train stimuli and the parameter values of the chosen stimuli are presented in Fig. 5.2. The gain optimization procedure is explained in Appendix B.

5.4 Experiment 1

The present experiment employed 20 noise-pulse train (NT) stimuli with varying AMTF and $\mathrm{FDCC_{new}}$ values in a serial-recall experiment, in order to observe the independent impact of the temporal and spectral features of irrelevant sounds on the ISE. The experiment was designed with the expectation that the spectrally modified noise-pulse trains would produce an ISE and the magnitude of serial-recall disruption would be reflected in the $\mathrm{FDCC_{new}}$ values, while the temporally modified stimuli should not disrupt serial-recall performance.

5.4.1 Method

Participants

A total number of 10 participants (four females and six males, age range between 18-50 years) from the Philips Research Laboratories in Eindhoven participated voluntarily. All participants were employees of Philips Research and stated that they had healthy vision and hearing and no history of memory related disorder. The eligibility criteria were cross checked prior to the experiment, before they signed the informed consent forms. The experimental procedure was evaluated and approved by the Internal Committee Biomedical Experiments (ICBE) of Philips Research.

Stimuli

There were 20 NT stimuli selected for the experiment: 10 NT stimuli with AMTF values of 0.58, 0.61, 0.66, 0.70, 0.74, 0.77, 0.81, 0.91, 0.96, 1, and FDCC_{new} values of 1; 10 NT stimuli with FDCC_{new} values of 0.06, 0.08, 0.19, 0.37, 0.52, 0.58, 0.60, 0.74, 0.95, 1 and AMTF values of 1. The parameter values of the selected 20 stimuli are presented in Fig. 5.2. Alongside the noise conditions, there was a silence condition (SLNC) which served as a control condition for the experiment. The average sound level of the stimuli in the experiment was calibrated to 60 $\text{dB}_{\text{LAeq1min}}$.

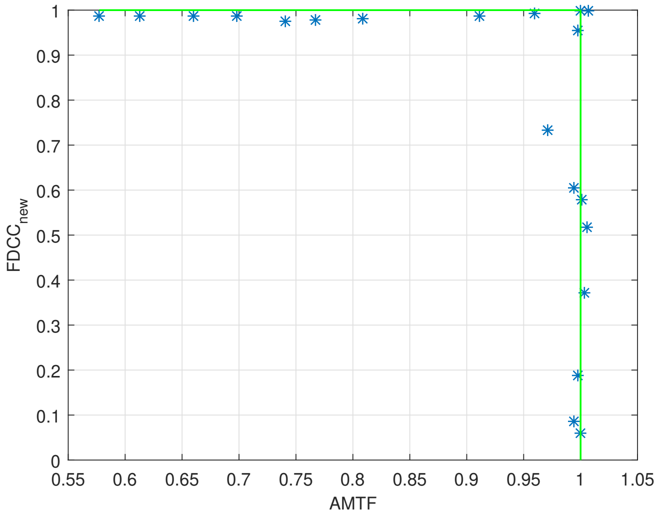


Figure 5.2: The FDCC_{new} and the AMTF values of the 20 audio stimuli. The x-axis shows the AMTF values and the y-axis shows the FDCC_{new} values. Each point represents a noise-pulse train (NT) stimulus.

Apparatus

The apparatus was same as the one reported in Sec. 2.3.1.

Procedure

The serial-recall is the focal task in this experiment and the MATLAB script used in Sec. 2.3 was used to conduct the experiment.

The serial-recall task 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: one dummy trial (NT stimulus, AMTF = 1, $\text{FDCC}_{\text{new}} = 1$), one SLNC and 20 NT stimuli. In each trial, a different background stimulus was presented 3 seconds before the first digit's appearance until the participant pressed 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.

5.4.2 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 performance was measured as error rate (%) out of nine digits.

The mean error rates as a function of the AMTF values of the NT stimuli and SLNC are presented in Fig. 5.3 and the test scores as a function of the FDCC_{new} values of the NT stimuli are presented as mean error rates, alongside SLNC in Fig. 5.4. The mean error rate for SLNC (28 %) is in line with what has been reported in the literature (e.g., Schlittmeier *et al.*, 2012; Park *et al.*, 2013; Liebl *et al.*, 2016) and in Chapters 2 and 4. However, it can be clearly seen from the two figures that there is no impact of NT stimuli on recall performance regardless of the type of modification applied. This observation was confirmed by a one way repeated measures ANOVA, $F(20, 180) < 1$. Due to the lack of any significant effect of sound on recall performance, no further analyses were conducted.

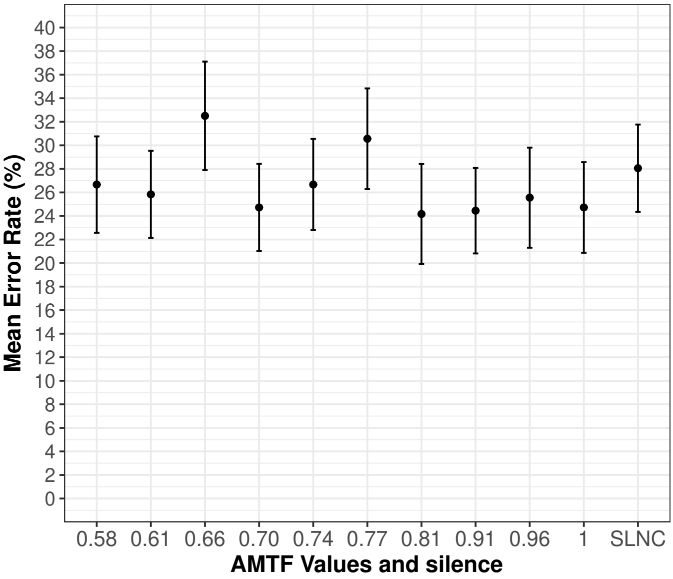


Figure 5.3: Mean error rates (%) as a function of the AMTF values of noise-pulse train (NT) stimuli and silence (SLNC). The error bars represent the standard error of the mean (SEM).

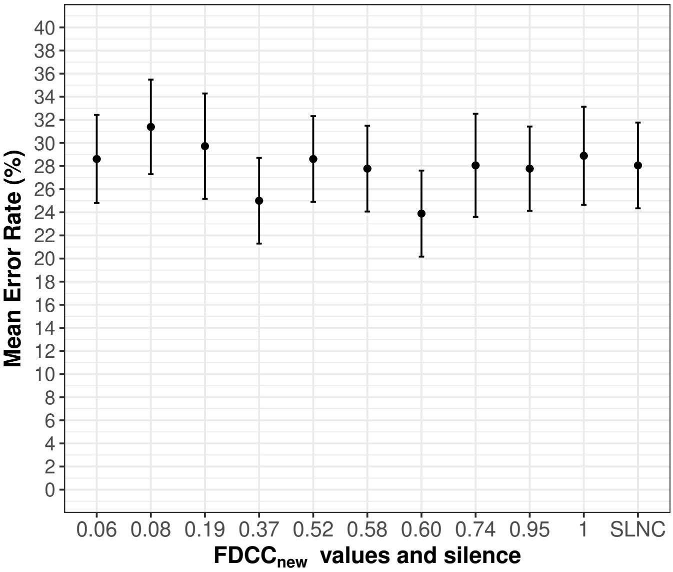


Figure 5.4: Mean error rates (%) as a function of the FDCC$_{\text{new}}$ values of noise-pulse train (NT) stimuli and silence (SLNC). The error bars represent the SEM.

5.4.3 Discussion

Results show that the NT stimuli did not create an ISE. The lowest error rate (highest score) was expected to be achieved in the SLNC condition (28 %), which, in fact, was not the result: All modifications, on average, had similar scores as the SLNC and the lowest mean error rate was obtained using the NT stimulus with an FDCC_{new} value of 0.60 (23 %). Nevertheless, the data did not show any systematic change in the error rate as a function of the two parameters.

The temporal variation was not expected to create an ISE: The temporal modifications which would create a change in AMTF, such as, changing the intensity of the tokens within the sound sequences (Tremblay and Jones, 1999), altering the time gap between tokens in a sequence (Tremblay and Jones, 1999) or modulating a speech source with random and fixed envelope (Jones *et al.*, 1992), were shown to be ineffective in terms of creating a serial-recall disruption. The results in this experiment are in line with what is reported in the aforementioned studies: There was no impact of temporally modified NT stimuli on serial-recall performance.

On the other hand, the experiment was designed with the hypothesis that the spectrally modified noise-pulse sequences would lead to a decrease in serial-recall performance. The expectation was derived from literature, where the studies showed that band-pass filtered noise bursts changing in center frequency (Tremblay *et al.*, 2001) and sine tones changing in frequency (Jones and Macken, 1993) disrupted serial-recall significantly, while repeated noise-bursts and sine tones did not. Furthermore, the magnitude of the disruption was expected to change in relation to the magnitude of the spectral variation, as evidenced in literature by employing sine tone sequences and pitch shifted vowels (Jones *et al.*, 1999), masked speech (e.g., Park *et al.*, 2013; Liebl *et al.*, 2016), noise-vocoded speech (Ellermeier *et al.* 2015; Chapter 2), sine-wave speech (Tremblay *et al.*, 2000; Viswanathan *et al.*, 2014), and low-pass filtered speech (Jones *et al.*, 2000). Clearly, the expectation of observing a larger performance decrease for spectrally modified stimuli, when compared with temporally modified noise-pulse train stimuli, was not realized in this experiment.

The results of the experiment showed that there was no impact of NT stimuli on serial-recall performance, regardless of the modification applied. Such an outcome prevents us from reaching solid conclusions about the impact of the temporal and spectral features of the irrelevant sound on serial-recall. There are four possible explanations about the lack of an ISE observed in the experiment.

First, the limited statistical power caused by the modest sample size in the present study (N = 10) may be the main reason behind the lack of an ISE observed. Due to the high number of acoustic conditions, each condition was only presented four times to each subject. Therefore, each sound condition was presented 40 times in total throughout the experiment (four repetitions of each sound condition per participant), and this number is very low when compared to that of the typical ISE studies. Despite this low power in our experiment, we investigate and discuss potential causes for the lack of an ISE in the next paragraphs in order to eliminate those while designing the second experiment.

Second, the irrelevant sound set did not consist of a type of stimulus which is known to produce an ISE. Therefore, it is not possible to judge the validity of the experimental procedure.

The third possible explanation comes from the regularity of the NT stimuli: The continuous alternation of the two pulses, ...P1-P2-P1-P2..., might be perceived as two independent sound sources of repeated tokens, instead of a stream of sound tokens changing in acoustic properties. A stream can be defined as the percept of organizing simultaneous or successive sound elements into one coherent sequence, as they all originate from the same sound source (van Noorden, 1975; Bregman, 1990). Depending on the acoustic similarities / differences of the sound elements, a rapid sequence of sounds can be perceived as one (fusion) or more than one stream (fission or stream segregation).

It has been shown that stream segregation can occur without needing the full attention of the listener. This was demonstrated in two ISE studies: Jones *et al.* (1999) employed four sine tone sequences where the alternating tones were zero, two, five or 10 semitones apart. The serial-recall performance was degraded as a function of the tone distance, and reached its minimum when the two tones were five semitones apart. The serial-recall performance increased remarkably in the 10 semitones difference condition, indicating that the participants perceived two repetitive streams instead of one changing in frequency. In the study of Macken *et al.* (2003), the presentation rate of the sequences was varied and it was observed that the serial-recall performance decreased up to the medium rate condition (100 ms separation), and the performance improved when the rate of presentation was increased from medium to high (10 ms separation).

In many studies streaming was induced by differences in the frequency of the sine tones (van Noorden, 1975; Bregman, 1990); differences in the center frequency of band-limited white noise bursts (Dannenbring and Bregman, 1976; Bregman *et al.*, 1999); and increasing the presentation rate (with alternation rates in the range 2–10 per second), by reducing the offset-to-onset interval (Bregman *et al.*, 2000). It was shown that when the difference in the frequencies was small, a higher rate of presentation was needed, and if the presentation rate was low, a larger frequency difference was required to elicit fission. The typical offset-to-onset intervals used in the stream segregation studies (0-120 ms in the studies where the presentation rate was not investigated) are shorter than that of the NT stimuli (200 ms), but longer than what was demonstrated to elicit fission in the study of Macken *et al.* (2003), 10 ms. The fission studies typically employ tasks in which the subjects are instructed to pay attention to the sounds delivered, and make judgments based on what they perceive. When the presented sound is unattended, as in the serial-recall tasks, the required offset-to-onset interval was even shorter, hence making it unlikely that the NT stimuli elicit fission. However, it has been shown that the streaming effect builds up rapidly in 10 seconds, and continues to build up gradually up to 60 seconds (Anstis and Saida, 1985), unless an abrupt change in the stimulus was produced, such as silence (Beauvois and Meddis, 1997; Cusack *et al.*, 2004). The rapid build up time, 10 s, coincides with the presentation of the seventh digit in the serial-recall task employed in this experiment, and it is known that an ISE can be generated until the end of the retention period. So, it is not possible to rule out the possibility that the NT stimuli have elicited fission, however, it should be noted that due to the relatively slow rate of presentation, this is unlikely.

Nevertheless, when the results of the experiment are interpreted together with the three possible explanations listed above, the NT stimuli with low spectral variation, with relatively high $\mathrm{FDCC_{new}}$ values, might have failed to generate an ISE due to the low number of repetitions of the sound conditions. For the NT stimuli with very low $\mathrm{FDCC_{new}}$ values, there is a possibility that participants might have perceived the spectrally modified NT stimuli as two repeating token sequences, resulting in an improved serial-recall performance.

Except for the possibility of fission, the lack of the impact of spectrally modified noise-pulses on serial-recall was especially surprising considering that when band-pass filtered noise bursts with different center frequencies were concatenated to form a changing-noise sequence, it was shown that serial-recall disruption was significantly higher than that of the steady-state noise sequence ($p < .001$) (Tremblay *et al.*, 2001). The noise-pulses used in the present experiment differ from the noise bursts used in that study in terms of frequency characteristics. The band-pass filtered white noise bursts change in center frequency from one token to another, so each token has energy in one band only. On the other hand, the reference noise-pulse used in this experiment comprises equal energy in seven octave bands. The spectral modification was obtained by applying gains with values between 0 and 1, to each band of the reference noise-pulse. The seven gain values were generated with an optimization procedure (see Appendix B) designed to find the optimum gain values that satisfy both a desired $\mathrm{FDCC}_{\mathrm{new}}$ and a constant AMTF value. So the spectral modification can be obtained by altering the energy in one band drastically, or applying moderate gains to each octave band. In the latter scenario, the perception of the frequency domain changes between the subsequent noise-pulses would be very different from the perception of the change in the center frequency of successive noise bursts used in the study of Tremblay *et al.* (2001).

The definition of the $\mathrm{FDCC}_{\mathrm{new}}$ also supports this possibility: The $\mathrm{FDCC}_{\mathrm{new}}$ only quantifies the change / similarity between the power spectra of the successive tokens, and does not take the absolute tone / center frequency difference, tonal distance, into account. When the center frequencies of the successive noise bursts are wider than one-third octave band, as used in the study of Tremblay *et al.* (2001), the corresponding $\mathrm{FDCC}_{\mathrm{new}}$ value would be very close to 0. In fact, a sequence of alternating band-pass filtered noise bursts with the center frequencies of 250 Hz and 500 Hz yields an $\mathrm{FDCC}_{\mathrm{new}}$ value of 0.09. When the bands are two octaves apart (250 Hz and 1000 Hz), the $\mathrm{FDCC}_{\mathrm{new}}$ value is 0.0001. If the center frequencies were within one-third octave bandwidth, it would be close to 1.

The changing-state noise used in the same study was regenerated here and the $\mathrm{FDCC}_{\mathrm{new}}$ value for the changing-noise sequence is 0.05, which is lower than the lowest $\mathrm{FDCC}_{\mathrm{new}}$ value obtained from the NT stimuli used in this experiment (0.06). It can be clearly seen that there is an important property of the changing-state hypothesis that the $\mathrm{FDCC}_{\mathrm{new}}$

is unable to detect and quantify. It is highly possible that this is the tonal distance.

However, the other three issues, the lack of adequate statistical power, the lack of a stimulus which is known to create an ISE, and, although unlikely, the possibility of elicited fission should be eliminated in order to make a solid conclusion about the results of the first experiment. Therefore, a second serial-recall experiment was conducted where all of the aforementioned issues were addressed with a twofold objective: investigating the impact of the temporal and the spectral features of irrelevant sounds on the ISE and investigating the impact of the type of spectral variation on the ISE.

5.5 Experiment 2

It was shown in stream segregation studies that when the B tone in ABAB (van Noorden, 1975) and ABA- sequences(Bregman *et al.*, 2000) was delayed, it was possible for participants to discriminate the temporally asynchronous stimulus more easily when the sound was perceived as one stream. The required duration of the delay depends on the tonal distance for pure tones and the difference in center frequencies for more complex, less tonal stimuli. A similar approach was followed for the present experiment: The asynchrony was produced by extracting the offset-to-onset information from short speech samples by the token selection stage of the FDCC$_{\text{new}}$ algorithm, and arranging the positions of tokens in NT stimuli based on those time gaps. Moreover, since the P1-P2-P1-P2 sequence of the NT stimuli was preserved, we have also introduced short silence segments (0.5-1.5 s) (Beauvois and Meddis, 1997; Cusack *et al.*, 2004) at the end of each short speech-positioned pulse train (1.5-2.5 s) to prevent a build up of the fission effect, in the case the obtained irregular offset-to-onset structure was inadequate. In addition to this, the amplitude and the duration of each token were also modified based on the information derived from the speech samples, so the new NT stimuli were converted into more speech-like stimuli.

A similar approach had been followed in an ISE study, mostly focusing on the rhythmic organization of the irrelevant sound tokens rather than the duration and the amplitude of each one. In the study of Jones and Macken (1995a), a single utterance of “*ah*” was used to create two sequences: An irregular inter stimulus interval (ISI) was created by placing the 300 ms long “*ah*” utterances with an ISI randomly selected from the range of 0, 100, 200, 800, 900 and 1000 ms, and the periodic sequence was generated by keeping an ISI of 500 ms between the subsequent utterances. The letter-recall task (N = 20) showed that the irregular ISI stimuli created a significantly higher mean error rate (M = 23 %) when compared with regular ISI (M = 19.7 %, $p < .05$) and silence (M = 18.5 %, $p < .01$). There was no significant difference between the regular ISI and the control condition.

At first glance the results disagree with the suggested prominence of the role of frequency domain changes in the changing-state hypothesis, however, the authors stated that the irregularity might have created a temporal conjunction of the utterances and this might have served as a basis for the changing-state effect: Two utterances might have been too close to each other, resulting in being perceived as one.

