Auditory temporal resolution and integration - stages of analyzing time-varying sounds

Pedersen, Benjamin

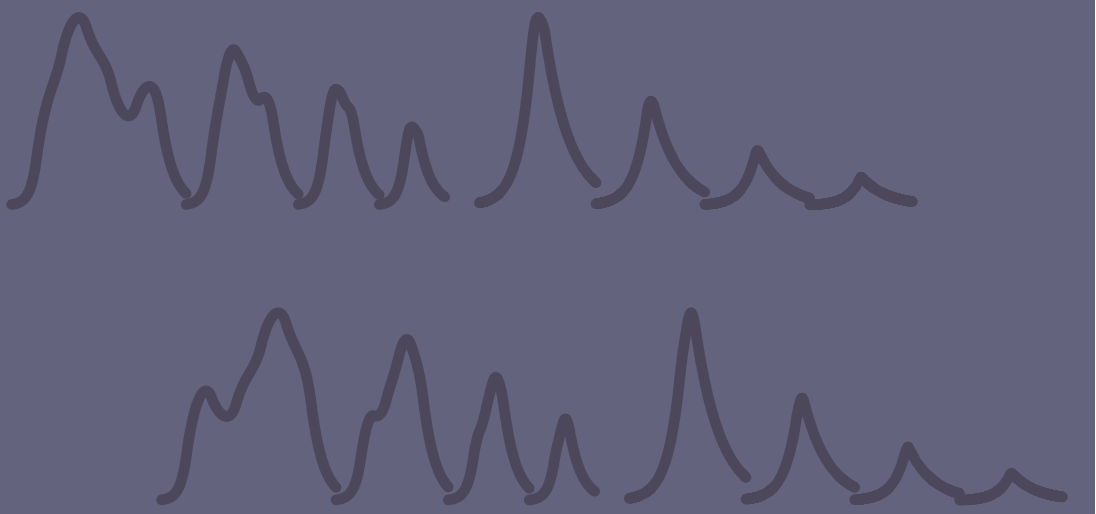
Publication date:
2007

Document Version
Publisher's PDF, also known as Version of record

Auditory Temporal Resolution and Integration

Stages of Analyzing Time-Varying Sounds

Benjamin Pedersen



Auditory Temporal Resolution and Integration

Stages of Analyzing Time-Varying Sounds

PhD thesis
by
Benjamin Pedersen

September 2006

Sound Quality Research Unit
Department of Acoustics
Aalborg University

Preface

This thesis has been submitted to the Faculty of Engineering, Science and Medicine at Aalborg University for partial fulfillment of the requirements for the award of the PhD degree. The research was carried out at the Sound Quality Research Unit at the university's Department of Acoustics in the period from September 2002 to October 2006.

I would like to thank all people of the Sound Quality Research Unit for mutual and fruitful exchange of ideas. I would also like to thank people of the Department of Acoustic who contributed with technical assistance or valuable feedback. Part of this research was carried out at the Psychoacoustics Laboratory (directed by Neal Viemeister) at the Department of Psychology, University of Minnesota, USA. I would like to thank the people of the laboratory for their hospitality and scientific feedback during the research visit.

The work of the Sound Quality Research Unit was partially funded by the companies Brüel & Kjær, DELTA Acoustics & Vibration, and Bang & Olufsen. Further financial support came from the Ministry for Science, Technology, and Development (VTU), and from the Danish Research Council for Technology and Production (FTP).

Especially I would like to thank Wolfgang Ellermeier for his supervision during the entire PhD project. His influence has positively and significantly shaped the outcome of the project and the development of my professional skills.

Finally, I would also like to thank the members of the PhD assessment committee for constructive remarks and suggestion concerning the content of this thesis.

Benjamin Pedersen
Aalborg, September 2006

The paper presented in Chapter 2 was published in a revised version in the Journal of the Acoustical Society of America after the publication of the thesis:

Pedersen, B. and Ellermeier, W. (**2008**). "Temporal weights in the level discrimination of time-varying sounds.", J. Acoust. Soc. Am. **123**, 963–972.

Summary

An important property of sound is its variation as a function of time, which carries much relevant information about the origin of a given sound. Further, in analyzing the "meaning" of a sound perceptually, the temporal variation is of tremendous importance. In spite of its perceptual importance, much is still unknown of how temporal information is analyzed and represented in the auditory system. Specifically, a large body of research has been concerned with identifying the acuity with which the temporal information is represented in the sensory system, and this has lead to some seemingly paradoxal observations: In binaural experiments (different sounds at the two ears) listeners are able to rely on temporal cues in the difference between the input from the two ears with a very fine resolution ($\sim$ 10 μs), whereas, when the same stimulation is provided to both ears, the listeners' ability to rely on temporal cues is much worse ($\sim$ 3 ms). For temporal integration of sound at levels close to the threshold of hearing, critical time-coefficients for integration seem to be as long as 100 ms to 200 ms. Furthermore, the temporal "acuity" also varies greatly over auditory tasks of different nature (temporal masking, gap detection, stream segregation, amplitude modulation detection, temporal order detection, etc.). The listening experiments presented in this thesis are all related to temporal resolution and integration in diotic listening (same sound to both ears). The purpose of the experiments is to clarify some of the apparent discrepancies by probing the auditory system in tasks of different "nature" in an effort to identify how different stages of perception might be responsible for the performance in the different tasks.

Specifically, the auditory tasks of the experiments in this thesis may be considered as falling into two categories: (1) Temporal integration when listeners have to judge the overall loudness of relatively long (compared to the temporal resolution of the auditory system) sounds fluctuating in level, and (2) temporal pattern recognition where listeners have to identify properties of the actual patterns of level changes.

In two experiments (falling into the first category) listeners had to judge sounds, with a duration of one second and randomly varying in level, as being either "loud" or "soft". From these judgments, temporal weighting curves were derived and showed that listeners generally emphasized onsets and offsets of the sounds in their judgments, but in idiosyncratic ways. Additionally, the temporal weighting changed if listeners were provided with feedback. In the second experiment, a spectral change was introduced in the center of sounds, leading to a perceptual emphasis of the temporal location of the spectral change in loudness judgments. These observations lead to the conclusion that loudness integration is not adequately described by a simple "summation" procedure as assumed in several models of loudness integration, but rather, auditory attention seem to be an important aspect when interpreting the results.

In two further experiments (falling into the second category) listeners had to discriminate temporal patterns in the envelope of noise samples being either ascending or descending in level. The duration of these patterns was varied to identify the temporal limit where discrimination was no longer possible. The limit was found to be in the order of 1 ms. The task was varied by adding flanking noise on both sides of the pattern to be identified, which dramatically changed the limit for the discrimination (to approximately 30 ms). The analysis of the results suggests that a key to understand this difference might be that, without the flanking noise, the patterns can be discriminated based on onset/offset cues,

which are absent in the case of flanking noise being present. Thus, the underlying hypothesis suggests that especially onsets of sounds have a particular elaborate representation in the sensory system. In two further conditions, examining the sensory processing of temporal variation, the pattern to be discriminated was repeated several times within a fixed time-frame (0.75 s). In the two conditions either the envelope only or the temporal fine-structure of the patterns was repeated. For relatively long durations of the patterns, the performance of the listeners was very similar in the two conditions, but for relatively short durations of the patterns the performance of the listeners seemed to be fundamentally different in the two conditions. When repeating the envelope only, listeners' performance was very similar to the case where the pattern was not repeated, but when the fine-structure was repeated, listeners were able to discriminate patterns with a much finer resolution. In the case of repeated fine-structure, no absolute lower temporal limit was found even though the duration of a single patterns was as short as 60 μs. Further, in the condition where the fine-structure was repeated, adding flanking noise seemed to improve rather than impair performance of the listeners for the shortest durations of the patterns. This shows that the concept of "energetic masking", which is often used to explain the performance of listeners, may be inadequate as it predicts that adding noise should worsen performance. Further, it might be noted that a temporal resolution of 60 μs is far better than what is normally considered the temporal limit of the auditory sensory system in the case of diotic stimulation.

The effects observed in the experiments presented in this thesis are too diverse to be adequately described by a single stage responsible for temporal processing. Therefore, in the thesis, several stages are suggested and an attempt is made to identify properties of their critical operating range. This partially explains the diversity of the measures of temporal resolution obtained in research concerned with auditory temporal processing, but much is still left to be explained. Hopefully the research presented in this thesis will help in disentangling different effects observed in listening experiments concerned with temporal processing.

Resumé (Summary in Danish)

En vigtig egenskab ved lyd er dens variation som en funktion af tid, hvilken indeholder vigtig information om lydens oprindelse. Yderligere er den tidsafhængige varians af største vigtighed for perceptuel analyse af lydens "mening". Trods den afgørende betydning for perception er meget stadig uvist om, hvordan lyd repræsenteres og analyseres tidsmæssigt i det auditive system. Meget forskning har beskæftiget sig med at identificere med hvilken opløsning tidsvarians er repræsenteret i det sensoriske system, hvilket har ledt til en række tilsyneladende paradokser: I binaurale forsøg (forskellig lyd ved de to ører) er forsøgspersoner i stand til at detektere meget små tidsmæssige forskelle ($\sim$ 10 μs). Derimod er de slet ikke i stand til at detektere tidsvarians med en sådan opløsning, når det samme input gives til begge ører ($\sim$ 3 ms tidsmæssige forskelle kan detekteres). For integration af lyde med niveauer tæt ved høretærsklen er der tegn på, at tidsperioden for integration er så lang som 100 ms til 200 ms. Den tidsmæssige opløsning varierer også stærkt ved opgaver stillet i forskellige typer lytteforsøg (forsøgskategorier på engelsk: Temporal masking, gap detection, stream segregation, amplitude modulation detection, temporal order detection, etc.). Lytteforsøgene, der præsenteres i denne afhandling, relaterer alle til tidsmæssig opløsning (temporal resolution) og integration i diotisk hørelse (samme input til begge ører). Formålet med forsøgene er at afklare nogle af disse tilsyneladende uoverensstemmelser ved at teste det auditive system i opgaver af forskellig "natur". Dette hjælper til at identificere, hvorledes forskellige perceptuelle niveauer er involveret, når forsøgspersonerne løser opgaver i forskellige typer forsøg.

Forsøgene, som beskrives i denne afhandling, kan betragtes som faldende i to kategorier: (1) Tidsmæssig integration når forsøgspersoner skal bedømme den samlede lydstyrke (loudness) af relativt lange lyde (i sammenligning med det auditive systems evne til at detektere tidsmæssig varians), som fluktuerer i lydstyrke, og (2) genkendelse af mønstre, hvor forsøgspersoner skal identificere mønstre bestemt ved deres variation i lydstyrke som funktion af tid.

I to forsøg (som falder i den første kategori) skulle forsøgspersoner bedømme lyde af et sekunds varighed og tilfældigt varierende i lydstyrke, som værende enten "høje" eller "lave". På grundlag af disse bedømmelser, blev tidsmæssige vægtningskurver beregnet, og de viste at forsøgspersonerne i deres svar generelt lagde ekstra vægt på en lyds begyndelse og slutning, men på meget individuelle måder. Yderligere ændredes vægtningen når forsøgspersonerne fik feedback. Et skift i lydens spektrum i midten af en lyd blev introduceret i en ny forsøgsbetingelse, hvilket førte til at forsøgspersonerne også lagde ekstra vægt på skiftet i spektrum. Disse observationer (samt andre) fører til den konklusion, at integration af lydstyrke (loudness integration) ikke passende kan beskrives som en simpel summeringsproces, som det antages i flere modeller for integration af lydstyrke, men snarere synes et begreb som auditiv opmærksomhed at være vigtig for at forstå resultaterne.

I yderligere to forsøg (faldende i den anden kategori) skulle forsøgspersoner skelne tidsmæssige mønstre i et støjsignals lydniveau som værende enten opad- eller nedadgående. Den tidsmæssige udstrækning af sådanne mønstre blev varieret for at finde den tidsmæssige grænse for, hvor mønstrene kunne skelnes. Grænsen var cirka ved 1 ms. Opgaven blev varieret ved at tilføje ikke-informativ støj på begge sider af mønstret, som skulle genkendes. Dette ændrede den tidsmæssige grænse kraftigt (nu ca. 30 ms). Analyse af resultaterne antyder, at nøglen til at forstå denne forskel ligger i, at uden den ikke-informative støj kan mønstrene skelnes på grundlag af deres forskelighed ved start og slut (onset/offset

cues). Dette er ikke muligt i det tilfælde, hvor støj er tilstede ved lydens start og slutning. Den grundliggende hypotese er, at specielt lyds begyndelse har en detaljeret repræsentation i sansesystemet. I to yderligere forsøgsbetingelser, der undersøger den sensoriske behandling af tidsmæssig variation, blev mønstrene, som skulle skelnes, gentaget flere gange indenfor et bestemt tidsinterval (0.75 s). I de to forsøgsbetingelser blev et enkelt mønster gentaget enten i detaljeret grad (fine-structure), eller også blev kun omridset (envelope) af et mønster gentaget. For relativt lang varighed af et mønster var forsøgspersonernes evne til at skelne mønstre næsten ens i de to betingelser, men for relativt korte mønstre var der stor forskel. Når kun omridset blev gentaget, var der næsten ingen forskel i forsøgspersonernes ydelse fra det tilfælde, hvor der ingen gentagelse var. Hvis derimod mønstret blev gentaget i detaljeret grad, var personerne i stand til at skelne mønstrene i meget finere grad. I tilfældet hvor mønstret blev gentaget i detaljeret grad blev der ikke fundet nogen nedre grænse for, hvornår mønstrene kunne skelnes, selvom udstrækningen af et enkelt mønster var så lav som 60 μs. Når mønstret blev gentaget i detaljeret grad, syntes tilføjelsen af ikke-informativ støj snarere at forbedre end at forringe forsøgspersonernes ydelse ved kort mønstervarighed. Dette viser at begrebet "energimæssig maskering" (energetic masking), som normalt bruges til at forklare forsøgspersoners ydelse, kan være fundamentalt forkert, da det forudsiger, at tilføjelsen af støj generelt skulle forværre forsøgspersonernes ydelse. Til sidst kan det bemærkes, at den observerede tidsopløsning på 60 μs er langt finere, end hvad der normalt betragtes som den nedre grænse for det auditive sensoriske system ved diotisk stimulation (samme input til begge ører).

Observationerne for tidsmæssig integration præsenteret i denne afhandling er så forskelligartede, at de vanskeligt kan beskrives med en enkelt enhed, som er ansvarlig for den tidsmæssige behandling. Derfor bliver det forslået at flere enheder er aktive, og egenskaber ved enhedernes virkeområde forsøges afgrænset. Dette er i en hvis grad i stand til at forklare forskelligartetheden af de mål, der i forskning er opnået for tidsintegration, men meget er stadig tilbage at forstå. Forhåbentlig vil resultaterne, som er fremlagt i denne afhandling, hjælpe til at adskille og forstå observationer opnået i lytteforsøg, der beskæftiger sig med tidsintegration i hørelsen.

Contents

Chapter 1

Introduction

Hearing is crucial for humans in that it enables them to communicate and react to their environment. Ultimately the life of an individual depends on whether he or she is able to analyze the surroundings and make the right actions. Further, communication is paramount for the development of social relations and for development of the individual's understanding of the world.

Hearing has the exquisite task of transforming sound, which in itself merely consists of vibrations of air molecules, into something meaningful. This requires extensive processing, and the full complexity of the working of the auditory sensory system is far from understood. Such knowledge is required, however, to be able to help people with hearing deficits, or in understanding how sounds of the environment (noise for example) influence humans, or in designing high-quality systems for sound reproduction or recording.

It is helpful to always have the basic purposes of hearing in mind, when trying to understand its functioning. In a first step toward a better understanding it may be beneficial to realize that the auditory sensory system must solve several tasks of different nature: (1) The ear has to act as a "microphone" transforming the acoustic vibration into neural electric/chemical activity, and (2) decode (identify properties and features) and (3) convey (from periphery to higher stages of perception) information of the acoustical input. Further, (4) the auditory system must also act as a "semantic analyzer" in interpreting the auditory input (in the analysis of language, but also to generally understand sound in relation to its context, that is, to obtain the "meaning" of the sound).

Already at this point some of the statements made may not be uncontroversial: At which stage is the auditory stimulation decoded; as in step (2) in the above, or is a complete trace of the recorded vibrations conveyed and decoded at more central cognitive stages? Further, the auditory system acting as a unified entity may invalidate the notion of independent processes taking place. If more detailed descriptions of the mentioned processes are desired, the case quickly gets complicated. For example, as suggested the electric/chemical activity is decoded, but into what (simple properties like pitch or loudness or more complicated aspects such as the location of the sounds origin in space)? And for the "semantic analyzer"; how is "meaning" derived from the input provided? A "passive" listener may record acoustic activity, without any requirement of meaning, but to which extent does perception require meaning? Does "meaning" influence how the stimulation is handled perceptually (is the sound perceptually handled according to assigned attributes for example)? In the extreme case, a listener cannot react in any consistent way to information that has absolutely no meaning to him or her. In an attempt to understand this complex process it therefore seems reasonable to identify some of the more fundamental limitations and capabilities of the auditory system.

Within the field of psychoacoustics, the hearing system is typically probed in listening test where participants are asked to respond to a given auditory task, as to obtain deeper knowledge of how the sensory systems carries out and to which degree it is able to perform specific tasks. It is important to realize that a listener is only able to respond to a given task to the extent that he/she can make sense of

the provided auditory information. Therefore listeners' performance in "simple" listening tasks may be thought to reflect the more basic aspects of sound perception, and the results of such experiments thought to describe some of the basic limitations of hearing, such as the amount of acoustic fine-structure which is recorded.

A fundamental understanding of the auditory system is important when trying to model the working of the system. The results of listening experiments may be used to determine critical parameters of such models. A given model may be evaluated in term of how well it can "generate a curve" which follows the outcome of a given experiment. It is important to realize that a wide range of mathematical functions are able to generate similarly looking graphs. In the specification of such functions very different "world views" may be adopted. Even though only one "world view" may be correct, the mathematical functions of different ones may be able to "fit" the performance of listeners equally well.

1.1 Frequency and temporal analysis

Physically, sound reaching the eardrum is changes in the pressure of the air as a function of time. So, fundamentally the acoustic input to the ear is given as a function of time. However, it is generally acknowledged that the temporal nature of sound is quickly turned into a frequency representation in hearing. This processes can be physically observed in that different portions of the basilar membrane of the cochlea are more sensitive to different frequencies, so different frequencies of the acoustical input are directly mapped to areas of maximum displacement of the basilar membrane Moore (2003a). This suggests that the description of the sound in terms of its *frequency spectrum* is fundamental to hearing. It should be noted however, that the ear does not transform the sound in the same way as an signal can be mathematically transformed into its frequency representation. A mathematical transform spans a certain time range, and thus, is of little use in describing *when*, in the time domain, specific events take place (at least when only the amplitude spectrum is concerned). Further, a full frequency transform requires integration of the signal over all time (Heisenberg-Gabor uncertainty principle). However, both aspects related to the frequency and the temporal details of a sound are relevant in sound perception. For example, in speech perception frequency analysis is important to identify different phonemes, but also analysis of the temporal order of the phonemes is important to decode the phonemes' "meaning". And on the larger scale, the order of the words in a sentence is of course crucial for the meaning of the sentence. Also in perception of music is pitch important, but as important is the temporal variation of pitch as to give a specific melody.

In listening tests the capabilities of the sensory system are often measured by probing the ability in the auditory system to both discriminate frequency and temporal details. The topic of this thesis is entirely on the capabilities of the sensory system to analyze temporal variation in sound.

1.2 Probing the hearing system's capabilities in temporal analysis

Assuming a lower limit exists for the temporal details that the sensory system is able to detect, this limit may be probed in listening tests by gradually decreasing the temporal extent of possible cues in stimuli, which listeners are asked to discriminate. When listeners are no longer able to discriminate different stimuli, the reason for this may be thought to be because the "temporal resolution" of their hearing is not as fine as that of the temporal details of the stimuli. But as suggested earlier, the sensory system has to carry out tasks of different nature. It could for example be reasoned that the detection of the temporal order of phonemes of a word is quite different from the detection of the order of words in a sentence, where in the first task the temporal order is crucial for identifying particular words and in the second task the importance of the identification is to arrange the word in the right order as to

obtain the right meaning. This suggests that different measures of temporal resolution may be obtained depending on the task the listener has to perform. If there was only one temporal limit of the sensory system, it would be expected that in all listening tests, independent of the nature of the task, this limit would always be found. In reality, measured temporal limits can vary extremely depending on the task.

Still it may be beneficial to apply the concept of an absolute lower limit of temporal details that the auditory sensory system is able to detect. Finer temporal variations than this limit are simply not present anywhere in the sensory system, and hence listeners will never be able to detect such fine details in any task. Knowing this limit would be helpful for understanding which possible cues listeners can rely on when performing different auditory tasks. The lower limit has been termed the "temporal resolution" of hearing, and a broad range of listening experiments try to arrive at a measure of this limit. It should be noted here that being able to discriminate sounds according to their amplitude spectrum is not considered a valid measure of temporal resolution, even though it is indirectly related to the temporal fluctuations of stimuli, and it is not valid because the amplitude spectrum contains no information about *when* in time events take place. As for example stated by Moore (2003a), temporal resolution is often taken as the ability of the sensory system to discriminate envelope patterns. But also, there is some controversy as to which extent the auditory system can rely on temporal fine-structure not present in the envelope.

Temporal resolution is often explained in terms of a temporal summation or smoothing process (which are conceptually somewhat similar). The "resolution" of this process is then considered a critical factor in temporal analysis (Moore, 2003a; Oxenham and Moore, 1994).

1.3 Diverging measures of temporal resolution

The evidence for and the location of such a limit of resolution is however not unequivocal. For loudness, for example, a critical time coefficient for temporal integration seem be in the order of 100 ms to 200 ms (Buus *et al.*, 1997; Glasberg and Moore, 2002; Moore, 2003a), even though these values have been a matter of debate, and fundamentally it may be considered problematic that people often are reported to have difficulty in comparing the loudness of sounds of different duration. Similar time constants are found in experiments where listeners have to detect sounds at levels close to the threshold of hearing (Viemeister and Wakefield, 1991; Gerken *et al.*, 1990; Neubauer and Heil, 2004; Zwislocki, 1960) and in experiments examining temporal masking (Moore, 2003a; Zwicker and Fastl, 1999), which is a phenomenon that is not completely understood (typically explained in terms of adaption in the auditory nerve or persistence of neural activity after the end of the stimulation). In tasks where listeners are asked to detect a silent gap in noise, the duration of the gap typically has to be in the order of a few milliseconds (Oxenham and Moore, 1994; Moore, 2003a). This limit is often denoted the "temporal resolution" of the auditory system. In experiments where binaural stimuli are used, listeners can utilize cues contained in differences across the ears, where time differences as small as 10 μs can be detected (Klumpp and Eady, 1956; Blauert, 1999). As indicated, different measures of the temporal capabilities of the auditory sensory system vary to a large extent, which may be hard to reconcile in a unified description of hearing. It is quite clear that a common explanation covering all observed effects cannot be given.

1.4 Goals and arrangement of the thesis

The primary goal of this thesis is to demonstrate that different types of temporal processing are involved in hearing depending on the task a listener is given. This will be demonstrated via four different listening tests, which supposedly probe several stages of temporal processing in the auditory system, starting with an experiment in which listeners have to temporally integrate sounds to discriminate their

loudness. The temporal variation of the stimuli of this experiment will occur over a range of one second, which is far longer than the temporal resolution of hearing. The stimulus is designed to reveal how listeners combine the loudness of different parts of a sound as to arrive at judgments of overall loudness. Two experiments in Chapter 2 explore this type of integration, and the interpretation of these results is further discussed in Chapter 3. Chapter 4 explores to which extent listeners are able to identify the shape of the envelope of stimuli, and how their performance is influenced by the addition of non-informative (masking) noise. Using similar stimuli as in Chapter 4, Chapter 5 demonstrates that listeners are able to rely on the fine-structure of sound under appropriate circumstances.

1.5 General results

The outcome of the four experiments and the main conclusions of the different chapters of the thesis are briefly summarized, and their implication for the main goals of the thesis will be explained. This summary is not meant to give a complete description of the entire work, and thus, for a full understanding the reader will have to consult the specific chapters of the thesis addressing the relevant topics.

1.5.1 Results and conclusions of Chapter 2

Based on two listening experiments, Chapter 2 (Pedersen and Ellermeier, 2006) focus on how listeners apply weighting to ten temporal segments of sounds with a total duration of 1 s when judging overall loudness of the sound. The outcome thus is a temporal weighting curve for each listener showing the "importance" of segments at different temporal locations for the judged loudness. Based on the results of the first experiment of the chapter, it is shown that some listeners emphasize onsets and offsets in their temporal weighting of the sound, but that the actual weighting varies to a large degree between listeners. Some listeners weight adjacent temporal segments very differently which shows that loudness integration is not a simple "smoothing" process as assumed in some models of loudness integration. Further, it is demonstrated that listeners change their pattern of temporal weighting if they are provided with feedback, and thus it may be concluded that the temporal integration is under the listeners' control to a certain extent. In addition, by introducing a spectral change in the middle of a sound, in a second experiment, it was shown that also the onset of a new "spectral event" is weighted more heavily. That listeners pay special attention to salient events within sounds may be a plausible explanation of this behavior. All in all this suggests that for temporal variation over far longer periods than the temporal resolution of hearing, the temporal variation is available in the sensory system, but to arrive at overall judgments of properties of the sound (loudness), this information is weighted and analyzed in complex ways, which is not adequately described as a simple summation process.

1.5.2 Results and conclusions of Chapter 3

In Chapter 3 it is further elaborated on, how the performance of the listeners in the first experiment of Chapter 2 may be understood when going beyond interpreting temporal weighting curves. It is first shown that listeners' judgments of loudness do not only depend on the one sound they are asked to judge, but also on sounds of previous trials. Further, if they are given feedback, they seem to respond in a way which is compatible with the random distribution used in the sound generation, for example: A listener not receiving feedback may hesitate to give the same response many times in a row, while a listener receiving feedback learns that sometimes such behavior is actually "correct". This demonstrates that the decision process is also based on experience in a complex way and not only on "integrated loudness".

Further, a model for temporal integration of loudness as suggested by Glasberg and Moore (2002) is applied to the stimuli of the listening experiment. The temporal properties of the model do not

readily predict the behavior of the listeners.

Following, a broad range of alternative ways of predicting the "loudness integration" of single sounds are suggested. Out of these suggestions, the most important (besides temporal weighting) aspects seem to be a non-linear dependence of loudness on the levels of the sound segments. This non-linearity is too large to be explained by the known non-linear relationship between loudness and the level of steady-state sounds. Alternatively it can be explained by assuming the listeners' attention is focused on relatively loud sound segments only.

In a short addendum, it is analyzed how loudness judgments and response time interrelate. A loudness judgment is the direct outcome of a decision process, while the response time may express aspects of the actual process. The general trend is that the response time for "loud" judgments is shorter the higher the levels of the segments and for "soft" judgments it is shorter the lower the segment levels. This may also suggest that loudness integration is not a "sampling" of a summed loudness, but rather a complex weighting of the available information about level, and this process takes longer time when the discrimination task is "hard".

Finally, results of an earlier study where listeners had to compare loudness of two temporally varying sounds are reanalyzed. The main finding is that listeners also weight onsets and offsets in a comparison task, but generally the last sound receives relatively grater weight. The reason for this can be thought to be caused by for example memory effects (recency) or distribution of attention. These two concepts are not easily disentangled when interpreting the results. However, the results suggest that the two sounds are individually integrated (same weighting curves, where both the onset of the first and the second sound are emphasized). That is, the auditory system does not seem to integrate the two sounds as a continuous stream, but rather identifies and independently integrates the relevant components.

1.5.3 Results and conclusions of Chapter 4

The results of a third experiment are described in Chapter 4 (Pedersen, 2006a). The main questions addressed in the chapter concern how listeners are able to identify envelope fluctuations, by finding temporal limits for identification of envelope patterns. Also suggestions for the cognitive processes which may hinder the performance are given. To that end listeners were asked to identify if a 3-segment pattern was either ascending or descending in level. The task was varied by adding flanking noise segments on both sides of the pattern: 0, 1, 3, and 9 non-informative noise segments on each side of the "target patter" respectively.

Adding one noise segment on each side had almost no effect on the listeners' performance, while adding three segments severely influenced their performance. In summary, to be correctly identified at a rate of 75%, the duration of one segment had to be 1 ms, 1 ms, 23 ms, and 30 ms when 0, 1, 3, or 9 segments were added respectively. The envelopes of the patterns in the different condition are shown in Figure 1.1 where the segment durations are set to the described limits. It is apparent in the figure that the added noise dramatically changes the temporal limit at which the pattern can be identified, and such a big change is not readily explaied by concepts such as energetic masking or the envelope being "smoothed" by a temporal window. Also, the performance does not change smoothly when more segments are added, but changes rather abruptly when adding three segments rather than one. To understand this it is suggested that onsets and offsets of sounds have an especially elaborate representation in the sensory system and that this may be the reason for the good performance when the "target pattern" is part of the onset or offset as opposed to the situations where onsets and offsets primarily contain non-informative noise.

So, in relation to the overall goals of the thesis, this shows that the sensory system may apply different temporal processing for onsets and offsets as compared to the analysis of an ongoing sound.