The generation of the irregular ISI in the present experiment does not allow a randomized temporal token grouping, because the information regarding the time gap between successive pulses is extracted from speech stimuli by the token selection stage of the $\mathrm{FDCC_{new}}$. The $\mathrm{FDCC_{new}}$ attempts to divide the speech into syllables, so the time gap between the two syllables would be preserved to a certain extent and it would reduce the possibility of a temporal conjunction effect. As a result, an aperiodic speech-positioned noise-pulse train, without temporal and spectral modifications, is not expected to create an ISE.

Four speech-positioned noise-pulse train (SNT) conditions are formed: a reference SNT ($\mathrm{SNT_{ref}}$, AMTF = 1, $\mathrm{FDCC_{new}}$ =1), a temporally modified SNT (TM, AMTF = 0.7, $\mathrm{FDCC_{new}}$ = 1), and two spectrally modified SNT stimuli ($\mathrm{SM_{Hi}}$, AMTF = 1, $\mathrm{FDCC_{new}}$ = 0.4, $\mathrm{SM_{Lo}}$, AMTF = 1, $\mathrm{FDCC_{new}}$ = 0.7, where the abbreviations of the $\mathrm{SM_{Hi}}$ and the $\mathrm{SM_{Lo}}$ are used to emphasize the magnitude of the spectral distinctiveness). A set of continuous speech samples and the reference NT stimulus (AMTF = 1, $\mathrm{FDCC_{new}}$ = 1) used in the first experiment are also introduced in the second experiment, in order to asses the validity of the experimental procedure and to observe the behaviour of the new $\mathrm{SNT_{ref}}$, respectively.

The hypothesis of the second experiment is similar to that of the first experiment: The spectrally modified noise stimuli are expected to create higher serial-recall disruption than the other noise conditions, and SLNC. The $\mathrm{SM_{Hi}}$ condition is expected to yield similar serial-recall disruption as the continuous speech stimuli and the $\mathrm{SM_{Lo}}$ condition is expected to degrade the performance less than the $\mathrm{SM_{Hi}}$ and speech conditions but more than the others. The TM condition is expected to yield similar mean error rates as the SLNC, reference NT and the SNT$_{\mathrm{ref}}$. To summarize, the serial-recall scores are expected to reflect the FDCC$_{\mathrm{new}}$ values, and not the AMTF values.

The generation of the SNT stimuli, spectral and temporal modifications and the detailed explanations about the acoustic conditions in the second experiment are presented in the following sections.

5.5.1 Speech-positioned noise-pulse train (SNT)

A reference SNT stimulus was generated by using the white-noise pulse, P1, as defined in Eq. 5.2. The information regarding the duration, the amplitude, and the ISI of the reference SNT stimuli was derived from short matrix speech samples (2-4 s) in Dutch (Houben *et al.*, 2014).

The token extraction stage of the FDCC$_{\mathrm{new}}$, which was explained in detail in the second paragraph of Sec. 3.2 in Chapter 3, was used to extract the information regarding the temporal structure of the short speech samples. The information about the duration and the amplitude of each feasible token, as well as the time gap between the successive tokens of the speech samples were extracted.

The reference noise-pulse (Eq. 5.2) was modified based on the aforementioned information obtained from a speech sample for each token, and the resulting Hanning-windowed amplitude and width modified pulses were concatenated based on the time gap information derived for corresponding peaks. An example of two short speech samples and the associated reference SNT stimulus is presented in Fig. 5.5.

It can be observed that the objective of creating a speech-like timing for the noise-pulse train was achieved, and the temporal structure of the short speech sample was reflected to a certain degree. Just like the first experiment, the temporal and spectral modifications were applied only to the even-indexed pulses, while the odd-indexed pulses were kept unmodified.

This technique allowed us to generate a sequence of noise-pulses with a similar approach as in 5.3. The stimuli used in the first experiment consisted of identical noise pulses positioned in an odd and even-indexed fashion, hence were labeled as P1 and P2. For the SNT, this is not the case anymore. However, the same terminology will be kept throughout the rest of the chapter to emphasize the unmodified / modified pulse distinction within a sequence and for the purpose of simplicity.

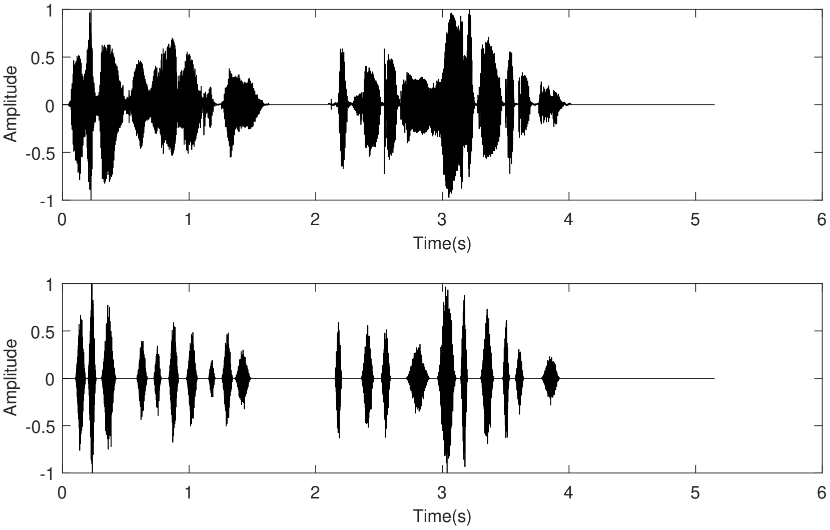


Figure 5.5: Time-domain plot of two short speech samples concatenated (top figure), and the short speech-positioned noise-pulse train (SNT) stimulus generated by the information derived from the same two short speech samples (bottom figure).

Modifying the temporal features of SNT

The temporal features of the SNT were modified by following the same approach explained in Sec. 5.3.1: The width of every P2 was increased in order to change the AMTF value. However, since the time gap between the tokens varies for each P2 in the case of SNT, the width of the P2 was increased to the maximum possible size: The shortest time gap between the P2 and the two neighboring tokens was taken as the maximum half-width size of the modified P2. An example of the reference SNT and the temporally modified SNT is presented in Fig. 5.6.

In experiment 2, 10 long SNT stimuli (40-45 s) were generated by concatenating the temporally modified short SNT stimuli. Only one temporal modification condition was generated, which was obtained by maximizing the width of the P2 of each long SNT stimulus.

Modifying the spectral features of SNT

Two target $\mathrm{FDCC}_{\mathrm{new}}$ values were chosen to be used in experiment 2: the $\mathrm{FDCC}_{\mathrm{new}}$ values of 0.4 and 0.7. The $\mathrm{FDCC}_{\mathrm{new}}$ value of 0.4 was chosen because the speech stimuli investigated in previous chapters typically yielded a value around 0.4. The value of 0.7 was chosen to include an additional level of spectral modification as an acoustic condition in

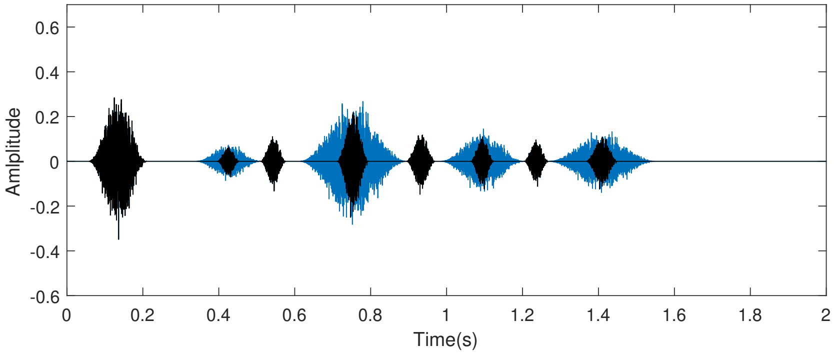


Figure 5.6: Time-domain plot of the reference speech-positioned noise-pulse train (SNT) stimulus (in black), and the temporally modified version of the same short reference SNT stimulus (in blue).

experiment 2.

The gains obtained from the optimization process explained in Sec. 5.3.2 were applied to 800 reference SNT stimuli where the pulse information for each one of them was extracted from a different short speech sample. The AMTF and the FDCC_{new} values of the resulting spectrally modified SNT stimuli were monitored and the ones which satisfy the conditions of a constant AMTF value, with a tolerance of 0.1 (0.9 to 1.1), and the target FDCC_{new} value were stored. For each target FDCC_{new} value, ten 40 to 45 second long SNT stimuli were generated by concatenating 20-22 of the obtained short SNT stimuli. The AMTF and the FDCC_{new} values of the 20 long SNT stimuli were cross-checked after the long SNT stimuli were formed. Each short SNT stimulus was used only once and only for one target FDCC_{new} value.

5.5.2 Method

Participants

Twenty-five participants (14 females and 11 males, age range between 18-50 years) participated who were recruited via the JF Schouten subject database of the Eindhoven University of Technology, Eindhoven, The Netherlands. All participants were students at the University and spoke English as second language. 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 English as a foreign language. They signed the informed consent forms before the experiment began, and the eligibility criteria were double checked.

They were paid a small compensation fee determined by the university department board. The experimental procedure was examined and approved by the Human Technology Interaction department, Eindhoven University of Technology.

Stimuli

In addition to the silence (SLNC), used as the control condition in this experiment, there were six acoustic conditions presented as irrelevant sounds to the participants: reference NT (the reference noise-sequence used in the first experiment) and reference SNT, both with the AMTF and FDCC_{new} values of 1, temporally modified SNT (AMTF = 0.7, FDCC_{new} = 1), two spectrally modified SNT stimuli (AMTF = 1 for both sets with the mean FDCC_{new} values of 0.4 and 0.7), and continuous speech in English. The average sound level of each stimulus was calibrated to 65 $\text{dB}_{\text{LAeq1min}}$.

***Reference NT* ($\mathbf{NT_{ref}}$)**

The NT_{ref} stimulus was the noise-pulse sequence generated by regularly spacing 50 ms long unmodified reference white-noise pulses and was the same as the one used in experiment 1. There was no modification applied to P2, so the AMTF and FDCC_{new} values were 1.

***Reference SNT* ($\mathbf{SNT_{ref}}$)**

The SNT_{ref} condition consisted of 10 unmodified SNT stimuli, where each SNT was generated by concatenating unmodified short speech-positioned noise-pulse trains. The information regarding the temporal structure of each short SNT was extracted from a different short matrix speech sample in Dutch (Houben *et al.*, 2014). Ten 40 to 45 second long SNT_{ref} stimuli were created for the experiment, where each one had a different temporal structure.

***Temporally modified SNT* (TM)**

The TM condition was made up of 10 temporally modified SNT stimuli where the pulse widths of each P2 were increased to the maximum possible, without overlapping the neighboring pulses. As in the SNT_{ref} condition, each short SNT was generated by deriving information from a different short matrix speech sample in Dutch (Houben *et al.*, 2014), and the short SNT stimuli were concatenated to form 10 long (40-45 s) SNT stimuli. In total, 10 long SNT were created and modified by increasing the width of each P2 pulse in each long SNT. The average AMTF value of the 10 long TM stimuli were computed to be 0.7 and the average FDCC_{new} value was 1.

***Spectrally modified SNT* (SM)**

The SM conditions were formed by 10 spectrally modified SNT stimuli with the average $\mathrm{FDCC}_{\mathrm{new}}$ value of 0.4, and 10 with the average $\mathrm{FDCC}_{\mathrm{new}}$ value of 0.7. Each short SM stimulus was generated by using a different short speech sample (Houben *et al.*, 2014) as a source of temporal structure information, and 20 long SM stimuli were formed by concatenating the two groups, based on the $\mathrm{FDCC}_{\mathrm{new}}$ values, of short spectrally modified SNT stimuli. The average AMTF values for each one of the 20 long SM stimuli were between 0.9 and 1.1.

***Continuous speech* (SPCH)**

The speech stimuli were taken from the set of speech samples used in the study of Park *et al.* (2013), and were the recorded samples from "National Public Radio" (http:// www.npr.org), which contained monologues or conversations in English between at most two persons on various topics. The recordings were sliced into 10 long (45-50 s) speech samples. The average $\mathrm{FDCC}_{\mathrm{new}}$ value computed for the 10 long English speech samples was 0.44.

Apparatus

The apparatus used for the current experiment were same as the one reported in Sec. 2.4.1.

Procedure

The GUI used for the serial-recall and the experimental procedure regarding the visual part of the experiment was same as reported in previous experiments, as well as that of experiment 1 in this chapter.

The serial-recall task design consisted of six blocks. The first block was the training block and consisted of 14 trials without any irrelevant sound (silence). When the training block ended, the responsible researcher checked if the participant had fully understood the focal task and if he/she had any questions or problems about the test procedure or the environment before continuing the experiment. The remaining five blocks also consisted of 14 trials each (2 trials for each acoustic condition) and the presentation order of 14 trials was randomized in each block. Six blocks were presented with 2 min breaks between the blocks and each block took approx. 10 min to complete. One experimental session was typically completed in about 65-70 min, including the breaks.

Chapter 5

5.5.3 Results

The evaluation of participants' experimental performance was the same as reported in previous experiments. The acoustic conditions and the test scores, represented as error rates (%), are shown in Fig. 5.7.

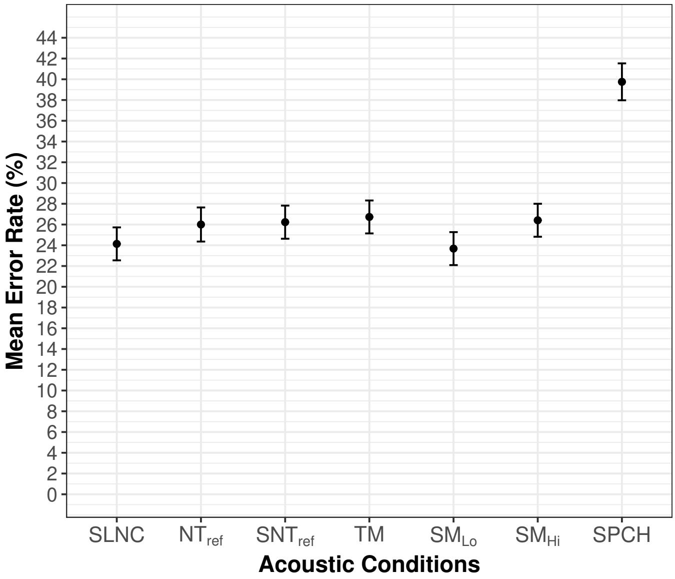


Figure 5.7: Recall performance of 25 participants are represented as mean error rates (%) for the seven acoustic conditions: silence (SLNC), reference noise-pulse train (NT$_{ref}$), reference speech-positioned noise-pulse train (SNT$_{ref}$), temporally modified SNT (TM), two spectrally modified SNT conditions (SM$_{Lo}$ and SM$_{Hi}$) and continuous speech (SPCH). Error bars represent the SEM.

The difference between the mean error rates for SLNC (24.15 %) and SPCH (39.75 %) is slightly higher than what is reported in previous chapters, but still within the range reported in the literature (approx. $\Delta 8-20\%$). The mean error rate for the silent condition is slightly lower than what was observed in experiment 1 (28 %), which are both not uncommon values in the ISE studies.

However, it is visible in Fig. 5.7 that there is no systematic variation in the error rates as a function of the acoustic conditions; the scores in the NT$_{ref}$, SNT$_{ref}$ and in the three modified SNT conditions are almost identical with the result in the SLNC condition. As it appears, the recall performances did not vary in any systematic manner for noise stimuli. The results of the statistical analyses confirmed the lack of effect of the noise-pulse trains, regardless of the modifications applied.

A one way repeated measures ANOVA determined that there was a highly significant impact of background sound on serial recall, $F(6, 144) = 16.44$, $p < .001$, $\eta^2 = 0.04$. Post hoc tests using the Bonferroni correction ($p = 0.005$ for 10 pairs) revealed that the mean error rate for SPCH (M = 39.75, SD = 27.75) was significantly higher than the one for all other acoustic conditions ($p < 0.001$), and there were no other significant differences between the pairs.

5.5.4 Discussion

The results of the experiment showed that there are no significant serial-recall disruption observed except for SPCH. The difference in mean error rates between the SLNC and the SPCH is in line with the range reported in the literature. The $\mathrm{NT}_{\mathrm{ref}}$, the $\mathrm{SNT}_{\mathrm{ref}}$ and the TM conditions did not create a disruption as expected.

On the other hand, the SM conditions did not produce an ISE, which supports the assumption that the tonal distance between the noise-pulses did not reach a sufficient level to produce serial-recall disruption. The $\mathrm{SM}_{\mathrm{Hi}}$ was expected to yield a mean error rate similar to that of the continuous speech, since the $\mathrm{FDCC}_{\mathrm{new}}$ values of both conditions are similar (≈ 0.4), and the $\mathrm{SM}_{\mathrm{Lo}}$ was expected to degrade the performance less than the $\mathrm{SM}_{\mathrm{Hi}}$ and SPCH.

Finally, it can be clearly seen that the values of the metrics, the AMTF and the $\mathrm{FDCC}_{\mathrm{new}}$, were not reflected in the serial-recall results. In fact, it was observed that there is no impact of the noise stimuli on serial-recall performance. The reasons behind this outcome is discussed in the following section in detail, together with the results of the first experiment.

5.6 General discussion

Two experiments were conducted where a reference noise-pulse stimulus was processed specifically to produce a systematic variation in the temporal and the spectral features of the irrelevant sounds, in order to investigate the impact of both features on serial-recall performance, independently. The first experiment employed a periodic alternating NT stimulus and the results showed that the NT stimuli did not generate an ISE, regardless of the modification applied. For the second experiment, the NT stimuli were further processed and converted into a more speech-like sound, SNT stimuli, and a second serial-recall experiment was conducted. The result was the same as that of the first experiment: The noise stimuli used in these experiments did not create any ISE.

As mentioned earlier, the idea of choosing white noise as the basis for the irrelevant sound stimuli in the two experiments was derived from the major findings from the literature: The irrelevant background sounds disrupt serial-recall performance when acoustic properties of the individual tokens vary from one segment to the next, and instrumental music, sine tones, pitch-shifted vowels, as well as native, foreign and reversed speech were shown to be capable of producing an ISE. And most relevant for the stimuli used in this study, band-pass filtered noise bursts disrupted serial-recall performance significantly when the alternating noise bursts had different center frequencies (Tremblay *et al.*, 2001).

The results of the first experiment turned out to be contradictory to the aforementioned findings. The reasons behind the lack of observing an ISE were investigated and four possible explanations were discussed in Sec. 5.4.3: the low sample size, the inability of validating the experimental procedure due to the lack of a stimulus which is known to produce an ISE, the stream segregation possibility due to the periodic structure of the noise-pulses and the lack of sufficient tonal distance between the noise-pulses due to the way the reference pulse was generated. The first three issues were addressed in the second experiment by increasing the number of participants, including a continuous speech condition and generating a speech-like noise stimulus where segments of short silences were integrated. However, the reference pulse used in the first experiment was preserved in order to isolate the possible explanation based on the tonal distance, since it reveals a major limitation of the spectral metric, the FDCC_{new}.

The NT stimuli were turned into a more speech-like sound by deriving information from short speech samples by the peak detection stage of the FDCC_{new}. It was shown in Chapter 4 that when the speech stimuli were segmented into tokens by the FDCC_{new}, the resulting segmented speech created serial-recall disruption similar to that of the continuous speech. So, if FDCC_{new} were an adequate prediction model for the ISE, the SM_{Hi} condition ($\text{FDCC}_{\text{new}} \approx 0.4$) should have created serial-recall disruption to a similar degree as the continuous speech condition ($\text{FDCC}_{\text{new}} \approx 0.4$) used in the same experiment.

The results of the second experiment showed that the spectrally modified SNT stimuli, with FDCC_{new} values of 0.4 and 0.7 did not create any serial-recall disruption. On the other hand, the continuous speech stimuli significantly disrupted the serial-recall which indicates that the experimental procedure was efficient. This outcome supports that the observed lack of any ISE in the first experiment may be due to the insufficient tonal distance between the successive noise-pulses and the $\mathrm{FDCC}_{\mathrm{new}}$ does not quantify that. This is an important limitation of the spectral estimator, especially when the irrelevant sounds are generated based on simple sounds such as sine tones or white noise.

When the speech and SNT stimuli employed in the second experiment are examined, it can be seen that they have similarities: a speech-like rhythm and token-to-token amplitude variation, the magnitude of the spectral similarity between tokens (averaged across tokens), and relatively broader frequency spectrum when compared to two sine tones and band-pass filtered white noise used in the ISE literature. As mentioned earlier, two successive band-limited sound tokens result in an $\mathrm{FDCC}_{\mathrm{new}}$ value of 0 if they don't have energy in common frequency bands. For both the speech and $\mathrm{SM}_{\mathrm{Hi}}$ stimuli, each one of the extracted successive tokens comprises energy in some common frequency bands: The power spectrum varies gradually, instead of making distinct and detached tonal shifts. From this perspective, it can be stated that the impact of the spectral variation within the continuous speech on serial-recall is different than that of the $\mathrm{SM}_{\mathrm{Hi}}$.

Considering that both the SPCH and $\mathrm{SM}_{\mathrm{Hi}}$ conditions did not yield distinct tonal shifts, the tonal distance can not explain the ISE observed for SPCH in the second experiment. The $\mathrm{FDCC}_{\mathrm{new}}$, as a value for the average magnitude of spectral similarity between tokens, produced promising results for SPCH in terms of ISE, but failed to do so for the SNT.

When the speech and the SNT stimuli are re-examined, it is clear that they also possess differences, such as contained semantic information and token set size. The token set size here refers to the number of different tokens in terms of frequency characteristics: Each speech token has a different frequency spectrum while the SNT stimuli were generated by concatenating pulses which have identical frequency spectra, in alternating fashion. However, neither the semantics (Jones *et al.*, 1990), nor the token set size (Tremblay and Jones, 1998) were shown to have any effect on serial-recall performance in previous studies.

The results of the second experiment, when interpreted with a focus on the spectral variation, yield discrepant and confusing outcomes: A magnitude of change in the power spectra between tokens of the speech stimuli is sufficient to create an ISE while the same magnitude of change is not relevant in the ISE, when the stimulus is noise-based. The lack of distinct tonal shifts looks like not important for the ISE when the irrelevant sound is speech, however, there are no other reasons known to us which can explain the lack of disruptive effects observed in $\mathrm{SM}_{\mathrm{Hi}}$ condition.

However, the changing-state hypothesis refrains from ascribing a special role to any stimulus within the context of the ISE, including speech. For instance, there was no difference between the mean error rates observed for pitch-shifted syllable and pitch-shifted sine tone sequences in the study of Jones and Macken (1993, Exp. 5): Speech and non-speech were shown to be equal in terms of their ability to disrupt serial-recall performance. This property of the changing-state hypothesis was labeled as the equipotentiality hypothesis and it was discussed in great detail in the study of LeCompte *et al.* (1997). According to the authors, there were two major issues with the experiments which led to the equipotentiality hypothesis: (1) The sample sizes of Exp. 2 (N = 24) and Exp. 5 (N = 20) in the study of Jones and Macken (1993) were too low to reach a statistical power of 80 %, (2) the use of random permutations of letters *C*, *H*, *J*, *U* in experiment 2, and the use of a syllable, *ah*, in experiment 5, make the speech condition a prototypical example of speech, rather than a meaningful one. A new experiment was designed in the study of LeCompte *et al.* (1997) by addressing these two issues: The sample size was increased to 59 and the irrelevant speech condition was created by using the four words –*hey*, *you*, *me*, *no*– in a randomized order in the sequence. The results, with reported statistical power of approx. 90 %, showed that the irrelevant speech disrupted serial-recall significantly more than the sine tone sequence ($p < 0.001$). A similar outcome was also reported in the study of Jones *et al.* (1999, Exp. 3), where the pitch-shifted vowel *i* disrupted serial-recall more than pitch-shifted sine tones. The mean error rate difference between the conditions was not significant, but authors stated that with sufficient power, the difference might have reached statistical significance.

The difference in serial-recall disruption between the irrelevant speech and tones was explained with the argument that the speech tokens exhibit greater token-to-token acoustic variation, which generates a larger changing-state effect than the simple tones do (Tremblay *et al.*, 2000). It should be noted that the concept of speech tokens was defined as the discrete speech items within the irrelevant sounds, such as words or syllables. Jones and Macken (1993) explicitly treated one utterance as one segment of sound, for instance, the letters were never treated as tokens in a word when there was more than one word presented in the sequence. In addition to this, it can also be argued that ascribing a special definition to the concept of a token for a certain type of sound, based on the acoustic complexity, would also attribute a special role to that particular sound within the context of ISE, at least if a prediction model is of major interest. Nevertheless, the token selection stage of the $\mathrm{FDCC}_{\mathrm{new}}$ was modified to be sensitive enough in terms of segmenting the words into syllables (see Sec. 3.2) and it was shown to be successful when it comes to extracting the parts that comprised disruptive properties of the irrelevant background speech (see Chapter 4).

Regardless of the discussion about the potential stimulus-specific behaviour of the spectral features of the irrelevant sounds on the ISE, it can be clearly seen that the $\mathrm{FDCC}_{\mathrm{new}}$ does not account for a global acoustic determinant of serial-recall disruption: The computed $\mathrm{FDCC}_{\mathrm{new}}$ values of the speech and noise stimuli do not follow the experimental results observed in the second experiment, and the lack of an ISE observed for noise conditions in both experiments prevents us from making a general statement regarding the impact of the temporal and spectral features of the irrelevant sounds on the ISE.

5.7 Conclusion

1. The noise-pulse train and speech-positioned noise-pulse train stimuli did not reduce serial-recall performance when compared to silence in the two experiments.

2. The modification of the temporal features of the stimuli did not produce an ISE: The systematic decrease in the AMTF values of the noise-pulse train stimuli in the first experiment, as well as the only temporally modified speech-positioned noise-pulse train condition in the second experiment failed to disrupt serial-recall performance.

3. The modification of the spectral features of the stimuli did not produce an ISE: The systematic decrease in the $\mathrm{FDCC}_{\mathrm{new}}$ values of the noise-pulse train stimuli in the first experiment, as well as the two spectrally modified speech-positioned noise-pulse train conditions in the second experiment did not produce any ISE.

4. The continuous speech condition employed in the second experiment created statistically significant serial-recall disruption as expected.

5. The $\mathrm{FDCC}_{\mathrm{new}}$ failed to make accurate estimates for the magnitude of the disruption for the noise and speech stimuli: The noise stimuli with low $\mathrm{FDCC}_{\mathrm{new}}$ values did not generate an ISE in the two experiments, and the two conditions, continuous speech and the spectrally modified speech-positioned noise-pulse train ($\mathrm{FDCC}_{\mathrm{new}} = 0.4$), produced significantly different serial-recall performances, but yielded very similar $\mathrm{FDCC}_{\mathrm{new}}$ values.

6 | Summary and conclusions

The studies reported in this thesis focused on investigating the role of spectral features of irrelevant background sounds on serial-recall performance, known as the irrelevant sound effect (ISE) (for a review, see Banburry *et al.*, 2001). In these studies, we had the opportunity to quantify the magnitude of spectral variation within irrelevant sounds using a psychoacoustic metric, the frequency domain correlation coefficient (FDCC) (Park *et al.*, 2013), which was designed to predict the detrimental impact of background sounds on the ISE. The metric was defined as a spectral similarity measure and was inspired by the changing-state hypothesis. The changing-state hypothesis states that the background sound should be segmentable into perceptually distinct tokens and each token should be different form the one that preceded it in order to produce an ISE (Jones and Macken, 1993). The metric attempts to transform the definition of the hypothesis into a psychoacoustic parameter: The FDCC is derived by segmenting the irrelevant sounds into tokens and computing the correlation between the successive tokens in the frequency domain, which was shown to be an important acoustic feature of the irrelevant sounds (e.g., Jones and Macken, 1993; Jones *et al.*, 2000; Ellermeier *et al.*, 2015; Senan *et al.*, 2018). The FDCC was evaluated and modified throughout the thesis using various types of stimuli in each chapter: noise-vocoded speech (NVS), a large set of stimuli from the literature (e.g., native, foreign and babble speech, music, office noise, traffic sounds, artificial animal sounds, etc.), various types of segmented stimuli, selectively reversed NVS, periodic and speech-positioned noise-pulse train stimuli.

The second chapter reported two studies where a set of distorted speech stimuli, NVS, was employed in serial-recall experiments which allowed us to evaluate the role of spectral variation on the ISE using the FDCC_{old}. The results showed that the FDCC_{old} was able to follow the trend of the serial-recall disruption up to an extent (6-band NVS, $\text{FDCC}_{\text{old}} = 0.72$) but failed beyond that. Three other metrics from the literature, the fluctuation strength (FS), the speech transmission index (STI) and the normalized covariance measure (NCM) were also evaluated in the same study and it was shown that the STI and the NCM were as successful as the FDCC_{old} in terms of following the trend of serial-recall performance: The noise-vocoding technique used in generating the NVS stimuli in the experiments produced a systematic change in both the frequency and the time domain hence the parameter values of the FDCC_{old} and the two temporal metrics, the STI and the NCM, varied systematically as the number of frequency bands increased. However, the serial-recall results did not follow the systematic change observed in the parameter values of the three metrics and none of the metrics was successful in predicting the ISE for the NVS conditions.

In the Chapter 3, the token selection stage of the FDCC_{old} was modified and the change was evaluated employing a large set of data (N = 91) from the ISE literature. The parameter values of the two versions of the FDCC, the FDCC_{old} and the FDCC_{new}, as well the FS were computed for the stimuli. Pearson correlation values between the experimental results and the values of the three parameters were compared. The results showed that the change in the token selection stage of the FDCC resulted in an improved Pearson's r ($r = 0.70$, $p < 0.01$) when compared to the FDCC_{old} ($r = 0.59$, $p < 0.01$). The correlation between the error rates of the total dataset and the FDCC_{new} values was significantly higher than what was observed for the FS ($r = 0.50$, $p < 0.01$). In addition to that, the FDCC_{new} and the FS were evaluated using the speech / masked speech stimuli only (N = 21) and the correlation between the FDCC_{new} ($r = 0.53$, $p < 0.05$) and the normalized mean error rates was higher than the correlation computed using the FS values ($r = 0.08$, $p > 0.05$). It was concluded that for the total set of behavioral measurements used in the study, the two metrics, the FDCC_{new} and the FS, were equally successful in predicting the ISE and for the masked / degraded speech stimuli, the FDCC_{new} parameter produced more promising values than the FS.

The change in the token selection stage of the FDCC_{new} was further evaluated in Chapter 4 using a serial-recall task where the segmented versions of the irrelevant sounds from the literature (Schlittmeier *et al.*, 2012; Senan *et al.*, 2018) were employed alongside a continuous speech condition. The serial-recall results observed for the segmented music, 1-

band and 6-band NVS, as well as the office-noise stimuli were compared with experimental results reported in the associated literature for the continuous versions of these conditions. The segmented and the continuous speech conditions were within-subjects variables. The results showed that the normalized mean error rates for the segmented (17 %) and continuous speech (18.7 %) conditions were similar. In addition to that, the serial-recall results observed for the segmented 1-band and music conditions were also similar to the results reported in the literature ($p > 0.05$). However, the segmented office-noise (normalized median error rate = 2.2 %) and the segmented 6-band NVS (normalized mean error rate = 1.8 %) produced different results when compared to the results reported in the literature for continuous office noise (normalized median error rate = 7 %) and 6-bands (normalized mean error rate = 8.05 %) conditions. In the second experiment, a specific type of NVS stimuli were collected from the ISE literature and were evaluated with respect to the experimental results reported in the original study (Dorsi *et al.*, 2018). The original NVS stimuli were modified by temporally reversing the lower two-thirds of the frequency bands of the NVS and this had resulted with an increase in serial-recall performance when compared to the original NVS. The authors had concluded that the reduction in speech fidelity was the reason behind the improvement in the performance. When the two sets of NVS stimuli were run through the $\mathrm{FDCC}_{\mathrm{new}}$ algorithm, it was observed that the reversal technique changed the position of one of the tokens and the single token with energy only in high frequency bands was integrated into a token with broader frequency spectrum which eventually increased the $\mathrm{FDCC}_{\mathrm{new}}$ values. It was concluded that the role of speech fidelity on the serial-recall performance might be lower than what was claimed to be in the original study and the token selection stage of the $\mathrm{FDCC}_{\mathrm{new}}$ was sensitive enough to detect that.

The results observed in Chapter 2 were further investigated by modifying the temporal and the spectral features of a noise-pulse train (NT) stimulus and employing the set of NT stimuli in a serial-recall task. The spectral modification was monitored by the $\mathrm{FDCC}_{\mathrm{new}}$ and the temporal modification was controlled by another metric, the average modulation transfer function (AMTF), which is a parameter similar to the STI. The first experiment showed that when the temporal and spectral features of the periodically positioned NT stimulus were modified independently from each other, there was no change in the serial-recall performance when compared to the silence. The second experiment in the same study varied the temporal positions of the noise pulses using the temporal information of speech stimuli extracted by the token selection stage of the FDCC$_{\text{new}}$ and employed the speech-positioned noise-pulse train (SNT) stimuli in a serial-recall experiment alongside a continuous speech condition. The results showed that while the continuous speech created a significant serial-recall disruption (normalized mean error rate = 15.6 %, $p < 0.001$), the SNT stimuli failed to disrupt the performance, even though the FDCC$_{\text{new}}$ values of one of the SNT conditions (SM$_{\text{Hi}}$, FDCC$_{\text{new}}$ = 0.4, normalized mean error rate = 2.28 %), and the continuous speech (FDCC$_{\text{new}}$ = 0.44) were very close. It was concluded that the magnitude of the spectral variation quantified by the FDCC$_{\text{new}}$ parameter could not account for the lack of ISE observed for the SNT conditions.

As summarized above, the spectral estimator was evaluated and modified throughout this thesis and certain advantages and limitations were observed. The advantages of the use of the FDCC$_{\text{new}}$ are discussed in the next section and the limitations, as well as further research proposals to overcome those limitations, are presented in Sec. 6.2.

6.1 Advantages of the FDCC$_{\text{new}}$

The main advantage of the spectral metric is that it was proposed as an ISE predictor solely and it attempts to provide a mathematical basis for the changing-state hypothesis. The lack of an quantitative estimator for the changing-state effect has been mentioned as a limitation in the literature (e.g., Schlittmeier *et al.*, 2008, 2012) and the FDCC$_{\text{new}}$ attempts to fill this gap by following the definition of the changing-state hypothesis. With the use of the FDCC$_{\text{new}}$ in the ISE studies, the variations introduced to the experimental stimulus can be measured quantitatively and the reasons behind the serial-recall disruption or the lack of it can be classified and / or distinguished from each other.

An example of the strength of this analytical approach was given in the second experiment reported in Chapter 4: The selective-reversal technique used in generating the reversed NVS stimulus simultaneously modified the spectro-temporal features of the NVS and this change was quantified by the FDCC$_{\text{new}}$. The use of the spectral metric allowed us to observe that the reversal technique used in the original study might not be the ideal method for investigating the role of the speech fidelity on the ISE, at least when used in combination with that particular stimulus.

A second example from this thesis comes from the results of the second experiment reported in Chapter 5: The SNT stimuli with an $\mathrm{FDCC}_{\mathrm{new}}$ value of 0.4 and the continuous speech with the same parameter value produced significantly different serial-recall disruption. It can be clearly stated that the role of the spectral variation on the ISE, as quantified by the $\mathrm{FDCC}_{\mathrm{new}}$, is not the same for the SNT and the original speech stimulus. For the SNT stimulus, the change in frequency from one pulse to the next one can not explain the lack of ISE by itself. Another example where the $\mathrm{FDCC}_{\mathrm{new}}$ can be put into good use as an analytical basis in auditory distraction research comes from the literature. For instance, the studies where a deviant sound (female spoken letter "B") had been added into a changing-state sequence of male spoken letters ("$\mathrm{ABAB}_{\mathrm{female}}\mathrm{AB}$"), it had shown an increase in the serial-recall disruption when compared to the changing-state sequence with no deviant sound (Hughes *et al.*, 2007). The computation of the $\mathrm{FDCC}_{\mathrm{new}}$ values for the two stimuli would give a clear idea regarding the difference between the magnitude of the spectral variation of the two stimuli and therefore might serve as a basis to distinguish the effect of the deviant sound from the changing-state sequence.

Another advantage of the metric is that it can provide a value for any sound that is segmentable. Similar to the FS model and in contrast to the STI and the NCM models, it does not require a reference signal. When compared with the FS model, it was also observed that the $\mathrm{FDCC}_{\mathrm{new}}$ produced better results for masked / degraded speech stimuli and although being far from perfect, it can be used as an ISE predictor in room acoustics research where noise masking systems are used to improve the performance of office workers in open-plan office settings (e.g., Haka *et al.*, 2009; Park *et al.*, 2013).

6.2 Limitations of the $\mathrm{FDCC}_{\mathrm{new}}$ and ideas for future research

The studies reported in this thesis allowed us to observe several limitations of the model with respect to its ability to predict the ISE. The critical limitations of the metric as well as some ideas for future research to overcome these limitations are presented below.