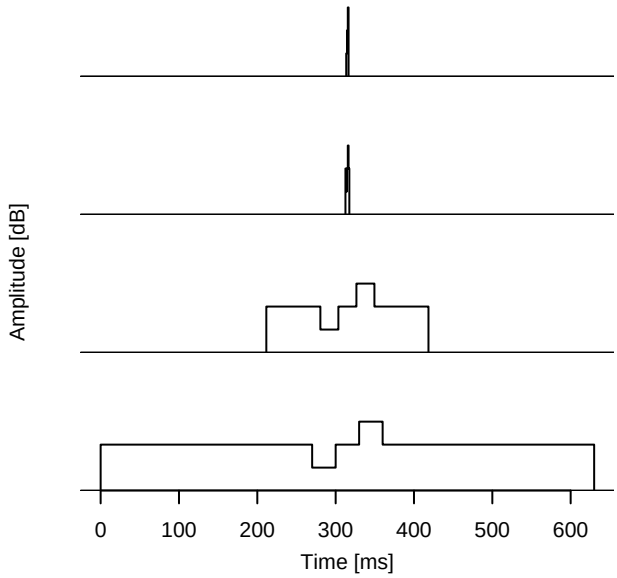


Figure 1.1: Envelopes of stimuli for which it is equally hard to identify the ascending pattern in the central part. In the top row only the target pattern is presented and in the following rows 1, 3, and 9 non-informative segments are added on each side of the pattern respectively.

1.5.4 Results and conclusions of Chapter 5

Chapter 5 (Pedersen, 2006b) extends the work outlined in Chapter 4, by continuously repeating temporal patterns within a fixed time frame. The same basic patterns are used as in the previous work, where patterns were defined in term of their envelope (ascending or descending). When repeating the patterns there are two cases, which are both explored: (1) Continuous repetition of the fine-structure of a single pattern, or (2) repetition of the envelope only. The distinction between these two types of repetition will be helpful in examining to which extent temporal processing works on the envelope and to which extent the auditory system can rely on fine-structure cues. As the results show, the repetition of envelope provides only little benefit for the listener in discriminating ascending and descending patterns over the case where only one single repetition of the pattern is presented. The case is quite different when the fine-structure is repeated: For relatively long durations of the patterns, performance of the listeners is almost identical in the three conditions (no repetitions, repeating envelope or repeating fine-structure), but for short pattern durations people are able to discriminate patterns in the fine-structure condition while it is impossible for them in the two other conditions. However, the performance varies greatly across listeners and for individual listeners the performance is not simply decreasing as a function of the duration the pattern. Some listeners are able to discriminate the patterns even at the shortest duration of the patterns used (segment duration: 20 μs). For the shortest duration of the patterns, adding non-informative noise segments at the sides of the patterns did generally not lead to a decreased performance, rather most listeners actually performed better.

Typically, the influence of noise on performance is explained in term of temporal "energetic" masking. Consequently it would be expected that performance would be generally worse when noise is added. As the opposite was observed for the shortest durations of the pattern when the fine-structure was repeated, this suggests that the masking observed in the experiment is of a different origin (informational masking for example). Further, the results suggest that different mechanisms for temporal processing are responsible for the performance in the different cases and that they vary over different critical time ranges. Specifically it is argued in Chapter 5 (Pedersen, 2006b) that separate processing mechanisms may exist for analyzing the envelope, analyzing onsets and offset, and for analyzing the temporal fine-structure. The demonstrated variability of listeners' "temporal resolution" over different

tasks, shows that it is vital to have an idea as to which stages of perception are crucial for the performance when drawing conclusions of the working of the sensory system. Especially when relating measured limits to models, it is crucial to understand which parts of perception are being measured and thereby realize the limitations of the model.

1.6 General conclusions

In this thesis auditory temporal processing in tasks of different nature (loudness integration or pattern discrimination) was explored over different time-ranges. The results suggest that auditory temporal processing as required in the different tasks cannot be described by a single "integrator device" in the sensory system. Rather it seems that different types of processing are responsible in the different tasks. In this section, an attempt is made to interpret the results across all studies presented in this thesis.

It was observed in the experiment of Chapter 5 that fine-structure can provide cues, which can be used for discrimination under the right circumstances. No absolute lower limit of temporal resolution could be found for all listeners and some listeners were able to discriminate patterns based on cues with an extremely short duration (in the range of 20 μs). This is incompatible with concepts typically adopted such as temporal "smoothing" or energetic masking, which thus do not seem to provide an adequate description of the functioning of the peripheral parts of the sensory system. However, a question arises: Why then are listeners not always able to perform tasks which, in terms of temporal resolution, ought to be much easier? For example, if smoothing of the envelope does not occur, why does the addition of non-informative noise impair performance? The suggestion given in Chapter 5 is that listeners may not be aware of the fine-structure (or the exact pattern of level-fluctuations) itself, only indirectly via "features" of the sound extracted at relatively low levels at which they have not yet reached a level of awareness. For the extraction of such features, the actual fine-structure may or may not be important, for example: A sound may be perceived as fluctuating in level without the actual pattern of level-fluctuation being available.

The concept of stimulus envelope is crucial in almost all models of temporal processing in audition, but it still remains a question whether such an "envelope" is actually extracted in the sensory system and, if the answer is positive, what is the nature of the envelope extraction process - is it smoothing by a temporal window as has been suggested? In interpreting the results presented in Chapter 4, where listeners identified temporal envelopes, a temporal window model was only marginally useful in explaining the results. It was suggested that temporal patterns without flanking noise segments can be discriminated based on onset/offset cues. This in term suggests that the experimental conditions containing stimuli with flanking noise may be thought to give a better picture of how listeners are able to utilize the envelope of stimuli in their judgments. As listeners' performance was relatively poor when flanking noise was present, this suggests that the representation of the envelope is relatively crude compared to the temporal capabilities of the peripheral parts of the auditory system. This questions the validity of measures of temporal resolution as obtained in gap detection experiments, where listeners have to identify a dip in the envelope. The outcome of such experiments may be considered a measure of the sensory system's ability to analyze the envelope of sound only, rather than its capabilities in analyzing fine-structure.

The loudness integration task, as explored in Chapter 2, may well be considered a task of analyzing the envelope of stimuli. It was shown in Chapter 4 that listeners were generally able to discriminate the temporal patterns with flanking noise present with a resolution in the range of 30 ms. This shows that the fluctuations in the stimuli of Chapter 2 are actually resolved temporally as level change occurred every 100 ms only. The question then is, how do the listeners arrive at loudness judgments? Is it a simple summation of the envelope, or is the envelope evaluated in a complex decision process? The answer is crucial for the design of a model of loudness integration. Current models assume that loudness integration is a summation process to a large extent, while the very different weighting curves

found for different listeners suggest that the envelope is evaluated in more complex ways as to judge its overall level. The interpretation of a "time-coefficient" (100 ms to 200 ms) for temporal loudness integration may thus be considered: It was shown that listeners were not "bound" by such a time coefficient in their judgments, but rather, they were able to put weight on certain segments only. As for example BP was able to resolve patterns with a resolution of 10 ms, a time-coefficient for integration should not be longer than this. An approach to understand how people integrate loudness may be to carefully consider which decision strategies they adopt when evaluating the overall level of a fluctuating pattern. As this may be very individual, as suggested by the results of Chapter 2, it may also be necessary to thoroughly consider the definition of loudness for temporally varying sound. Loudness is typically defined as "perceived impression of intensity", but if each person's "impression" is different the concept becomes almost meaningless. As earlier stated, loudness judgments may be the outcome of a decision process, so it is relevant to ask whether it is reasonable to include a decision process in the loudness concept. Rather, one may define loudness as the intensity perception *underlying* the decision process. However it might be a matter of debate to which extent such an "underlying intensity" percept exists. And of course the alternative definition complicates the measurement of loudness in listening experiments dramatically as listeners' judgments cannot be taken at face value, since a decision process will always underlie the judgments to some extent.

To better understand the process of perceptual loudness integration future studies may try to more clearly identify how loudness is related to various perceptual properties (for example: fluctuation rate, ramping, properties of onset/offset) and not only to intensity perception alone. It has been found that just noticeable differences in intensity depend on overall intensity (a special case off Weber's law, see for example Hellman and Hellman (2001) for a discussion of the topic). This may be used to examine the relation between loudness and intensity perception: For example Stecker and Hafter (2000) showed that sounds with slow attack and fast decay were perceived louder than sounds with fast attack and slow decay. Consequently it may be assumed that the just noticeable difference in overall level is larger for sounds with slow attack and fast decay. Whether this is actually true may be tested in a listening test. In a listening experiment, just noticeable difference in intensity may also be examined for each of the ten segments of the stimuli as used in Experiment 1 of Chapter 2. This may help to reveal to which extent the derived weighting curves reflect perceived intensity or a "decision rule" at a higher perceptual stage. It may be noted that a study somewhat similar to this has already been described by Stellmack *et al.* (2005). They used relatively short (50 ms) sounds and found that intensity differences were detected especially well at the onset. This suggests that heavier weighting of onsets in loudness judgments is caused by an increased sensitivity at the onset, rather than the onset being perceived louder, as a relatively large just noticeable difference would be expected in the case of the latter, the reason being: If the first segment is perceived louder, then the just noticeable difference in loudness should be larger for this segment.

How the auditory system more fundamentally processes time-variance was examined in the two experiments where listeners had to discriminate temporal envelope/fine-structure. This studies may be extended in several ways: In the study of repeating patterns the temporal separation of single patterns was always 25 segments. This may, however, be varied, which is especially interesting in the condition where the fine-structure is repeated. Increasing the separation between the repeated patterns may help to identify over how large intervals the auditory system is able to analyze fine-structure. In the extreme case, where the separation is very large, it may be assumed that the sensory system is not able to "see" that the fine-structure is identical in the repetitions, and in this case the performance of listeners may be expected to be identical in both the case where fine-structure is repeated and in the case where only the envelope is repeated. It may also be of interest to examine how the auditory system is able to utilize temporal cues across frequency-bands. The present studies only included broad band signals, however, the stimuli of the experiment may be filtered in frequency bands, to identify the importance of temporal cues across frequency bands. It was hypothesized that onsets cues play an important role in the case

where flanking noise was not present. This notion may be further explored: When flanking noise is present, the position of the temporal pattern may be varied as to be positioned both close to and far from the onset. However, further modification of the stimuli may be needed to avoid any spectral cues which may be present when descending and ascending patters cannot be generated by time-reversal.

The suggestions given for future experiments indicate that much is still to be discovered, and that there are limitations in interpreting the results of the experiments presented. All in all, the results of this thesis demonstrate that different levels of auditory temporal processing appear to be responsible for different tasks. Suggestions are given in identifying several such stages of processing. Hopefully, this may help in focusing future experiments on specific stages as to obtain a more complete description of their functioning.

Chapter 2

Paper 1: Temporal weighting in loudness judgments of level-fluctuating sounds

The paper presented in this chapter was published in a revised version in the Journal of the Acoustical Society of America after the publication of the thesis:

Pedersen, B. and Ellermeier, W. (**2008**). "Temporal weights in the level discrimination of time-varying sounds.", J. Acoust. Soc. Am. **123**, 963–972.

Temporal weighting in loudness judgments of level-fluctuating sounds *

Benjamin Pedersen† and Wolfgang Ellermeier

Sound Quality Research Unit (SQRU), Department of Acoustics, Aalborg University, Fredrik Bajers Vej 7-B5, 9220 Aalborg Øst, Denmark

(Dated: September 12, 2006)

To determine how listeners weight different portions of the signal when making loudness judgments, they were presented with 1-s noise samples the levels of which randomly changed every 100 ms by repeatedly, and independently, drawing from a normal distribution. A given stimulus could be derived from one of two such distributions, a decibel apart, and listeners had to classify each sound as belonging to the "soft" or "loud" group. Subsequently, logistic regression analyses were used to determine, to what extent each of the 10 temporal segments contributed to the overall loudness judgment. In Experiment 1, a non-optimal weighting strategy was found that emphasized the beginning, and, to a lesser extent, the ending of the sounds. When listeners received trial-by-trial feedback, however, they approached optimal, equal weighting of all stimulus components. In Experiment 2, a spectral change was introduced in the middle of the stimulus sequence, changing from low-pass to high-pass noise, and vice versa. It was shown that the temporal location of the stimulus change was strongly weighted, much as a new onset. These findings are not accounted for by current loudness models, but are consistent with the idea that temporal weighting in loudness judgments is driven by salient events.

PACS numbers: 43.66.Fe, 43.66.Cb, 43.66.Ba, 43.66.Mk

I. INTRODUCTION

A. Weighting level information in auditory discrimination tasks

When evaluating the loudness of a sound, the auditory system may be assumed to integrate information both across spectral regions and over time. A powerful tool to study such integration processes has been the *analysis of weights* given to the stimulus components defined in the experiment. Pioneered by COSS analysis (Berg, 1989), a number of related methodologies have evolved (e.g. Lutfi, 1995), all of which have in common that the listener does not have to be explicitly queried as to his or her weighting of the informational elements. Rather, all but a *global* judgment of pitch (Berg, 1989), loudness (Willihnganz *et al.*, 1997), or lateralization (Saberi, 1996; Stecker and Hafter, 2002) is required, from which, via statistical analysis or the construction of psychometric functions, its relation to the particular informational components is derived.

1. Spectral weights

Most of the few studies applying the analysis-of-weights methodology to loudness, have been concerned with the determination of *spectral* weights in level-discrimination tasks (Doherty and Lutfi, 1996, 1999; Kortekaas *et al.*, 2003; Willihnganz *et al.*, 1997). To that end, in a two-interval, forced-choice paradigm, random, independent level perturbations were added to each of a number of tonal components of different frequency, and the effect of these frequency-specific perturbations on the listener's overall decision yielded the spectral weights in question. Typically, these were found to be relatively flat, though sometimes with greater emphasis given to the highest or lowest frequency components (see Kortekaas *et al.*, 2003).

2. Temporal weights

There have been hardly any studies on the weighting of level information as a function of time. Buus (1999) investigated the detectability of a series of six adjacent 25-ms, 1-kHz tone pulses in masking noise. By adding independent level perturbations to the pulses, he was able to construct conditional psychometric functions relating detectability to the random level variations, separately for each of the six temporal pulse locations. From the slopes of these psychometric functions, much like in COSS analysis, relative weights were derived specifying the contribution of each temporal position in the pulse sequence to overall detectability. Analyzing three listeners in a number of experimental conditions, Buus found their weighting functions to be nearly optimal, i.e. giving equal weight to each of the (equally informative) six pulses, with small, but statistically significant departures favoring the middle portion of the pulse sequence (see his Figure 3).

*Parts of this work were presented at the 149th meeting of the Acoustical Society of America, Vancouver, Canada, May 2005 and at the joint meeting of the German and the French acoustical societies (CFA/DAGA), Strasbourg, France, March 2004.

†Electronic address: bp@acoustics.aau.dk

Lutfi's (1990) studies of sample discrimination contained one condition in which sequences comprised of up to 12 tones had to be discriminated on the basis of an overall level difference between target and standard sequence. COSS analysis (performed on the data of a single listener, see Lutfi's Figure 9) showed the weights assigned to the elements in the sequence to be approximately equal.

In a study involving one of the present authors (Ellermeier and Schrödl, 2000), using a 2IFC paradigm, on each trial listeners compared two 1-s samples of broadband noise (one of which was incremented relative to the other by 1 dB) with respect to their overall loudness. The noise samples were divided into 10 segments of 100 ms each onto which small, random level perturbations were imposed. Using COSS analysis (Berg, 1989), weights were derived for the 10 temporal segments. They exhibited a bowl-shaped pattern with the beginning of the noise sequence, and (to a lesser extent) the end being emphasized.

B. Memory effects

Further evidence for an unequal weighting as a function of time comes from studies investigating performance effects supposedly related to the functioning of auditory memory. These studies, however, looked at the discriminability of tone patterns in which *frequency* (or pitch) changes rather than level changes had to be tracked. McFarland and Cacace (1992) found strong primacy and recency effects in tone patterns being between 7 and 13 elements long, i.e. significantly better discrimination at the beginning or end of the sequence.

Surprenant (2001) varied the inter-stimulus interval (ISI) between the sequences to be discriminated, and found strong recency effects, with additional primacy effects emerging as the ISI was increased. Whether such memory effects are obtained for the discrimination of level changes as well, remains an open question.

C. Rationale

Given the scarce and equivocal evidence regarding temporal weighting in level discrimination (or loudness integration), it appears worthwhile to reinvestigate the issue. In contrast to earlier investigations, that shall be done using a *one-interval task* much like in the original study illustrating the weights technique (Berg, 1989). In the present implementation, subjects will be presented with a single stimulus on each trial, and will simply have to classify it as belonging to the "loud" or "soft" set defined by the experiment. This task is conceptually much simpler than a 2IFC task (see Kortekaas *et al.*, 2003), and it does not require assumptions about the memory processes involved, such as making different predictions depending on the length of the inter-stimulus interval (Surprenant, 2001).

Furthermore, since it is conceivable that the contradictory outcomes of some of the studies of temporal weighting may be due to different degrees of practice with the task, or to different strategies used, in Experiment 1, the opportunity to acquire an optimal weighting shall be experimentally manipulated by giving one group of listeners explicit trial-by-trial feedback as to the "correct" response alternative, while another group receives no such feedback, and thus no chance to optimize their strategy.

Finally, since those authors motivated by theories of memory have speculated on the "distinctiveness" of certain events in the temporal sequence, such as the beginning and end of a sound (Neath *et al.*, 2006; Surprenant, 2001), in Experiment 2 additional distinct events shall be experimentally induced by abruptly changing the spectral content of the sound to be judged. In particular, noise sequences will be designed that instantaneously shift from a low-pass to a high-pass characteristic (and vice versa) in the middle of the temporal sequence. Potentially, the spectral shift might constitute a new "distinct" event, e.g. signaling a new "onset", and thereby altering the weight pattern when compared to a control sequence of non-changing broadband noise.

II. EXPERIMENT 1 - LOUDNESS OF SINGLE SOUNDS

A. Method

1. Listeners

Ten listeners (1 female, 9 male) including the authors ("WE" and "BP" in the figures) participated in the experiment. The mean age of the participants was 26 years (range: 18 to 46 years). All were audiometrically screened, and no one was found to have significant hearing loss (more than 20 dB hearing loss at more than one frequency of 0.125, 0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4, 6, and 8 kHz). Except for the authors, the participants were students with little or no experience in listening experiments.

2. Apparatus

Stimuli were generated digitally on the PC controlling the experiment. A Tucker Davis Technologies System 3 was used for digital-to-analog conversion (RP2.1 unit), setting appropriate levels (two PA5 attenuators), and for powering the headphones (HB7 unit). Signals were presented diotically via headphones (Beyerdynamic DT 990 PRO), at a sample rate of 50 kHz and with 24 bit resolution.

The listeners were seated in a double walled listening cabin during the experiment and made responses using two buttons marked "soft" and "loud" on a special button box connected to the Tucker Davis RP2.1 unit. The

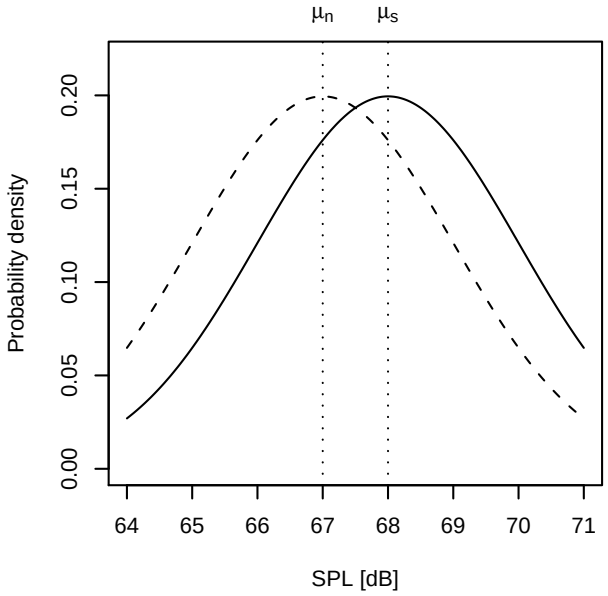


Figure 1. "Noise" (broken line) and "signal" (solid line) distributions from which sound levels were drawn.

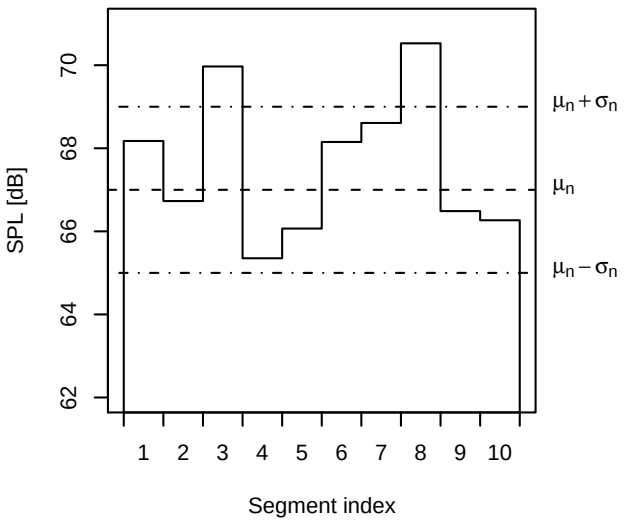


Figure 2. Temporal envelope of a sound sample (here: "noise").

box was also used for providing feedback using red and green lights.

3. Stimuli

The sounds used in the experiment were samples of white noise having 1 s duration. Their overall level was randomly varied every 100 ms, thus producing a stepwise level-fluctuating sound consisting of 10 segments (see Figure 2). The overall level of each segment was picked randomly from one of two normal distributions denoted "signal" and "noise", with the "signal" distribution having a higher mean value. The "signal" distribution had mean value $\mu_s = 68$ dB SPL and a standard deviation of $\sigma_n = 2$ dB. The "noise" distribution had a mean value $\mu_n = 67$ dB SPL and a standard deviation of $\sigma_s = 2$ dB. The two distributions are schematically depicted in Figure 1.

The setup was calibrated using an artificial ear (Brüel & Kjær 4153) with a microphone (Brüel & Kjær 4134). When sound pressure levels are used throughout this article, they refer to the RMS sound pressure level of a continuous broad-band noise as would be measured in the artificial ear at the given presentation level.

4. Experimental procedure

Participants were instructed that the sounds "were randomly generated", and came from a "soft" or a "loud" set of levels with equal probability. A one-interval two-alternative forced-choice paradigm was used. On each trial, the listener heard a single sound and was asked to judge it as being either "soft" or "loud". In the sequence of trials "noise" and "signal" sounds were presented in random order.

Listeners were divided into two groups in one of which the listeners received trial-by-trial feedback. If the generated sound was from the "noise" distribution and the listener responded "soft" or if the sound was from the "signal" distribution and the response was "loud" the feedback was a green light, in the other cases it was a red light. No such feedback was given to the other group.

After the completion of each block of 130 trials, overall feedback was given by telling the participants the percentage of "correct" responses they had obtained, i.e. responses which agreed with the "noise" or "signal" property of the stimulus. This type of overall feedback was given to all listeners. It helped to motivate the listeners, however based on this type of feedback, it was impossible to change a decision strategy based on trial-by-trial learning.

The first two and a half sessions were used for training. During training the difference between the "noise" and "signal" means, μ_s and μ_n, was successively decreased from 3 dB over 2 dB to a final 1-dB difference.

5. Data collection

The experiment was arranged in blocks of 130 trials of which only the trials 10 to 130 were analyzed, leaving the first 9 trials for building up a decision criterion. Five such blocks made up one session, which lasted approximately 40 minutes. Each listener proceeded through 10 sessions.

6. Determination of temporal weights

In making an overall loudness judgment, listeners are assumed to base their responses on a decision variable, D, defined as:

$$D(\mathbf{x}) = \left(\sum_{i=1}^{10} w_i x_i\right) - c \ , \qquad (1)$$

where $\mathbf{x}$ is a vector of the ten segment levels constituting a given sound. x_i refers to the sound pressure level in decibels of each of the 10 segments and w_i is a perceptual weight given to the i'th segment. It is assumed that the weighted sum of the segment levels is compared to a fixed decision criterion c. So the strength of the decision variable is given by the difference between the magnitude of the weighted sound levels and the fixed decision criterion.

A logistic function was employed to statistically relate the binary dependent variable (judgments of "loud" and "soft") to the strength of the decision variable:

$$\Psi(D) = p(\text{“}loud''\text{)} = \frac{e^D}{1+e^D} = \frac{1}{1+e^{-D}} \ , \qquad (2)$$

where Ψ describes the probability, p, of a "loud" response. Note that sometimes other functions (e.g. normal ogives, Berg, 1989) are used to characterize Ψ, but it has been shown, and is true for the present data, that the estimated weights are to a great extent insensitive to the choice of function (Tang *et al.*, 2005).

Insertion of Equation 1 in Equation 2 gives:

$$\Psi(\mathbf{x}) = p(\text{“}loud'' \mid \mathbf{w}, c, \mathbf{x}) = \frac{1}{1+e^{c-\sum_i w_i x_i}} \qquad (3)$$

The outcome of the experiment is a sequence of "loud" and "soft" responses with associated values for $\mathbf{x}$. The values of $\mathbf{w}$ and c which are most likely to yield the results, under the given model, can be estimated by maximum likelihood optimization. For the logistic function, as applied here, this is also known as logistic regression. Standard test statistics for the validity of the model can be applied and furthermore the logistic regression has the benefit of being directly applicable to binary ("loud" and "soft") data (see for example Cohen, 2003). These are the main reasons for choosing logistic regression over alternative methods used in other studies estimating weights (for example Berg, 1989; Ellermeier and Schrödl, 2000; Lutfi, 1995). Though conceptually different, the various methods at hand give very similar estimates for perceptual weights in practice.

It is seen from Equation 3, that the regression coefficients ($\mathbf{w}$ and c) are not linearly related to the predicted probability of "loud". The non-linear relationship is generally true for logistic regression. In this work however, the logistic function is used as a psychometric function, and the regression coefficients are linearly related to the strength of the underlying decision variable as stated in Equation 1.

In Equation 1, a linear relationship between the decision variable, D, and the segment levels, $\mathbf{x}$, is assumed. Generally, however, the loudness of steady-state sounds is not linearly related to the sound pressure level in decibels, but within the range of levels used in the present experiment (60 dB to 75 dB SPL) the relationship is close to linear (see Moore, 2003).

B. Results of Experiment 1

1. Weighting curves

In total 4598 trials per listener (38 blocks × 121 trials) were used to derive weighting curves. The individual weighting curves are seen in Figure 3. The weights are the scaled regression coefficients of the logistic regression (w_i in Equation 1), which provided the most likely fit to the listeners responses given the segment levels (x_i). The coefficients (w_i) are scaled by a factor so the sum of the ten weights is 1. This normalization makes the *relative* importance of each segment (the weighting curve) comparable across listeners. Different scaling values for different listeners reflect individual differences in sensitivity to level changes, which imply that the overall sensitivity is not reflected in the scaled weighting curves.

Figure 3 contrasts the derived weighting curves for listeners receiving feedback (bottom row) with those of listeners not receiving feedback (top row). For each segment weight, the error bar indicates the 95%-confidence interval. Comparing the size of the error bars to the weight differences between segments, it is clear that the shape of the weighting curve is meaningful for a given listener and not a product of random processes. It is also clear that the weighting curves are highly individual, consider "BJ" vs. "CP" for example: "CP" heavily weights the beginning of the sound, while "BJ" put most weight on the end. For most listeners either the beginning or ending of the sound is weighted more heavily. Exceptions from this are "EH" and "JV", who do not show pronounced weighting of specific segments.

The effect of feedback can be inspected by comparing listeners in the upper row of Figure 3 to those in the lower one. Mean weights for the feedback and no-feedback conditions are shown in Figure 4. Comparing the two mean weighting curves it looks as if feedback did influence the overall shape of the weighting curves. The tendency to emphasize the beginning or the end of a sound seems to be more pronounced in the group of listeners who did *not* receive feedback.

An estimate of the statistical significance of this apparent influence of feedback is not easily made, since (a) all weights are normalized to sum to 1, and (b) the weights for the 10 segments are not statistically independent for a given listener. Therefore the following testing strategy is suggested: If a listener does not receive feedback, ei-

ther the beginning or the ending of the sound is weighted more heavily. In any case (beginning, ending or both being weighted more heavily), the central part of the sound must receive less weight due to the normalization. A score for each listener's weighting of the central part of the sound can be obtained by calculating the sum of the "central" weights 4 to 8. One score is thus obtained for each of the listeners in each group, and the scores in the groups can be compared using a two-sample t-test. It turned out to be highly significant, $t(7.16) = 5.30$ [1]; $p = 0.001$ indicating that the central weights in the no-feedback group were lower than in the feedback group. This in turn means that the curves in the feedback and no-feedback conditions do indeed have different shapes. If non-normalized weights are used, the certainty is even greater, however, this merely implies that listeners receiving feedback performs better than those not receiving feedback.

The same approach can be used to compare the mean weighting curve in each group to "flat" weights (all weights being equal to 0.1). When, in the no-feedback group, the central weights are compared to a value of 0.1, a one sample t-test results in $t(4) = 10.01$; $p < 0.001$, and in the feedback group: $t(4) = 0.54$; $p = 0.62$. That is, the central part of the mean curves is significantly different from optimal weighting for the no-feedback group only. However, from the 95%-confidence intervals in Figure 3 it is clear that some weights are significantly different from the optimal 0.1 for individual listeners both in the feedback and no-feedback group.

C. Discussion

Global loudness judgments of level-fluctuating noise samples produced evidence for a non-optimal temporal weighting in that onsets (and to a lesser extent offsets) were weighted more heavily in contributing to overall loudness. Trial-by-trial feedback significantly reduced this emphasis, effectively resulting in an approximately equal (i.e. optimal) weighting of all segments of the sounds.

The present experiment thus provides support both for equal (as in Buus, 1999; Lutfi, 1990) and unequal (as in Ellermeier and Schrödl, 2000) temporal weights, and though all previous studies used some form of feedback it may be speculated that it may have been implemented more or less efficiently. The fact, however, that those participants receiving feedback in the present study were able to "optimize" their performance to approximate ideal weights, suggests that there is considerable potential for "perceptual learning" in the temporal weighting patterns.

It thus appears, as has been shown for spectral weights (Lutfi, 1995; Southworth and Berg, 1995), there is considerable liberty in how listeners weight the components of perceptual information available, and that, depending on the task requirements, different weighting patterns may emerge. The considerable individual differences evident in the present data also argue for a certain flexibility in the assignment of weights.

Both the individuality of the weighting patterns, and their susceptibility to feedback argue against the observed temporal weighting being solely due to some low-level integration process, as assumed in most models of loudness and temporal integration (e.g. Buus *et al.*, 1997; Glasberg and Moore, 2002). Rather, cognitive processes (allocation of attention, memory, modification of decision rules) appear to be involved.

The outcome of Experiment 1, however, does not specify the nature of the processes very well. It remains open, for example, whether the emphasis of beginning and ending observed in the unbiased listening condition is due to memory effects (primacy and recency), or simply to the perceptual salience of onsets and offsets.

III. EXPERIMENT 2 - LOUDNESS OF TWO-EVENT SOUNDS

A. Introduction

To further clarify the issues raised by Experiment 1, a second experiment was performed, in which sounds of the same duration and temporal structure as those used in Experiment 1 were subjected to a sudden spectral change in the middle of the temporal sequence. The spectral change thus constitutes a salient event which is not tied to primacy or recency, and the effect of which on the temporal weighting pattern may be observed.

B. Method

1. Listeners

Six naive listeners took part in the experiment, none of whom had participated in Experiment 1. Their hearing was screened, and no one was found to have significant hearing loss (more than 20 dB hearing loss at more than one frequency of 0.125, 0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4, 6, and 8 kHz). The participants were five males and one female with an average age of 24 years (range: 22 to 28 years).

2. Apparatus

In Experiment 2, different hardware was used for signal generation: Signals were digitally generated using a sound card (RME HDSP9632) and subsequently converted to an analog signal via a digital-to-analog converter (Tracer Technologies Big DAADI), using 16 bit resolution and a sample rate of 44.1 kHz. The resulting signal was fed to a headphone amplifier (Behringer HA4400) and diotically played over headphones (Beyerdynamic DT 990 PRO).

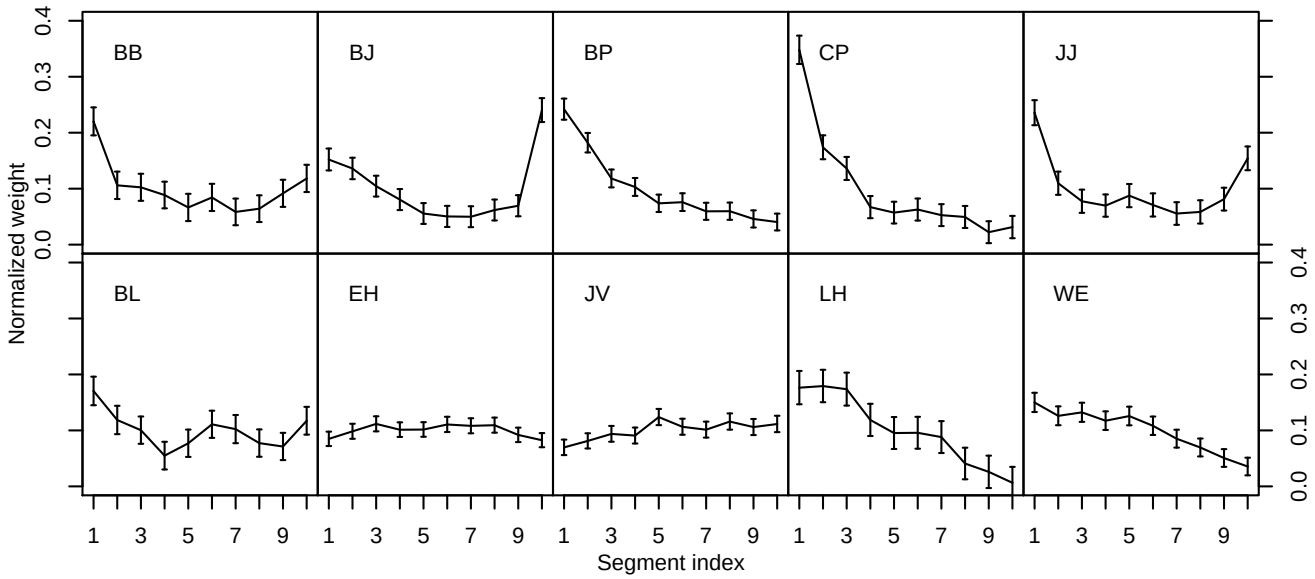


Figure 3. Temporal weights for Experiment 1. Top row: With trial-by-trial feedback. Bottom row: No feedback. The error bars indicate the 95%-confidence intervals for the weights as calculated from the logistic regression.

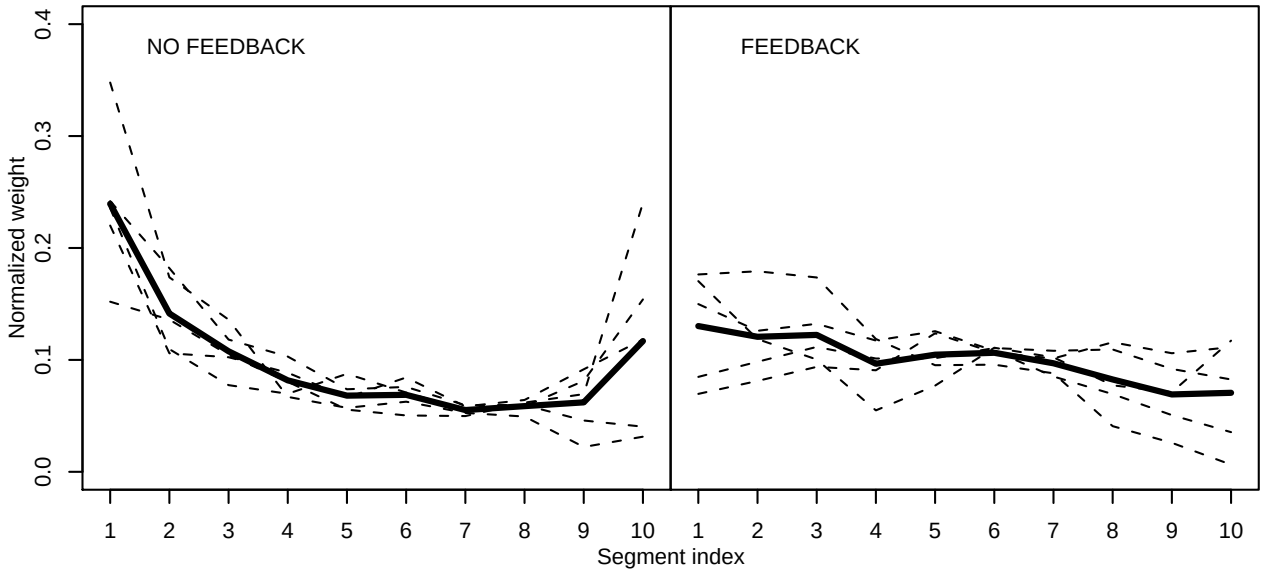


Figure 4. Average weights (Experiment 1) for the two feedback conditions. The thick lines indicate the mean weights across listeners. The broken lines indicate the individual weights from figure 3.

3. Stimuli

As in Experiment 1, all sounds were of one second duration and the levels of the ten temporal segments were chosen from random distributions having the same parameters as in Experiment 1. The only difference was the spectral content of the sounds. In one condition of Experiment 2 the first half of the sound (i.e. the first five segments) was low-pass filtered and the last part (the last five segments) high-pass filtered (see Figure 6). This type of sound is denoted "LH", indicating the change from low to high frequency content. In a different condition the segments were filtered in the opposite order, denoted "HL", i.e. changing from high-pass to low-pass filtered noise. The cut-off frequency was 1 kHz for both high- and low-pass filters. The filtering was done us-

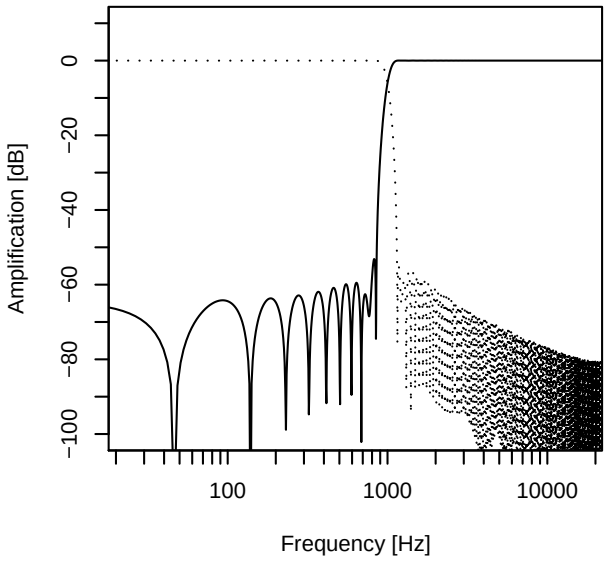


Figure 5. Amplitude responses of the filters used in Experiment 2. The solid curve is the response of the filter used for high-pass filtering and the dotted curve for the filter used for low-pass filtering.

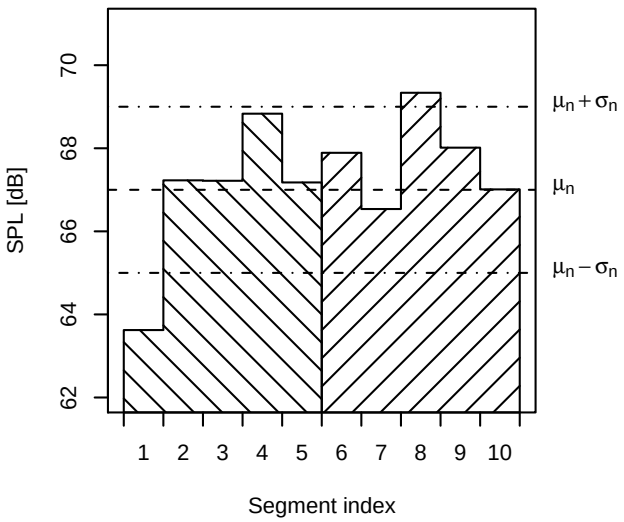


Figure 6. Temporal envelope of a sound sample (here: "noise") as used in Experiment 2. The different shading is to illustrate the change in frequency content in the "LH" and "HL" conditions.

ing digital finite impulse response filters (FIR) filters, the computed amplitude responses of which are shown in Figure 5. The phase response of each filter was linear. The two filtered blocks were aligned so no silent interval occurred. A third condition, where no spectral change occurred, was included for comparison with Experiment 1. In this condition white noise was used as in Experiment 1 (denoted "WN").

4. Experimental procedure

The listeners' task was the same as in Experiment 1. After hearing a single sound, the listener responded whether it was "loud" or "soft". No trial-by-trial feedback was given. After each block of 200 trials the percentage of "correct" responses based on the distribution from which the sounds were drawn was communicated to the participants. Because of the difference in quality of the filtered blocks, the listeners were specifically instructed to judge the composite sound as one whole.

Before data collection started all listeners learned the task in a similar way as in Experiment I. The difference in mean between the "noise" and "signal" distributions was slowly decreased (from 4 to 1 decibels). The training blocks contained fewer trials (50) and both "LH", "HL" and "WN" blocks were included. Feedback on the percentage of "correct" responses, helped listeners to realize whether they where on the right track.

5. Data collection

The experiment was arranged in blocks of 200 trials each. A given block contained either filtered noise (both "LH" and "HL") or broad-band noise ("WN"). In blocks containing filtered noise, "LH" and "HL" trials were presented in a random sequence. A total of 1200 trials per condition ("LH", "HL", or "WN") was presented. Each session, lasting approximately 40 minutes, contained three blocks, one in which unchanging white-noise stimuli were presented ("WN"), and two containing spectral changes ("LH" and "HL"). The order of the blocks was counterbalanced within listeners, and across the six sessions used for data collection. 1140 trials were used per condition and listener in the regression analysis, since the first 9 trials in each block were discarded for practice.

6. Loudness calibration

In order to present the filtered noises at equal loudness, all listeners initially performed individual loudness matches before proceeding to the experiment proper. An adaptive two-interval forced choice one-up/one-down paradigm was used to match samples of either low-pass or high-pass filtered noise to the fixed white-noise reference at 67.5 dB SPL. All sounds had a duration of 0.5 s and there were no random fluctuations of the segment levels.

The resulting loudness matches varied somewhat

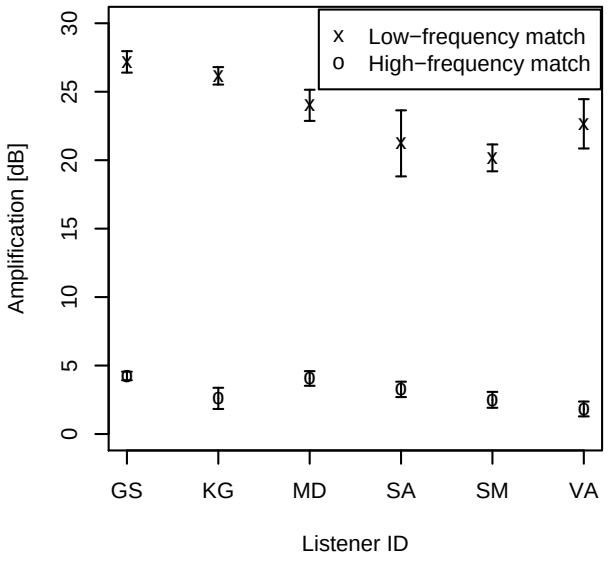


Figure 7. Loudness matches for Experiment 2. Amplification required for the low-pass and high-pass noises to match a 67.5 dB SPL white-noise reference. Individual outcomes for each of the 6 listeners are depicted. The error bars indicate the 95%-confidence interavals.

across listeners (up to ca. 7 dB for the low-pass, and 3 dB for the high-pass noise, see Figure 7). They required the low-pass noise to be raised in level by approximately 23 dB on average to be equally loud as the broadband noise. The high-pass noise required 3 dB amplification on average to achieve the same loudness. In the experiment proper *individual* matches were used for calibration of the filtered blocks.

C. Results of Experiment 2

As in Experiment 1, weighting curves were derived for each individual listener, using logistic regression, separately for the white noise ("WN"), low-high ("LH"), and high-low ("HL") conditions. The estimated weights are depicted in Figure 8, the mean weights across listeners in Figure 9.

The results of the white-noise condition may be compared to those of Experiment 1, in which identical stimuli were used. A similar trend as in the "no feedback" condition of Experiment 1 is found (compare top rows of Figure 8 and 3), with relatively greater weights being assigned to the initial sound segments. The results of the two experiments are very similar (see the mean weights depicted in Figure 4 and 9), except that the emphasis on the initial segments is even greater, and there is no evidence for higher weighting of the ending of the sound in the new experiment. As in Experiment 1, the weighting patterns greatly vary across listeners.

When a spectral change is introduced in the middle of the sound ("LH" and "HL" in the center and bottom rows of Figure 8), the weighting curves show distinctly different patterns. For most listeners the sixth segment (for which the spectral change occurs) receives greater weight in the "LH" and "HL" conditions. It also appears that the order of the high- and low-pass filtered blocks makes a difference for the weighting strategy applied by the listeners, though in idiosyncratic ways, consider "GS" for example: In the "LH" condition his decision is based almost exclusively on the first segment (beginning of low-frequency block), whereas in the "HL" condition both the first and the sixth segment contribute significantly to the decision. Thus, the start of the low-frequency block is always heavily weighted by this listener, but the beginning of the high-frequency block is only weighted heavily if it is also the onset of the entire sound. Listener "SM" almost shows the reverse behavior with respect to the weighting in the two spectral conditions. Finally, "SA" almost seems to ignore the high-frequency part of the sound in both the "LH" and "HL" conditions.

As can be seen from the size of the 95%-confidence intervals depicted in Figure 8, some listeners were clearly more consistent in their weighting than others. Nevertheless, all listeners performed significantly better than chance.

A statistical test as to whether the spectral change ("LH" or "HL") made a difference compared to the non-changing ("WN") condition was performed in the following way: The 6th and 7th segments were defined as reflecting the onset of the spectral change. By summing each listener's weights for these two segments, a score for the weighting of the spectral change was calculated for each listener in each condition. Using these scores, two-tailed, repeated measures t-tests were performed, between the spectral-change conditions and the non-changing condition. They revealed the weights for the critical segments (6 and 7) to be significantly greater in the spectral-change conditions, both when comparing "LH" with "WN": $t(5) = 3.02$, $p = 0.03$, and when comparing "HL" with "WN": $t(5) = 2.65$, $p = 0.045$. Thus, the increased weighting given to the onset of a new spectral event (see Figure 9) appears to be statistically significant.

D. Discussion

Experiment 2 showed the temporal location at which a spectral change occurred to receive just as much weight as the initial onset of the composite sound. This is consistent with the idea of perceptual weighting being guided by salient events. These may be onsets, offsets, spectral shifts, or qualitative changes yet to be investigated such as changes in spatial location, etc.

The results of Experiment 2 are not easily reconciled with a memory explanation based on primacy and recency effects, at least not one that requires the entire

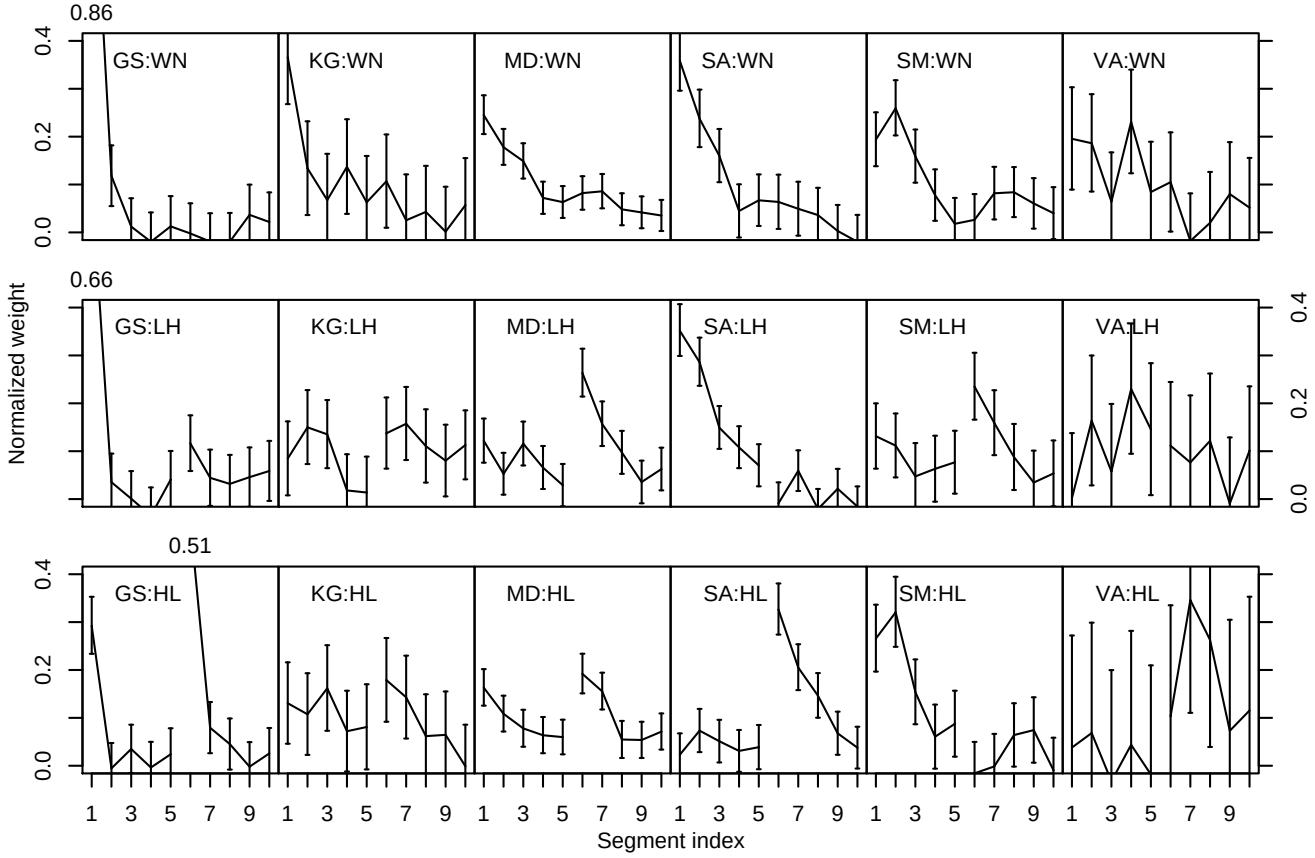


Figure 8. Weighting curves for all listeners in Experiment 2. Columns represent listeners while rows contain the different experimental conditions. First row: Broad-band noise with no frequency change. Second and third row: Two-event conditions; low-high and high-low changes respectively. The onset of the spectral change is indicated by a break in the weighting curve. The error bars indicate 95%-confidence intervals for the weights as calculated from the logistic regression.

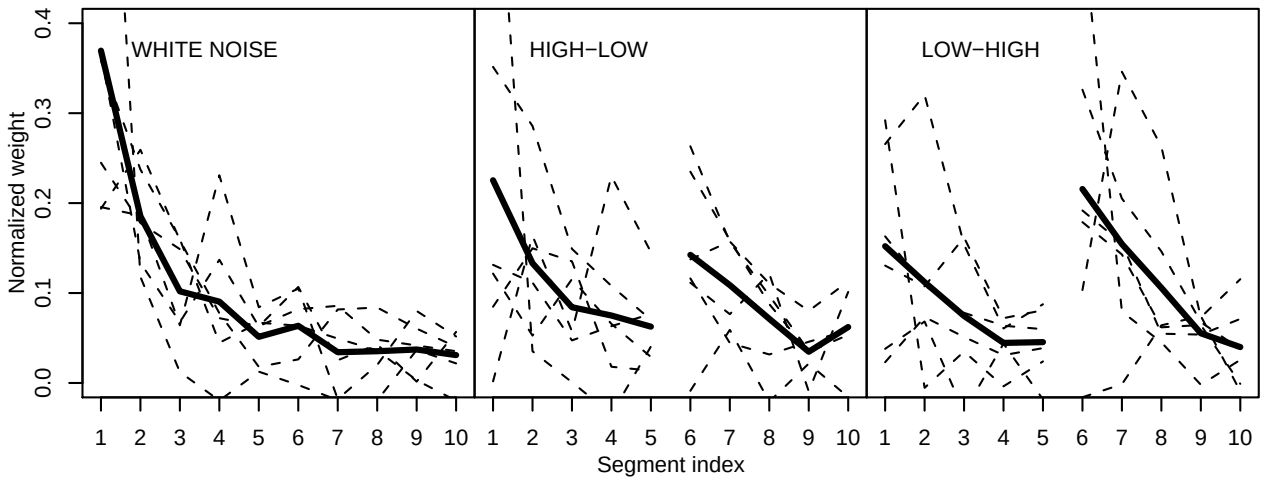


Figure 9. Average weights (Experiment 2) for the three spectral conditions. The thick lines indicate mean weights. Broken lines are individual weights from Figure 8.

sound to be stored in memory in a simple sequential way. Whether assumptions about "resetting" the onset detector, or separate storage of the two spectral events might remedy the situation, is doubtful.

IV. FINAL DISCUSSION AND CONCLUSION

The results of the present experiments appear incompatible with the notion of an automatic, accumulative integration process as hypothesized by most loudness models (e.g. Glasberg and Moore, 2002; Zwicker, 1977). A major outcome of these loudness models is to generate a continuous loudness curve, which is to account for the results of e.g. temporal masking experiments and subjective loudness matches of modulated sounds. But to predict how listeners arrive at global loudness judgments requires further stating how this continuous curve is "integrated" to produce a single judgment. The present data address both of these stages.

In the calculation of a continuous loudness curve, all current models operate with some sort of temporal summation with a critical time coefficient in the range from 20 ms to 50 ms depending on whether the loudness curve is rising or falling (Glasberg and Moore, 2002; Grimm *et al.*, 2002; Zwicker, 1977). It is therefore impossible for loudness determined by these models to fluctuate any faster than the time coefficients allow. The fact that in the present experiment, for some listeners, adjacent segments were weighted very differently (see Figure 3), implies that their "continuous loudness" must fluctuate at least as rapidly as the segment duration of the sounds (100 ms) or else a particular segment could not be "singled out" receiving extra weight. Thus, though the time coefficients of the models are not in direct contradiction with the observed weighting patterns, there is some indication that the integration taking place is not a simple "smoothing" process. In their loudness model, Glasberg and Moore (2002) introduce a further stage of determining "long-term" loudness, with integration coefficients of approximately 100 ms for rising and 2000 ms for falling loudness. These long time coefficients are not compatible with the results of the present experiments.

When it comes to integrating the sensory information into a loudness judgment, the present experiments provides further evidence against the operation of simple loudness integration:

(1) Weights derived for the 10 temporal segments defined were not uniform, but rather, in the unbiased, non-feedback conditions of Experiment 1 and 2, provided evidence for perceptual emphasis of onsets and offsets. That is not predicted by any of the current loudness models. Nor is it predicted by practical measurement rules (e.g. Zwicker and Fastl, 1999) that assume values close to the maximum (e.g. the 4th percentile Grimm *et al.*, 2002; Zwicker and Fastl, 1999) to determine the loudness of a time-varying sound. All of these rules would, for the randomly varying sounds used in the present experiments, imply "flat" weighting curves to result.

(2) When trial-by-trial feedback was provided in Experiment 1, listeners adapted their temporal weights to approach an optimal, uniform weighting of all stimulus segments. Such a "learning effect" is hard to reconcile with the notion of an automatic integration process operating in the auditory periphery with a relatively long time coefficient. Rather, the listener must have access to some representation of the segment loudnesses (prior to integration) with a finer resolution than the segment duration in order to modify weights to maximize the percentage of "correct" responses.

(3) When a qualitative change was introduced into the noise sequence by switching the spectrum from a low-pass to a high-pass characteristic (or vice versa) in Experiment 2, listeners strongly weighted the onset of the "new" sound feature, thus boosting weights in the central portion of the composite stimulus. That is inconsistent with temporal wideband energy integration which would be "blind" to the spectral change; it is also inconsistent with a memory explanation based on a "primacy" and "recency" advantage.

(4) All of the weighting patterns observed exhibited considerable inter-individual variability. That in itself argues against a low-level integration mechanism, which one would not assume to leave degrees of freedom for individual idiosyncrasies. Rather it suggests some cognitive process to be involved, which can be controlled by the listener to some extent.

What then are the alternatives for understanding the weighting of level information, and its adaptability to various listening conditions? It appears that, in the time range of several hundred milliseconds investigated here, different stimulus segments must be individually accessible, granting the listener "multiple looks" (Viemeister and Wakefield, 1991) on a temporal loudness pattern. Depending on the task requirements (Experiment 1) or on stimulus features (the spectral changes in Experiment 2), these "looks" may be weighted differently, under implicit control by the listener. The particular salience of onsets and offsets, as well as qualitative changes in the stimulus, may be due to mechanisms of memory, or more likely to the "distinctiveness" (Neath, 1993; Neath *et al.*, 2006) of these events in relation to other stimulus components, thereby attracting greater perceptual weight.

How could these hypotheses be put to further tests? If memory was a factor, one might expect the timing of the event sequence to play a crucial role. Furthermore, to explore the "distinctiveness" concept, salient changes other than spectral ones (e.g. spatial lateralization) might be explored, or an event could be generated by switching

from coherent to incoherent noise samples of the carrier signal across the two ears. Potentially, the segment levels could also be different across the ears, providing a means to examine both temporal and binaural loudness summation.

Hopefully, based on such research, a clearer picture will emerge, on how the loudness of time-varying sounds is determined by both perceptual and cognitive processes.

V. ACKNOWLEDGMENTS

We would like to thank Florian Wickelmaier for helpful hints regarding the statistical analysis of the results. This research was carried out as part of the "Centercontract on Sound Quality" which establishes participation in and funding of the "Sound Quality Research Unit" (SQRU) at Aalborg University. The participating companies are Bang & Olufsen, Brüel & Kjær, and DELTA Acoustics & Vibration. Further financial support comes from the Ministry for Science, Technology, and Development (VTU), and from the Danish Research Council for Technology and Production (FTP).

(1) Non-integer degrees of freedom results since a Welch-Satterthwaite approximation was used, which means that same variance for the weights in the feedback and no-feedback group is not assumed

Berg, B. G. (**1989**). "Analysis of weights in multiple observation tasks", J. Acoust. Soc. Am. **86**, 1743–1746.

Buus, S. (**1999**). "Temporal integration and multiple looks, revisited: Weights as a function of time.", J. Acoust. Soc. Am. **105**, 2466–2475.