The results of the first experiment showed that the $\mathrm{FDCC}_{\mathrm{old}}$ values continued to decrease while there was no further serial-recall disruption observed beyond the 6-band NVS condition. Similar results were observed in the literature where NVS was used as irrelevant sound stimulus (Ellermeier *et al.*, 2015) and the serial-recall disruption did not follow the number of frequency bands used in the NVS beyond a critical point. It indicates that there may be some perceptual features of the changing-state hypothesis that the FDCC_{new} is unable to capture. In addition to the perceptual limit that was observed in the NVS studies, the FDCC_{new} is also not able to detect the auditory stream segregation effect (see Chapter 5 for a detailed discussion) which was shown to have an impact on the ISE (Jones *et al.*, 1999). The lack of perceptual aspects of the metric was known to us and one of the critical features of the changing-state hypothesis, the role of segmentation on the ISE, was investigated in Chapter 4. Moreover, an attempt was also made to quantify the just noticeable differences of the metric (see Appendix C), however, the experiment was unsuccessful and therefore the results were not taken into account in this thesis. Nevertheless, the metric would benefit most from the line of research that aims at improving the perceptual aspects of the spectral parameter. Furthermore, the spectral metric currently works as a monaural prediction model and the studies from the literature delivered evidence that the stream segregation (Jones and Macken, 1995a) as well as stream segmentation (Jones and Macken, 1995b) can occur based on spatial cues and both have an impact on the ISE (Jones and Macken, 1995a).

The results reported in Chapter 4 revealed another limitation of the metric: The FDCC_{new} does not possess a criterion for the temporal distance between the successive tokens and treats each token equally in terms of its disruptive capacity. The segmented office noise stimuli used in the first experiment in Chapter 4 consisted of long segments of steady-state noise which might have an impact on the serial-recall performance. This may not be a limitation for speech and speech-like stimuli but for sounds with long steady-state segments, this may be be a problem. The token distance criterion can be implemented after investigating the minimum duration of silence that is required to increase the serial-recall performance and the FDCC_{new} values can be adapted if such long segments of silence are detected in the irrelevant sounds.

The biggest limitation of the metric was observed in Chapter 5: It is not capable of quantifying a tonal distance between successive segments of the sound. If the change in frequency between the successive sine tones is more than one third-octave apart, the FDCC_{new} results in a value of 0 and if it is less than one-third octave apart, the FDCC_{new} produces a value close to 1. This limitation is mostly derived from the use of the correlation as a method to quantify the similarity of the two magnitude spectra: The correlation value between the magnitude spectra of the two sine tones with different frequencies is 0. It is very unlikely to improve this limitation by modifying the current definition of the spectral metric since quantifying the tonal distance between the successive tokens would require a more elaborate approach. An example for such an approach can be to follow a very different method and integrate an auditory model into the second stage of the algorithm which was recently proposed as a perceptual metric for assessing the similarity of musical sounds (Osses, 2018).

6.3 General conclusion

The major objective of the this thesis was to evaluate the FDCC as an ISE predictor and this was carried out by designing serial-recall experiments accompanied by sets of irrelevant stimuli which were crafted with the aim of revealing the advantages and the limitations of the parameter. Due to the definition of the metric, the studies reported in the thesis focused on the spectral variation within the irrelevant sounds and its relation to the ISE. Particular interest was concentrated on generating stimuli which allowed us to combine and distinguish acoustic and speech-specific features of the irrelevant sounds, such as noise-vocoded speech and SNT stimuli. The motivation for this choice was to gain further insights into the role of spectral variation on the ISE as well as to investigate the boundaries of the metric, and if possible, to improve it. The reported studies revealed that the relation between the magnitude of the spectral variation in the irrelevant sounds and the ISE is not linear (Chapter 2), the change in the token selection stage of the algorithm improved the accuracy of the metric (Chapter 3), the FDCC$_{\text{new}}$ can be a promising ISE predictor for speech and degraded / masked speech stimuli (Chapter 3) and the metric can be used as a quantitative measure in the ISE studies (Chapter 4 and 5).

References

ANSI **(1997)**. "Methods for calculation of the speech intelligibility index," S3.5–1997 (American National Standards Institute, New York).

Anstis, S. M., and Saida, S. **(1985)**. "Adaptation to auditory streaming of frequency-modulated tones," J. Exp. Psychol. Hum. Percept. Perform. **11**(3), 257–271.

Baddeley, A. D. **(1997)**. *Human memory: Theory and practice.* (Allyn & Bacon, Massachusetts)

Baddeley, A. D. **(2000)**. "The episodic buffer: a new component of working memory?," Trends. Cogn. Sci. **4**(11), 417–423.

Baddeley, A. D. **(2003)**. "Working memory: looking back and looking forward," Nature Rev. Neurosci. **4**(10), 829–839.

Baldwin, C. L. **(2016)**. *Auditory cognition and human performance: Research and applications.* (CRC Press, Florida).

Banbury, S., and Berry, D. C. **(1998)**. "Disruption of office-related tasks by speech and office noise," Br. J. Psychol. **89**(3), 499–517.

Banbury, S. P., Macken, W. J., Tremblay, S., and Jones, D. M. **(2001)**. "Auditory distraction and short-term memory: Phenomena and practical implications," Hum. Factors **43**, 12–29.

Beauvois, M. W., and Meddis, R. **(1997)**. "Time decay of auditory stream biasing," Percept. Psychophys. **59**(1), 81–86.

Biley, F. C. **(1994)**. "Effects of noise in hospitals,". Br. J. Nurs. **3**(3), 110–113.

Bregman A. S. **(1994)**. *Auditory scene analysis: The perceptual organization of sound.* (MIT Press, Massachusetts).

Bregman, A. S., Colantonio, C., and Ahad, P. A. **(1999)**. "Is a common grouping mechanism involved in the phenomena of illusory continuity and stream segregation?," Percept. Psychophys. **61**(2), 195–205.

Bregman, A. S., Ahad, P. A., Crum, P. A., and O'Reilly, J. **(2000)**. "Effects of time intervals and tone durations on auditory stream segregation," Percept. Psychophys. **62**(3), 626–636.

Broadbent, D. E. **(1982)**. "Task combination and selective intake of information," Acta Psychol. (AMST) **50**(3), 253–290.

Buchner, A. **(1996)**. "On the irrelevance of semantic information for the llIrrelevant speech effect." Q. J. Exp. Psychol. **49A**(3), 765–779.

Chen, F., and Loizou, P. C. **(2010)**. "Contribution of consonant landmarks to speech recognition in simulated acoustic-electric hearing," Ear Hear. **31**(2), 259–267.

Chen, F. **(2011)**. "Intelligibility prediction for distorted sentences by the normalized covariance measure," Int. J. Speech Technol. **14**(3), 237–243.

Chen, F., and Loizou, P. C. **(2011a)**. "Predicting the intelligibility of vocoded speech," Ear Hear. **32**(3), 331–338

Chen, F., and Loizou, P. C. **(2011b)**. "Predicting the intelligibility of vocoded and wideband Mandarin Chinese," J. Acoust. Soc. Am. **129**(5), 3281–3290.

Colle, H. A., and Welsh, A. **(1976)**. "Acoustic masking in primary memory," J. Mem. Lang. **15**(1), 17–31.

Colle, H. A. **(1980)**. "Auditory encoding in visual short-term recall: Effects of noise intensity and spatial location," J. Mem. Lang. **19**(6), 722–735.

Cusack, R., Decks, J., Aikman, G., and Carlyon, R. P. **(2004)**. "Effects of location, frequency region, and time course of selective attention on auditory scene analysis," J. Exp. Psychol. Hum. Percept. Perform. **30**(4), 643–656.

Dannenbring, G. L., and Bregman, A. S. **(1976)**. "Stream segregation and the illusion of overlap," J. Exp. Psychol. Hum. Percept. Perform. **2**(4), 544–555.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., and McGettigan, C. **(2005)**. "Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences," J. Exp. Psychol. Gen. **134**(2), 222–241.

Dorman, M. F., Loizou, P. C., and Rainey, D. **(1997)**. "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," J. Acoust. Soc. Am. **102**(4), 2403–2411.

Dorsi, J. **(2013)**. "Recall disruption produced by noise-vocoded speech: A study of the irrelevant sound effect," Master thesis, State University of New York, New Paltz, NY, http://hdl.handle.net/1951/63031.

Dorsi, J., Viswanathan, N., Rosenblum, L. D., and Dias, J. W. **(2018)**. "The role of speech fidelity in the irrelevant sound effect: Insights from noise-vocoded speech backgrounds," Q. J. Exp. Psychol. A. **71**(10), 2152–2161.

Ellermeier, W., and Hellbrück, J. **(1998)**. "Is level irrelevant in "irrelevant speech"? Effects of loudness, signal-to-noise ratio, and binaural unmasking," J. Exp. Psychol. Hum. Percept. Perform. **24**(5), 1406–1414.

Ellermeier, W., and Zimmer, K. **(2014)**. "The psychoacoustics of the irrelevant sound effect: A review," Acoust. Sci. and Technol. **35**, 10–16.

Ellermeier, W., Kattner, F., Ueda, K., Doumoto, K., and Nakajima, Y. **(2015)**. "Memory disruption by irrelevant noise-vocoded speech: Effects of native language and the number of frequency bands," J. Acoust. Soc. Am. **138**(3), 1561–1569.

Fastl, H. **(1982)**. "Fluctuation strength and temporal masking patterns of amplitude-modulated broadband noise," Hear. Res. **8**(1), 59–69.

Fastl, H., and Zwicker, E. **(2007)**. *Psychoacoustics: Facts and Models*, 3rd ed. (Springer, Berlin), Chap. 10, pp. 247–256.

Fisher, R. A. **(1921)**. "On the probable error of a coefficient of correlation deduced from a small sample," Metron **1**, 3–32.

Francart, T., Van Wieringen, A., and Wouters, J. **(2008)**. "APEX 3: a multi-purpose test platform for auditory psychophysical experiments," J. Neurosci. Methods. **172**(2), 283–293.

Galbrun, L., and Ali, T. **(2012)**. "Perceptual assessment of water sounds for road traffic noise masking," in *Proceedings of the joint SFA-IOA Acoustics 2012 Nantes Conference, 23-27 April*, (Nantes, France), pp. 2147–2152.

Goldsworthy, R., and Greenberg, J. **(2004)**. "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," J. Acoust. Soc. Am. **116**(3), 3679–3689.

Greenwood, D. D. **(1961)**. "Auditory masking and the critical band," J. Acoust. Soc. Am. **33**(4), 484–502.

Greenwood, D. D. **(1990)**. "A cochlear frequency-position function for several species–29 years later," J. Acoust. Soc. Am. **87**(6), 2592–2605.

Haapakangas, A., Hongisto, V., Hyönä, J., Kokko, J., and Keränen, J. **(2014)**. "Effects of unattended speech on performance and subjective distraction: The role of acoustic design in open-plan offices," Appl. Acoust. **86**, 1–16.

Haka, M., Haapakangas, A., Keränen, J., Hakala, J., Keskinen, E., and Hongisto, V. **(2009)**. "Performance effects and subjective disturbance of speech in acoustically different office types–a laboratory experiment," Indoor Air, **19**(6), 454–467.

Hermann, T., and Hunt, A. **(2011)**. *The Sonification Handbook*, J. G. Neuhoff (Ed.). (Logos Verlag, Berlin), pp. 399–425.

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

Houben, R., Koopman, J., Luts, H., Wagener, K. C., Van Wieringen, A., Verschuure, H., and Dreschler, W. A. (2014). "Development of a Dutch matrix sentence test to assess speech intelligibility in noise," Int. J. Audiol. **53**(10), 760–763.

Houtgast, T., and Steeneken, H. J. M. **(1973)**. "The modulation transfer function in room acoustics as a predictor of speech intelligibility," Acta Acust. united Ac. **28**(1), 66–73.

Houtgast, T., and Steeneken, H. J. M. **(1985)**. "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acoust. Soc. Am. **77**(3), 1069–1077.

Hughes, R. W., Tremblay, S., and Jones, D. M. **(2005)**. "Disruption by speech of serial short-term memory: The role of changing-state vowels," Psychon. Bull. Rev. **12**(5), 886–890.

Hughes, R. W., Vachon, F., and Jones, D. M. **(2005)**. "Auditory attentional capture during serial recall: Violations at encoding of an algorithm-based neural model?," J. Exp. Psychol. Learn. Mem. Cogn. **31**(4), 736–749.

Hughes, R. W., Vachon, F., and Jones, D. M. **(2007)**. "Disruption of short-term memory by changing and deviant sounds: support for a duplex-mechanism account of auditory distraction," J. Exp. Psychol. Learn. Mem. Cogn. **33**(6), 1050–1061.

Hughes, R. W., Hurlstone, M. J., Marsh, J. E., Vachon, F., and Jones, D. M. **(2013)**. "Cognitive control of auditory distraction: impact of task difficulty, foreknowledge, and working memory capacity supports duplex-mechanism account," J. Exp. Psychol. Hum Percept. Perform. **39**(2), 539–553.

Hughes, R. W. **(2014)**. "Auditory distraction: A duplex-mechanism account," Psych J. **3**(1), 30–41.

International Electrotechnical Commission **(2003)**. "Sound System Equipment-Part 16: Objective rating of speech intelligibility by speech transmission index," International Standard IEC 60268–16.

ITU-T **(2000)**. "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU-T Recommendation P. 862.

ITU-T **(2005)**. "P.862: Revised Annex A - Reference implementations and conformance testing for ITU-T Recs P.862, P.862.1 and P.862.2," url: http://www.itu.int/rec/T-REC-P.862-200511-I!Amd2/en

Jahncke, H., Hongisto, V., and Virjonen, P. **(2013)**. "Cognitive performance during irrelevant speech: Effects of speech intelligibility and office-task characteristics," Appl. Acoust. **74**(3), 307–316.

Jones, D. M., and Miles, C., and Page, J. **(1990)**. "Disruption of proofreading by irrelevant speech: Effects of attention, arousal or memory?," Appl. Cogn. Psychol. **4**(2), 89–108.

Jones, D., Madden, C., and Miles, C. **(1992)**. "Privileged access by irrelevant speech to short-term memory: The role of changing state," Q. J. Exp. Psychol. A. **44**(4), 645–669.

Jones, D. **(1993)**. "Objects, streams, and threads of auditory attention,", in *Attention: Selection, awareness, and control: A tribute to Donald Broadbent*, edited by A. D. Baddeley and L. Weiskrantz (Clarendon Press/Oxford University Press, New York), pp. 87–104.

Jones, D. M., Macken, W. J., and Murray, A. C. **(1993)**. "Disruption of visual short-term memory by changing-state auditory stimuli: The role of segmentation," Mem. Cognit. **21**(3), 318–328.

Jones, D. M., and Macken, W. J. **(1993)**. "Irrelevant tones produce an irrelevant speech effect: Implications for phonological coding in working memory," J. Exp. Psychol. Learn. Mem. Cogn. **19**(2), 369–381.

Jones, D. M., and Macken, W. J. **(1995a)**. "Organizational factors in the effect of irrelevant speech: The role of spatial location and timing," Mem. Cognit **23**(2), 192–200

Jones, D. M., and Macken, W. J. **(1995b)**. "Auditory babble and cognitive efficiency: Role of number of voices and their location," J. Exp. Psychol-Appl. **1**(3), 216–226.

Jones, D. M., Beaman, C. P., and Macken, W. J. **(1996)**. "The object-oriented episodic record model," in *Models of short-term memory*, edited by S. E. Gathercole (Psychology Press, London), pp. 209–238.

Jones, D. M. **(1999)**. "The cognitive psychology of auditory distraction: The 1997 BPS Broadbent Lecture," Br. J. Clin. Psychol. **90**(2), 167–187.

Jones, D. M., Alford, D., Bridges, A., Tremblay, S., and Macken, W. J. **(1999)**. "Organizational factors in selective attention: The interplay of acoustic distinctiveness and auditory streaming in the irrelevant sound effect," J. Exp. Psychol. Learn. Mem. Cogn. **25**(2), 464–473.

Jones, D. M., and Tremblay, S. **(2000)**. "Interference in memory by process or content? A reply to Neath," **(2000)**. Psychon. Bull. Rev. **7**(3), 550–558.

Jones, D. M., Alford, D., Macken, W. J., Banbury, S. P., and Tremblay, S. **(2000)**. "Interference from degraded auditory stimuli: Linear effects of changing-state in the irrelevant sequence," J. Acoust. Soc. Am. **108**(3), 1082–1088.

Kates, J. **(1992)**. "On using coherence to measure distortion in hearing aids," J. Acoust. Soc. Am. **91**, 2236–2244.

Kates, J., and Arehart, K. **(2005)**. "Coherence and the speech intelligibility index," J. Acoust. Soc. Am. **117**, 2224-–2237.

Klapp, S. T., Marshburn, E. A., and Lester, P. T. **(1983)**. "Short-term memory does not involve the" working memory" of information processing: The demise of a common assumption," J. Exp. Psychol. Gen. **112**(2), 240–264.

Lange, E. B. **(2005)**. "Disruption of attention by irrelevant stimuli in serial recall," J. Mem. Lang. **53**(4), 513–531.

LeCompte, D. C., and Shaibe, D. M. **(1997)**. "On the irrelevance of phonological similarity to the irrelevant speech effect," Q. J. Exp. Psychol. A. **50**(1), 100–118.

LeCompte, D. C., Neely, C. B., and Wilson, J. R. **(1997)**. "Irrelevant speech and irrelevant tones: The relative importance of speech to the irrelevant speech effect," J. Exp. Psychol. Learn. Mem. Cogn. **23**(2), 472–483.

Levitt, H. **(1971)**. "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **29**(2B), 467–477.

Liebl, A., Assfalg, A., and Schlittmeier, S. J. **(2016)**. "The effects of speech intelligibility and temporal–spectral variability on performance and annoyance ratings," Appl. Acoust. **110**, 170–175.

Ljung, R., Sörqvist, P., Kjellberg, A., and Green, A. M. **(2009)**. "Poor listening conditions impair memory for intelligible lectures: implications for acoustic classroom standards," Building Acoustics **16**(3), 257–265.

Loizou, P. C., Dorman, M., and Tu, Z. **(1999)**. "On the number of channels needed to understand speech," J. Acoust. Soc. Am. **106**(4), 2097–2103.

Ma, J., Hu, Y., and Loizou, P. C. **(2009)**. "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," J. Acoust. Soc. Am. **125**(5), 3387–3405.

Macken, W. J., Tremblay, S., Houghton, R. J., Nicholls, A. P., and Jones, D. M. **(2003)**. "Does auditory streaming require attention? Evidence from attentional selectivity in short-term memory," J. Exp. Psychol. Learn. Mem. Cogn. **(29)**(1), 43–51.

Marsh, J. E., Hughes, R. W., and Jones, D. M. **(2008)**. "Auditory distraction in semantic memory: A process-based approach," J. Mem. Lang. **58**(3), 682–700.