Buus, S., Florentine, M., and Poulsen, T. (**1997**). "Temporal integration of loudness, loudness discrimination, and the form of the loudness function.", J. Acoust. Soc. Am. **101**, 669–680.

Cohen, J. (**2003**). *Applied multiple regression/correlation analysis for the behavioral sciences*, 3rd edition (Lawrence Erlbaum, Mahwah, NJ).

Doherty, K. A. and Lutfi, R. A. (**1996**). "Spectral weights for overall level discrimination in listeners with sensorineural hearing loss.", J. Acoust. Soc. Am. **99**, 1053–1058.

Doherty, K. A. and Lutfi, R. A. (**1999**). "Level discrimination of single tones in a multitone complex by normal-hearing and hearing-impaired listeners.", J. Acoust. Soc. Am. **105**, 1831–1840.

Ellermeier, W. and Schrödl, S. (**2000**). "Temporal weights in loudness summation", in *C. Bonnet (Ed.), Fechner Day 2000. Proceedings of the 16th annual meeting of the International Society for Psychophysics*, 169–173 (Université Louis Pasteur, Strasbourg).

Glasberg, B. R. and Moore, B. C. J. (**2002**). "A model of loudness applicable to time-varying sounds", J. Audio Eng. Soc. **50**, 331–342.

Grimm, G., Hohmann, V., and Verhey, J. L. (**2002**). "Loudness of fluctuating sounds", Acust. Acta Acust. **88**, 359–368.

Kortekaas, R., Buus, S., and Florentine, M. (**2003**). "Perceptual weights in auditory level discrimination.", J. Acoust. Soc. Am. **113**, 3306–3322.

Lutfi, R. A. (**1990**). "Informational processing of complex sound. II. Cross-dimensional analysis.", J. Acoust. Soc. Am. **87**, 2141–2148.

Lutfi, R. A. (**1995**). "Correlation-coefficients and correlation ratios as estimates of observer weights in multiple-observation tasks", J. Acoust. Soc. Am. **97**, 1333–1334.

McFarland, D. J. and Cacace, A. T. (**1992**). "Aspects of short-term acoustic recognition memory: modality and serial position effects.", Audiology **31**, 342–352.

Moore, B. C. J. (**2003**). *An Introduction to the Psychology of Hearing*, 5th edition (Academic Press, San Diego, CA).

Neath, I. (**1993**). "Distinctiveness and serial position effects in recognition.", Mem. Cognit. **21**, 689–698.

Neath, I., Brown, G. D. A., McCormack, T., Chater, N., and Freeman, R. (**2006**). "Distinctiveness models of memory and absolute identification: evidence for local, not global, effects.", Q. J. Exp. Psychol. **59**, 121–135.

Saberi, K. (**1996**). "Observer weighting of interaural delays in filtered impulses.", Percept. Psychophys. **58**, 1037–1046.

Southworth, C. and Berg, B. G. (**1995**). "Multiple cues for the discrimination of narrow-band sounds", J. Acoust. Soc. Am. **98**, 2486–2492.

Stecker, G. C. and Hafter, E. R. (**2002**). "Temporal weighting in sound localization.", J. Acoust. Soc. Am. **112**, 1046–1057.

Surprenant, A. M. (**2001**). "Distinctiveness and serial position effects in tonal sequences.", Percept. Psychophys. **63**, 737–745.

Tang, Z., Richards, V. M., and Shih, A. (**2005**). "Comparing linear regression models applied to psychophysical data", J. Acoust. Soc. Am. **117**, 2597.

Viemeister, N. F. and Wakefield, G. H. (**1991**). "Temporal integration and multiple looks.", J. Acoust. Soc. Am. **90**, 858–865.

Willihnganz, M. S., Stellmack, M. A., Lutfi, R. A., and Wightman, F. L. (**1997**). "Spectral weights in level discrimination by preschool children: synthetic listening conditions.", J. Acoust. Soc. Am. **101**, 2803–2810.

Zwicker, E. (**1977**). "Procedure for calculating loudnesss of temporally variable sounds.", J. Acoust. Soc. Am. **62**, 675–682.

Zwicker, E. and Fastl, H. (**1999**). *Psychoacoustics: Facts and models* (Springer, Berlin).

Chapter 3

Modeling level discrimination: Non-linear and across-trial effects

3.1 Introduction

Previous analyses (Pedersen and Ellermeier, 2006) showed how listeners weight the temporal components of a sound when judging its overall loudness. The analysis required two assumptions: (1) When making a loudness judgment, the listeners consider each trial in isolation, and (2) the loudness judgment is based on a linear weighting of the ten segments constituting a single sound. There are several reasons why these assumptions may not by justified:

- A loudness judgment does not just depend on the stimulus presented on a given trial.
- Specifically, a judgment may depend on the "history" of the sounds encountered (DeCarlo and Cross, 1990): Listeners may learn and change behavior throughout the experiment or they may be unable to consider a sound of a given trial in complete isolation from the sounds presented on previous trials.
- Furthermore, listeners' decisions may be influenced certain statistics of the sound exposure: The maximum sound level of a sound may determine its judged loudness or the variability of the sound pressure level within a trial may have an effect.

This chapter will address questions like these. The chronology of the chapter is to examine first to which extent loudness judgments do not depend on the sound of a single trial only. To that end, several aspects are considered: (1) Effects which do not directly depend on the sound level (bias toward either "loud" or "soft" or special response patterns (alternating "loud"/"soft" responses for example)). (2) It will be demonstrated how the loudness judgment on a given trial depends on the levels presented on previous trials. In following sections it will be examined how listeners may process information of single trials before arriving at a loudness judgment. To that end, first some empirical observations are summarized, where the statistical probability of a given response as a function of the segment levels of a given sound is analyzed. This will help to identify important aspects, which a model of the decision process must be able to account for. In the following, several suggestions for modeling the decision process are given. First, the ability of two of the more established theories for temporal integration in hearing (model of temporal loudness integration by Glasberg and Moore (2002) and temporal window model by Oxenham and Moore (1994)) to predict the behavior of listeners is examined. As theses two models do not seem to explain the behavior of the listeners, a range of models are formulated mathematically and fitted to the responses of the listeners. The ability of the fitted models to predict the listeners' behavior are then compared to identify which models include important aspects of the decision process.

It has been demonstrated that reaction time is related to loudness (Wagner *et al.*, 2004), therefore the response time of the listeners is also analyzed. This reveals a strong correspondence between sound level and response time, but not in a way which can be simply related to the listeners' actual loudness judgments. In spite not being directly related to loudness, the response time analysis gives insight into the complexity of the decision process.

The chapter ends with analyzing temporal weights in a task where listeners had to compare the loudness in pairs of level-fluctuating sounds (Ellermeier and Schrödl, 2000). This extends the analysis to also examine how listeners handle several integrated loudness impressions in a comparison task.

If not stated otherwise, all results presented in this chapter are based on the data obtained in Experiment 1 of the present thesis (Pedersen and Ellermeier, 2006). Potentially, the same analyses could also be applied to the results of Experiment 2 of the thesis, but this has been avoided to maintain a clear focus throughout the chapter and because of the larger amount of data collected for each listener in Experiment 1 allowing for greater statistical power.

The overall goal of this chapter is to arrive at conclusions about the factors involved in the temporal integration of loudness, extending the study presented in the previous chapter by not assuming that loudness judgment is based on a simple linear (weighted) summation of segment levels.

3.2 Effects across trials

In the previous chapter it was assumed that a listener's "soft" or "loud" response depended only on the ten segment levels of the trial in question. There are, however, several reasons why listeners may not always behave in accordance with this assumption:

- Response category bias - When a listener is uncertain about the response he/she may use only one type of answer like an "I don't know" category.
- Persistence of response bias - A listener may feel uncomfortable pressing the same button many times in a row even though a long series of the same response might be correct.
- Unstable decision criterion - Listeners need a reference to discriminate "loud" and "soft", and this reference may for example depend on the levels of the sounds of previous trials.
- Effects of learning - The listener may get better throughout the time-course of the experiment or change his/her decision criterion.

In the following, the listeners' behavior will be analyzed in an attempt to clarify the questions raised.

3.2.1 Level-independent effects

Response category bias

All listeners are biased toward answering either "loud" or "soft" as summarized in Table 3.1. The table simply gives the overall percentage of "loud" responses. When assuming a binomial distribution of the responses, then the estimated probability of "loud" is identical to the percentage of "loud" judgments and the confidence intervals for its estimated value can be calculated. The bias may not always be independent of level, if for example a listener applies a relatively low "reference level" in his/her judgments, the amount of "loud" responses will be higher than the amount of "soft" responses. This type of bias is accounted for in the modeling as described in section 3.5, and only level-independent bias is of concern here. Discriminating these two types of bias is not easily done, since listeners did not give any reason why they made a given decision.

	BB	BJ	BP	CP	JJ	BL	EH	JV	LH	WE
Percent "loud"	54.0	56.5	53.9	55.9	48.3	47.5	52.3	53.5	60.2	54.0
p-value	$***$	$***$	$***$	$***$	0.024	0.001	0.002	$***$	$***$	$***$

Table 3.1: Response category bias. Each listener's percentage of "loud" responses in the first row and, in the bottom row, the probability that this could have occurred if the real probability of "loud" was 50% (p-value, two-tailed binomial test). "$***$" indicates $p < 0.001$.

Persistence of response bias

A listener may feel uncomfortable about pressing the same button many times in a row, or may continuously press the same button if he/she is not certain of the responses. In either case a given response will depend on the listener's earlier responses. Calculating the autocorrelation sequence of a listener's responses at different lags can be used to estimate this effect. For example at a lag of one, the autocorrelation shows how likely it is that a listener's current response will be the same as the previous, where the correlation coefficient ranges from -1 (the listener's responses are always opposite in two consecutive trials) to 1 (the listener's responses are always the same in two consecutive trials). A more detailed description is for example given by Venables and Ripley (2002). The autocorrelation sequence for the responses is shown for each listener in Figure 3.1. It should be noted that some trials in the calculation of the autocorrelation sequence occurred in different blocks in the listening experiment. For example, at a lag of one, the first trial of a block is compared to the last trial of the previous block. This comparison may not be relevant, but it happens relatively seldom that the two trials compared fall into different blocks.

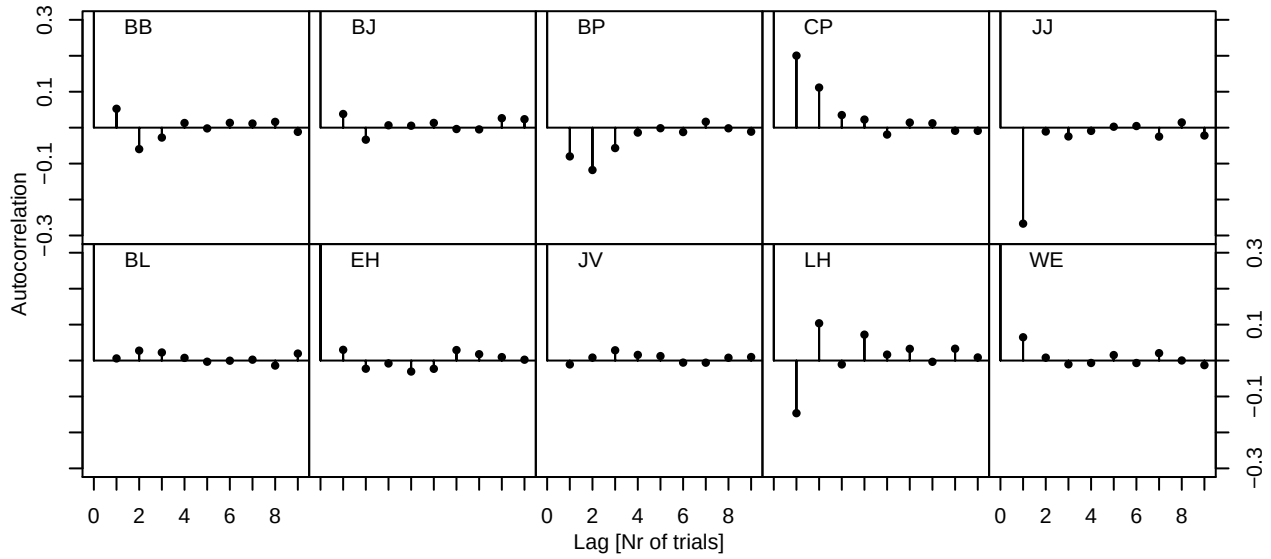


Figure 3.1: Autocorrelation sequences for the responses of each individual listener. The lag is indicated on the x-axis, and at a lag of 1 the value on the y-axis indicates how a listener's previous response correlates with the response of the current trial. Top: Listeners not receiving feedback; bottom: Listeners receiving feedback.

The order of generated "loud" and "soft" sounds was random in the experiment, meaning that the autocorrelation sequence of the generated stimuli is 0 for any lag different from 0. Thus, ideally, the autocorrelation sequence of a listener's responses should also be 0. From Figure 3.1 it is seen that this is not the case for all listeners. On a given trial listener JJ is likely to give the opposite response as in the previous trial (large negative correlation coefficient at a lag of 1) and listener CP is likely to give the

same response in two (or even three) consecutive trials (positive correlation coefficients at a lag of 1,2, and 3). Listener LH seems to give alternating responses. It could be hypothesized that listeners benefit from feedback to avoid this behavior, because the feedback would "tell" them that sometimes multiple sounds of the same category occur in a row. Further they may notice how often a given type of sound ("soft" or "loud") occurs in a row and how often this happens and adapt their behavior accordingly. It is therefore of interest to determine to which extent listeners are able to adapt to the "statistics" of the signal generation. This may be done in the following way: All responses of a listener are considered in their presented order. On this list it can be counted how often the listener gives sequences of the same response in a row. To understand this consider the following series of responses:

```
loud, soft, soft soft, loud, soft
```

On this list there is one occurrence of "soft" in a sequence of 3, one occurrence of "soft" in a sequence of 1 and 2 occurrences of "loud" in a sequence of 1. The responses of all listeners were analyzed this way and the results are shown in Figure 3.2. Each count was multiplied by the length of the relevant sequence, so for example the 3 "soft" in a row in the example, would contribute with a value of 3 to the y-value at the "3-soft-in-a-row" category. This makes the total sum of the y-values equal to the number of trials. The figure also shows (as a solid line) the behavior of a model giving a "loud" response when the mean level of the segments of a sound is bigger than the overall mean and "soft" response if it is less. The line thus represents the "statistics" of the signal generation. From the figure it is seen that listeners in the feedback group generally behave in accordance with the "statistics of generation" curve except for LH's "soft" responses for which it is seen that LH hesitates to press "soft" in longer sequences. In the no-feedback group people generally deviate from the "statistics" curve. BP, CP, and JJ deviate most where JJ is reluctant to give the same response many times in a row and CP gives the same response in a row too often. BP often gives the same response two times in a row. BB and BL give continuous "loud" responses too often. In summary, people in the feedback group seem to "statistically" adapt to the feedback. It should be noted that this form of adaption does not give any benefit in obtaining better performance scores since the segment levels on different trials are uncorrelated.

3.2.2 Level dependent effects across trials

Weighting contrasts

For each response a listener gives he/she needs a reference for the judgment. For example the listener may have a good impression of the overall mean level of all sounds or the listener could simply compare the level of the sound of the current trial to that of the previous. The latter can be tested for the weighting model, if the prediction of a response is based on segments across several trials. Consider the following model for example:

$$L(\mathbf{x}) = \sum_{i=1}^{10} w_{0,i}x_{0,i} + \sum_{i=1}^{10} w_{1,i}x_{1,i} + \sum_{i=1}^{10} w_{2,i}x_{2,i}$$

The weights $w_{0,i}$ are weights put on the segments of the current trial while $w_{1,i}$ are weights put on the previous trial and $w_{2,i}$ are weights put on the trial before that. The weights were fitted as described by Pedersen and Ellermeier (2006). The fitted weights are shown for all listeners in Figure 3.3.

Ideally, only the weight put on the segments of the current trial should be different from 0, since the task of the listener is to judge only the sound of a single trial. Listener BP and JJ deviate most from this in that they seem to contrast levels of previous trials with levels of the current, maintaining the same shape of the weighting curve, but with opposite sign. It can be observed that, in absolute measures, some heavily weighted segments of previous trials are more important than weakly weighted segments

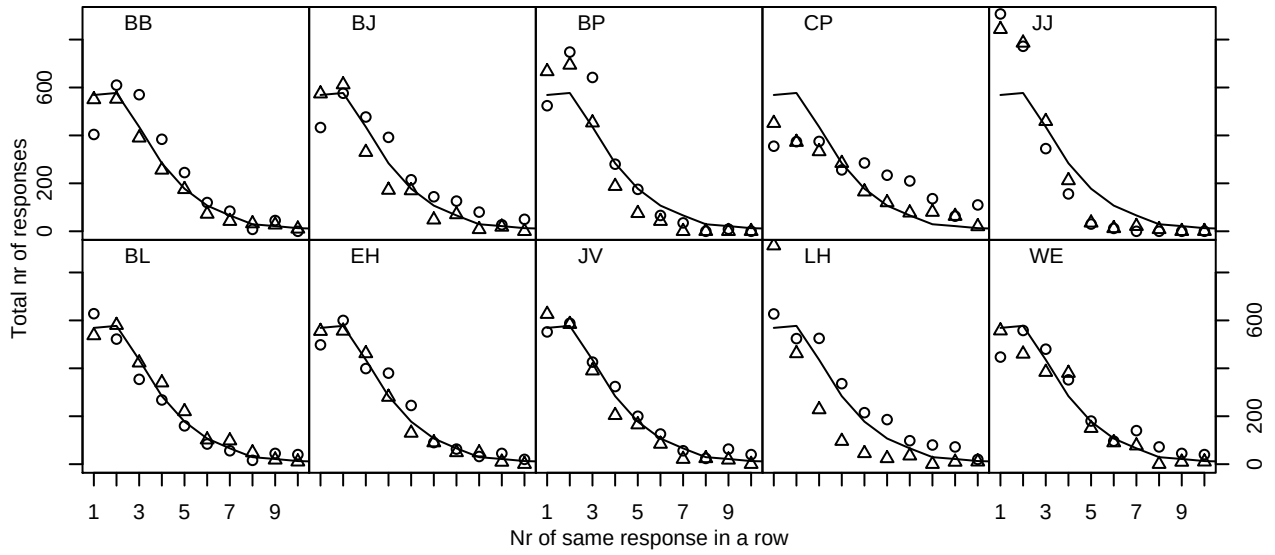


Figure 3.2: Frequencies of same response in a row for individual listeners. The x-axis indicates the possible numbers of same response in a row. Circles correspond to "loud" responses in a row and triangles "soft". The y-axis shows the number of trials falling in series of length as indicated on the x-axis. If for example a listener gives three "loud" responses in a row this will contribute with a count of 3 for the circle at an x-value of "3 in a row". The sum of all "circle and triangle"-values is equal to the total number of trials. The line shows the "behavior" of a simple model making judgments based on the mean value of the levels of the segments. Listeners not receiving feedback in the top row and listeners receiving feedback in the bottom row.

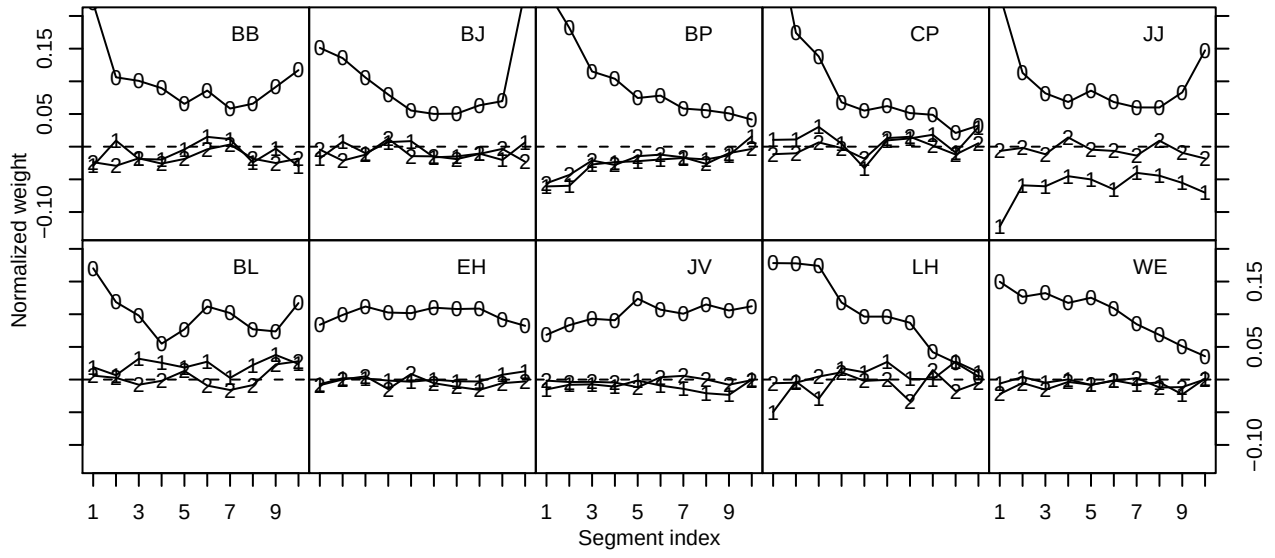


Figure 3.3: Weighting contrasts for individual listeners across three trials. All weights are scaled by the same factor chosen so the sum of the weights of curve "0" is 1. The numbers on the curves indicate the lag in relation to the current trial.

of the current trial. In the feedback group this behavior is not observed to the same extent. This could be an indication that feedback is helpful for maintaining a fixed reference level for the judgments.

3.2.3 Changes in weighing curve as a function of time

It may be of concern whether the listeners maintain the same weighting curves throughout the entire experiment, or they change over time. This is especially interesting for the listeners in the feedback group as it may seem plausible that they slowly changed their weighting according to the feedback. In an attempt to graphically examine this, weighting curves were derived from data obtained in different phases of the experiment as a function of time from the beginning of the experiment toward the end. The obtained data for each listener were divided into ten overlapping portions. The results of 1000 trials were included in each portion. The total amount of trials per listener was 4598 so approximately 5 independent estimates of each listener were obtained for different phases of the experiment. All trials were sorted according to their index in time in the total series of trials, and from the sorted list of trials the portions were picked. 10 different estimates of the weighting curves were made for each listener as a function of index, so for example the first estimate of weights is based on the 1000 first judgments made by a given listener and the next estimate is based on trial number 400 to 1399. The weighting curves for all listeners are shown in Figure 3.4, where different listeners are shown in different columns and their weighting curves for the different phases of the experiment are shown in different rows. The weighting curves were calculated as described by Pedersen and Ellermeier (2006).

The main conclusion of inspecting Figure 3.4 is that listeners to a large extent maintain the same weighting pattern throughout the experiment. Only one listener, JV, behaves according to the hypothesis that listeners in the feedback group get closer to "optimal" weighting throughout the experiment. In the beginning listener JV emphasized the end of the sound, where as, at the end of the experiment, JV's weights are nearly "flat". The relatively stable weighting patterns indicate that the weighting curve is settled before the experiment proper, potentially in the initial training.

3.3 Modeling the decision process

In our earlier work on temporal weights (Pedersen and Ellermeier, 2006), listeners were always assumed to give loudness judgments based on the weighted sum of the 10 segments of a given sound. This analysis revealed that listeners apply very different individual weights. It has, however, not been shown that the temporal weighting is the only factor involved in the decision process. There have been different suggestions in the literature for how listeners arrive at an overall judgment of loudness: Moore *et al.* (2003) propose a simple mean of the loudness curve (similar to the case where all temporal weights are equal), and Zwicker and Fastl (1999) propose that the level exceeded a certain percentile of the time is decisive for loudness judgments, further elaborated on by Grimm *et al.* (2002). More complex aspects have been found to be involved in listening experiments: Stecker and Hafter (2000), for example, found that ramped sounds are perceived louder that damped sounds, and Oberfeld and Plank (2005) found that, in judging loudness of sounds slowly fading in, listeners emphasized the temporal segment where fading ended.

Because of the random segment variation of the sounds used in the experiment, the sounds posses, and to different degrees, many of the properties found to be influential. Therefore it seems worthwhile to examine how different concepts help in explaining the performance of the listeners.

3.3.1 Model-independent observations

Before turning to the description of specific models, there are a number of observations, which can be made independently from assumptions about the decision process, and which provide insight into

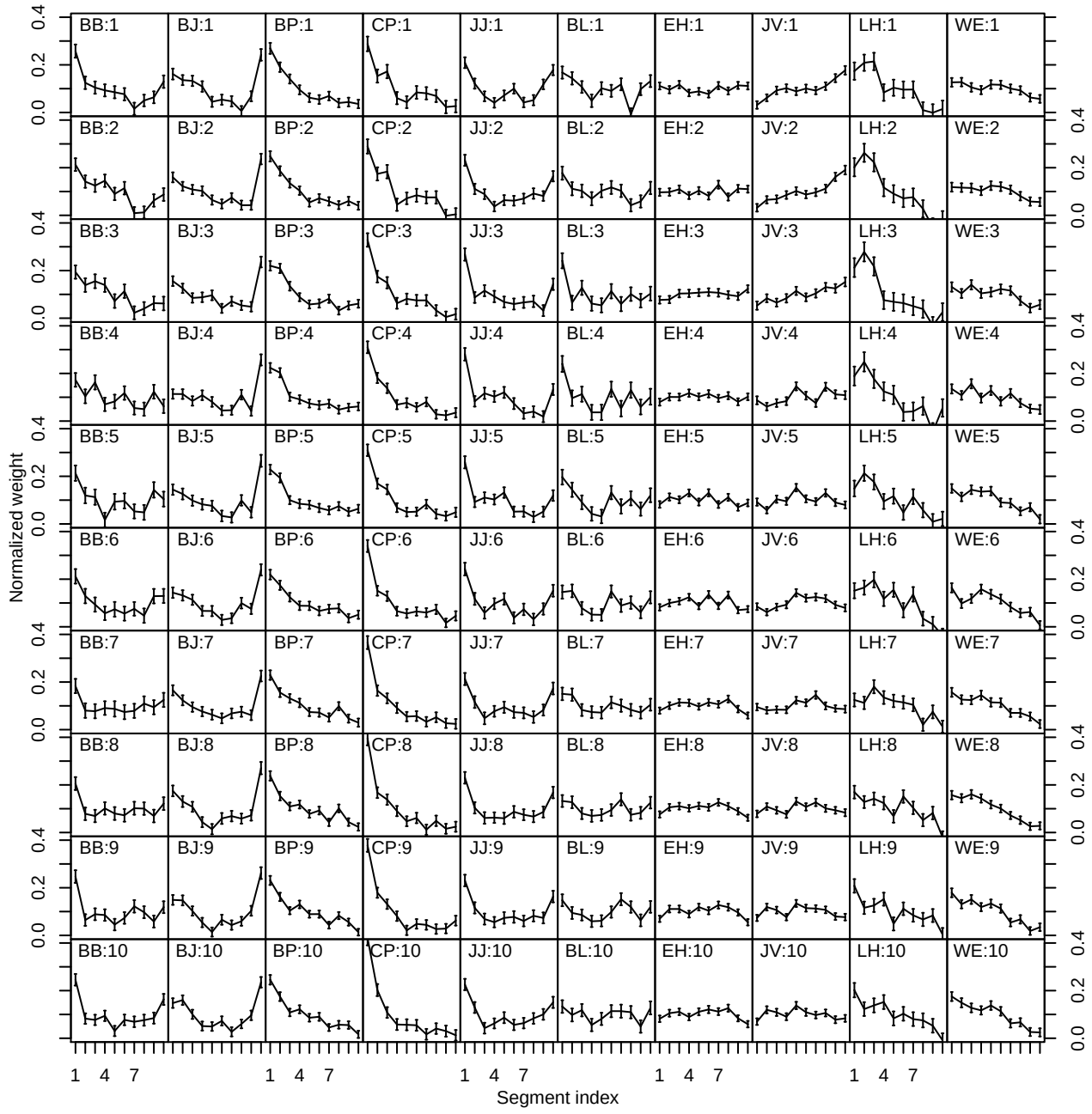


Figure 3.4: Changes in weighting curves for all listeners in different columns. The weighting curves are fitted to responses in different phases of the experiment as indicated in the upper left corner of the graphs and changing across the rows. “1” indicates the first phase of the experiment and “10” the last. The first 5 listeners, from left to right, are the listeners in the no-feedback group. Error bars indicate plus/minus one standard deviation as calculated in the logistic regression.

which aspects a model must be able to account for.

Smoothing curves

For visual presentation of results a smoothing function will be extensively used throughout this chapter, so the following short-hand notation is introduced:

$$Y(X)_{\overline{nr\_intervals,nr\_in\_interval}}$$

Here Y depends on X, where Y is a vector of observations at physical levels given by the values of the vector X. The expression tells that Y and X have been smoothed according to the (integer) subscripts (*nr_intervals* and *nr_in_interval*) in the following way:

(1) X is sorted in ascending order.

(2) If elements were reordered in X, a similar reordering is made of the elements of Y.

(3) X is divided into *nr_intervals* (potentially overlapping) intervals, so there are *nr_in_interval* elements in each interval. The intervals are equidistant in term of their *index* in the sorted vector X. The first interval contains the first elements of the sorted X and the last interval contains the last elements of the sorted X.

(4) For each interval the mean value of the *nr_in_interval* elements is calculated. The calculated values are the "smoothed" X.

(5) Similarly, mean values are calculated for intervals of Y, which are then the "smoothed" Y.