Marsh, J. E., Hughes, R. W., and Jones, D. M. **(2009)**. "Interference by process, not content, determines semantic auditory distraction," Cognition **110**(1), 23–38.

Marsh, J. E., and Jones, D. M. **(2010)**. "Cross-modal distraction by background speech: What role for meaning?" Noise Health **12**(49), 210–216.

Martin, R. C., Wogalter, M. S., and Forlano, J. G. **(1988)**. "Reading comprehension in the presence of unattended speech and music," J. Mem. Lang. **27**(4), 382–398.

Miles, C., Jones, D. M., and Madden, C. A. **(1991)**. "Locus of the irrelevant speech effect in short-term memory," J. Exp. Psychol. Learn. Mem. Cogn. **17**(3), 578–584

Moore, B., and Glasberg, B. **(1993)**. "Suggested formulas for calculation auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. **74**, 750–753.

Nairne, J. S. **(1990)**. "A feature model of immediate memory," Mem. Cognit. **18**(3), 251–269.

Neath, I. **(2000)**. "Modeling the effects of irrelevant speech on memory," Psychon. Bull. Rev. **7**(3), 403–423.

Osses Vecchi, A. A. **(2018)**. "Prediction of perceptual similarity based on time-domain models of auditory perception." (Doctoral dissertation). Eindhoven: Technische Universiteit Eindhoven.

Parizet, E., Ellermeier, W., and Robart, R. **(2014)**. "Auditory warnings for electric vehicles: Detectability in normal-vision and visually-impaired listeners," Appl. Acoust. **86**, 50–58.

Park, M., Kohlrausch, A., and van Leest, A. **(2013)**. "Irrelevant speech effect under stationary and adaptive masking conditions," J. Acoust. Soc. Am. **134**(3), 1970–1981.

Perham, N., and Vizard, J. **(2011)**. "Can preference for background music mediate the irrelevant sound effect?," Appl. Cogn. Psychol. **25**(4), 625–631.

Plomp, R., and Mimpen, A. M. **(1979)**. "Improving the reliability of testing the speech reception threshold for sentences," Audiology **18**, 43–52.

van de Poll, M. K., Ljung, R., Odelius, J., and Sörqvist, P. **(2014)**. "Disruption of writing by background speech: The role of speech transmission index," Appl. Acoust. **81**, 15–18.

Reinten, J., Braat-Eggen, P. E., Hornikx, M., Kort, H. S., and Kohlrausch, A. **(2017)**. "The indoor sound environment and human task performance: A literature review on the role of room acoustics," Build. Environ. **123**, 315–332.

Rix, A., Beerends, J., Hollier, M., and Hekstra, A. **(2001)**. "Perceptual evaluation of speech quality (PESQ) – A new method for speech quality assessment of telephone networks and codecs," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. **2**, pp. 729–752

Roberts, B., Summers, R. J., and Bailey, P. J. **(2011)**. "The intelligibility of noise-vocoded speech: spectral information available from across-channel comparison of amplitude envelopes," Proc. R. Soc. Lond., B, Biol. Sci. **278**, 1595–1600.

Salamé, P., and Baddeley, A. **(1982)**. "Disruption of short-term memory by unattended speech: Implications for the structure of working memory," J. Verbal Learning Verbal Behav. **21**, 150–164.

Schlittmeier, S. J., Hellbrück, J., and Klatte, M. **(2008)**. "Does irrelevant music cause an irrelevant sound effect for auditory items?" Eur. J. Cogn. Psychol. **20**(2), 252–271.

Schlittmeier, S. J., Weißgerber, T., Kerber, S., Fastl, H., and Hellbrück, J. **(2012)**. "Algorithmic modeling of the irrelevant sound effect (ISE) by the hearing sensation fluctuation strength," Atten. Percept. Psychophys. **74**, 194–203.

Senan, T. U., Park, M., Kohlrausch, A., Jelfs, S., and Navarro, R. F. **(2015)**. "Spectral and temporal features as the estimators of the irrelevant speech effect," in *Proceedings of Euronoise 2015*, edited by C. Glorieux (Maastricht, The Netherlands), pp. 1925–1930.

Senan, T. U., Jelfs, S and Kohlrausch, A. **(2018)**. "Cognitive disruption by noise-vocoded speech stimuli: Effects of spectral variation," J. Acoust. Soc. Am. **143**(3), 1407–1416.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. **(1995)**. "Speech recognition with primarily temporal cues," Science **270**(5234), 303–304.

Sörqvist, P. **(2010)**. "High working memory capacity attenuates the deviation effect but not the changing-state effect: Further support for the duplex-mechanism account of auditory distraction," Mem. Cognit. **38**(5), 651–658.

Sörqvist, P. **(2014)**. "On interpretation and task selection in studies on the effects of noise on cognitive performance," Front. Psychol. **5**(1249), 1–4.

Steeneken, H. J. M., and Houtgast, T. **(1980)**. "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. **67**, 318–326.

Tremblay, S., and Jones, D. M. **(1998)**. "Role of habituation in the irrelevant sound effect: Evidence from the effects of token set size and rate of transition," J. Exp. Psychol. Learn. Mem. Cogn. **24**(3), 659–671.

Tremblay, S., and Jones, D. M. **(1999)**. "Change of intensity fails to produce an irrelevant sound effect: Implications for the representation of unattended sound," J. Exp. Psychol. Hum. Percept. Perform. **25**(4), 1005–1015.

Tremblay, S., Nicholls, A. P., Alford, D., and Jones, D. M. **(2000)**. "The irrelevant sound effect: Does speech play a special role?," J. Exp. Psychol. Learn. Mem. Cogn. **26**(6), 1750–1754.

Tremblay, S., MacKen, W. J., and Jones, D. M. **(2001)**. "The impact of broadband noise on serial memory: Changes in band-pass frequency increase disruption," Memory, **9**(4-6), 323–331.

van Noorden, L. P. A. S. **(1975)**. *Temporal coherence in the perception of tone sequences* (Doctoral dissertation). Eindhoven: Technische Universiteit Eindhoven.

Vachon, F., Hughes, R. W., and Jones, D. M. **(2012)**. "Broken expectations: violation of expectancies, not novelty, captures auditory attention," J. Exp. Psychol. Learn. Mem. Cogn. **38**(1), 164–177.

Viswanathan, N., Dorsi, J., and George, S. **(2014)**. "The role of speech-specific properties of the background in the irrelevant sound effect, " Q. J. Exp. Psychol. A. **67**(3), 581–589.

Wojcicki, K. **(2011)**. "MATLAB wrapper for the PESQ binary (https://nl.mathworks.com/matlabcentral/fileexchange/33820-pesq-matlab-wrapper)." MATLAB Central File Exchange. Retrieved January 18, 2018.

Yang, W., and Kang, J. **(2005)**. "Soundscape and sound preferences in urban squares: a case study in Sheffield," Journal Urban Design, **10**(1), 61–80.

Zwicker, E. **(1961)**. "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)," J. Acoust. Soc. Am. **33**(2), 248.

List of figures

List of tables

Appendices

The following appendices are included in the next pages:

A. **FS, STI, NCM, CSII and PESQ**
The appendix explains the metrics used in Chapter 2. In addition to the FS, the STI and the NCM, two other metrics, the CSII, and the PESQ, are also explained in detail. The parameter values of the CSII and the PESQ measures for NVS stimuli are presented.

B. **Gain optimization procedure for generating spectrally-modified noise-pulse train stimuli**
The gain optimization procedure used for generating the experimental stimuli employed in Chapter 5 and in Appendix C is described.

C. **An attempt to quantify the perceptual sensitivity to changes in the $\mathbf{FDCC_{new}}$**
This appendix presents a three-interval forced choice experiment which was conducted with the aim of quantifying the perceptual sensitivity to changes in the spectral metric, the $\mathrm{FDCC_{new}}$.

A | FS, STI, NCM, CSII and PESQ

This appendix presents detailed explanations regarding implementations of the metrics, the fluctuation strength (FS), the speech transmission index (STI) and the normalized covariance measure (NCM), which were analyzed in Chapter 2 alongside the frequency domain correlation coefficient (FDCC_{old}). Two additional metrics, the coherence-based speech intelligibility index (CSII) and the perceptual evaluation of speech quality (PESQ), are also introduced in order to further investigate the impact of intelligibility of noise-vocoded speech (NVS) on the serial-recall results.

The results of Chapter 2 showed that there was a systematic decrease in serial-recall performance as a function of the number of frequency bands used in the NVS stimuli, up to a critical point, where the NVS became highly intelligible. This outcome was discussed in detail in Sec. 2.6 and two objective speech intelligibility metrics were analyzed, the STI and the NCM. The parameter values of the two metrics increased as a function of the number of frequency bands, indicating a relation between intelligibility and serial-recall results. However, the parameter values continued to increase beyond the critical point, 6-band NVS, beyond which there was no further systematic decrease in performance. In addition to this discrepancy, the STI value computed for 1-band NVS is unrealistically high, 0.54, while the corresponding NCM value is 0.35.

The limitations of the STI when used with nonlinearly processed speech such as tone- and noise-vocoding, has been discussed in the literature and alternative metrics were proposed alongside the NCM: CSII and PESQ (Goldsworthy and Greenberg, 2004; Chen and Loizou, 2011a). These two metrics were shown to produce promising values as intelligibility predictors for tone- and noise-vocoded speech (Chen and Loizou, 2010; Chen, 2011; Chen and Loizou, 2011a), and are thus analyzed as ISE predictors in this appendix.

The definitions of the parameters used in Chapter 2, as well as the

CSII and the PESQ, are explained individually in the following sections and the parameter values of the CSII and the PESQ, derived from NVS stimuli, are presented in the final section.

FS

Two different kinds of hearing sensation occur when sounds are amplitude modulated. For amplitude modulations below 20 Hz, the hearing sensation is called fluctuation strength (FS) and when the modulation frequency rises above 20 Hz the hearing sensation of roughness takes place (Fastl and Zwicker, 2007, p. 247 - 256). The 20 Hz limit is not a strict border from one sensation to another, instead, the transition between the two is smooth.

The unit of FS is vacil, and 1 vacil is defined as the sensation caused by a 60-dB, 1-kHz tone 100 % amplitude modulated at 4-Hz. The FS shows a bandpass characteristic as a function of modulation frequency and at 4 Hz it reaches its maximum: A 4 Hz modulation frequency, which is also the syllable rate in fluent speech, creates a large fluctuation strength whether the modulated sounds are broad-band or narrow-band.

An FS model was proposed in the study of Fastl (1982):

$$F \sim \frac{\Delta L}{(f_{mod}/4\,Hz) + (4\,Hz/f_{mod})} \tag{A.1}$$

where f_{mod} is the modulation frequency and ΔL is the temporal masking depth. It should be noted that the temporal masking depth, ΔL, is different from the magnitude of the physical modulation depth due to forward masking. The denominator in Eq. A.1 underlines the importance of the 4 Hz modulation frequency.

For amplitude modulated broad-band noise, the magnitude of temporal masking depth was shown to be independent of the center frequency. However, for amplitude modulated and frequency modulated tones there is a frequency dependency. For amplitude or frequency modulated tones fluctuation strength may be approximated by integrating the temporal masking depth, ΔL, along the critical-band rate:

$$F = \frac{0.008 \int_0^{24\,\text{Bark}} (\Delta L/\text{dB Bark})\, dz}{(f_{mod}/4\,Hz) + (4\,Hz/f_{mod})}\ \text{vacil} \tag{A.2}$$

While for these synthetic sounds the values of ΔL are available in the literature (Fastl, 1982; Fastl and Zwicker, 2007) this is typically not the situation for real life sounds. It was mentioned in the last paragraph of the corresponding chapter of the book (Fastl and Zwicker, 2007, Ch. 11) that a computer program was developed which uses differences in specific loudness instead of the ΔL. Several commercial software packages offer computation of FS values (Pulse by Brüel & Kjaer, ArtemiS by Head Acoustics GmbH, PAK by Müller BBM) while the information regarding the actual calculations is not revealed.

The FS was for the first time used as the basis of an ISE prediction model in the study of Schlittmeier *et al.* (2012) for the first time. Later, several studies analyzed FS values of the irrelevant sounds and reported the outcomes (Ellermeier *et al.*, 2015; Liebl *et al.*, 2016). The FS values reported in these studies as well as those reported in Chapter 2 and Chapter 3 of this thesis, were computed using Artemis Suite (Head Acoustics, Herzogenrath, Germany) except for the values derived from the study of Schlittmeier *et al.* (2012), which were computed using PAK software (Müller-BBM VibroAkustik Systeme, Planegg, Germay) by the authors of the study.

STI

The STI was proposed as an ISE predictor in the study of Hongisto (2005) by describing values of a large set of sounds using a sigmoid function. The study concluded that the magnitude of disruption reaches a maximum around the STI value of 0.6 and stays constant beyond that. The lowest critical value was found to be 0.2, where there was no disruption expected for STI values lower than that.

The metric was evaluated in several ISE studies (e.g., Park *et al.*, 2013; Ellermeier *et al.*, 2015; Liebl *et al.*, 2016), as well as in Chapter 2 of this thesis. The STI values for the NVS stimuli used in Chapter 2 were computed by applying the following procedure to both the reference (original speech) and the test (NVS) signals.

The computation of the STI begins with filtering the reference and the test signals with octave-wide band-pass filters with center frequencies ranging from 125 Hz to 8 kHz. The intensity envelope of each band is extracted by squaring and low-pass filtering the signals with a cut-off frequency of 30 Hz. The output is analyzed for each modulation frequency through a one-third octave-wide band-pass filter with center frequencies ranging from 0.63 to 12.5 Hz (see Sec. B in Houtgast and Steeneken (1985)). The modulation index for each octave band and for each modulation frequency is computed by taking the root-mean-square of each modulation-band-specific envelope and normalizing by the mean of the envelope. This results in two 7 × 14 matrices, one for the test and one for the reference signal, containing the modulation index for each octave band at each modulation frequency. The modulation index matrix of the test signal is compared with the modulation index matrix of the reference signal in order to compute the modulation reduction matrix, m. The m is converted into a corresponding signal-to-noise ratio (SNR) by:

$$SNR_{STI}(i,j) = 10\ log_{10}\left[m(i,j)\ /\ (1 - m(i,j))\right] \tag{A.3}$$

where i represents the octave-band and j denotes the modulation frequency band. The resulting SNR values in the matrix are limited to [-15, 15] dB before computing the octave-band specific mean, which is done by averaging the 14 SNR values derived from one octave band without multiplying with a modulation frequency band weighting factor. The octave-band-specific SNR values are summed, taking into account the weighting factors of the seven octave-bands:

$$\overline{SNR} = \sum_{i=1}^{7} w_i \times\ SNR_{STI}(i) \tag{A.4}$$

The values of w_i, weighting factors for seven octave bands, are 0.13, 0.14, 0.11, 0.12, 0.19, 0.17, and 0.14. The computation is finalized by converting the average SNR values to the index, STI:

$$STI = \frac{(\overline{SNR} + 15)}{30} \tag{A.5}$$

The STI takes values between 0 and 1. The STI value of 1 indicates the maximum intelligibility of speech, while 0 indicates maximum degradation of the reference speech signal. The STI values reported in Chapter 2 were computed by the MATLAB script used in the study of Park *et al.* (2013).

NCM

The NCM is an STI-variant measure with a major difference: Instead of quantifying the change in modulation depth between the input and output signals' envelopes using the modulation transfer function, it computes the covariance between the reference and test envelope signals computed in each frequency band.

The NCM values are computed as follows: First, the reference and test signals are divided into a number of bands (20 bands, spanning the bandwidth of 300 - 3400 Hz for the present study), using 4th order Butterworth filters, with the center frequencies determined by a cochlear-frequency position function (Greenwood, 1990). The envelope of each frequency band is extracted using a Hilbert transform and downsampled to 25 Hz, in order to limit the modulation frequencies to 0 - 12.5 Hz. The normalized covariance in the i-th frequency band is calculated as:

$$p_i = \frac{\sum_t (x_i(t) - \overline{x_i(t)})(y_i(t) - \overline{y_i(t)})}{\sqrt{\sum_t (x_i(t) - \overline{x_i(t)})^2}\sqrt{\sum_t (y_i(t) - \overline{y_i(t)})^2}} \tag{A.6}$$

The $x_i(t)$ and $y_i(t)$ represent the downsampled envelopes of the reference and test signals in the i-th band, while $\overline{x_i(t)}$ and $\overline{y_i(t)}$ are the mean values of the corresponding envelopes, respectively. The SNR in each band is computed as

$$SNR_i = 10\, log_{10}(\frac{p_i^2}{1 - p_i^2}) \tag{A.7}$$

and limited to the range of [-15, 15] dB. The transmission index (T) of each band is computed by mapping SNR values to the range 0 to 1:

$$T_i = \frac{(\overline{SNR_i} + 15)}{30} \tag{A.8}$$

The computation is finalized by averaging the transmission index values across all frequency bands:

$$NCM = \frac{\sum_{i=1}^{20} T_i \times W_i}{W_i} \tag{A.9}$$

where W_i are the weights applied to each of the 20 bands (see Table A.1).

The NCM values for the NVS stimuli reported in Chapter 2 were based on our own implementation in MATLAB. Later, the MATLAB script used in the study of Chen (2011) was provided to us by the author, and it was used to crosscheck our results.

CSII

The CSII was proposed by Kates and Arehart (2005) as one of the extensions of the magnitude squared coherence (MSC) function (or the normalized cross-spectral density of two signals) which has been used to analyze

Table A.1: AI weights (ANSI, 1997) with corresponding center frequencies of each band, used in the implementation of the NCM (left) and the CSII (right) for sentence materials.

	NCM		CSII	
Band	Hz	Weight	Hz	Weight
1	325	0.0772	150	0.0192
2	377	0.0955	250	0.0312
3	435	0.1016	350	0.0926
4	500	0.0908	450	0.1031
5	571	0.0734	570	0.0735
6	650	0.0659	700	0.0611
7	737	0.0580	840	0.0495
8	834	0.0500	1000	0.0440
9	941	0.0460	1170	0.0440
10	1060	0.0440	1370	0.0490
11	1191	0.0445	1600	0.0486
12	1337	0.0482	1850	0.0493
13	1498	0.0488	2150	0.0490
14	1676	0.0488	2500	0.0547
15	1873	0.0493	2900	0.0555
16	2092	0.0491	3400	0.0493
17	2334	0.0520	-	-
18	2602	0.0549	-	-
19	2898	0.0555	-	-
20	3227	0.0514	-	-

effects of hearing aid distortion on speech intelligibility (Kates, 1992). The CSII has been investigated as a speech intelligibility metric for degraded, masked speech stimuli as well as noise and tone vocoded speech in several studies (Ma *et al.*, 2009; Chen and Loizou, 2011a,b). The metric generated promising values for the intelligibility of tone-vocoded stimuli, while failing to predict the intelligibility scores for NVS stimuli (Chen, 2011).

The computation of the MSC begins with dividing the reference and test signals into a number (K) of overlapping windowed frames (using 30-ms Hanning windows with 50 % overlap in this study), computing the cross power spectrum for each frame, and then taking the average across all frames. For K number of data frames, the MSC at frequency bin, w, is given by:

$$MSC(w) = \frac{|\sum_{k=1}^{K} X_k(w) Y_k^*(w)|^2}{\sum_{k=1}^{K} |X_k(w)|^2 \sum_{k=1}^{K} |Y_k(w)|^2} \tag{A.10}$$

The asterisk sign stands for the complex conjugate while $X_k(w)$ and $Y_k(w)$ denote the spectra of $x(t)$ and $y(t)$. For the present study, $x(t)$ corresponds to the reference and $y(t)$ corresponds to the test signal. The MSC measure takes values between 0 and 1.

The resulting MSC value is used to compute the signal-to-distortion (SDR) ratio by using the coherence between the reference and test signals by:

$$SDR_{CSII}(i,k) = 10\,log_{10}\frac{\sum_{n=1}^{N} G_i(w_n) \times MSC(w_n) \times |Y_k(w_n)|^2}{\sum_{n=1}^{N} G_i(w_n) \times [1 - MSC(w_n)] \times |Y_k(w_n)|^2} \tag{A.11}$$

where $G_i(w)$ represents the rounded exponential filter (Moore and Glasberg, 1993) centered around the i-th critical band, and N is the FFT size. The resulting value is limited to [-15, 15] dB, and mapped linearly to the range 0 to 1 using:

$$T_{CSII}(i,k) = \frac{(SDR_{CSII}(i,k) + 15)}{30} \tag{A.12}$$

Finally, the CSII measure is calculated by:

$$CSII = \frac{1}{K}\sum_{k=0}^{K-1}\frac{\sum_{i=1}^{N} T_{CSII}(i,k) \times W(i,k)}{\sum_{i=1}^{N} W(i,k)} \tag{A.13}$$

where $W(i,k)$ is the weight placed on the i-th frequency band (i.e., band importance function, (ANSI, 1997)). The center frequencies and the associated weightings of the 16 bands used in the current study are presented in Table A.1.

The CSII values for the NVS stimuli were computed by our own implementation in MATLAB and are presented in the last section of this appendix. The MATLAB implementation was crosschecked with the script used in the study of Chen (2011), which was provided by the author.

PESQ

The PESQ measure was originally developed as a speech quality predictor for narrow-band handset telephony and narrow-band speech codecs (ITU-T, 2000; Rix *et al.*, 2001). It was used as an intelligibility metric for noise suppressed speech stimuli, corrupted by four different maskers (car, babble, train, and street interferences) at two different SNR levels (0 and -5 dB) in the study of Ma *et al.* (2009). The Pearson's r between the speech recognition scores and the PESQ values yielded a moderately high correlation value, 0.79.

Later, the PESQ measure was evaluated in the study of Chen (2011) with tone- and noise-vocoded speech, and the trend of the PESQ values followed the increase in the number of frequency bands employed in the NVS stimuli. It was observed that the prediction accuracy was lower for the tone-vocoded when compared to noise-vocoded speech. The Pearson correlation between the PESQ values and the intelligibility scores observed for the different number of frequency bands in the vocoded stimuli was not reported based on the type of the carrier. For all the vocoded stimuli, the Pearson's r was reported as 0.46.

The computation of PESQ begins with equalizing the levels of the test and the reference signals to a standard listening level. Afterwards, the two signals are passed through a filter with a response similar to that of a telephone handset. Time delays are compensated for by time aligning the signals and the loudness spectra are computed. The loudness difference between the test and the reference signals is computed and averaged over time and frequency in order to generate a subjective quality rating. The PESQ values are within the range of -0.5 to 4.5, while for most cases the output stays between 1.0 and 4.5. High values indicate higher quality.

The PESQ values for the NVS stimuli were computed by MATLAB scripts compiled from the PESQ version 2.0 binary (ITU-T, 2005; Wojcicki, 2011).

Parameter values of the CSII and the PESQ for NVS stimuli

The NVS stimuli were analyzed with the CSII and the PESQ metrics. The CSII values as a function of the acoustic conditions are presented in Fig. A.1.

The CSII was shown to produce high prediction accuracy with noise-suppressed speech (Ma *et al.*, 2009), broadband (non-vocoded) Mandarin Chinese (Chen and Loizou, 2011a) and tone-vocoded English (Chen and Loizou, 2011b). However, it can be seen that the CSII values did not follow the increase in the number of frequency bands in NVS stimuli. In fact, there is no variation in the parameter values, except for the original speech.

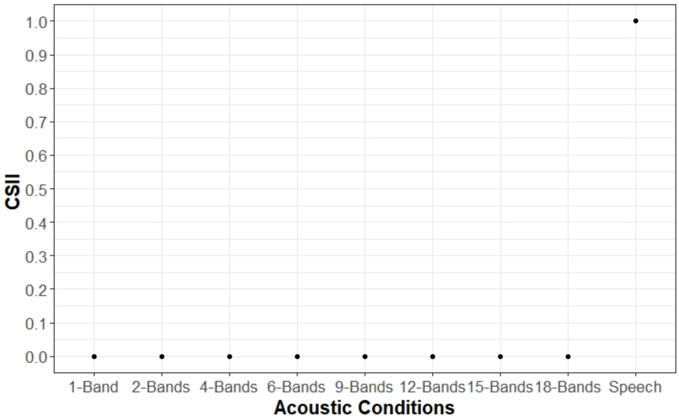


Figure A.1: Parameter values of the CSII metric for all acoustic conditions used in the two experiments in Chapter 2.

This outcome is due to the computation of the CSII: The CSII measure mainly makes use of the spectral-envelope information to asses intelligibility and could not account for the degradation regarding the temporal envelope of the NVS stimuli. The CSII values were not well correlated with the increase in the number of frequency bands, subsequently with the intelligibility of the NVSS, as well as with the serial-recall results collected in Chapter 2.

It should be noted that, in the study of Chen (2011), the CSII values for the tone-vocoded speech increased as a function of the number of frequency bands, although not in a systematic way: The CSII values increased between 2 to 4 bands and then stayed constant up to 8 bands, where the values continued to increase again for higher number of bands.

The PESQ values for the NVS stimuli and the original speech condition can be seen in Fig. A.2. The PESQ measure was shown to perform modestly well on predicting the intelligibility of consonants and sentences in noise (Ma *et al.*, 2009) and for predicting the intelligibility of tone-vocoded English (Chen and Loizou, 2010), as well as NVS (Chen, 2011). The PESQ values for the NVS stimuli reported in the study of Chen (2011) increased gradually from 2-band (approx. 0.95) to 8-band NVS stimuli (approx. 1.8), which is a relatively small increase when the range of the parameter (-0.5 - 4.5) is considered.

It can be seen that there is a slight increase in PESQ values as a function of the number of frequency bands for the NVS stimuli used in Chapter 2. However, while 1-band NVS yielded a PESQ value of

1.07, the 18-band NVS resulted in a value of 1.20. The increase in the PESQ values is not reflecting the increase in the number of frequency bands when compared with the results in Chen (2011). The authors had stated that the PESQ measure was well correlated with the intelligibility scores of tone-vocoded speech when the channel number was fixed at 8, while the curve of the intelligibility scores between 2 to 5 bands was not reflected in the PESQ values.

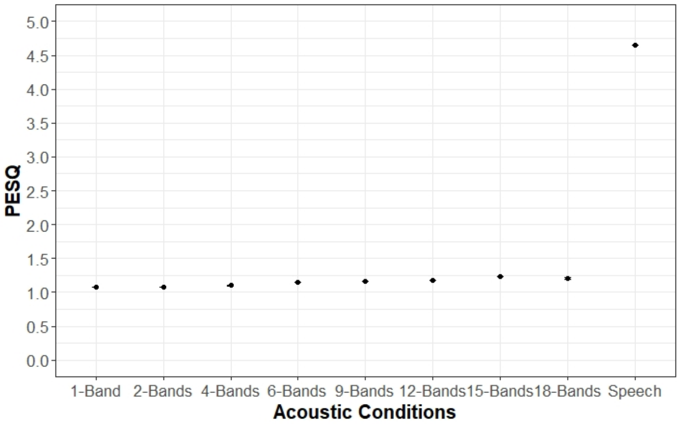


Figure A.2: Parameter values of the PESQ measure for the NVS and original speech stimuli used in the two experiments reported in Chapter 2.

Finally, it should be emphasized that there was no subjective intelligibility test conducted for the stimuli used in Chapter 2. Instead, the relation between the number of frequency bands and the perceived intelligibility of the NVS stimulus is based on experimental results reported in the literature (e.g., Davis *et al.*, 2005; Chen, 2011; Ellermeier *et al.*, 2015). Based on the common findings derived from these studies, it can be concluded that the CSII metric failed to account for the intelligibility, while the PESQ values demonstrated a small increase when the number of frequency bands increased. More important to this study, neither one of the metrics is capable of predicting the serial-recall results reported in Chapter 2.

B | Gain optimization procedure for generating spectrally-modified noise-pulse train stimuli

The impact of temporal and spectral features of irrelevant sounds on the ISE was investigated using a reference noise-pulse train stimulus (see Sec. 5.3) in a serial-recall task after modifying the temporal and spectral features of every second-indexed pulse, P2, in Chapter 5. The temporal modification was obtained by increasing the pulse width of P2 which changed the AMTF values systematically. The spectral modification was obtained by applying gains to seven octave bands (125 Hz - 8 kHz) of P2 which varied the FDCC$_{\text{new}}$ values. Because this would result in a change in AMTF as well, an optimized gain structure was developed.

The present appendix describes the gain optimization procedure developed for generating spectrally modified noise-pulse train (NT) stimuli used in the two experiments reported in Chapter 5. The procedure explained here was also used for modifying the spectral features of the experimental stimuli employed in Appendix C and the only difference between the two processes is the different acoustic characteristics of the reference stimuli used in the optimization procedure explained below.

Gain optimization procedure

The process began with investigating the behaviour of the modulation index in each octave band with respect to the gain values applied. A total of eleven gains were applied in each octave band of P2, ranging from $\theta = 0$ to $\theta = 1$ in steps of $\Delta = 0.1$. The duration of the width of P2 was kept at 50 ms.

The modulation index values for all octave bands and for each modulation frequency were calculated and it was observed that the patterns of the modulation index values of each modulation frequency were similar in all octave bands. Therefore, the patterns were averaged across octave bands and the modulation index values as a function of gains are presented in Fig. B.1.

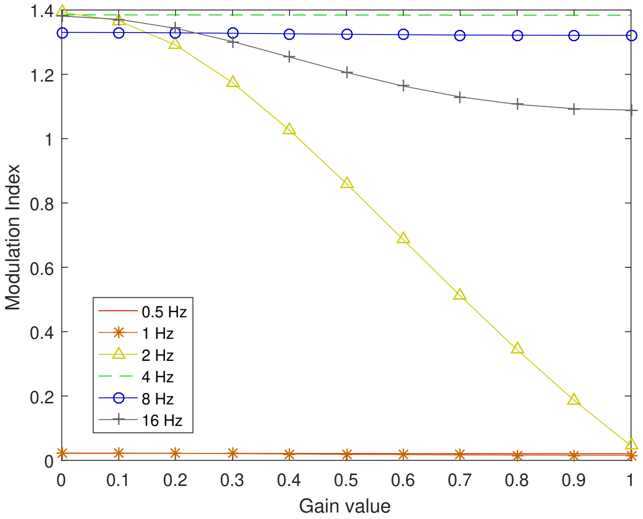


Figure B.1: The mean modulation index values for each modulation frequency as a function of the eleven gains.

Here it should be reminded that in order to compute the MTF a reference signal is needed. The modified pulses were used to generate a spectrally modified 1 min NT stimulus for each gain level and the MTF for each NT stimulus was computed using the unmodified NT stimulus as the reference. The resulting MTF values of each modulation frequency as a function of the gains are presented in Fig. B.2. It can be seen that when the gain of each octave band increased, the MTF decreased at the 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 remained constant. The behaviour of the MTF in each modulation frequency band with respect to octave band gains is formulated as a quadratic function:

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

where α_j, β_j, γ_j are the coefficients of the quadratic function in the j-th modulation frequency, and the θ_i is the gain applied to the i-th octave band. The coefficient values for each modulation frequency band were approximated using the built-in *polyfit* function in MATLAB (The

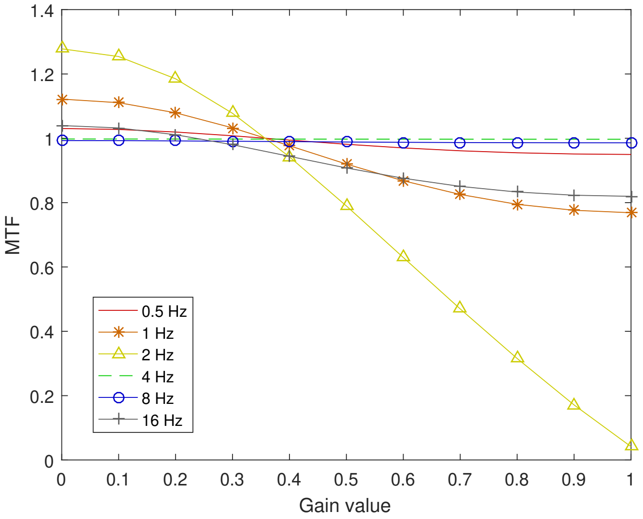


Figure B.2: The MTF values as a function of octave band gains for each modulation frequency. The x-axis presents the gain values applied to P2 and the y-axis shows the corresponding MTF values. Each line represents a different modulation frequency.

MathWorks Inc., Natick, MA), and are presented in Table B.1.

Table B.1: Coefficient values for α, β and γ.

	0.5 Hz	1 Hz	2 Hz	4 Hz	8 Hz	16 Hz
α	0.037219	0.11496	-0.50495	0	0	0.092307
β	-0.13014	-0.52378	-0.8472	0	0	-0.365651
γ	1.0378	1.1536	1.332	1	1	1.1066

The MTF values as a function of octave band gains for each modulation frequency were computed using the coefficient values presented in Table B.1 in the quadratic function (Eq. B.1) and the resulting values are presented in Fig. B.3.

The quadratic function relates the gains with the MTF, thus the computation of the AMTF in a given octave band gain value was realized by:

$$\overline{M} = \frac{1}{N\ K} \sum_{i=1}^{K=7} \sum_{i=1}^{N=6} M_{ij}(\theta_i) \tag{B.2}$$

where N represents the number of the modulation frequency bands and K is the number of octave bands.

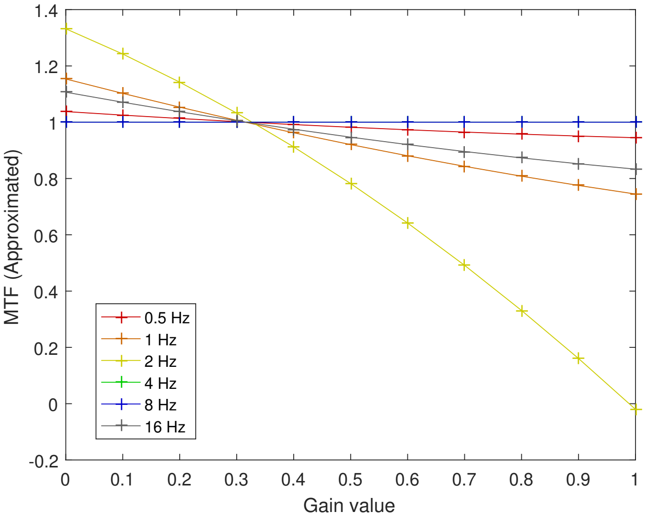


Figure B.3: The MTF values computed using the coefficient values presented in Table B.1, as a function of octave band gains for each modulation frequency are presented. The x-axis presents the gain values applied to P2 and the y-axis shows the approximated MTF values. Each line represents a different modulation frequency.

Equation B.2 was used to generate a set of octave band gains which satisfies the condition of a desired FDCC$_{\text{new}}$ value and a constant AMTF value by optimizing a cost function:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & Q(\Theta) = (F(\Theta) - F_0)^2 - (\overline{M}(\Theta) - \overline{M_0})^2 \\ \text{subject to} \quad & 0 \leq \Theta \leq 1 \end{aligned} \tag{B.3}$$

where Θ denotes the vector containing gain values for seven octave bands, F, F_0, $\overline{M}$ and $\overline{M_0}$ represent calculated FDCC$_{\text{new}}$, the target FDCC$_{\text{new}}$, calculated AMTF and the target AMTF values, respectively.

The cost function (Eq. B.3) was presented as a constrained minimization problem and the MATLAB function *fmincon* was used to find the optimal Θ which satisfies both the desired values for the FDCC$_{\text{new}}$ and the AMTF.

This optimization method was tested for the target FDCC$_{\text{new}}$ values ranging from 0 to 1, in steps of 0.1, while the AMTF was fixed at 1. The optimum gain values obtained from the algorithm were checked by computing the actual estimators in order to observe if the following conditions were satisfied:

$$\begin{aligned} F_{tol} &= |F - F_0| \leq 0.05 \\ M_{tol} &= |M - M_0| \leq 0.1 \end{aligned} \tag{B.4}$$

where F_{tol} and M_{tol} stand for the tolerated deviations from the desired $\mathrm{FDCC_{new}}$ and ATMF values, respectively.

The initial gain values were randomized and the computation was repeated again if the conditions expressed above were not satisfied after 20 trials. A maximum number of 500 iterations was allowed for one desired $\mathrm{FDCC_{new}}$ value: If the conditions (Eq. B.4) were still not met, the process restarted with the next desired $\mathrm{FDCC_{new}}$ value.

Several trials revealed that gains associated with the $\mathrm{FDCC_{new}} \leq 0.2$ were suitable values to be used as initial values to begin with the optimization process. Three trials were run with three different initial values, and 30 sets of gains were obtained. The gains resulted in the $\mathrm{FDCC_{new}}$ values ranging from 0.06 to 1, while the AMTF values obtained from the same gains were kept at 1, within a tolerance of 0.1.