The confidence interval of the estimated mean of the values of Y in a given interval can be calculated using a simple t-test. If the values have a binomial nature a confidence interval can be calculated based on the binomial distribution.

Non-linearity as a function of overall level?

For the segment levels used in the experiment (60 dB to 75 dB SPL), loudness is almost linearly related to the sound pressure level in decibels for steady-state sounds (see for example Moore, 2003a). Still, it may be of concern whether the overall (mean) level of a sound is linearly related to the perceived level. Therefore the probability of a "loud" response has been depicted as a function of the mean level of the ten segments of a given sound in Figure 3.5. In the figure the gray-shaded area of 95% confidence shows the actual "smoothed" responses for each listener. The solid line shows the best fitted curve (more carefully described in Chapter 3.5) relating the listener's performance to a simple mean value of the ten segment levels. A non-linear relation between response and segment levels would be seen as a systematic deviation of the solid curve from the "smoothed" performance. Seemingly there is no significant disagreement between the fitted curve and the listeners' actual performance, thus it may be concluded that there is not a significant non-linear term in the relationship between overall level and perceived level.

Conditioned on single segment (COSS) functions

The graphical illustration of the relation between response and overall level in Figure 3.5 is "blind" to effects concerned with single segments such as the first segment being more significant for the judgment or there could be more complicated interaction effects. Inspired by the method described by Berg (1989), the importance of each segment levels is examined. The method uses COSS function

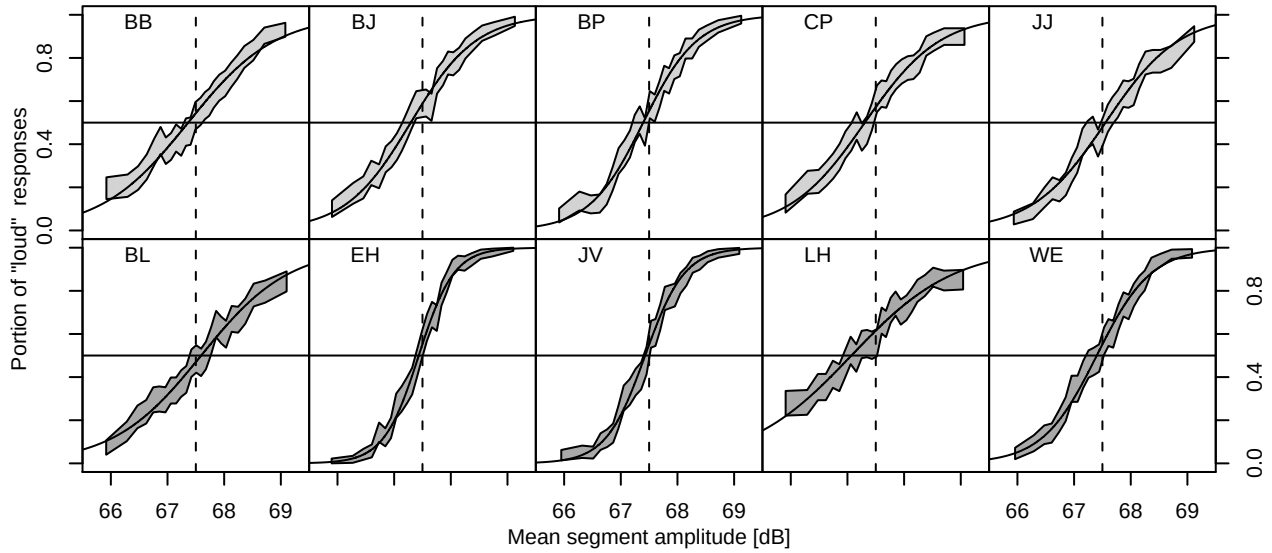


Figure 3.5: Probability of "loud" as a function of mean level of the individual ten segment levels of a sound. $Response(x_{mean})_{\overline{25,250}}$

(Conditional On Single Stimulus) to relate the level of each segment to the probability of a "loud" response. In the present work this is implemented by generating plots similar to those in Figure 3.5, but instead of having only one plot per listener, one plot per listener per segment is generated, where the x-axis denotes the sound pressure level of the relevant segment only. In Figure 3.6 such COSS functions are plotted for all listeners.

If listeners behaved according to a "linear" rule the COSS functions would be approximately straight lines with a slope indicating its significance relative to other segments (strictly speaking this is only true when the COSS functions are transformed an accordance with the statistics of the signal generation). This does indeed seem to be the case for some listeners, EH and JV for example. For others (BB and BL for example) it does not seem to be the case. Especially for low segment levels many listeners seem to respond "loud" more often as would be assumed according to a linear model. As a rough summary of the results it can be stated that, when the level of a given segment is above a certain value (67.5 dB for example) it seems to contribute to the overall loudness in a way which is compatible with a linear model, but when the level is below this value, it does not contribute as much as would have been expected from a linear model, the reason being: When a segment level is below the overall mean (67.5) it is lower (on average) than other segments and the lower it is, it should make a "soft" response more likely. Some listeners, however, responds "loud" too often, which implies that the low segment level did not contribute as much as expected as to give a "soft" response. There is even some weak evidence that for the lowest segment levels the probability of a "loud" response is higher than for moderately higher levels (see segment 2 of BB in Figure 3.6 for example). In some cases this would imply that decreasing the level of a segment increases the probability of a "loud" response. A model, which is able to account for this is considered later.

Summary

In summary the presentation of the results in Figure 3.5 and 3.6 showed:

- In describing overall performance of listeners (Figure 3.5), a logistic curve (solid line in the figure) followed the listeners responses well - no overall deviation was identified.

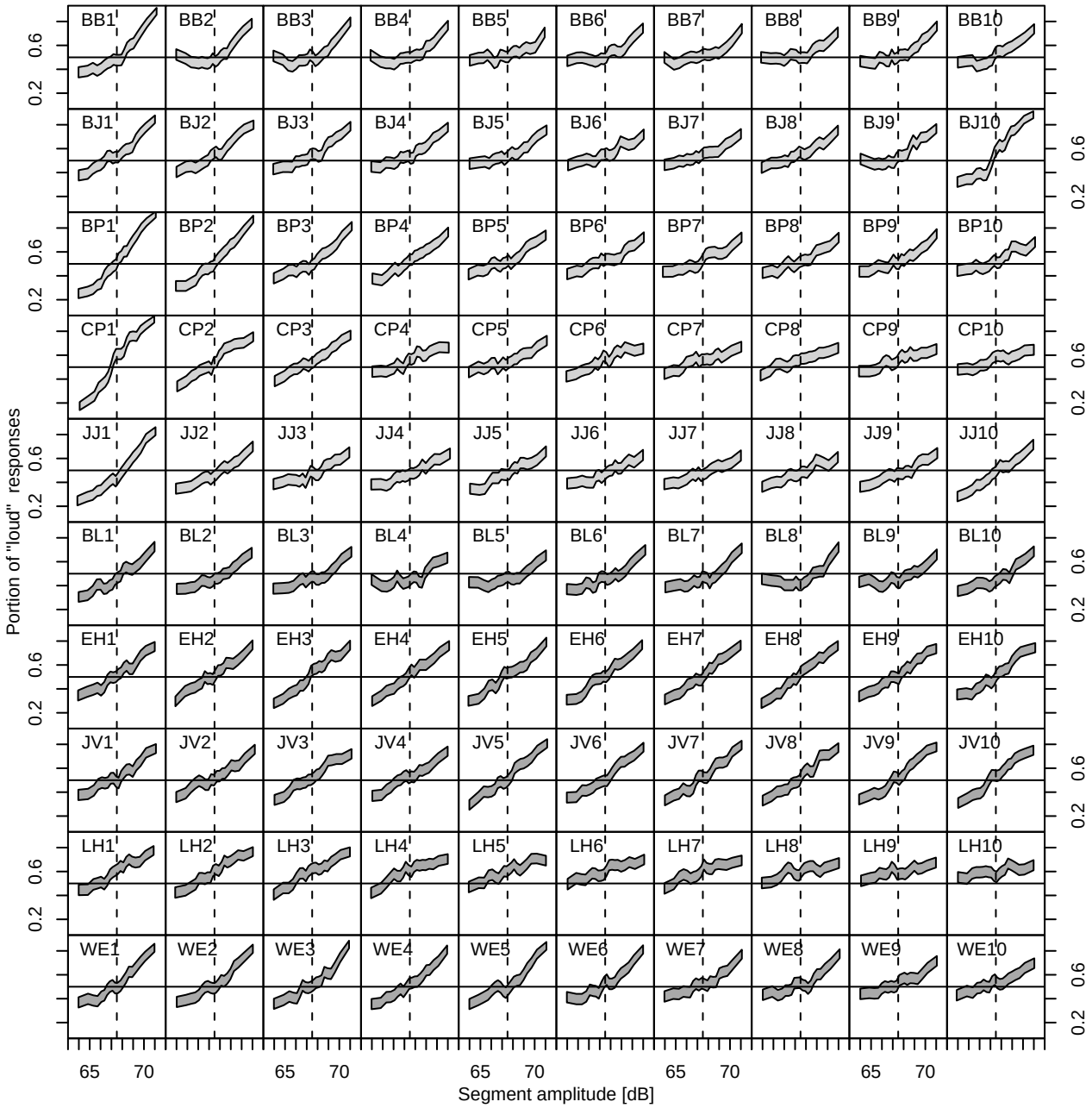


Figure 3.6: Probability of "loud" as a function of level of single segments for all listeners, $Response(x_i)_{\overline{20,500}}$. Regions of 95%-condidence are indicated. In the upper half (light-grey graphs) are listeners in the no-feedback condition and in the lower half (dark-grey) the listeners in the feedback condition. Listener ID and segment number indicated in the upper left corner of each figure. The overall mean of all segment levels (67.5 dB) is indicated by a vertical broken line.

- Figure 3.6 revealed that for low segment levels, listeners responded "loud" more often (to different degrees) than expected from a linear model.

The fact that non-linear effects were identified to a larger extent for individual segments than for the overall mean, suggests that the explanations lie in how loud a given segment is in relation to the levels of other segments of a given sound. If for example all segment levels are low (giving a low mean value), listeners do not seem to press "loud" more often than would be expected from a linear model, but it *is* the case when only single segments are considered. The COSS functions for the nine last segments of listener BB (Figure 3.6) show that when a given segment's level is below the overall mean (67.5 dB) he gives equally many "loud" and "soft" responses. This suggests that when a segment's level is sufficiently low, that segment is simply ignored by BB. Similar effects can be observed for other listeners, but to a smaller extent.

Explaining derived COSS functions

There are a number of reasons, which can be considered responsible for the non-linearity of the COSS functions:

(1) Temporal masking - If a high level of a segment persists in the sensory system over following softer levels, it could be argued that this causes the listener to press "loud" more often as expected. This is however hard to reconcile with two observed facts: (1) Not all listeners responded "loud" too often for low segment levels (EH and JV for example) and consequently there is no masking in their sensory system, and (2) the weighting curves derived by Pedersen and Ellermeier (2006) showed that adjacent segments can be weighted very differently, which could not be the case if there was a "smoothing" effect. Also, if the masking is mainly thought to be due to forward masking, only little (backward) masking would be expected for the first segment. This is indeed the case to some extent, but the first segment of for example WE does not appear to be less "masked" than following segments.

(2) Loudness percentile - The loudness exceeded a certain percentile of the time determines the listeners' response, and consequently the segment levels below this critical level has no influence on the judgment.

(3) Attention - Louder segments (in relation to other segments of the same sound) catch more attention, and are thus weighted more heavily, which causes the listener to respond "loud" more often, as the softer segments do no catch the same attention.

The first of these point is examined in greater depth for the loudness model by Glasberg and Moore (2002) in Section 3.4. The two latter points are examined in Section 3.5, where various paradigms are applied in an attempt to predict listeners' loudness judgments.

3.4 Applying a temporal loudness model

In relation to the discussion of temporal masking given in the previous section there may be two types of models to be considered: (1) Models predicting the loudness of time-varying sounds (Glasberg and Moore, 2002; Zwicker, 1977), and (2) models of temporal masking (as for example Oxenham and Moore, 1994). The two types of models will be addressed in the two following sections.

3.4.1 Evaluating models of time-varying loudness

There are only few models, which are able to predict loudness of time-varying sounds. The earliest appears to be described by Zwicker (1977). In this work Zwicker mainly gives a description of the main principles of the model, and critical parameters (as for example the time coefficient for temporal integration) are only given by reference to other work (suggestions for time coefficients are for example given by Vogel (1975)). More recently, a computer program was described by Widmann *et al.* (1998), but the program is not a full loudness model, and only supposed to account for post-masking. The lack of a unified description of Zwicker's model makes it hard to use it in practice. Therefore the focus will be on the model given by Glasberg and Moore (2002), which also contains many of the elements introduced by Zwicker.

Glasberg and Moore's model of time-varying loudness

Glasberg and Moore (2002) suggest how a model should integrate loudness over time to generate a curve of "continuous" loudness. Their suggestion for loudness integration is fundamentally an iterative process, which they formulate as (Equation 1 and 3 of their paper):

$$S'_n = \begin{cases} \alpha_a S_n + (1-\alpha_a) S'_{n-1} & \text{if } S_n > S'_{n-1} \\ \alpha_r S_n + (1-\alpha_r) S'_{n-1} & \text{if } S_n \leq S'_{n-1} \end{cases} \tag{3.1}$$

This equation works on discretized "instantaneous" loudness (S_n), which in their formulation has been sampled at a rate of 1 kHz (T_i = 1 ms). Subindices in the equation refer to index in the discrete sequence. S' is the integrated "short term" loudness sequence, and its definition is of a recursive nature as it appears from the equation where S' is present on both sides of the equality. α_a and α_r of the equation indicate the integration rate, when the "short term" loudness curve is either rising ($S_n > S'_{n-1}$) or falling ($S_n \leq S'_{n-1}$), and they are related to critical time coefficients of integration (T_a and T_r): $\alpha_a = 1 - e^{-T_i/T_a}$ and $\alpha_r = 1 - e^{-T_i/T_r}$. Their suggested values are: α_a = 0.045 and α_r = 0.02, which gives: T_a = 22 ms and T_r = 49 ms.

Further, Glasberg and Moore introduce the concept of "long term" loudness, which in its definition is identical to Equation 3.1, except that "instantaneous" loudness, S, is substituted with "short term" loudness S', "short term" loudness is substituted with "long term" loudness, S'', and different integration rates are used (α_{al} = 0.01 and $\alpha_{rl} = 0.0005$). This corresponds to the time coefficients: T_{al} = 99 ms and T_{rl} = 2 s.

Glasberg and Moore suggest that "short term" loudness may account for aspects such as temporal masking or persistence of neural activity in the auditory sensory system, while they use "long term" loudness to account for the results of experiments in which the loudness of modulated sounds was judged by listeners. It is not clear which type of loudness, "short term" or "long term", would be more relevant in describing the performance of the listeners in the task described by Pedersen and Ellermeier (2006), therefore both "loudness" types will be applied to the stimuli of the experiment in the following.

Conversion to sones

Before temporal integration is possible an "instant loudness" curve is needed. In the experiment described by Pedersen and Ellermeier (2006) the sound pressure levels were within the range from 60 dB to 75 dB SPL. In this relatively narrow range the relationship between loudness in sone and sound pressure level in dB is close to linear. Consequently, Equation 3.1 might by directly applied to the stimuli of the experiment, where "instantaneous" loudness is given by the level of the ten individual segments of a sound. But to stay as close as possible to Glasberg and Moore's formulation of the model as possible, all segment levels were transformed from values in dB to sone. The transformation was

made via the standard ANSI S3.4 (2005), which describes a procedure for the calculation of loudness of stationary sounds given their amplitude spectra. A computer program for the calculation of loudness according to the standard is freely available [1] and was used for the transform. The amplitude spectrum of a gives segment was supplied to the program by its spectrum level and bandwidth. The bandwidth was always the full bandwidth supported by the program (20 Hz to 18 kHz) and in this range the spectrum level was assumed flat. The actual spectrum level was estimated by: SPL$-10log(w)$, where "SPL" is the calibrated sound pressure level of a given sound segment as used in the experiment and w is the bandwidth of the noise (22 kHz) (Kinsler, 2000). The program for loudness calculations was used in "diffuse field" mode, as the headphones were "diffuse field" compensated as described by the manufacturer.

Loudness integration of segment patterns

The result of loudness integration of the stimuli used by Pedersen and Ellermeier (2006) may be best understood by graphical inspection of the generated loudness patterns, which are depicted in Figure 3.7 ("short term") and Figure 3.8 ("long term"). The curves were generated by sampling the "instantaneous" loudness at a rate of 1 kHz and applying Equation 3.1.

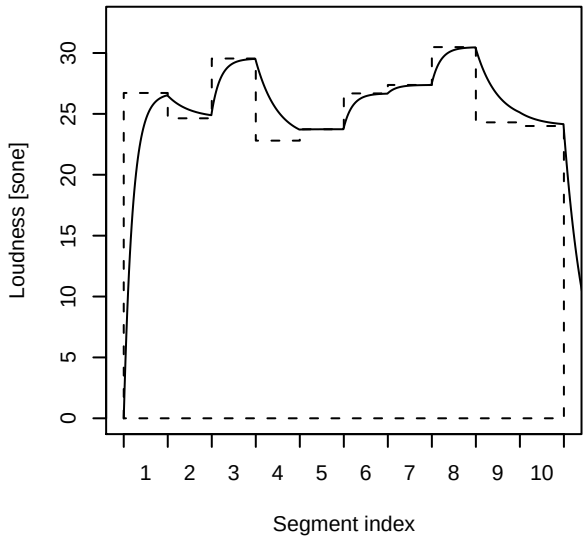


Figure 3.7: Example of "short term" integrated loudness. "Instantaneous" loudness indicated by the broken lines.

The graph in Figure 3.9 is identical to that of Figure 3.8, but Figure 3.9 shows the graph over a wider time range. It is clear from the figures that "short term" loudness follows the "instantaneous" segment pattern quite closely, while in the case of "long term" loudness the curve continues for a long period after the end of the sound. "Long term" loudness is able to follow the "instantaneous" pattern to some degree when loudness is increasing, but not when it is decreasing.

Predicting loudness judgments

It still remains a question how listeners should derive their judgments of overall loudness based on the continuous "short" or "long term" loudness curves. To be able to compare the loudness model to the

[1] PC program for the calculation of loudness according to ANSI S3.4-2005 can be downloaded from `http://hearing.psychol.cam.ac.uk/Demos/ansiloud.zip`

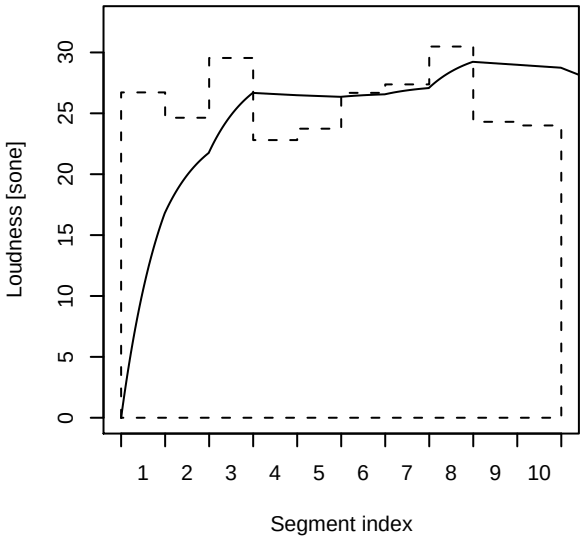


Figure 3.8: Example of "long term" integrated loudness. "Instantaneous" loudness indicated by the broken lines. See Figure 3.9 for a view over a wider time range.

real behavior of listeners, a simple rule is proposed, which integrates the loudness curves within a fixed time window starting at the onset of the sound. Typically listeners gave their response 0.1 s to 0.4 s after the sound of a given trial had ended (see Figure 3.17 for example). This may seem as a paradox for the "long term" loudness curve as it continues for a long time outside this range and would actually interfere with the sound of the following trial.

Three different time windows for the integration were tested (1 s, 1.5 s and 11 s) and for both "short term" and "long term" loudness. To examine what temporal weighting pattern such integration and decision rules would predict, the model was applied to the dataset of an actual listener (BB) of the experiment. The model was then used to predict loudness judgments, where "loud" responses were generated when the integrated "short" or "long term" loudness exceeded the overall mean of all loudnesses predicted by the model, and "soft" responses were generated in the other cases. The predictions of the model were used to derive "weighting curves" for the model, which can be compared to the weighting curves of the actual listeners as presented by Pedersen and Ellermeier (2006). The model "weighting curves" are shown in Figure 3.10 for different integration windows and for "short" and "long term" loudness. In the figure it is seen that "short term" loudness predicts "flat" weighting curves, independent of the overall integration time, which is not surprising since the loudness curve is very close to zero outside the integration window in all cases. The "long term" weighting curves varies as a function of the length of the integration window, and for a short window predicts an emphasis on the first segments of the sound, and for a long window, on the last segments of the sound (to a smaller degree). It should be noted that the "long term" model would never be able to put very different weight on adjacent segments, because their corresponding levels in the "long term" loudness curve cannot be very different because of the long time coefficients used for integration (see Figure 3.8). However, very different weighting of adjacent segments was observed for the real listeners (Pedersen and Ellermeier, 2006). Further, the "long term" model may suggest that the weighting curve is a function of response time, as listeners with a long response time would be able to "integrate" the loudness curve for a longer period. If this was indeed the case, listeners with a long response time should emphasize the last segments of a sound and listeners with a short response time the first segments. Response times are found in Figure 3.17, and relatively slow listeners are for example CP, JJ, and BJ while LH, BP, and EH are relatively fast. BJ did indeed weight the last segment more heavily, but CP showed pronounced weighting of the first segment. BP did indeed weight the first segments more heavily, but EH did

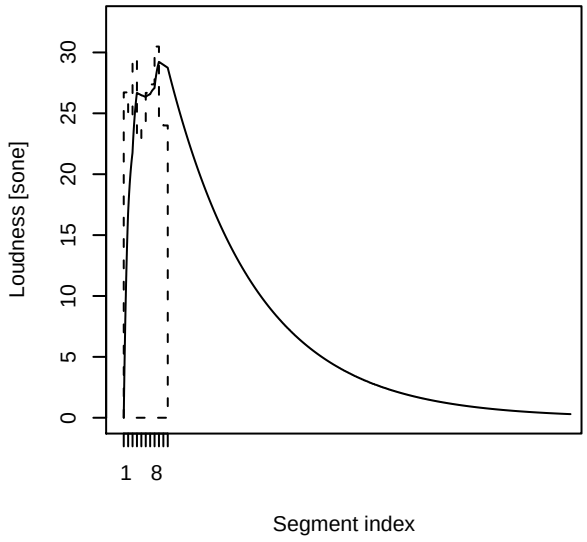
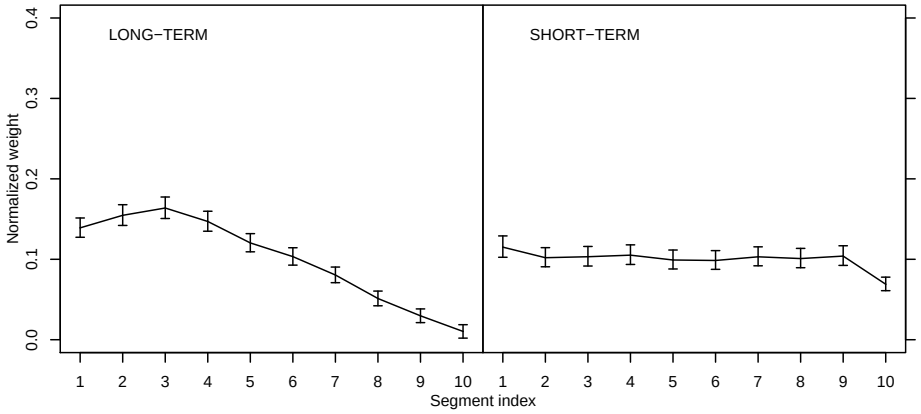


Figure 3.9: Example of "long term" integrated loudness. "Instantaneous" loudness indicated by the broken lines. See Figure 3.8 for a view over a narrower time range.

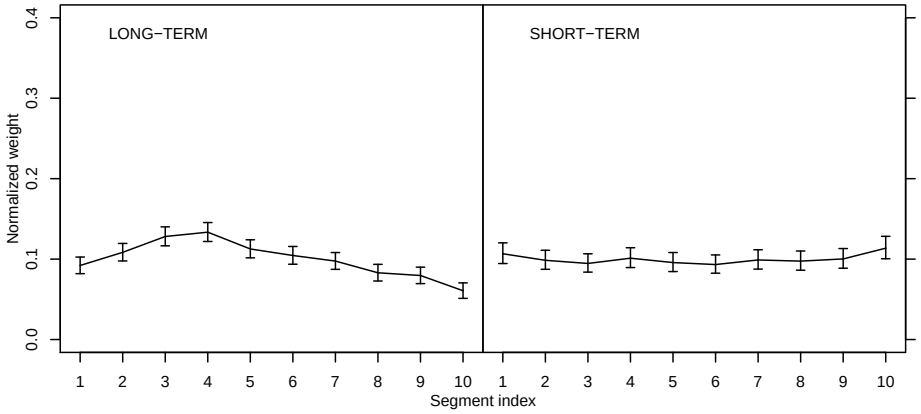
not show pronounced weighting of any segments. JJ weighted both the first and final segments more heavily. In conclusion, even though the loudness model is able to predict different shapes of the weighting curve, this seems to be unrelated to the behavior of the real listeners.

Because loudness predictions were made for the data set for an actual listener (BB), the predictions of the model may be compared to his actual performance. However, it is hardly feasible to do it for all listeners because of the rather long computation time for the temporal integration. The listener's and the model's performance were compared using the same method as is introduced in Section 3.5, if the predictions of loudness are substituted for L in Equation 3.3. It was found that the best predictions was made by the "short term" model, which was able to predict 69.5% of listener BB's judgments. For a comparison, a model based on a simple mean value of the ten segment sound pressure levels is able to predict 68.1% of his responses, and if a second-order polynomial non-linear relation between segment level and loudness is introduced the model is able to predict 74.8% of his responses. This demonstrates that the loudness model hardly gives better predictions of loudness judgments than a simple mean value of the sound pressure level. The polynomial model performs significantly better, which suggests that it is the slightly non-linear relationship between Sone and sound pressure which caused the loudness model to perform slightly better than the simple mean.

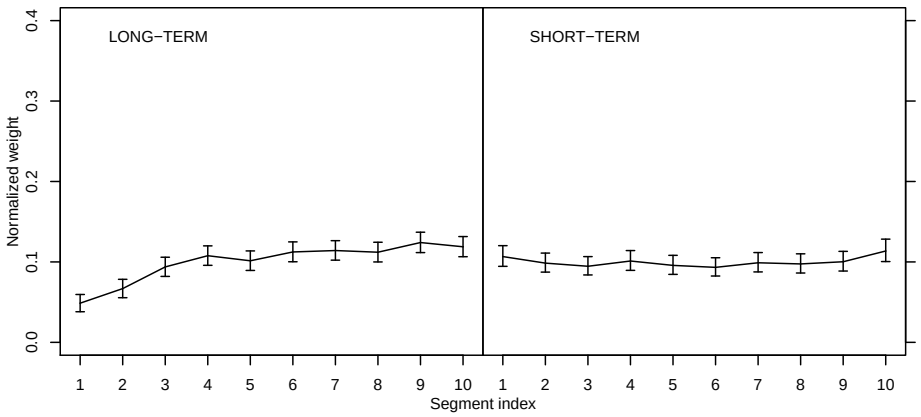
It may be argued that listeners do not integrate the continuous "long term" or "short term" curves, but rather, "sample" the instantaneous value of the curve at the time when they make their judgment. However, this alternative decision rule has some implications for the weighting curves, which are clearly at odds with the actual behavior of the listeners: If the "short term" loudness is "sampled" in loudness judgments, only the last segment has any significant influence on the judgment since the "short term" curve will be very close to the level of the last segment at the moment of the judgment. This is seen in Figure 3.11, where weighting curves are depicted in the case where loudness judgments are based on "sampling" of the loudness curve. For "long term" loudness the first segments receive relatively little weight, which was never the case for actual listeners. For the weighting curves in the figure, the loudness curve was sampled 0.5 s after the end of the sound, but varying the delay between the end of the sound and the moment of loudness "sampling" has relatively little influence on the shapes of the weighting curves. It thus appears, that "integration" of the loudness curve is better than "sampling" the loudness curve when explaining the behavior of actual listeners.



(a) 1 s total integration time



(b) 1.5 s total integration time



(c) 11 s total integration time

Figure 3.10: Derived weights if listeners based judgments on overall loudness as described by Glasberg and Moore's model using long- (first column) and short-term (second column) time coefficients respectively. Different time windows for the overall integration are used in different rows.