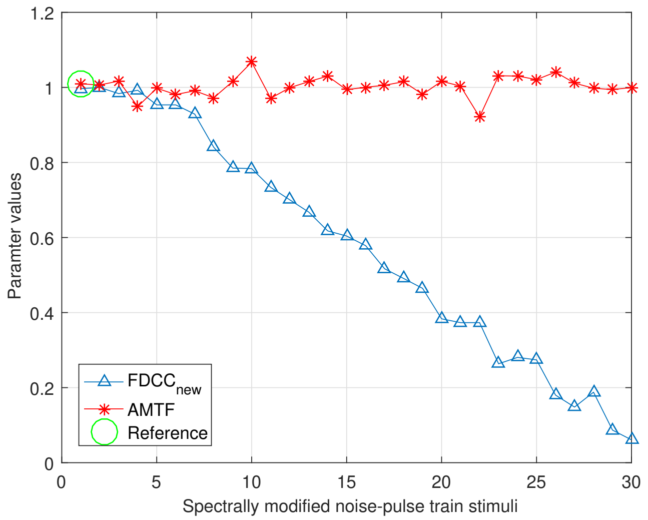


Figure B.4: The $\mathrm{FDCC_{new}}$ and the AMTF values of the 30 spectrally modified noise-pulse train stimuli. The x-axis shows the spectrally modified noise-pulse train stimuli and the y-axis shows the parameter values of the $\mathrm{FDCC_{new}}$ (blue triangle markers) and the AMTF (red asterisk markers) of each stimulus.

The results show that the intended purpose of modifying the $\mathrm{FDCC_{new}}$ while keeping the AMTF constant was achieved by the aforementioned technique. Thirty spectrally modified NT stimuli were generated and the parameter values for the set of NT stimuli are presented are Fig. B.4.

C | An attempt to quantify the perceptual sensitivity to changes in the $\mathrm{FDCC}_{\mathrm{new}}$

C.1 Introduction

The role of spectral variation on the irrelevant sound effect (ISE) has been investigated using a spectral metric, the frequency domain correlation coefficient ($\mathrm{FDCC}_{\mathrm{new}}$), in different experiments throughout this thesis. The experimental stimuli were generated by techniques which produced a systematic spectral variation in the stimuli in order to evaluate the accuracy of the $\mathrm{FDCC}_{\mathrm{new}}$ as a prediction model for the ISE.

These experiments not only allowed us to observe the strengths and the limitations of the $\mathrm{FDCC}_{\mathrm{new}}$, but also provided a deeper understanding regarding the relation between the spectral features of the sounds and the ISE. It was shown that the serial-recall results resembled the $\mathrm{FDCC}_{\mathrm{new}}$ values up to an extent when the speech stimuli were manipulated in the frequency domain in a systematic way. On the other hand, the results reported in Chapter 5 clearly showed that when the experimental stimuli were generated by modifying simpler non-speech stimuli, the $\mathrm{FDCC}_{\mathrm{new}}$ failed to predict the lack of an ISE: The noise-pulse train (NT) stimuli, which comprised a similar degree of spectral variation as the speech stimuli, did not produce an ISE.

This was considered to be a surprising outcome since a similar type of stimulus, a sequence of band-pass filtered noise bursts, was shown to create an ISE (Tremblay *et al.*, 2001). It was discussed that the lack of ISE observed in Chapter 5 might be due to the signal processing technique used to generate the NT stimuli: The reference noise pulse was generated by summing seven band-limited noise-pulses and the change in the $\mathrm{FDCC}_{\mathrm{new}}$ values was generated by applying gains to seven octave bands. This is different than concatenating band-pass filtered noise bursts with different center frequencies (Tremblay *et al.*, 2001), because the spectra of the two successive band-limited noise bursts with different center frequencies would not comprise energy in shared frequency bands if the quality factor (Q) is high enough. This would result in an FDCC$_{\text{new}}$ value of 0 if the distance between the two center frequencies is larger than one-third octave.

The results of Chapter 5 revealed a critical limitation for the FDCC$_{\text{new}}$: The FDCC$_{\text{new}}$ can not account for the tonal distance between the successive tokens of the sounds. The FDCC$_{\text{new}}$ value would not change if the center frequencies of the two band-pass filtered noise bursts were two, three or four octaves apart. The observed limitation of the metric produced a new question about the lack of ISE observed in the same experiment: Did the participants hear any differences between the NT stimuli with different FDCC$_{\text{new}}$ values?

In the present study, we expand on this question and further investigate the perceptual sensitivity to changes in the FDCC$_{\text{new}}$ by measuring just noticeable differences (JNDs) of the different value ranges of the spectral metric values. The JNDs were measured with an alternative forced choice (AFC) paradigm (Levitt, 1971), in which subjects had to listen to trials of three sound intervals. One of these intervals was a spectrally modified version of the other two (reference) intervals and had to be identified by the participants. The NT stimuli used in the first experiment in Chapter 5 were chosen for the task in order to further investigate the lack of ISE observed in the results. The stimuli and the experimental procedure are explained in the following sections. The results and the discussion are presented in sections C.4 and C.5.

C.2 Noise-pulse train (NT)

A half-second long reference noise-pulse train was generated by concatenating the two Hanning-shaped reference noise-pulses employed in the experiments in Chapter 5 (see section 5.3, Eq. 5.2). The pulse width of the first pulse (P1) was kept the same as in Chapter 5, 50 ms, while the pulse width of the second pulse (P2) was increased to 150 ms. The amplitudes of P2 and P1 were equal and the peaks of the two pulses were separated by 250 ms. Four half-second noise-pulse sequences were concatenated and a two second long reference NT stimulus was formed. The reference NT stimulus is presented in Fig. C.1.

The reference NT stimulus has a flat spectrum. The frequency spectrum of P2 is identical to P1 and the FDCC$_{\text{new}}$ value is 1. The experi-

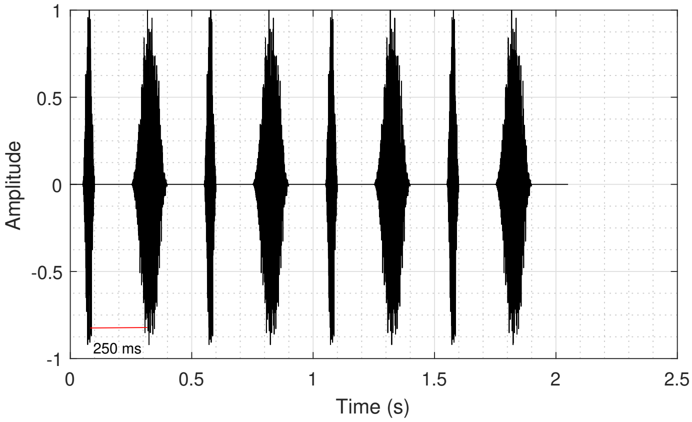


Figure C.1: Time course of a two second long section of the reference noise-pulse train stimulus. The two pulses, P1 and P2, alternate two times per second.

mental stimuli were generated by applying gains to the octave bands of P2 while keeping the temporal features unmodified which is controlled by computing the average modulation transfer function (AMTF, for a detailed explanation see 5.2.1) values constantly. The method used in the present study is identical with the method used in Chapter 5 and presented in Appendix B, however, an additional step of gain optimization was also introduced.

C.2.1 Modifying spectral features

The gain optimization procedure described in Appendix B was initiated by using the reference NT stimulus of the present study, with 21 target FDCC_{new} values between 0 and 1, with a stepsize of 0.05. It was observed that it was not possible to generate gains which would yield FDCC_{new} values lower than 0.25 without modifying the temporal features. Therefore, the lowest FDCC_{new} value used in the pilot and the main experiment in the present study was chosen to be 0.25.

An additional gain optimization procedure was required due to the definition of the FDCC_{new}: The FDCC_{new} quantifies the spectral variation from one token to the next one, P1 to P2 in this case, but does not reflect the overall spectrum of the stimulus. For instance, a desired FDCC_{new} value can be reached by applying seven similar gain values to the P2 or only altering one band drastically. The resulting two NT stimuli would be easy to distinguish from each other even though the FDCC_{new} values are the same, hence not suitable for an AFC task.

The aforementioned limitation was demonstrated by generating four sets of gains where three (Fig. C.2a, Fig. C.2b, Fig. C.2c) of those produce similar $\mathrm{FDCC}_{\mathrm{new}}$values while the fourth one (Fig. C.2d) produces a very different $\mathrm{FDCC}_{\mathrm{new}}$ value. When the gains presented in the left (a), the middle-left (b), the middle-right (c) and the right (d) graphs in Fig. C.2 were applied to P2 of the reference NT stimulus, the resulting $\mathrm{FDCC}_{\mathrm{new}}$ values were 0.5059, 0.5256, 0.5140 and 0.7939, respectively. The NT stimuli produced using the gains presented in Fig. C.2b ($\mathrm{FDCC}_{\mathrm{new}} = 0.5256$) and Fig. C.2c ($\mathrm{FDCC}_{\mathrm{new}} = 0.5140$) will be easily distinguished from the NT stimulus produced using the gains presented in Fig. C.2a ($\mathrm{FDCC}_{\mathrm{new}} = 0.5059$) in an AFC task. On the other hand, the NT stimulus generated using the gains presented in Fig. C.2d ($\mathrm{FDCC}_{\mathrm{new}} = 0.7939$) will not be easy to identify when it is presented next to the NT stimulus generated using the gains in the left graph (Fig. C.2a), even though the corresponding $\mathrm{FDCC}_{\mathrm{new}}$ values are further apart.

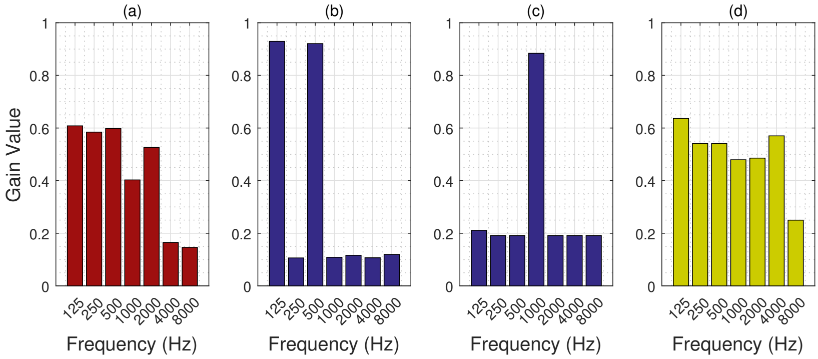


Figure C.2: Four sets of gain values are presented which were generated by the optimization procedure where the lower and upper ranges of the applicable gains were 0 and 1. The gain values presented in the left graph (a) yield an $\mathrm{FDCC}_{\mathrm{new}}$ value of 0.5059, the gains presented in the middle-left graph (b) produce an $\mathrm{FDCC}_{\mathrm{new}}$ value of 0.5256, the gains presented in the middle-right graph (c) generate an $\mathrm{FDCC}_{\mathrm{new}}$ value of 0.5140, and the gains presented in the right graph (d) generate an $\mathrm{FDCC}_{\mathrm{new}}$ value of 0.7939 when applied to P2.

In order to prevent this, the optimization procedure was modified in a way that it would have access to a smaller range of applicable gain values by decreasing the upper and increasing the lower boundaries of gain limits for each octave band. This limitation reduces the likelihood of obtaining an extreme gain value for one of the octave bands. An additional step of gain optimization was introduced to determine the new limits of the gain optimization procedure: 16 sets of octave-band gains between

0.25 and 1 were generated by using the optimization procedure which satisfy an ATMF value of 1 (with a tolerance of 0.1) and 16 FDCC$_{\text{new}}$ values between 0 to 1 (with a tolerance of 0.02), with a stepsize of 0.05. The average (μ) and the standard deviation (σ) of the 16 gains for each octave band were computed. The new lower boundary (θ_{min}) for each octave band was computed by $\mu_i - \sigma_i$ and the upper boundary (θ_{max}) was determined by $\mu_i + \sigma_i$, where i represents i-th octave band. The lower and upper boundaries for applicable gains in each octave band are presented in Table C.1.

Table C.1: The lower and upper boundaries of the applicable gain values.

	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	8 kHz
θ_{min}	0.35	0.04	0.34	0.02	0.05	0.04	0.04
θ_{max}	0.88	0.65	0.82	0.56	0.52	0.53	0.29

The results of introducing a new stage to the gain optimization procedure were investigated by running a pilot AFC test.

C.2.2 Pilot experiment

Three sets of gains were generated by running the optimization process three times with the same gain limits and same target FDCC$_{\text{new}}$ values. Each set of obtained gains was used to produce a set of spectrally modified NT stimuli by applying the gains to P2. Each desired FDCC$_{\text{new}}$ value was obtained three times, one in each set, with different gain structures. The expectation is that, when the JNDs are measured for the same FDCC$_{\text{new}}$ values generated by different sets of gains with identical limits, the results should be similar.

Method and procedure

An adaptive three-interval AFC (3-AFC) procedure was used in order to validate the aforementioned gain-range limitation technique. Two of the intervals always contained the reference stimulus, while the remaining interval contained the spectrally modified NT. The participants were asked to detect which interval was different from the other two.

The thresholds were determined with an adaptive 1-up, 2-down staircase procedure (Levitt, 1971), which meant that the absolute difference between the FDCC$_{\text{new}}$ values of the spectrally modified and the reference NT stimuli (ΔFDCC$_{\text{new}}$) was reduced when the participant correctly identified the spectrally modified interval in two consecutive trials. If the participant failed to detect the modified NT, the $\Delta\mathrm{FDCC_{new}}$ was increased. The participant was given a visual-feedback regarding his / her answer after each trial.

The stepsize, the change in the $\Delta\mathrm{FDCC_{new}}$ from one trial to another, of each trial varied according to the $\Delta\mathrm{FDCC_{new}}$ of the present trial: If the $\Delta\mathrm{FDCC_{new}}$ of the present trial was larger than 0.05, the $\Delta\mathrm{FDCC_{new}}$ was increased or decreased by a stepsize of 0.05 and if the $\Delta\mathrm{FDCC_{new}}$ value of the present trial was between 0.05 and 0.03, the stepsize was 0.01. For $\Delta\mathrm{FDCC_{new}}$ values between 0.03 and 0.01, the stepsize was 0.005 and if the participant reached the trials where the $\Delta\mathrm{FDCC_{new}}$ was 0.01 and lower, the stepsize was 0.0025. The target $\Delta\mathrm{FDCC_{new}}$ values of the pilot experiment were the same as those of the major experiment for the reference $\mathrm{FDCC_{new}}$ value of 0.5. The target values are presented in the first column of Table C.2.

The $\Delta\mathrm{FDCC_{new}}$ value of the first trial in each session was 0.2 while the minimum value reachable was 0 and the maximum value was determined by the $\mathrm{FDCC_{new}}$ value of the reference NT. For the pilot experiment, the $\mathrm{FDCC_{new}}$ values of the three reference NT stimuli were chosen to be 0.5. The $\mathrm{FDCC_{new}}$ values of the target stimuli were lower than the reference stimulus, 0.5, until the $\Delta\mathrm{FDCC_{new}}$ value of the present trial reached 0.25. For the $\Delta\mathrm{FDCC_{new}}$ values higher than 0.25, the $\mathrm{FDCC_{new}}$ values of the target stimuli were higher than the $\mathrm{FDCC_{new}}$ value of the reference stimulus. The values written in blue color in the third column of Table C.2 denote the $\Delta\mathrm{FDCC_{new}}$ values of the trials where the $\mathrm{FDCC_{new}}$ value of the reference stimulus is lower than that of the target stimulus. Here it should be reminded that the actual $\Delta\mathrm{FDCC_{new}}$ values in the third column of Table C.2 are the values used in the main experiment for the reference $\mathrm{FDCC_{new}}$ of the 0.5 condition, not in the pilot. However, the first column of the table can be used as a guideline for the stepsize structure of the pilot experiment.

If the participant reached the trial where $\Delta\mathrm{FDCC_{new}}$ had the maximum possible value and could still not detect the different stimulus three times in one session, the session was terminated and data was discarded. There were no preventive measures taken for the reversed scenario: If participant reached the trial where $\Delta\mathrm{FDCC_{new}}$ was 0, then it would eventually be counted as a reversal in every one time out of three, and this would be reflected in the measured JNDs. However, a visual and an auditory event was programmed to trigger a signal in the monitoring screen placed outside of the sound booth in order to make sure that the responsible researcher observes the situation in real time and stops the experiment if necessary. Each session was stopped after 10 reversals and the mean of the $\Delta\mathrm{FDCC_{new}}$ values observed in the last six reversals served as the threshold.

Stimuli, participants and apparatus

There were three acoustic conditions and each condition consisted of a set of spectrally modified NT stimuli generated by a different set of gains. For each condition, 20 spectrally modified NT stimuli, including the reference NT stimuli with the $\mathrm{FDCC_{new}}$ value of 0.5, were generated by applying the 20 octave-band gains to P2. The 19 spectrally modified NT stimuli yielded 19 different $\mathrm{FDCC_{new}}$ values which correspond to the stepsize structure explained in the previous section (see first column in Table C.2). Here it should be noted that there was a degree of variation allowed for these values and the NT stimuli which yielded the $\mathrm{FDCC_{new}}$ values closest to the target $\mathrm{FDCC_{new}}$ values were chosen.

Four normal-hearing subjects (three males and one female, age range = 18 - 28) participated in the experiment. All subjects were students of the Eindhoven University of Technology and reported normal hearing and normal or corrected vision. The participants were paid a small compensation fee.

The experimental procedure was structured as a block design: Each condition was presented in only one block and repeated once for each participant. The blocks were presented in randomized order and the presentation of the experimental blocks was preceded by a short training block. There were 3 min breaks between each block and one experimental session was completed in approx. 30 minutes.

The experiment took place in the same lab and the same hardware equipment was used as it was reported in Sec. 2.4.1. The 3-AFC test was conducted using APEX 3 software developed at ExpORL (Francart *et al.*, 2008). The sounds were presented through headphones in a diotic reproduction and the average sound level of each stimulus was calibrated to 65 $\mathrm{dB_{LAeq}}$.

Results

The results are presented in Fig. C.3 as mean values with standard errors (SEM) over four subjects. The JNDs measured for the three 0.5

conditions were slightly different from each other, however, the differences were not statistically significant. The lowest measured JND was 0.22 and the highest was 0.30.

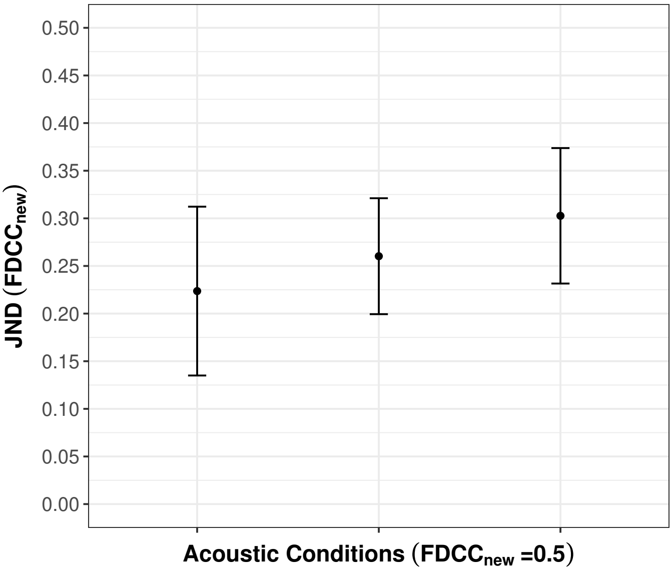


Figure C.3: JNDs for the three acoustic conditions where the reference stimulus used in each condition yielded an $\mathrm{FDCC}_{\mathrm{new}}$ value of 0.5. Error bars represent the SEM. (N = 4)

The objective of the pilot experiment was to observe if the gain-range limitation introduced to the gain optimization stage was successful to avoid the generation of stimuli with extreme gain values. The results showed that the measured JNDs for three conditions were statistically similar and the NT stimuli with similar $\mathrm{FDCC}_{\mathrm{new}}$ values generated similar JNDs regardless of the profile of the gains applied. The adapted gain optimization procedure was employed to generate the spectrally modified NT stimuli for the main 3-AFC experiment.