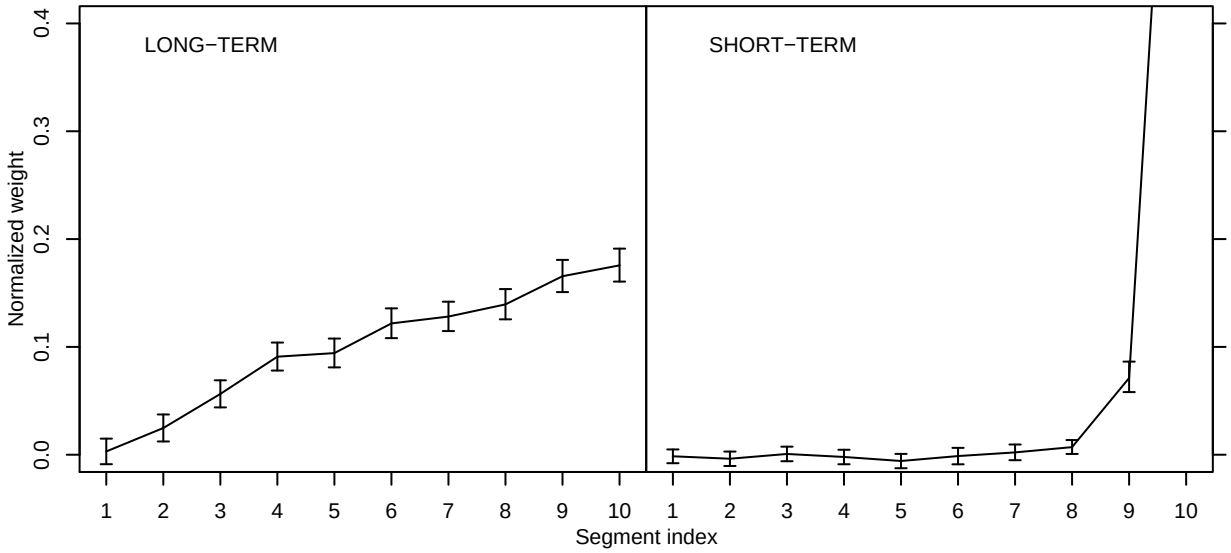


Figure 3.11: Derived weights if listeners "sampled" the continuous loudness curve as calculated from Glasberg and Moore's model for "long term" (left) or "short term" (right) loudness.

Summary

The application of the model for temporal integration of loudness as described by Glasberg and Moore (2002), was applied to stimuli used in the experiment by Pedersen and Ellermeier (2006). The loudness model was used to predict loudness judgments, but did not give significantly better predictions of an actual listener's behavior than could be obtained from a simple mean value of the sound pressure levels of the stimuli. The "long term" loudness model was able to predict different shapes of weighting curves, depending on the time window used for integration. However, the predicted weighting patterns could not be related to the weighting patterns of real listeners. In their weighting curves, real listeners were able to "single out" specific segments receiving greater weight, which is not possible for a "long term" loudness model. Further, some listeners weighted both the first and last segments of a sound more heavily, which is in conflict with predicted weighting patterns of the model.

All in all it appears that the parts of the loudness model, which describe the temporal integration, do not help in understanding the behavior of listeners. For the "static" part of the model, it seemed that the non-linear transformation from sound pressure to sone may have captured some aspects of the listeners' loudness perception.

3.4.2 Temporal window

A temporal window is often thought to play an important role in temporal processing, and is often modeled by a sliding window, which smoothes the envelope of a perceived sound (see for example Oxenham and Moore, 1994). Thus it may be hypothesized that the temporal window has an important influence on how listeners will judge the loudness of fluctuating sounds. In this chapter a sliding window will be analyzed and applied to the stimuli of the experiment, taking the following form:

$$w(t) = \begin{cases} He^{t/\tau_b} & \text{t} < 0 \\ He^{-t/\tau_f} & \text{t} \geq 0 \end{cases}$$

Here w is the temporal window as a function of time. It is constructed via two exponential functions with different time coefficients, τ_b and τ_f, for forward and backward masking respectively.

The envelope of one segment out of the ten constituting one sound in the experiment can be described as:

$$v_i(t) = \begin{cases} x_i & 0 \leq \mathrm{t} < \mathrm{T} \\ 0 & \text{otherwise} \end{cases}$$

Here x_i is the level of the relevant segment, and v_i is an expression for the envelope of the segment as a function of time. T is the duration of a single segment. One entire sound can be described by time shifting and adding ten such segments.

It needs to be stated how the output of the temporal window is integrated to arrive at a global loudness judgment. A simple model, which integrates the total output of the sliding window, is considered here. First the case of integrating a single segment is considered. The output of the sliding window at a given time, t, is the convolution ($*$) of the temporal window ($w(t)$) and its input ($v_i(t)$):

$$w * v_i(t) = \int_{-\infty}^{\infty} w(\lambda) v_i(t-\lambda) d\lambda = \begin{cases} x_i H \tau_b (e^{\frac{t}{\tau_b}} - e^{\frac{t-T}{\tau_b}}) & \mathrm{t} < 0 \\ x_i H (\tau_b (1 - e^{\frac{t-T}{\tau_b}}) + \tau_f (1 - e^{-\frac{t}{\tau_f}})) & 0 \leq \mathrm{t} < \mathrm{T} \\ x_i H \tau_f (e^{-\frac{t-T}{\tau_f}} - e^{-\frac{t}{\tau_f}}) & \mathrm{t} \geq \mathrm{T} \end{cases}$$

Integrating this expression over all time gives the total output of the sliding window over a single segment:

$$\int_{-\infty}^{\infty} w * v_i(t) dt = x_i H T (\tau_f + \tau_b)$$

Because of the distributive property of the convolution operator this is easily generalized to the case of ten segments, using time shifted segments as described above:

$$\int_{-\infty}^{\infty} w * v(t) dt = \sum_{i=1}^{10} x_i H T (\tau_f + \tau_b)$$

Here $v(t)$ is the sum of ten time shifted segments with ten different envelope amplitudes x_i. It is seen that the output would be identical to a model with linear temporal weights where all weights are equal ($HT(\tau_f + \tau_b)$). So if a loudness judgment is based on the integrated output of a sliding window model, all estimated weights would be identical, when a model with linear weights is fitted to the responses of the listeners (as Pedersen and Ellermeier (2006) for example did).

3.5 Alternative models of loudness judgments

3.5.1 Fitting procedure

The basic procedure for fitting models used in this chapter is similar to the procedure outlined by Pedersen and Ellermeier (2006), where the following equation was fitted to the listeners' judgments:

$$\Psi(D) = p(\text{“}\texttt{loud}\text{”}) = \frac{e^D}{1+e^D} = \frac{1}{1+e^{-D}} \tag{3.2}$$

Here Ψ expresses the probability, p, of a "loud" response as a function of the strength of the decision variable D. Earlier (Pedersen and Ellermeier, 2006) D was given by linear temporal weighting of the 10 levels of a sound. To investigate different decision strategies, different forms of D will be explored, in its basic form expressed as:

$$D(\mathbf{x}, c) = L(\mathbf{x}) - c \tag{3.3}$$

As in the earlier work, c is a fixed decision threshold. L is a function of the ten segment levels, $\mathbf{x}$, expressing the modeled perceived overall loudness. Different forms of D are introduced by varying the "loudness" function L. If for example L is equal to a weighted sum of the segment levels, the model is identical to the one presented by Pedersen and Ellermeier (2006).

Fitting algorithm

Pedersen and Ellermeier (2006) fitted Ψ to the listeners' responses using maximum likelihood optimization. This procedure requires the model to be of the *generalized linear model* form. As this is not the case for all models of this chapter, a different algorithm is used, which optimizes the model according to a *non-linear least-squares* fit. The main difference between the procedures is given by their names, where one finds the coefficients for a model, which most likely would have yielded the results, and the other minimizes the deviation of the predictions from the listeners' responses in term of their squared distance. For a deeper explanation, see for example the books by Venables and Ripley (2002) or Insightful (2005) for a description of `glm` and `nls`, which are the functions used for *maximum likelihood* and *non-linear least-squares* fitting respectively. In the cases where both methods have been applied in this work for the same model, they give almost identical coefficients.

Modeling scheme

The strategy will be to select a relatively broad range of expressions for the perceived overall loudness L, and compare their ability to predict the behavior of the listeners. The models can be selected both on statistical and behavioral grounds, meaning that a model can be of interest for two reasons: (1) To explore how a statistical property correlates with the listeners' responses (for example, does the variance of the segment levels within one sound affect the listeners' behavior), and (2) a hypothesis of the listeners' decision strategy can be set up and its agreement with their actual behavior examined.

In the following, 15 different models of L are described and tested. Some are variations of the same basic ideas, and others are fundamentally different. Some models have more parameters to be fitted than others, giving them greater freedom to fit the listeners' performance. If there are enough parameters, the model may be flexible enough to predict the behavior of the listener without incorporating "true" aspects (processes actually taking place in the auditory system) of the actual processes involved in loudness judgments. Therefore, first models of the decision processes are given using relatively few parameters, and then extended giving better fits, but with the danger that the increased power of the model is only due to the generally increased flexibility of the model. To be able to examine possible group effects and individual effects all models were fitted to each listener's individual data, pooled data within the feedback and no feedback groups, and the pooled data of all listeners.

The mathematical formulations of L for all models are given in appendix A where it is also described how the models interrelate and which models are generalizations of others. A brief summary of the basic ideas of the models is also given here, and their basic functioning summarized in Table 3.2:

Simple statistic

Some models use a simple statistic of the ten segment levels of a sound to predict the listener's response. Different models use the mean, maximum, minimum, and variance of the segment levels in making predictions. This will help to answer the following questions: Is the minimum as important as the maximum? If the listeners' responses are based on a simple linear combination of the segment levels this should be the case. Also, if responses are based on a simple linear combination, the variance of

Model	Short model description	Nr parameters
M1	Mean	2
M2	Maximum	2
M3	Minimum	2
M4	Variance	2
M5	Envelope profile	3
M6	Polynomial non-linearity	3
M7	Attention	3
M8	Moment of inertia	4
M9	Temporal weighting	11
M10	Polynomial non-linearity and temporal weighting	12
M11	Temporal weighting before attention	12
M12	Attention before temporal weighting	12
M13	Attention, power as a function of segment index	12
M14	Interaction and temporal weighting	20
M15	Polynomial non-linearity as a function of segment index and temporal weighting	21

Table 3.2: Overview of models for predicting loudness (L in Equation 3.3).

the segment levels should not be correlated with the responses of the listeners, because, in the signal generation, segment levels are uncorrelated.

Temporal weighting

Some models weight a given segment level with respect to its temporal index. In its simplest version this is similar to the temporal weighting described by Pedersen and Ellermeier (2006), but is extended to work in combination with other model schemes described in this chapter.

Non-linearity as a function of level

It is well known that the loudness of stationary sounds is not linearly related to their sound pressure level in decibels. However, within the segment levels used in this experiment, the relationship is close to linear. Anyhow, the possibility of a non-linear relationship is introduced by relating predicted loudness to the segment levels via a second order polynomial fit. This allows for non-linearity to a large extent, which can be examined by inspection of the coefficients of the polynomial fit.

Envelope profile

It has been shown that it is important for judged loudness whether the sound is increasing or decreasing in levels as a function of time (Stecker and Hafter, 2000; Oberfeld and Plank, 2005). This reasoning is introduced in a model by estimating whether the segment levels are increasing or decreasing for a given sound. This is done by a linear regression fit to the segment levels as a function of segment index. The slope of the fitted line and the intercept is then used to predict the responses of the listeners.

Moment of inertia

The physical concept of moment of inertia can be considered a measure of the centeredness of the distribution of mass in space. The concept has been adopted in the modeling, and in this context it is a measure of how the segment levels are distributed temporally. A high value of the moment of inertia means that high-level segments lie (temporally) relatively far from the “center of mass”, whereas a low

value means that the loud segments are close to the "center of mass". The "center of mass", in this context, is a measure of the temporal position of the center of the sound with respect to the segment levels weighted by their distance to the center, where "distance" is measured in temporal units (segment index). Intuitively the "center of mass" may be considered the "balance" point of the sound.

Attention

As it was shown earlier in this chapter, some listeners seemed to ignore a segment in his/her judgment if it was below a certain value. This may be understood in terms of attention. If the level of a sound segment is low compared to the other segments, the attention of the listener is only on the louder segments. Thus the soft segment should receive a relatively small weight, while the louder segments should be more heavily weighted. To model this, the concept of an "attention weight" is introduced:

$$w_{attention_i} = \frac{x_i^p}{\sum_{j=1}^{10} x_j^p} \tag{3.4}$$

Here $w_{attention_i}$ is the attention weight given to the i'th segment. x_i is the segment level of the i'th segment, and p is a fitted exponent, which determines the degree of attention weighting. If p approaches infinity, the maximum segment is weighted by a factor of 1 and the other segments with a factor of zero, and if p approaches minus infinity, the minimum segment is weighted by a factor of 1 and the other segments with a factor of zero. If $p = 1$, then the sum of the attention weighted segments equals their mean value.

Combining modeling paradigms

The basic ideas for the modeling are introduced one by one in the models M1 to M9. Models M10 to M12 combine temporal weighting with the simple polynomial and attention models. Both M11 and M12 combine the attention model and temporal weighting, the difference is in the order of which this is done, the reason being: If attention weighting and temporal weighting are both involved in the decision process, is then the attention weighting applied to the segment levels after they have been temporally weighted, or is the attention weighting carried out first and then temporally weighted? M11 and M12 examine both these cases. Model M13 and M15 allow the coefficient used for attention weighting and the polynomial non-linearity respectively, to be a function of segment index.

Interaction

If energetic temporal masking plays an important role it is expected that adjacent segments interfere to a large extent. Therefore one model (M14) includes interaction terms for adjacent segments.

3.5.2 Assessing model fits

To compare how successful the various models are in their prediction of listener performance, both numerical and graphical measures are used. A simple and intuitive way to estimate how good a model is, is to use the model to predict when the listeners respond "loud" and "soft". The predictions can then be compared to the listeners' real performance and the proportion of correct responses calculated. However, a model predicts the chance of a given outcome as a continuous variable, and this has to be compared to the binary ("loud"/"soft") responses of the listener. A comparison can be made if the model prediction is "loud", when the chance of a "loud" response is above 0.5, and respond "soft" when it is below 0.5. In this way a score can be obtained for each model, but care should be taking when comparing the scores, because of uncertainty in their estimation. The portion of correct responses can be considered the estimated probability, p, that the model makes a correct prediction. The estimate of p

thus lies within certain limits of confidence, which can be calculated based on the binomial distribution and depends on the total number of trials used to estimate p. The "resolution" of the scores can thus be estimated by calculating 95%-confidence intervals for example. The results of such calculations are summarized in Figure 3.12. The ranges of the confidence intervals are shown for different values of estimated p. The confidence intervals are shown when 4598, 22990, and 45980 trials are used to estimate p. Theses numbers were the available trials for each listener, each group (feedback and no-feedback), and the total number of trials respectively.

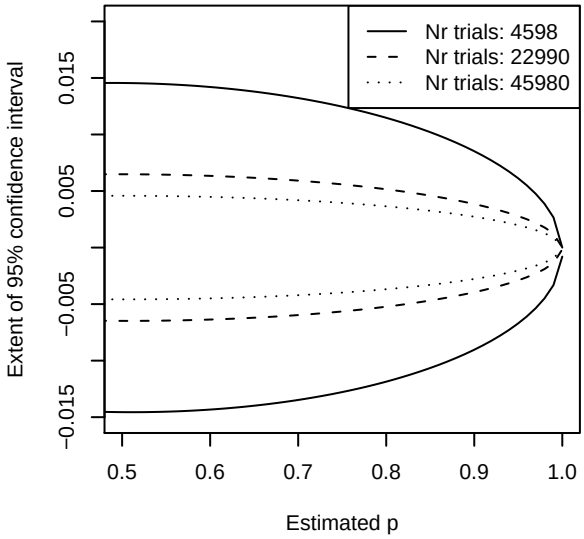


Figure 3.12: Range of 95%-confidence intervals for estimated p in a series of Bernouilli trials. The dotted curve reflects the distance of the upper and lower limits of the confidence interval from the estimated probability, when 4598 trials are used to estimate p. The dashed and dotted curve illustrate the limits when 22990 or 45980 trials are used.

In tables where performance of different models are summarized (Table 3.3 for example), "explained variance" is also indicated in a separate column. The values of this column were calculated from the residual sum of squares (SS_{model}). That is, the sum of the squared difference between each estimated probability of a given outcome and the actual outcome for all experimental trials. An unintelligent model would always predict the chance of a given outcome to be 0.5 ("loud" or "soft"). The sum of the squared errors for such a model (SS_{tot}) would be: $nr\_trials \cdot 0.5^2$, because the error on a given trial would always be -0.5 or 0.5. nr_trials is the total number of trials in the experiment. A measure of a given model's ability to explain the variation in the listeners' judgments can be calculated as: $(SS_{tot} - SS_{model})/SS_{tot}$, which gives a value in the range from 0 (model not better than chance) to 1 (all listener judgments predicted correctly). Though the explained variance is indicated in the tables, only the performance scores in term of the percentage of correct predictions will be used in the text because of its more intuitive interpretation.

Modeling pooled data

There are different goals in examining the fitted models, therefore fitting was carried out in three different ways: (1) The regression coefficients were estimated by fitting the models to the pooled data of all listeners. This will help to identify which model(s) best capture(s) general effects across listeners. The drawback is that individual effects and group effects (caused by feedback) cannot be observed. Further, the results may be dominated by listeners who gave the most consistent judgments. (2) Therefore the models were also fitted for each listener. This can potentially reveal whether different models capture

different aspects of the listeners' performance. (3) Finally, it is also of interest to compare models fitted to pooled data for the feedback and no-feedback group respectively. This may help to reveal in which way people performed differently in the two groups. Specifically it was observed by Pedersen and Ellermeier (2006) that temporal weighting is different in the two groups, and more pronounced in the no-feedback group. It might be that factors other than temporal weighting are more important in accounting for this difference.

Models fitted to pooled data of all listeners

The overall scores (fitted to the pooled data of all listeners) of each model are summarized in Table 3.3. To confidently say that a given model is better than another one, performance (estimated p) must differ by more than 1% (see Figure 3.12).

Model	Performance	Explained variance	Nr parameters
M1	73.2	0.285	2
M2	69.4	0.198	2
M3	60.2	0.065	2
M4	56.8	0.028	2
M5	73.7	0.295	3
M6	74.5	0.311	3
M7	74.6	0.312	3
M8	73.8	0.300	4
M9	73.9	0.301	11
M10	75.2	0.328	12
M11	75.4	0.330	12
M12	75.4	0.330	12
M13	75.3	0.329	12
M14	75.0	0.326	20
M15	75.2	0.328	21

Table 3.3: Overview of the performance of fitted models in predicting the pooled judgments of all listeners.

As expected, the performance of a model depends on the number of parameters, which allow for different degrees of flexibility. In understanding the decision process of the listeners, it is not only the models giving the best fits that provide important information. So, to start with the relatively simple models, it is observed that a model based on a simple mean value (M1) of the segment levels performs remarkably well. A model based on the peak segment level (M2), performs significantly worse. The peak segment also reflects the level that is reached 10% of the time (or less), and is therefore related to the decision rule suggested in the literature, which states that the loudness exceeded a certain percentile of the time determines the judged loudness (Zwicker and Fastl, 1999; Grimm *et al.*, 2002). So the available data do not support a "percentile" rule. Not surprisingly, a model based on the minimum segment level (M3) is worse than the model based on the maximum level. A linear model would not predict this, so it can be concluded that high-level segments play an important role compared to low-level segments. Since the segment levels were drawn independently from normal distributions, a linear model would predict that the statistical variance of the ten segment levels is not related to loudness. This is not entirely the case, which can be seen from model M4, which makes predictions of loudness based on the variance of the ten segment levels. It thus seems that interactions between segment levels are involved in the listeners' judgments. But at the same time it may be argued that the performance of the variance model is so inferior that it shows that listeners are able to "ignore" the variance (degree

of fluctuation) in their judgments to a large degree. Models that take the shape of the envelope into account (M5, M8) do not seem to perform especially well, given the number of parameters. It thus seems that it is of little importance whether the sound follows and ascending or descending pattern (M5). Further, it seems that it is not important whether the segment levels are "peaked" or "spread out" (M8). Finally it should be noted that the variance of the segment levels in the experiment was rather small, and thus may not be well suited to test "envelope effects".

The models discussed up to now (M1 to M8) have relatively few parameters, and have tested some of the basic ideas. Models M9 to M10 have relatively many parameters and are to a large extent combinations of some of the basic models. So generally they give better fits. It is seen that models M11 and M12 perform best, but other models in the range M10 to M15 are almost as good and the small differences in performance could have occurred by chance. However, it may be noted that M14 and M15 do not seem to benefit from their relatively many parameters. It thus looks like the polynomial "bend" of the probability of "loud" as a function of segment level, does not depend on segment index to a large extent (which M15 allows). Further, M14 suggest that interaction between adjacent segments does not give as accurate a description of the interaction effects as the model which more globally accounts for interaction (attention models). The temporal weights model (M9) actually performs worse that some of the simpler models (M6 and M7). This may seem in contrast with the results by Pedersen and Ellermeier (2006), who showed that temporal weighting strongly depends on segment index, and thus it may be expected that temporal weighting is crucial for a good model. The reason for the seemingly contradictory results may indicate that temporal weighting is performed in individual ways and cannot be generalized, and further, some listeners had close to "flat" weighting curves. Generally, it seems to be a trend that models which either include attention weighting or a polynomial dependency on segment level do best (attention: M7, M11, M12, M13; polynomial: M6, M10, M15). Both these groups of models predict a non-linear dependency between segment level and the probability of "loud". A non-linear dependency was also demonstrated earlier by non-linear COSS functions. It thus appears that the non-linearity is a more pronounced effect than temporal weighting, a least when considering effects across listeners.

From the performance estimates it is almost impossible to discriminate models with many parameters (M10 to M15), and the difference in performance is insignificant for these models. Table 3.4 shows how the predictions of the different models correlate. The correlation between two models was calculated as the correlation between their predictions of "probability of loud" for the data set that was used to derive the model coefficients. From the table it is clear why the performance of some of the models cannot be discriminated: Models M10, M11, M12, M13 and M15 give virtually identical predictions (the correlation coefficient is 1). This is also the case for model M6 (polynomial non-linearity) and M7 (attention weighting), so the validity of the different underlying concepts is not easily compared. It may also be noted that the predictions of all linear models (M1 and M9) are uncorrelated with the predictions of the variance model (M4), but the predictions of the variance model are not completely uncorrelated with the responses of the listeners.

Models fitted to pooled data within feedback and no-feedback group

As it was demonstrated by Pedersen and Ellermeier (2006), feedback influenced the performance of the listeners. Therefore it seems relevant to compare the models when they are fitted within each group (feedback or no-feedback). The performance of the different models, when fitted within group, is summarized in Table 3.5.

Considering the performance of the models within the no-feedback group, very much the same as when case of pooling the results of all listeners seems to hold. Comparing the feedback and no-feedback group, model scores are generally slightly lower in the no-feedback group, except for model M4 (variance). This indicates that the variance of the segment levels plays a more important role in the no-feedback condition. It is also observed that the benefit of temporal weighting is greater in

RE	1.00															
M1	0.53	1.00														
M2	0.44	0.64	1.00													
M3	0.25	0.62	0.22	1.00												
M4	0.15	0.00	0.56	-0.58	1.00											
M5	0.54	0.98	0.62	0.61	0.00	1.00										
M6	0.55	0.96	0.76	0.44	0.26	0.94	1.00									
M7	0.56	0.95	0.79	0.45	0.28	0.94	1.00	1.00								
M8	0.54	0.97	0.62	0.60	0.00	0.99	0.93	0.93	1.00							
M9	0.55	0.97	0.62	0.60	0.00	0.99	0.93	0.93	1.00	1.00						
M10	0.57	0.93	0.74	0.43	0.26	0.95	0.97	0.97	0.96	0.96	1.00					
M11	0.57	0.93	0.77	0.44	0.27	0.94	0.97	0.97	0.95	0.95	1.00	1.00				
M12	0.57	0.93	0.77	0.44	0.27	0.94	0.97	0.97	0.95	0.95	1.00	1.00	1.00			
M13	0.57	0.93	0.76	0.44	0.26	0.95	0.97	0.97	0.95	0.95	0.99	1.00	1.00	1.00		
M14	0.57	0.93	0.72	0.45	0.21	0.95	0.96	0.96	0.96	0.96	0.99	0.98	0.98	0.98	1.00	
M15	0.57	0.93	0.74	0.43	0.26	0.95	0.97	0.97	0.96	0.96	1.00	1.00	1.00	1.00	0.98	1.00
	RE	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15

Table 3.4: Correlation between the predictions of all pairs of models for the pooled data of all listeners. The first column "RE" is the correlation with the actual responses of the listeners.

the no-feedback group (compare for example the difference between M1 and M9 within each group), which is reasonable since temporal weighting was much more pronounced for listeners in the no-feedback group, though in very individual ways. When combining temporal weights and non-linearity as a function of segment level (M12 for example), the predictions of the models are almost equally good in the feedback and in the no-feedback conditions. This indicates that listeners in the feedback condition are not "better" than listeners in the no-feedback condition, but feedback may have changed their behavior, which causes a relatively better score for the simplest models.

The performance of the models may also be graphically inspected, and a procedure to do this as a function of segment index and level is proposed: The main idea is to analyze how many "loud" responses a given model predicts for a given segment within a certain range of the levels of that segment. Using the same data set as was used in the experiment, this can be compared to the actual performance of the listeners. Conceptually this can be understood as dividing the COSS functions of the listeners' performance (as for example depicted in Figure 3.6) by similar COSS functions constructed from the predictions of a given model. That is, within a certain interval (bin) of segment levels the percentage of "loud" responses given by the actual listeners is divided by the percentage of "loud" responses predicted by the model. The results of such analyses are shown in Figure 3.14 and 3.13 for the prediction of the performance of listeners in the feedback and no-feedback conditions respectively (similar plots for the pooled results of all listeners and for individual listeners are found in Appendix A.2).

When first considering Figure 3.13 (no-feedback), it is observed that models M2 and M3 (maximum and minimum) have strongly non-monotonic curves. This suggests that neither a "minimum" nor a "maximum" decision rule applies. Model M4 (variance) is the only model that overestimates the probability of "loud" for low segment levels. This may not be surprising since a low segment level causes a high value for the calculated variance (as is also the case for high segment levels), and this illustrates that listeners are not judging variance alone. Temporal weighting (M9) is able to reduce the variation of the curves as a function of segment index (different colors) at high segment levels (compare for example M1 and M9 in Figure 3.13). In general (except M4), models seem to underestimate the probability of "loud" at low segment levels and overestimate the probability of "loud" at high segment levels. Polynomial models or models applying attention weighting are able to reduce this to a certain extent for low segment levels (compare M6 and M7 to M1 for example). However, no model is able to eliminate this trend completely, but it could well be that it is mainly due to effects, which are impossible for the models to account for, such as: A constant error rate of the listeners, which is independent of

Model	Performance (NF)	Explained variance (NF)	Performance (FB)	Explained variance (FB)	Nr parameters
M1	71.6	0.258	74.9	0.314	2
M2	68.8	0.193	69.9	0.203	2
M3	59.0	0.053	61.3	0.079	2
M4	57.9	0.036	56.0	0.022	2
M5	72.5	0.278	75.0	0.318	3
M6	73.3	0.291	75.7	0.334	3
M7	73.4	0.293	75.8	0.335	3
M8	73.8	0.301	75.1	0.318	4
M9	73.8	0.304	75.0	0.319	11
M10	75.4	0.338	75.9	0.338	12
M11	75.5	0.341	76.0	0.340	12
M12	75.6	0.341	76.0	0.340	12
M13	75.5	0.340	75.9	0.339	12
M14	75.5	0.337	75.8	0.337	20
M15	75.5	0.340	75.9	0.339	21

Table 3.5: Overview of the performance of the models listed in Table 3.2 in predicting the pooled judgments in the feedback (left) vs. no-feedback group (right).

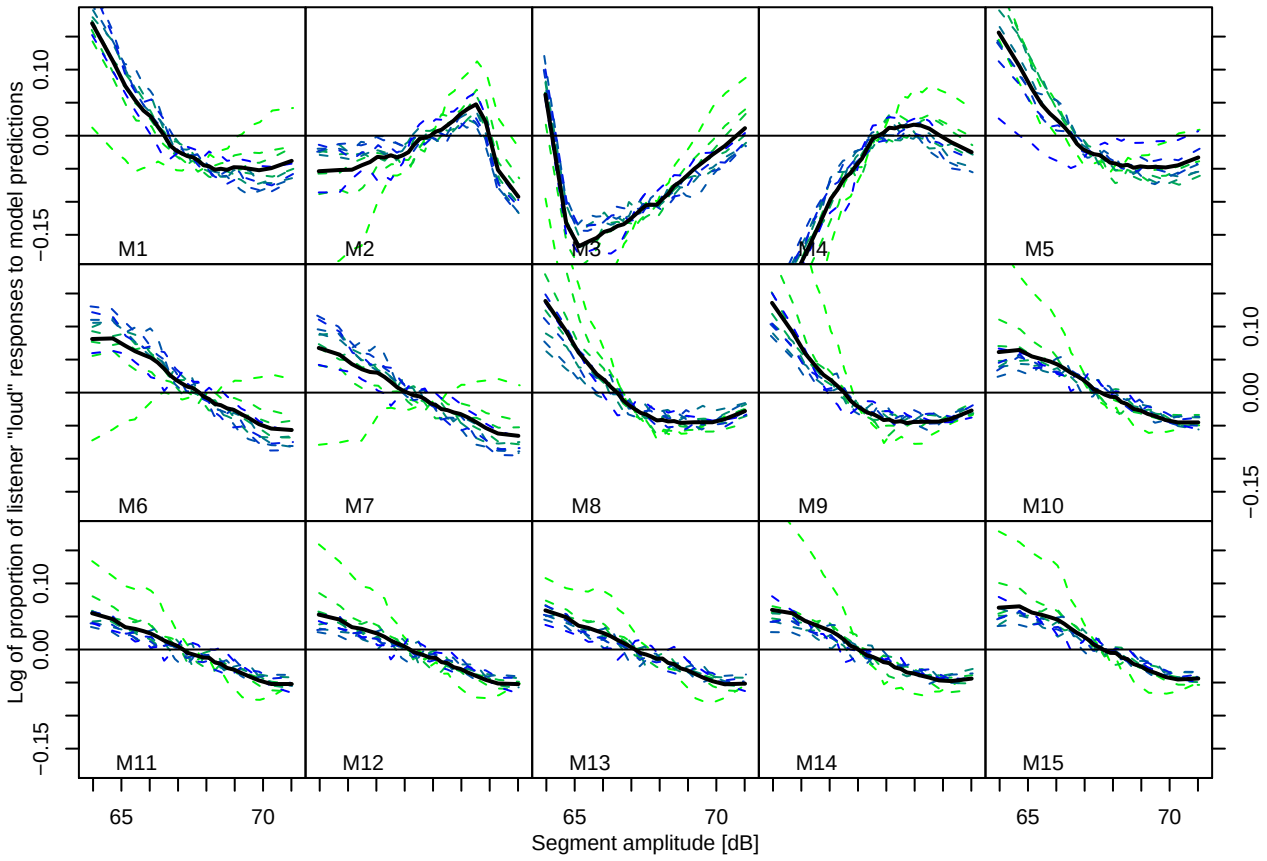


Figure 3.13: Percentage of listener “loud” responses divided by the percentage of model “loud” predictions as a function of segment level. Plotted for all ten segments as broken lines (segment 1 (green) to segment 10 (blue)) and the solid line is the mean over then ten segment curves. Model ID is indicated in the lower left corner of each panel.