C.3 Experiment

The perceptual sensitivity of the $\mathrm{FDCC}_{\mathrm{new}}$ was determined by measuring JNDs with six different reference $\mathrm{FDCC}_{\mathrm{new}}$ values, in order to obtain a set of JND measures as a function of the $\mathrm{FDCC}_{\mathrm{new}}$ values. The experimental procedure was identical with that of the pilot test, except for the number of sessions which was increased to 6, one for each condition, for each participant. The description of the experimental procedure is given below.

Table C.2: List of ΔFDCC$_{\text{new}}$ values for each acoustic condition. Each column represents a set of ΔFDCC$_{\text{new}}$ values for a group of spectrally modified noise-pulse train stimuli which belongs to one of the six acoustic conditions. The FDCC$_{\text{new}}$ value of the reference stimulus used in each condition is presented in the second row of the table and the target ΔFDCC$_{\text{new}}$ values for each row is denoted in the first column. The colored values denote the cases where the FDCC$_{\text{new}}$ value of the test signal is higher than the FDCC$_{\text{new}}$ value of the reference stimulus.

	ΔFDCC$_{\text{new}}$ values of each set of stimuli					
ΔFDCC$_{\text{new}}$	0.25	0.5	0.7	0.8	0.9	0.98
0.0025	0.0020	0.0010	0.0024	0.0069	0.0025	0.0031
0.0050	0.0047	0.0016	0.0062	0.0094	0.0049	0.0057
0.0075	0.0072	0.0031	0.0089	0.0109	0.0082	0.0083
0.0100	0.0087	0.0113	0.0118	0.0126	0.0111	0.0107
0.0150	0.0142	0.0160	0.0196	0.0136	0.0232	0.0156
0.0200	0.0200	0.0217	0.0203	0.0187	0.0252	0.0182
0.0250	0.0244	0.0240	0.0294	0.0252	0.0345	0.0254
0.0300	0.0283	0.0311	0.0361	0.0285	0.0438	0.0317
0.0400	0.0390	0.0404	0.0497	0.0454	0.0467	0.0401
0.0500	0.0530	0.0481	0.0562	0.0490	0.0698	0.0506
0.1000	0.1026	0.0947	0.0993	0.1003	0.1071	0.0990
0.1500	0.1599	0.1447	0.1454	0.1499	0.1550	0.1530
0.2000	0.1988	0.1968	0.1986	0.1904	0.2054	0.2019
0.2500	0.2573	0.2530	0.2653	0.2456	0.2595	0.2471
0.3000	0.2964	0.3010	0.2941	0.3093	0.3000	0.3046
0.3500	0.3492	0.3470	0.3391	0.3561	0.3553	0.3547
0.4000	0.3990	0.4032	0.3984	0.4008	0.4189	0.4056
0.4500	0.4480	0.4331	0.4286	0.4553	0.4505	0.4554
0.5000	0.5059	0.4811	-	0.4992	0.5105	0.5063
0.5500	0.5495	-	-	0.5289	0.5528	0.5531
0.6000	0.6042	-	-	-	0.6088	0.6199
0.6500	0.6554	-	-	-	0.6385	0.6649
0.7000	0.6806	-	-	-	-	0.7038
0.7500	0.7386	-	-	-	-	0.7510

C.3.1 Method

Participants

A total number of 14 participants (eight females and six males, age range between 18-50 years) were recruited via the JF Schouten subject database of the Eindhoven University of Technology, Eindhoven, The Netherlands. All participants stated that they had healthy vision, hearing and no history of memory related disorder. The eligibility criteria were cross checked prior to the experiment, before they signed the informed consent forms. The experimental procedure was examined and approved by the Human Technology Interaction department, Eindhoven University of Technology and the Internal Committee Biomedical Experiments (ICBE) of Philips Research.

Stimuli

Six reference $\mathrm{FDCC}_{\mathrm{new}}$ values were chosen: 0.25, 0.5, 0.7, 0.8, 0.9 and 0.98. Six sets of spectrally modified NT stimuli were generated by applying six sets of gains produced by a range-limited gain optimization procedure explained in the previous section, where the upper and lower boundaries of the applicable gains were identical for each set and the same as those presented in Table C.1. The gain values used to generate the six reference NT stimuli are presented in Table C.3. The set of stimuli for the 0.5 condition was regenerated with the same gain limitations, hence they were not the ones used in the pilot experiment. The $\Delta\mathrm{FDCC}_{\mathrm{new}}$ values of each spectrally modified stimulus in each condition are presented in Table C.2.

Table C.3: The gain values used to generate the six reference noise-pulse train stimuli $\mathrm{FDCC}_{\mathrm{new}}$ values.

Ref. $\mathrm{FDCC}_{\mathrm{new}}$	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	8 kHz
0.25	0.85	0.06	0.80	0.08	0.08	0.06	0.06
0.5	0.56	0.23	0.45	0.12	0.17	0.15	0.06
0.7	0.40	0.41	0.47	0.26	0.28	0.38	0.20
0.8	0.39	0.38	0.44	0.30	0.31	0.36	0.24
0.9	0.35	0.35	0.41	0.30	0.29	0.35	0.28
0.98	0.59	0.50	0.55	0.17	0.28	0.18	0.27

Apparatus

The apparatus was the same as the one used in the pilot experiment.

Procedure

The experimental procedure was identical to that of the pilot experiment: Each condition was repeated once per session and presented in one block only. The presentation order of the blocks was randomized and the actual experiment began after participants completed a short training block. One experimental session took approx. 60 mins to complete including the breaks.

C.4 Results

The measured JNDs are presented in Fig. C.4 as a function of the $\mathrm{FDCC}_{\mathrm{new}}$ values for 14 subjects. Two of the subjects could not detect the different stimulus when the $\Delta\mathrm{FDCC}_{\mathrm{new}}$ value of the active trial was the maximum possible in the 0.7 and 0.98 conditions. Corresponding experimental data were discarded.

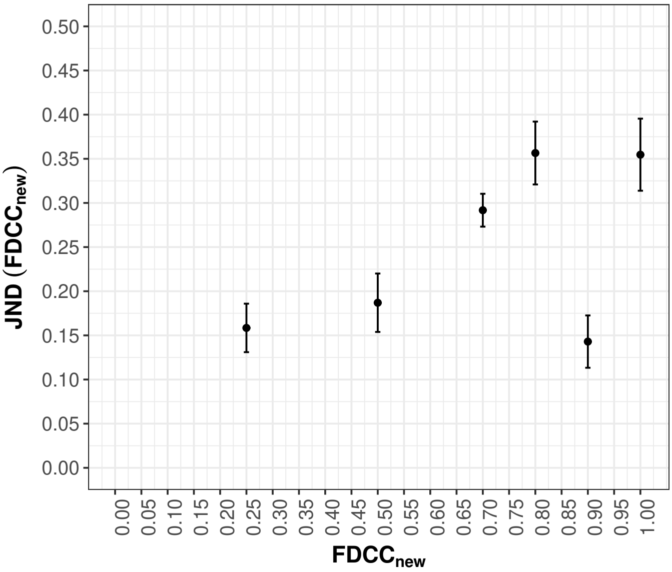


Figure C.4: JNDs for six acoustic conditions as a function of the $\mathrm{FDCC}_{\mathrm{new}}$ values. Error bars represent the SEM. (N = 14)

Each set of spectrally modified NT stimuli generated for a specific reference $\mathrm{FDCC}_{\mathrm{new}}$ value was treated as an acoustic condition and the effect of acoustic treatments on JNDs was confirmed by a one way repeated measures ANOVA, $F(5, 63) = 15.79$; $p < .001$. Post-hoc analyses were conducted using the Bonferroni correction ($p = 0.003$ for 15 pairs) and seven pairs of acoustic conditions were found to be significantly different ($p < .05$). The statistical results are summarized in Table C.4.

The results showed that the JNDs tend to increase as a function of the reference $\mathrm{FDCC}_{\mathrm{new}}$ values. Participants successfully detected the different NT stimulus when the $\Delta\mathrm{FDCC}_{\mathrm{new}}$ value reached approx. 0.15 in the lowest reference $\mathrm{FDCC}_{\mathrm{new}}$ value condition (0.25), while for the reference $\mathrm{FDCC}_{\mathrm{new}}$ value of 0.98 the JND was measured as 0.35. There was a steep increase observed in the JNDs between 0.5 (0.19) and 0.8 (0.36) conditions and the JND value for 0.7 (0.29) condition was in between. Noticeably, the JND for the reference $\mathrm{FDCC}_{\mathrm{new}}$ value of 0.9 did not follow the trend. In fact, the smallest JND value (0.14) was measured for the 0.9 reference $\mathrm{FDCC}_{\mathrm{new}}$, while the largest JND value (0.36)

Table C.4: Pairwise comparison of JNDs for the all possible pairs using Bonferroni correction. Only the seven significant pairs are presented.

Statistically significant pairs	JNDs ($\mathrm{FDCC}_{\mathrm{new}}$)	p values
0.25 - 0.8	0.16 - 0.36	$p < .001$
0.25 - 0.98	0.16 - 0.35	$p < .001$
0.5 - 0.8	0.19 - 0.36	$p < .01$
0.5 - 0.98	0.19 - 0.35	$p < .01$
0.7 - 0.9	0.29 - 0.14	$p < .05$
0.8 - 0.9	0.36 - 0.14	$p < .001$
0.9 - 0.98	0.14 - 0.35	$p < .001$

was measured for 0.8, one of its two neighboring conditions.

C.5 Discussion

The results demonstrated a well-defined behaviour of the JNDs as a function of the $\mathrm{FDCC}_{\mathrm{new}}$ values, except the statistically significant decrease of the JND observed for the 0.9 condition. A possible explanation can be that the results of the pilot experiment might be misleading and the gain limitation approach might have been unsuccessful. In that case, one of the NT stimuli would be much more easy to detect and the corresponding $\Delta\mathrm{FDCC}_{\mathrm{new}}$ value would appear more frequently in the list of reversals.

The $\Delta\mathrm{FDCC}_{\mathrm{new}}$ values corresponding to the reversals obtained from the 3-AFC sessions for 14 participants in the 0.9 condition are presented in Table C.5. When the $\Delta\mathrm{FDCC}_{\mathrm{new}}$ values of the reversals were investigated, it was observed that 21 times (denoted by red and green font colors) out of 98 reversals occurred when the $\Delta\mathrm{FDCC}_{\mathrm{new}}$ was 0.0698 (10th row in the 6th column in Table C.2). Participants successfully detected the NT stimulus seven times (green) out of 21 and 14 reversals (red) were incorrect answers. However, similar numbers of repetitions of one of the NT stimulus in the reversal lists as a correct response were also observed in the other acoustic conditions and the number of incorrect responses associated with this particular stimulus already eliminates a possible gain-limitation procedure problem.

Despite the unexpected result observed in the 0.9 condition, the behaviour of the JNDs supports the lack of ISE observed in the first experiment reported in Chapter 5: Participants might have failed to perceive any differences between the majority of the NT stimuli presented in the

Table C.5: The list of ΔFDCC$_{\text{new}}$ values corresponding to the last six reversals of each participant observed in the 0.9 condition are presented. The colored values show the reversals observed when the ΔFDCC$_{\text{new}}$ of the active trial was 0.0698. The values in red color represent the incorrect answers given by the participant and the green colors indicate that the participant successfully identified the stimulus as the different one.

ΔFDCC$_{\text{new}}$ values of each reversal for each participant						
Par1	0.5105	0.4189	0.4505	0.4189	0.4505	0.4189
Par2	0.1071	0.0467	0.0698	0.0438	0.0698	0.0467
Par3	0.0698	0.1071	0.0698	0.1071	0.0698	0.1071
Par4	0.0698	0.1071	0.0698	0.3553	0.2054	0.3553
Par5	0.1071	0.0698	0.1550	0.1071	0.1550	0.0698
Par6	0.0698	0.2595	0.0467	0.0698	0.0467	0.0698
Par7	0.0698	0.1071	0.0467	0.2054	0.0467	0.0698
Par8	0.2595	0.1550	0.3553	0.1550	0.3553	0.0698
Par9	0.0438	0.0467	0.0438	0.0467	0.0345	0.0698
Par10	0.1550	0.3000	0.2054	0.2595	0.1071	0.1550
Par11	0.0698	0.1550	0.0698	0.1550	0.0698	0.1550
Par12	0.0698	0.1071	0.0467	0.0698	0.0467	0.1071
Par13	0.0345	0.0438	0.0345	0.0467	0.0232	0.0345
Par14	0.4189	0.2595	0.3553	0.1550	0.2054	0.1550

experiment. It should also be noted that the 3-AFC task requires subjects to concentrate on the sounds presented to them, while the serial-recall task instructs the opposite. It might be possible that the required ΔFDCC$_{\text{new}}$ values to perceive any difference in the NT stimuli may be even larger in a serial-recall context than the JNDs measured in this study.

Nevertheless, the JNDs measured for the 0.9 condition prevent us from quantifying the perceptual sensitivity of the FDCC$_{\text{new}}$ and the present study only accommodates indications regarding the perceptual behaviour of the spectral metric for the NT stimuli. Therefore, the study is positioned in the appendix section and the results are not considered as a basis for any further conclusions and / or assumptions in the rest of the thesis.

Acknowledgements

This dissertation is the final product based on four years of work during which I was fortunate enough to share some great moments with some brilliant people. There were certain periods that I was in doubt about my work which eventually made me feel demotivated, but there was always someone around who pointed me in the right direction. I want to show my gratitude to all who helped to make this dissertation possible.

Firstly, I would like to express my sincere gratitude to my supervisors Prof. Armin Kohlrausch and Dr. Sam Jelfs for the continuous support through this process, for their patience, accessibility, involvement, motivation, and immense knowledge. The term “supervision” is not broad enough to reflect the true nature of the comprehensive mentorship I have received from them within the last four years. I am sure that I will experience the positive impact of their “sculpturing” in every aspect of my professional and personal life. I could not have imagined having better mentors for my Ph.D. study.

Second, I would like to thank all the senior researchers from our European project, BATWOMAN, for creating such an acoustics knowledge hub for us. Special thanks go to the junior researches of BATWOMAN for all the good times we had in many different places all around Europe.

My sincerest gratitude also goes to my friends and colleagues in the Brain, Behaviour and Cognition department of Royal Philips for making me feel at home. In particular, I would like to thank my manager Marieke van der Hoeven for keeping me away from the “real work” and helping me focus on my dissertation, my desk neighbor Inge Blei for uncountable favors including an express reimbursement scheme, Laura and Evelijne for helping me to understand the fundamentals of cognitive science, as well as Hanne and Hristo for all the coffees with laughter.

Besides my within-department friends, I would also like to thank Enrique, Ilapha, and Nemanja for the football, beers and the cheat-sheet on being a Marie-Curie fellow in Philips. I am very grateful to Patrick for helping me with my papers and Murtaza for the occasional stimulating conversations at the campus.

I am also thankful to my friends at the Human Technology Interaction group at the Tu/e. In particular, I would like to thank Mark, Giacomo, Anne, Indre, Elçin and Sima for the overall high quality of the sense of humor and all around friendliness. I am also thankful for the help that I got from Ellen, Anita, Daniël and Antal.

I want to reserve a special space for Alejandro, with whom I have shared countless scientific and non-scientific moments. I would like to particularly thank him for all the scripts, templates, and heated scientific discussions, as well as preventing me from spending a cold night in the Brussels train station.

I am grateful to my family in Turkey, who always supported my appetite for further education and provided me through moral, emotional and financial support in my life. Special thanks go to my sister for designing the cover of this book. I am also grateful to everyone in the family of Adriaansens who always showed sincere interest in my work and supported me along the way.

Finally, I would like to thank my girlfriend, flatmate, and partner-in-crime, Malou, whose name would be in the author's list of this dissertation if it was allowed. I feel obliged to mention some examples of her involvement, such as rehearsing presentations with me, proofreading my manuscripts and above all, recursively being exposed to the unpleasant sounds while participating in the home-made pilot sessions. This dissertation would still be in the preparation stage without her.

Toros Ufuk Senan
Eindhoven, January 2019

Curriculum Vitae

Toros Ufuk Senan was born on 24 August 1983 in Adana, Turkey. He was graduated with a bachelor of science degree in architecture in 2009 from the Yeditepe University in Istanbul. This was followed by obtaining a master of arts degree in sound engineering and design from Istanbul Technical University in Istanbul (2012) and a master of science degree in sound and music computing from Universitat Pompeu Fabra in Barcelona (2014). On 15 May 2014 he was employed by Royal Philips and the Human-Technology Interaction group at the Eindhoven University of Technology where he started a Ph.D. project under the supervision of Prof. Dr. Armin Kohlrausch and Dr. Sam Jelfs. The Ph.D. thesis titled "An evaluation of a psychoacoustic model of the changing-state hypothesis" was performed within the Initial Training Network BATWOMAN in the framework of a Marie Sklodowska-Curie Action, which aimed to stimulate multidisciplinary research in acoustics.

Publications

Senan, T. U., Jelfs, S and Kohlrausch, A. **(2018)**. "Cognitive disruption by noise-vocoded speech stimuli: Effects of spectral variation," J. Acoust. Soc. Am. **143**(3), 1407–1416.

Senan, T. U., Kohlrausch, A. and Jelfs, S. **(2017)**. "An Attempt to Predict ISE by a Spectral Estimator," in *Proceedings of ICBEN*, (Zurich, Switzerland).

Senan, T. U., Park, M., Kohlrausch, A., Jelfs, S., and Navarro, R. F. **(2015)**. "Spectral and temporal features as the estimators of the irrelevant speech effect," in *Proceedings of Euronoise 2015*, edited by C. Glorieux (Maastricht, The Netherlands), pp. 1925–1930.

Colophon

This thesis was typeset using LaTeX. The cover of this dissertation was designed by Nihal Işık Senan and the dissertation was printed by: Proefschriftmaken || www.proefschriftmaken.nl.