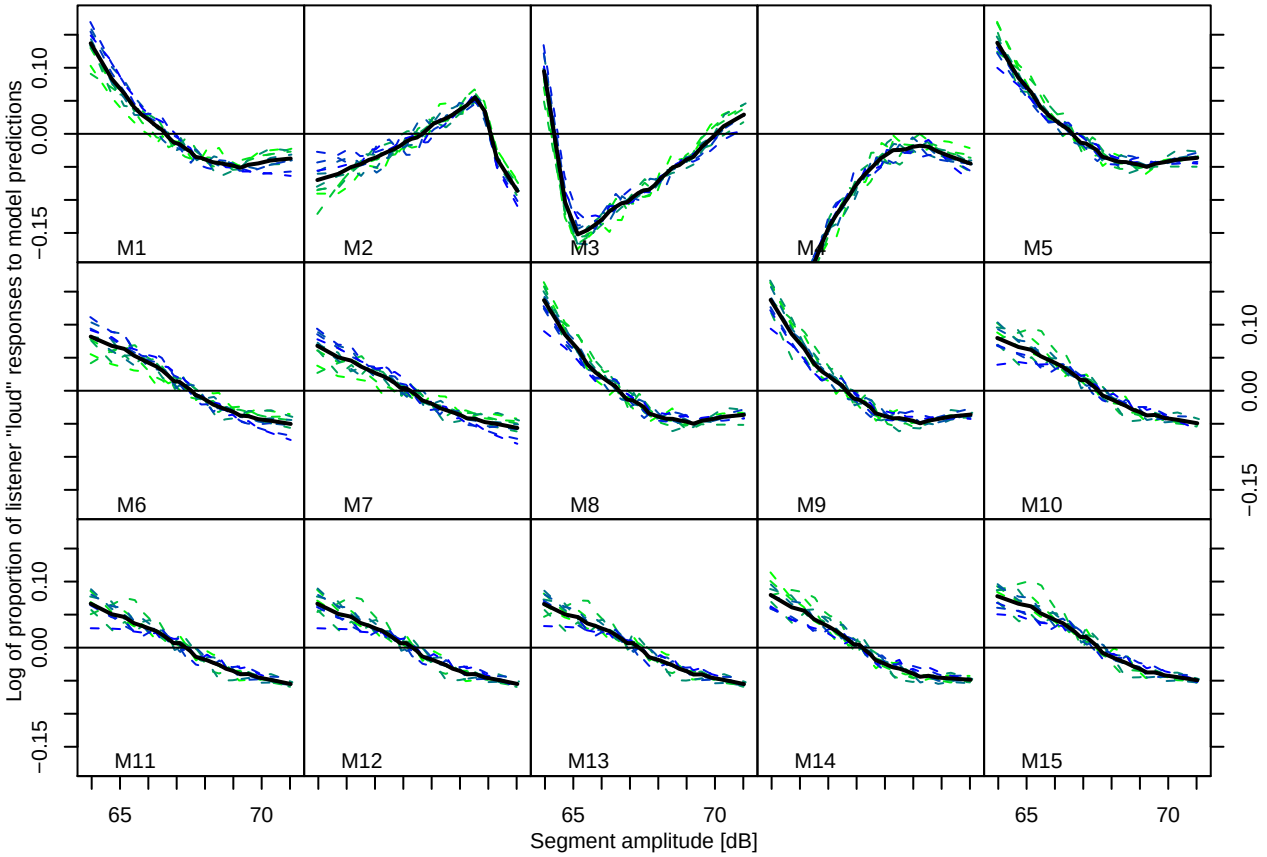


Figure 3.14: Same as Figure 3.13, but for listeners in the feedback group.

segment levels or across-trial effects (which was identified in earlier sections). Comparing Figure 3.13 and 3.14, the effect of feedback can be inspected. The main difference seems to be a greater variability as a function of segment index (curves of different colors in the figures) in the no-feedback group.

Models fitted for individual listeners

Models may capture different aspects of the decision strategy, which in term may vary across the listeners. Therefore all models were also fitted to the individual data of the listeners, which yielded performance scores for each model for each listener. Because of the many score-values to compare they are summarized in a figure (Figure 3.15) rather than a table.

It is observed that, for the well-performing models, the variation in performance is much greater across listeners than across different models. Not surprisingly, temporal weighting (M9) makes the largest difference for listeners, for whom pronounced differences in temporal weighting were shown by Pedersen and Ellermeier (2006). It may also be interesting to observe that for some listeners (BB, BL, and BJ) model M4 (variance) is better that M3 (minimum). This may support the notion that soft segment are ignored and the attention payed to other segments.

3.5.3 The “best” model

It is of course relevant to identify which model predicts the listeners’ performance best, but it may be of even greater interest to identify whether the concepts described by a model translate into similar processes occurring in the sensory system. The answer to the first question is, that in general M11 and M12, which both include attention and temporal weighting, give the best predictions (see Table 3.3).

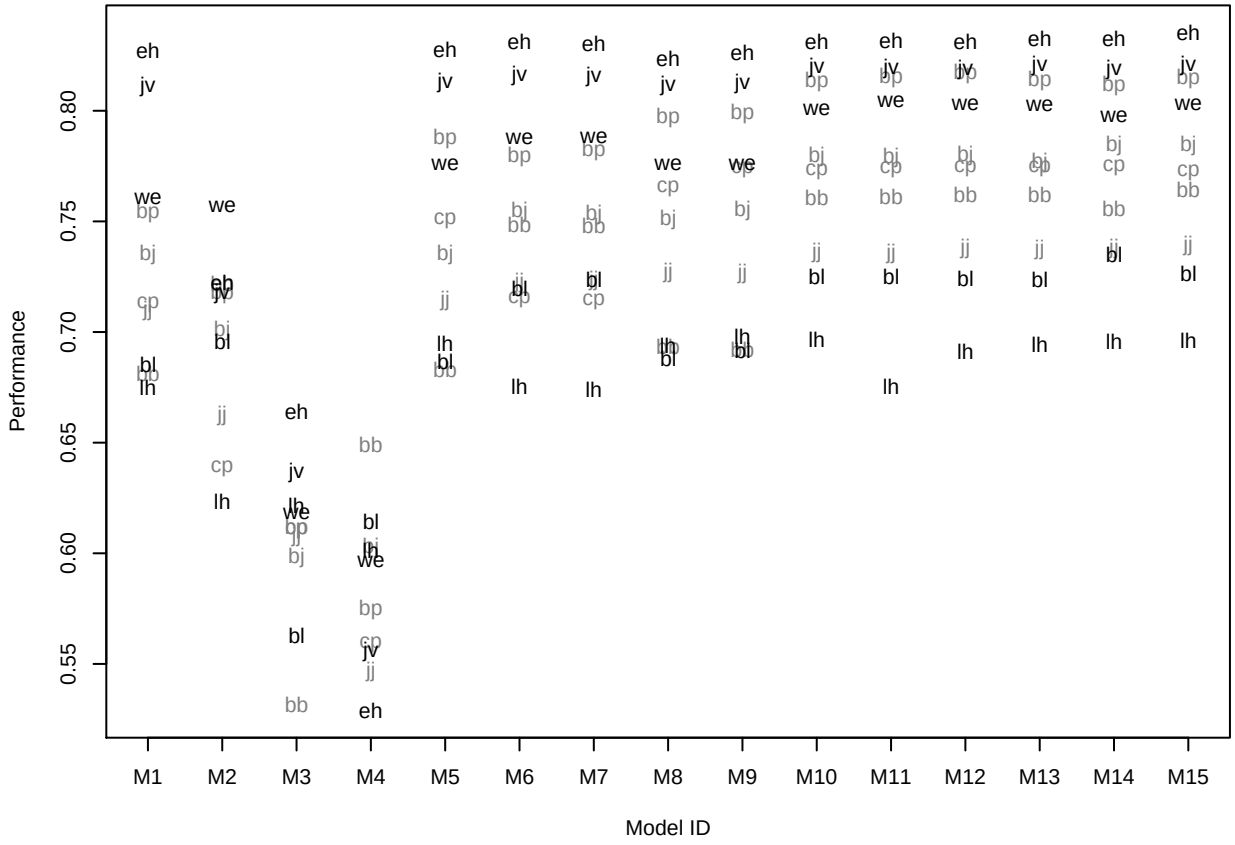


Figure 3.15: Performance of each model for each listener. The models are identified on the x-axis, and on the y-axis it is shown how well the respective models predict the responses of each listener identified by their ID.

But all models from M10 to M15 generally perform well, and their difference in performance is not significant. The second question is addressed in the following.

Non-linear level dependence or attention?

It seems that models which include either a polynomial non-linearity or attention weighting are performing best, but based on the performance scores they cannot be easily discriminated. One approach to try to discriminate them anyhow, is to consider their ecological validity. First, a non-linear relationship between sound pressure level and perceived loudness has often been found, which speaks in favor of the "polynomial" interpretation. And for example Lutfi and Jesteadt (2005) argue that non-linear COSS functions may reflect this. COSS functions for the present data were shown in Figure 3.6, where the non-linear relationship can be observed for a number of listeners. Earlier it was mentioned that some listeners seemed to ignore a given segment if its level was sufficiently low. This type of effect would only make sense for attention weighting. Further, there was weak evidence that if the segment level was sufficiently low, decreasing it even further would *increase* the probability of a "loud" response for some listeners. This is not compatible with the polynomial interpretation, but can be predicted by attention weighting: If attention is moved from a soft segment to a louder one, the soft segment has less influence on the judgment, and the lower the segment level of the soft segment is, the more attention is directed to the louder segment and the greater the probability of a "loud" response becomes (because now only the loud segments determine the response). Such behavior can actually be predicted by the attention weighting as it was mathematically formulated.

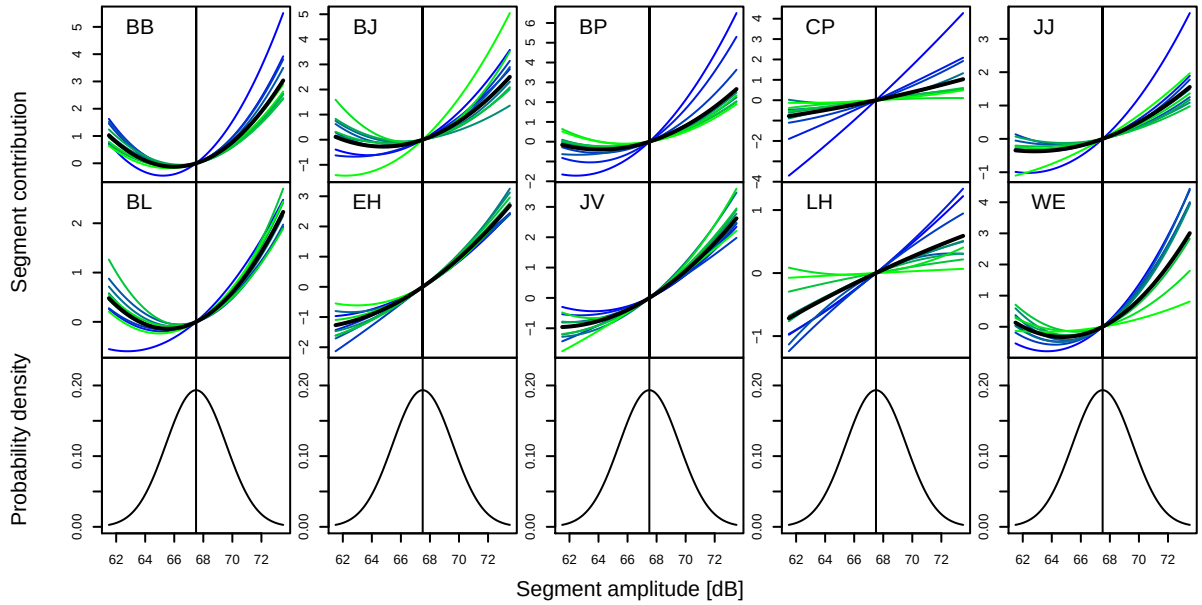


Figure 3.16: The contribution to overall loudness by different segments (segment 1 (green) to segment 10 (blue)) as a function of segment level. The contribution is calculated from the segment-dependent polynomial regression coefficients of model M15. In the bottom row the distribution of the segment levels in the experiment is shown.

Further, the actual non-linearity of the polynomial models may be inspected. To that end, in Figure 3.16, it is depicted how much each segment contributes to the predicted overall loudness as a function of its level. The curves in the figure are based on model M15 fitted to individual data. In model M15 different polynomial regression coefficients were allowed for different segments, which are indicated by different colors for different segments in the figure. The regression coefficients for a given segment were used to calculate how much this segment contributes to the overall loudness as a function of its level. The obtained parabolas were offset as to have a contribution of zero at the mean segment level. The panels in the bottom row of the figure indicate the distribution of segment levels used in the experiment and outside this range the model regression may not be valid. It is observed that the "bend" of the curves is quite different for different listeners. For two listeners (CP and LH) there is almost no curvature, but in general U-shaped patterns are observed, where the slope is greatest at high segment levels. It is not possible for the parabolas to stay at a "flat" level, but it seems to be a general trend that for low segment levels they have close to zero contribution. For some listeners the parabolas start increasing again for lower segment levels within the range of levels used in the experiment. This argues against an interpretation of the non-linearities as a simple non-linearity between level and perceived loudness. As attention weighting is able to predict such behavior, attention weighting seems to be a more plausible explanation of the listeners' non-linear behavior.

3.6 Response time and loudness

It has been hypothesized, and to a large extent confirmed, that loudness and reaction time are intimately related, which is for example demonstrated by Wagner *et al.* (2004); Pfingst *et al.* (1975); Kohfeld *et al.* (1981). The use of reaction time as a measure of loudness is for example relevant in experiments where listeners cannot directly respond to their perception of loudness as is for example the case for infants or animals. The response times of the listeners were recorded in all listening experiments presented in this thesis, and it seems worthwhile to more carefully investigate these data in the experiment where

the listeners had to make loudness judgments. It should be noted, however, that listeners were not asked to respond as quickly as possible, so the measured response times are not directly comparable to reaction time as measured by Wagner *et al.* (2004). The analysis may help to address relevant questions concerning the relationship between loudness and response time:

(1) Is response time related to the absolute level of loudness or only indirectly, depending on the listener's task? For example: If comparing the loudness of two sounds, is the response time then related to the difference in loudness between the two sounds or their absolute levels?

(2) If judging global loudness of multiple components, is the response time then a function of the loudness of individual components in a comparable way to the contribution of individual components to the global judgment of loudness?

Answering these questions would help to understand if response time is a valid measure of perceived loudness and to which extent the same sensory processes underlies the two.

3.6.1 Data collection

Data from Experiment 1 of the paper by Pedersen and Ellermeier (2006) were used in the analyses of response time in this chapter. In the experiment, response time was measured from the time where the sound of the trial ended until the listener pressed a button ("loud" or "soft"). Listeners were not instructed to respond at any particular pace. Trials with a response time longer than 1 s were discarded from the analyses, because longer response times possibly occurred when for example the listener took a short break.

3.6.2 Response time results

Response time as a function of overall mean level

The response times of different listeners vary to a large extent, so response times are only analyzed individually. In Figure 3.17, the response time is plotted as a function of the mean level of the ten sound segments of trial. The response time has been smoothed as $Response\_time(x_{mean})_{\overline{100,800}}$ (see section 3.3.1). The trials were analyzed for "soft" (dark gray in the figure) and "loud" (light gray) responses individually, giving monotonically rising and falling curves respectively.

In Figure 3.17 it is seen that, if a listener makes a "soft" judgment, the response time gets longer the higher the mean value of the segment levels is. The opposite is the case if the listener makes a "loud" judgment. This clearly shows that the response time does not depend on the absolute level of a sound, but rather its distance to a level not far from the overall mean segment level. In general it seems that the response time increases/decreases linearly as a function of the mean segment level. Further, it seems that this linear relationship depends on whether a "soft" or "loud" judgment was made. This can for example be observed for listener JV in the Figure 3.17, who generally gives faster "soft" responses, and the slope of the reaction time as a function of mean segment level depends on whether a "loud" or "soft" judgment was made. This may be explained by a bias for one of the categories ("loud" or "soft"), if for example the listener always had his/her finger ready on the "soft"-button before the response. However, this does not seem to provide a full explanation as for example both listener EH's fastest and slowest response times occur for "loud" judgments. If one category was favored over another, the response time for this category would generally be lower. Further, one would expect this type of bias to be more dominant for short response times (where moving the finger takes relatively long compared to the response time), but in fact the opposite is often observed: The largest difference in response time (between "loud" and "soft" responses) occurs for relatively long response times (see listener JV for example).

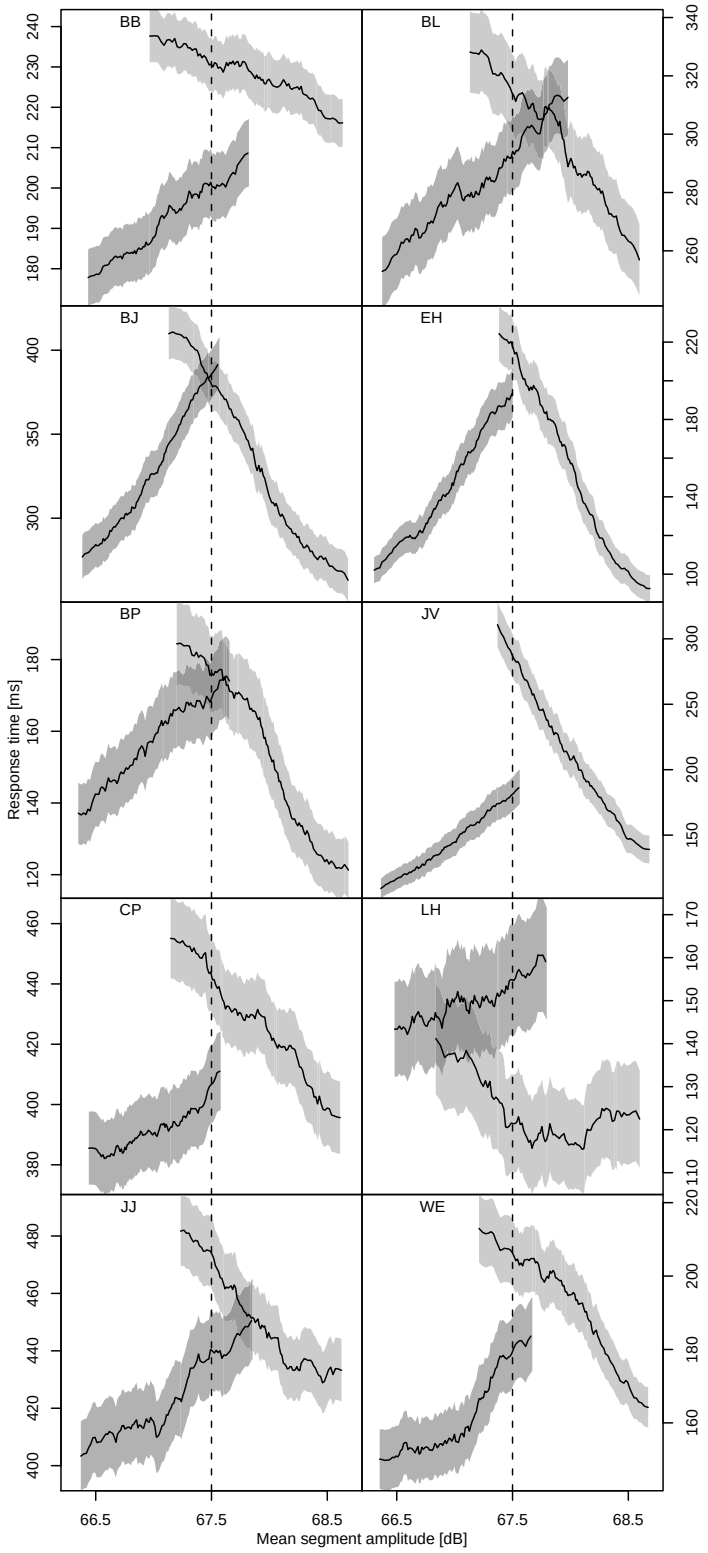


Figure 3.17: Response time (y-axis) as a function of the mean segment level of the ten segments of a sound (x-axis) for individual listeners. The dark gray graph is for trials where the listener responded “soft” and the lighter gray for the trials where he/she responded “loud”. The shaded areas illustrate the regions of 95%-confidence. Listeners in the right column received feedback and listeners in the left column did not receive feedback.

Response time as a function of segment levels

To determine how the levels of individual segments contribute to response time, the response time is plotted as a function of the level of each of the ten segments of a sound in Figure 3.18, smoothed as $Response\_time(x_i)_{\overline{20,800}}$, where x_i is the level of the i'th segment.

The main conclusion of inspecting the figure is: Generally there seems to be a linear relationship between response time and segment level. Further, the linear relationship varies over different segments. The linear relationship means that a multiple regression analysis can be made to determine how much each segment contributes to the response time, much in the same way as weighting curves were derived by Pedersen and Ellermeier (2006).

Weighting curves for response time

The previous section suggested that weights can be derived, which describe the influence of the level of each segment on the response time. This was done using a least-square multiple regression analysis, and was performed for "loud" and "soft" judgments independently. The regression coefficients except for the intercept (the weighting curve) are depicted in Figure 3.19, where the weighting curves for "loud" judgments are shown as dashed curves and for "soft" judgments as a solid curves.

The negative regression coefficient observed especially for "loud" judgments means that the higher the segment level, the shorter response time, and vice versa for "soft" judgments. Hence, one might expect the solid and dashed lines to be mirror-symmetric relative to a horizontal line. This indeed appears to be the case for many listeners, but especially listener BJ and WE seem the deviate from this behavior for the last segments. For "loud" judgments, they have positive weights for the last segments, which means that high levels of these segment cause a longer response time, which is in conflict with a simple interpretation of the response time reflecting loudness.

If response time was directly reflecting perceived loudness, the weighting curves would be expected to be similar to those derived by Pedersen and Ellermeier (2006). This is for example the case for listener CP, who in his loudness judgments heavily emphasized the onset. This is also the case for his weighting curves for response time. But the weighting curves for listener JV and EH, who had close to "flat" weights for loudness judgments, are not "flat", but get close to zero weight at the final segments.

3.6.3 Discussion

The analysis of listeners' response time showed that loudness and response time are related, in that response time clearly dependent on segment level(s). However the dependency is not straightforward:

- Response time was not a linear function of overall level. Rather, the response time decreased when the overall segment level increased and the listener judged the sound to be "loud". But if the listener judged it to be "soft", the response time increased when the overall segment level increased.

- A straightforward correspondence between weighting curves for loudness judgment and for response time could not be observed for all listeners.

- When judging "loud", some listeners (BB and WE) had regression coefficients of opposite sign for the onset- and offset-segments. The consequence is, if the segment level of an onset-segment decreases, response time also decreases. But if the offset-segment level decreases, response time increases. If response time directly reflected loudness, the consequence would be that increasing the level (of the offset-segments) causes the perceived loudness to decrease.

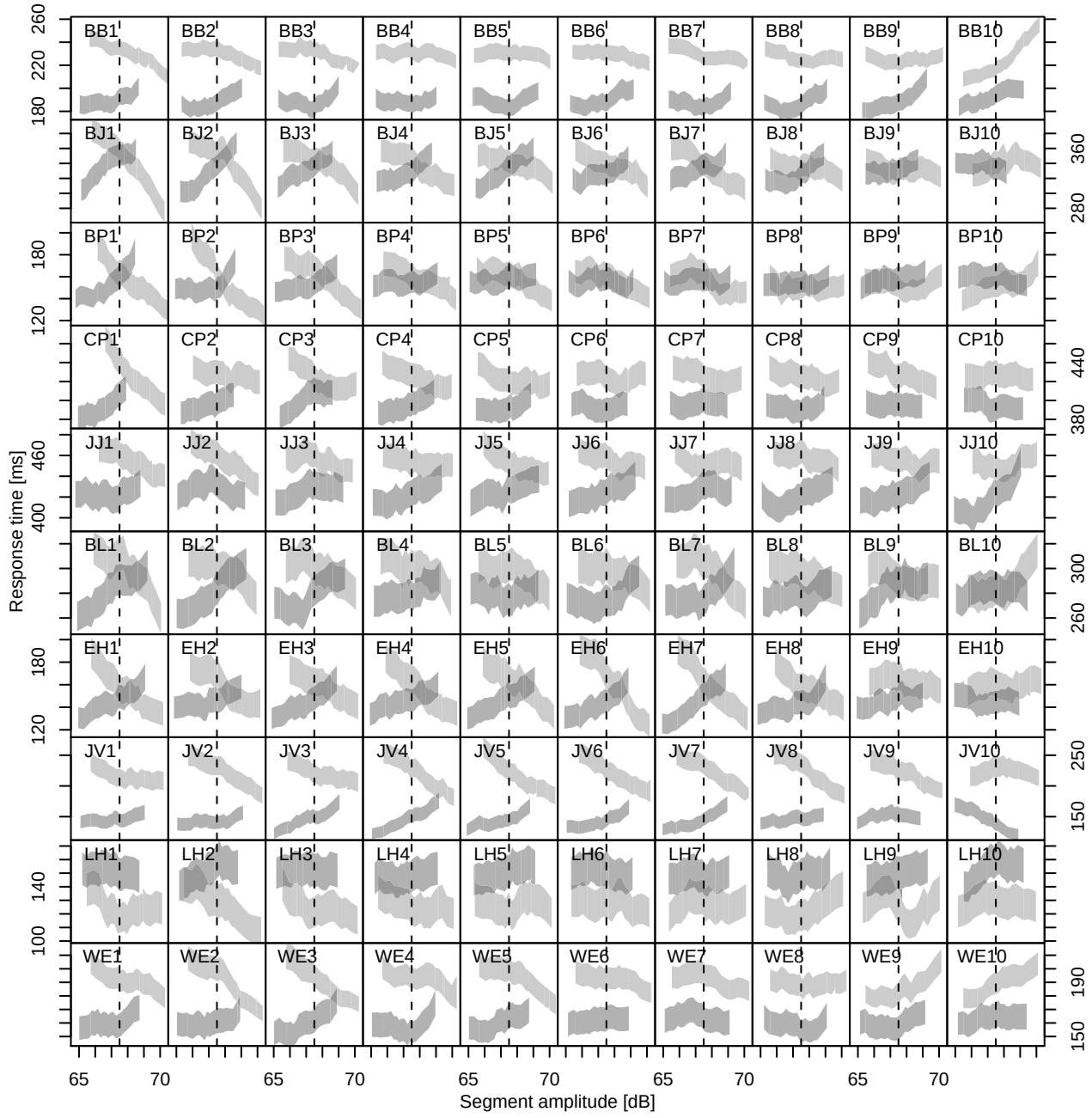


Figure 3.18: Response time (y-axis) as a function of segment level (x-axis) for individual listeners. The dark gray areas are for trials where the listener responded “soft” and the lighter gray for the trials where he/she responded “loud”. Analyses were made for all segments individually and are shown in different columns (1 to 10, where the column index corresponds to the segment index). The shaded areas illustrate the regions of 95%-confidence. Listeners in the five bottom rows received feedback and listeners in the top rows did not receive feedback.

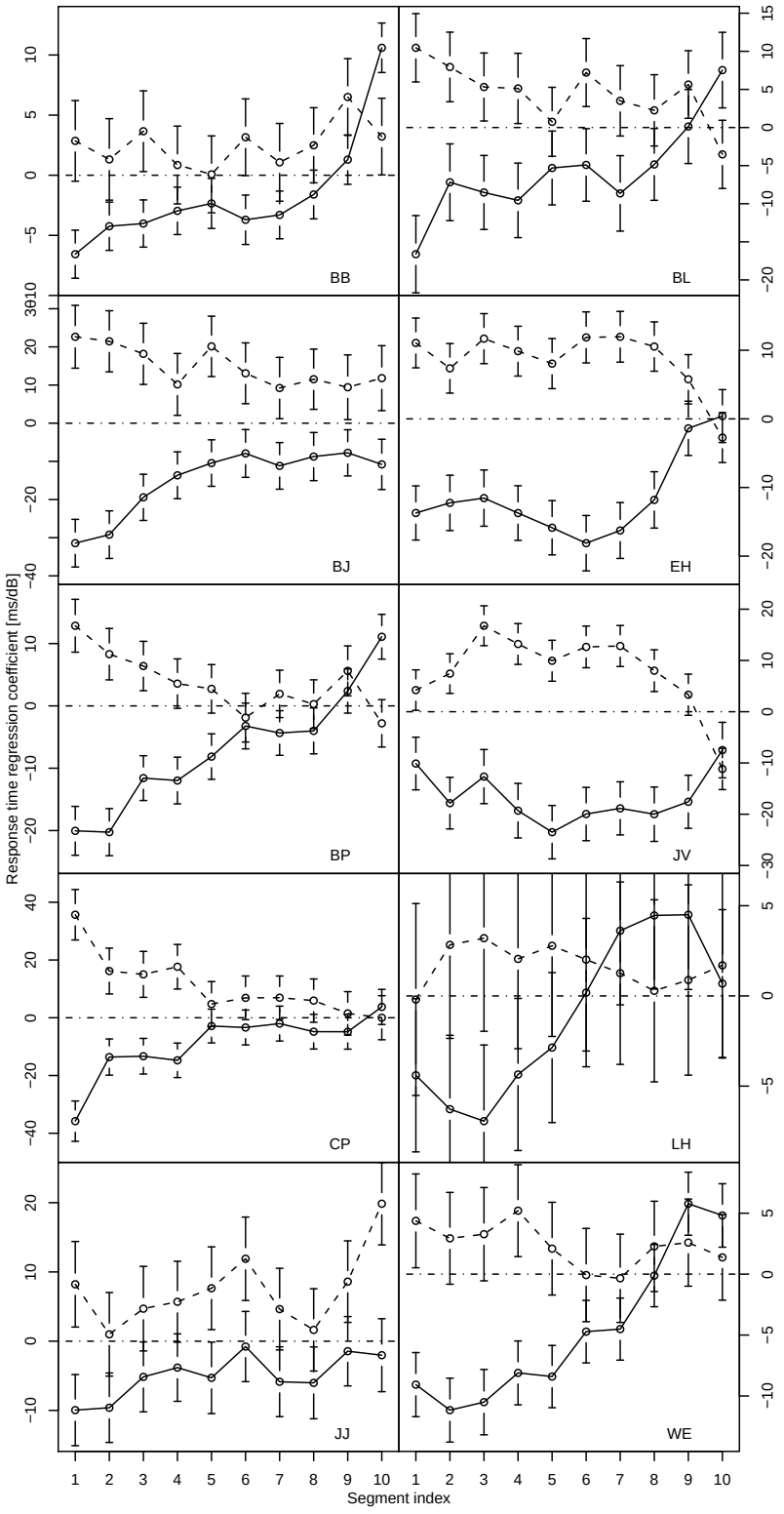


Figure 3.19: Weighting curves for segment levels for the prediction of response time. The solid line is the weighting curve for that cases where the listener responded “loud”, and the dashed line for the cases where he/she responded “soft”. Listeners in the right column received feedback and listeners in the left column did not receive feedback.

All in all, this suggests that response time not directly relates to loudness. Rather, the response time is related to the task the listeners have to perform, which in turn is related to loudness. It seems that, when the task is difficult for the listeners (at levels close to the mean segment level of the entire experiment), response time is long, and shorter when the judgment they must make is "obvious". In general, when interpreting response times in terms of loudness it is therefore important to have intimate knowledge about that "nature" of the listeners' task.

Further, more evidence was provided that temporal integration is not a simple smoothing process, because single segments must be "singled out" (perceptually) to receive less (or more) weight.

In this section both the term "reaction time" (initially used) and "response time" have been used, and in their context describe two slightly different things: "Reaction time" describes the measurement of time in tasks where listeners make their responses as fast as possible and "response time" in tasks where listeners give their response at a "natural" pace. The results of an experiment may of course depend on whether "reaction time" or "response time" is measured, but it is not obvious which would be more directly related to perceived loudness.

3.7 Temporal weights in a comparison task

The experimental design applied by Pedersen and Ellermeier (2006) required the listeners to consider only a single sound in their judgments of loudness. In a similar experiment Ellermeier and Schrödl (2000) derived temporal weights, but based on a two-interval forced-choice paradigm, where listeners were asked to judge which of two sounds was louder. They used similar stimuli as were used by Pedersen and Ellermeier (2006), but they did not include any spectral changes and trial-by-trial feedback was always provided. In their analysis and derivation of temporal weighting curves they assumed that the two sounds of a given trial were identically weighted and that the listeners' responses were determined by the differences between the two weighted sums of segment levels. It may however by criticized that it is not obvious that listeners weight the sounds of the two intervals equally, and further, it is not obvious either that the comparison process carried out by the listener is only a matter of comparing two summed values. By applying a one-interval paradigm, Pedersen and Ellermeier (2006) avoided these problems, but it also seems worthwhile reanalyzing the data of the study by Ellermeier and Schrödl (2000) where the critical assumptions are avoided. This potentially gives some insight into how listeners compare loudnesses.

3.7.1 Data collection

For the purpose of reanalyzing the result, the data of 7 listeners contributing 3000 trials each were used. A more detailed description of the procedure is given by Ellermeier and Schrödl (2000).

3.7.2 Results

Temporal weights were derived for individual listeners using the same procedure as described by Pedersen and Ellermeier (2006). However, on a given trial two sounds were presented, so the regression model is formulated slightly different:

$$D(\mathbf{x}) = \left(\sum_{i=1}^{10} w_{2,i}x_{2,i}\right) - \left(\sum_{i=1}^{10} w_{1,i}x_{1,i}\right) - c \tag{3.5}$$

Here $\mathbf{x}$ is a vector of the segment levels of both the sound of the first interval ($x_{1,i}$) and the second interval ($x_{2,i}$). The constant c allows for a bias toward the sound of the first or the second interval. D expresses the strength of the decision variable on which the listener is assumed to base his/her

response. $w_{1,i}$ are the weights put on the ten segments of the first interval while $w_{2,i}$ are the weights put on the segments of the second interval. As earlier, a logistic function was used to statistically relate the decision variable to the probability of a given outcome, which in this case is the probability of the sound of the second interval being judged as louder:

$$\Psi(\mathbf{x}) = p(\text{“2}^{\text{nd}}\ \texttt{interval louder”} \mid \mathbf{w}, c, \mathbf{x}) = \frac{1}{1 + e^{c - \sum_i w_{2,i} x_{2,i} + \sum_i w_{1,i} x_{1,i}}} \tag{3.6}$$

The weighting curves were obtained by maximum likelihood optimization for the regression coefficients with respect to the listeners' responses. Individual weighting curves are depicted in Figure 3.20, where the weighting curves for the first interval are given by $w_{1,i}$ and $w_{2,i}$ for the second interval. The weights of the second interval were normalized so their sum is one. The weights of the first interval were scaled by the normalization factor used for the second interval, so they do not sum to one in general. Using the same scaling factor for the two intervals allows for a comparison of the weighting curves across the intervals.

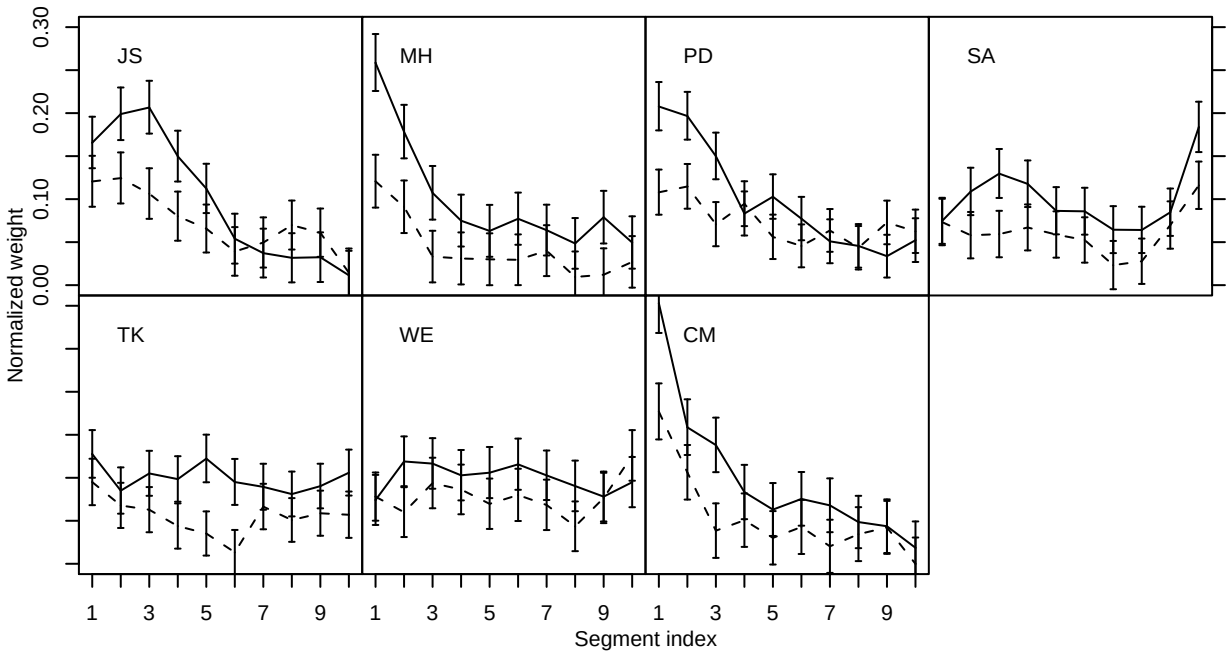


Figure 3.20: Individual weighting curves when comparing loudness of sounds in a two-interval task. The solid curve is the weighting applied to the sound of the second interval, and the dashed curve is the weights applied to the sound of the first interval. The ID of each listener is indicated in the upper left corner of each panel. 95%-confidence intervals are indicated by the errorbars. The total duration of each sound was 1 s.

As seen in the figure, the individual weighting curves generally have a similar shape in the two intervals, but the weights put on the second interval are larger. Most listeners emphasize either the onset or offset of sounds, which was also observed in the study by Pedersen and Ellermeier (2006), and considerable interindividual variability of the weighting curves is observed.

Mean curves for each interval and across listeners are shown in Figure 3.21. The mean weighting curves generally show the same tendencies as the individual weights: The second interval is weighted more heavily, but the shape of the weighting curves is similar for the two intervals.

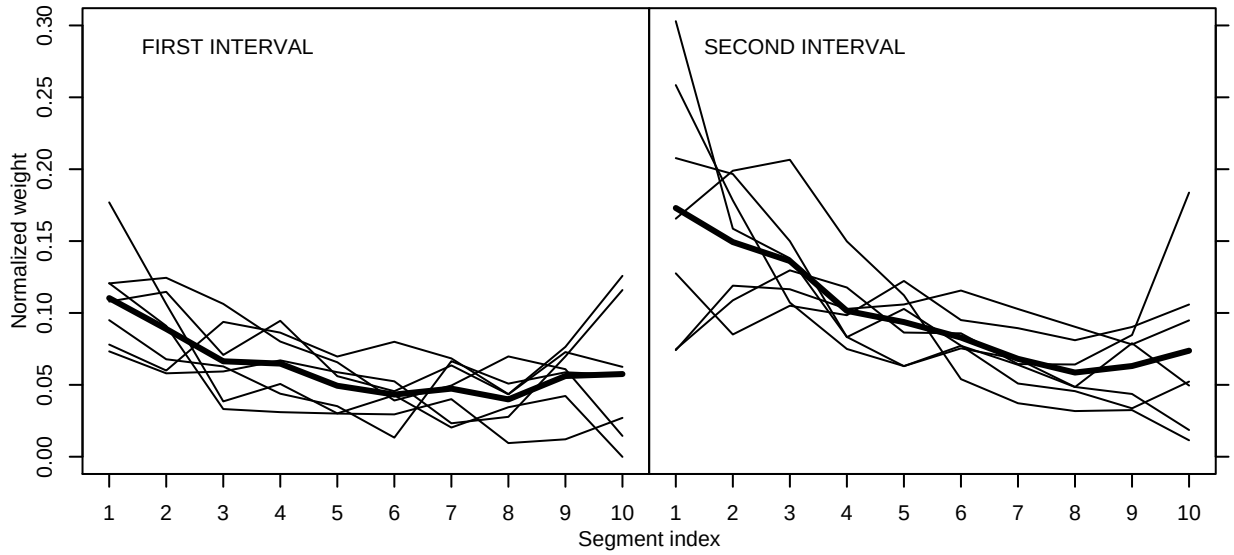


Figure 3.21: Mean weighting curves when comparing loudness of sounds in a two-interval task. The left panel shows the weighting of the first interval and the right panel shows the weighting of the second. The thin lines are the individual weights as also presented in Figure 3.20. The total duration of each sound was 1 s.

3.7.3 Stimuli of 200 ms duration

The study by Ellermeier and Schrödl (2000) also contained a condition where the total duration of the sounds was 200 ms, but these results were not presented in their paper. However, it is of interest how temporal weighting might be different when the duration is reduced. Therefore the data of this condition were also analyzed and the results are presented in Figure 3.22 for 8 individual listeners. The sounds were composed of ten segments as earlier, but the duration of a single segment was only 20 ms. Considering the results presented in the figure, the individual weighting curves again show that the second interval is weighted more heavily. The weighting curves are generally flatter than for the longer stimuli, but some listeners put significantly more weight on early segments (BS and RW). It is observed that the weighting of adjacent segment can be different (two first segments of RW for example). This suggests that perceptual emphasis of onsets and offsets is smaller for the shorter sounds, but underlying the listeners' judgments, individual segment loudnesses must be available for the listeners, or else single segments could not be singled out receiving greater weight.

For the sake of completeness the mean weighting curves are presented for the 200 ms sounds in Figure 3.23

3.7.4 Summary

When listeners had to judge and compare loudness of sounds in a two-interval task it was observed that:

- The sound of the second interval was weighted more heavily.
- The shapes of the weighting curves in the two intervals were similar.

When the duration of the sounds was reduced from 1 s to 200 ms, the emphasis on onset and offsets was reduced, but a few listeners emphasized the onsets, indicating that the temporal resolution of the perceptual processes underlying their judgments is finer than that of a single segment (20 ms).

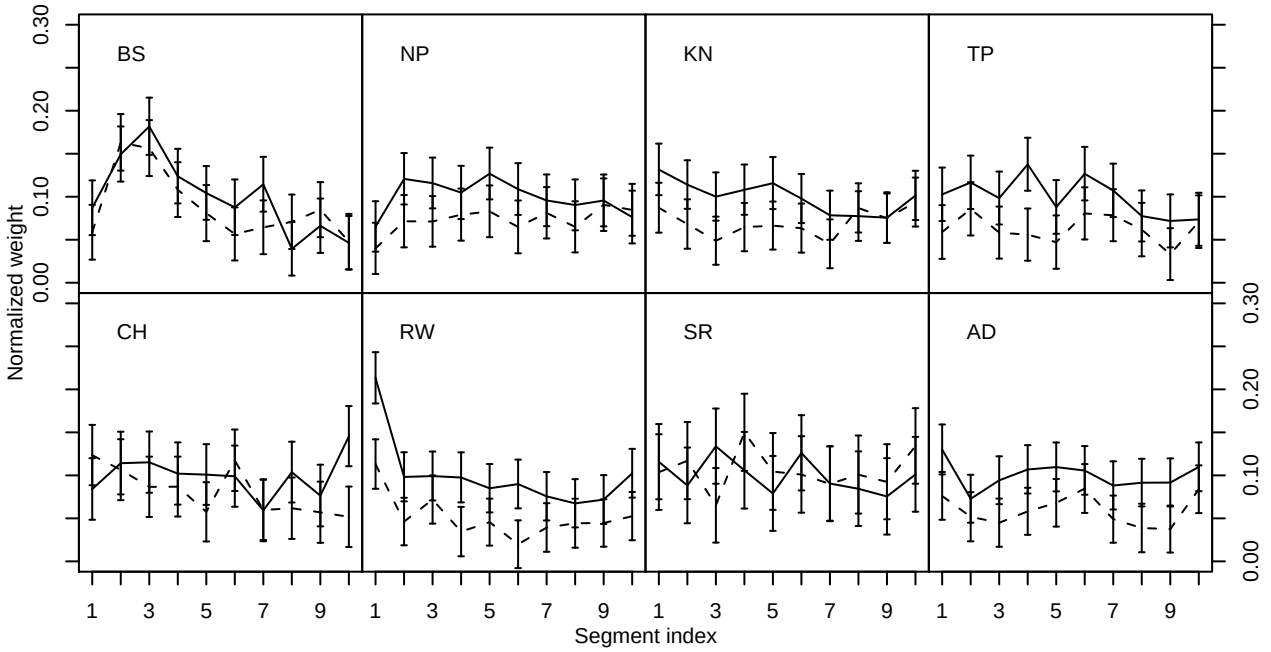


Figure 3.22: Same as Figure 3.20, but the total duration of each sound was 200 ms.

All in all this may suggest that the same perceptual process is responsible for separately integrating the sounds of each interval (similar shape of the weighting curves for first and second interval). The reason why the last interval is weighted more heavily may be explained in terms of memory effects, if the sound of the last interval is assumed to be more fresh in memory (recency effect), the response of the listener will depend on the last sound to a larger extent. However, it may also be a matter of attention for example. It is easy to imagine that instructing listeners to pay more attention to the first sound would increase its received weighting.

3.8 Conclusions

The main purpose of this chapter was to identify important factors in loudness judgment requiring temporal integration. This final conclusion will summarize the implications of the described results for the overall goal.

3.8.1 Temporal loudness integration is *not* a simple summation process

First, this is supported by the observed fact that judging the loudness of a sound on a given trial depends on the sounds of previous trials. This indicates that listeners arrive at their judgments considering the sound on a given trial in its relation to sounds of other trials in a process, which is significantly more complicated that "sampling" the summed loudness. Secondly, some listeners were observed to almost ignore specific segments of a sound when they were below a certain level (Figure 3.6), which would not be possible if all segments of a sound were included in a summation. Thirdly, the mathematical formulation of attention weighting was quite successful in predicting the listeners' performance. Attention cannot be paid to specific segments of a sound only, after a "summing" procedure has been applied, so to the extent that attention effects were observed this rules out simple loudness summation in the periphery of the sensory system. Further evidence that simple summation does not occur

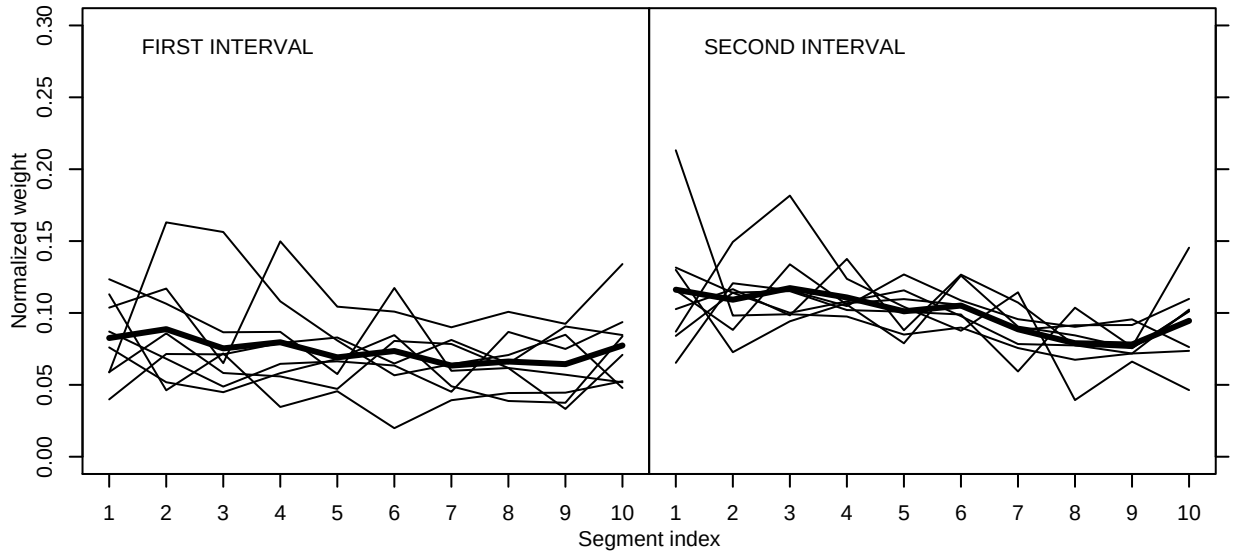


Figure 3.23: Same as Figure 3.21, but the total duration of each sound was 200 ms. The thin lines are the individual weights as also presented in Figure 3.22.

is given by Pedersen and Ellermeier (2006), who for example argue that it is hard to reconcile the observed fact that listeners are influenced (in their temporal weighting) by feedback with the notion of simple summation.

3.8.2 Temporal loudness integration is a *non-linear* process

The COSS functions of Figure 3.6 showed that the probability of a "loud" response is not linearly related to individual segment levels of a sound. Two conceptually different explanations for this observation were given: (1) There is a simple non-linear relationship between sound pressure and perceived loudness, or (2) what looks like a simple non-linearity is caused by a complex decision process where attention might play an important role. The two concepts were not easily disentangled when analyzing the results, but considering the ecological validity of the parameters of fitted models, evidence in favor of the "attention" explanation was found. Further support that the loudness integration is a complex decision task is given by the analysis of response time, which showed that response times are generally longer when the response is not obvious, the question is: Why do listeners take longer time to respond in this case? It is hard to imagine that a simple summation process would take longer time in this case, so rather it may be hypothesized that listeners consider their response more carefully in a process where the available information is carefully weighted. Finally, listeners' weighting curves in a task where they had to compare sounds of two intervals had similar shape for the two intervals. This may indicate that the two sounds are identified as separate events and weighted independently. This suggests that the loudness integration in a comparison task is a matter of identifying and thereafter weighting the relevant information.

Chapter 4

Paper 2:
Temporal masking in the auditory identification of envelope patterns

Temporal masking in the auditory identification of envelope patterns

Benjamin Pedersen*

Sound Quality Research Unit (SQRU), Department of Acoustics, Aalborg University, Fredrik Bajers Vej 7-B5, 9220 Aalborg Øst, Denmark

(Dated: September 12, 2006)

To probe processes responsible for temporal analysis in audition, possibly taking place at different stages of perception, listeners had to identify if temporal envelope patterns were either ascending or descending. A descending pattern was generated by decreasing the overall sound level of a broad-band noise carrier in three 10-dB steps. The duration of one step was varied over the range from 10 ms down to 0.2 ms. An ascending sound was constructed by time-reversal of a descending pattern, hence no overall cues were available for the discrimination in the amplitude spectrum. Continuous psychometric functions were derived for 5 listeners over the range of the segment duration. The temporal limit for 75%-correct identification was at approximately 1 ms segment duration. In three further conditions 1, 3, or 9 non-informative segments, at the same level as the central segment of the pattern, were added on each side of the target pattern to be identified. Adding 1 segment had little influence on listeners' performance, but when 3 or 9 segments were added the performance limit of the listeners was approximately 20 times higher (20 ms). Such large a decline in performance is hardly predicted by energetic temporal masking, hence alternative suggestions for the sensory processes responsible for the drop in identification perfromance are given.

PACS numbers: 43.66.Dc, 43.66.Fe, 43.66.Nm, 43.66.Mk

I. INTRODUCTION

Numerous studies have been concerned with measuring the ability and limitations of the auditory sensory system to analyze the temporal information present in sounds, and of these a broad range of experiments probe the hearing mechanism with the goal of determining its temporal resolution. Most studies arrive at an auditory resolution of a few ms (see for example Moore, 2003a,b, for a review). It is also broadly agreed that perception involves a chain of sensory processing at different cognitive levels before the listener is able to respond to the task defined by the experiment. Attempts have been made to explain the functioning of the separate levels, but at a closer look it becomes apparent that no model of hearing gives a unified description, which is broadly able to predict the behavior of listeners. This paper attempts to identify some of the reasons for this, by probing listeners' ability to identify temporal variation in a range of tasks, which models would typically predict to be of similar difficulty. Contrary to the model predictions it is demonstrated that listeners' performance is very different across the tasks. In an attempt to explain this discrepancy, it is explained why different stages of perception may be responsible for the performance in the different tasks.

*Electronic address: bp@acoustics.aau.dk

A. Stages in modeling hearing

Models of hearing typically follow the reasoning as for example outlined by Oxenham and Moore (1994) including the stages: (a) Stimulation reaching the ear is filtered into frequency bands resembling the frequency mapping taking place in the cochlea, (b) a non-linear level-dependent transformation accounting for non-linearities known to exist, (c) sliding integration using a window accounting for forward and backward temporal masking, commonly described by two exponentially decaying functions, and finally (d) a decision mechanism. The last stage, the decision process, is often prone to assumptions made by the experimenter and often includes some sort of simple overall integration closely following the concept of mathematical integration. However, the assumption of simple integration has been demonstrated to be problematic by Viemeister and Wakefield (1991), who, in a simple signal detection task, found that listeners' performance was incompatible with simple integration of masking noise. Rather, they hypothesized, listeners may be assumed to "integrate" over several "looks" - perceptual excerpts of the sound. Given the decision mechanisms' influence on the predictions of the overall model, it seems worthwhile to probe the assumptions via listening tests, whose interpretation is not bound by a specific model, this is however not so easy in practice.

In an attempt to more clearly identify processes, which may take place at difference perceptual stages, and thus should be modeled and interpreted accordingly, a series of two experiments is proposed, and this paper describe the first of the two. Four related tasks were included in the first experiment and were designed to probe the

supposedly higher levels of perception in tasks of varying difficulty, but all putting similar demands on temporal resolution in the sensory system. The variation was made by the addition of non-informative noise on the sides of a "target" pattern, which had to be identified (see Figure 1 for an overview). Consequently, there will be a focus on various aspects of the higher level decision processes in this paper while, in the second paper, it is demonstrated that extremely low measures of temporal resolution of hearing can be obtained by manipulating stimuli similar to the ones used in this work.

B. Neurophysiological and psychoacoustical evidence for different stages in hearing

Näätänen and Winkler (1999) give an extensive review of the neurophysiological evidence that different stages exist, and their findings are explained in somewhat different terms than the model for example described by Oxenham and Moore (1994). Näätänen and Winkler (1999) find evidence for and conclude that perception includes the following stages (in this particular order): (a) Feature extraction, where specific features of the stimulation arise, and (b) a feature trace, i.e. the extracted features are buffered for a while and remain available for (c) a unitary stimulus representation where different features are integrated and mapped onto a temporal dimension. Of these stages the first two are considered pre-representational and the last representational by the authors, and only the last stage carries integrated information and information about temporal order. Also, the auditory information is only available for conscious perception after the last stage. However, it is clear that specific features may be available for conscious perception before a fully integrated percept exists, for example the pitch of a level-fluctuating tone is available before the temporal properties of the level-fluctuations are completely mapped to a temporal axis. Hence, it may be hypothesized that there is a major difference in the performance of listeners in tasks where they can make judgments based on single static features alone or if they must be aware of the temporal properties of the stimulation in order to do their task. In the discussion of listening experiments in which listeners can perform with a very fine temporal resolution it is often observed that the listeners performed their task according to a difference in "quality" between the stimuli defined by the task. This is for example the case in the work described by Divenyi and Hirsh (1974) where listeners were asked to judge the order of three different tones in a sequence. The temporal separation of the tones was varied making the discrimination task increasingly difficult. This is interesting for the current discussion because the nature of the task seemed to change with the temporal separation: When tones were sufficiently spaced the listeners had a clear impression of a tonal sequence, this however changed when the separation decreased, making it harder to identify the temporal position of individual tones. Interestingly, for very short separations (in the order of a few ms), the listeners were still able to perform the task, but based their judgments solely on the perceived quality of the entire sequence, not being able to identify individual tones.

This has inspired the current work, as this type of task seems able to probe both perception of temporal order and perception of features, taking place at relatively higher (e.g. decision device) and lower stages (e.g. temporal windowing) of cognition.

Further support for the notion of feature extraction may be found in the psychoacoustic literature concerning auditory amplitude modulation detection (see Viemeister *et al.*, 2002, for a review). It is beyond the scope of this paper to review this large body of literature, but it is interesting to note how the concept of specialized modulation detectors (modulation detection filter banks) is introduced to explain the results of listening experiments. For the purpose of this paper, such detectors may well be considered specialized feature extractors. The notion of "hard-wired" modulation filter banks may be questionable, but the evidence that a "modulation" feature is present in perception is interesting for the current discussion.

Also, the literature concerning "feature extraction" in pitch perception and binaural hearing is interesting in this context, and is reviewed more carefully by Pedersen (2006).

C. Probing "low" or "high" cognitive stages in listening experiments?

It is often assumed that listeners' performance in relatively simple auditory tasks reflects peripheral stages of hearing. It is however not obvious that this assumption is always valid. The question is for example raised in a slightly different research area, namely in research in reading and writing disabilities. The main question is whether such disabilities may be caused by deficits in the temporal processing and maybe even related to peripheral parts of the auditory sensory system. Reviews of the evidence are given by Farmer and Klein (1995) and Studdert-Kennedy and Mody (1995). More recently Schäffler *et al.* (2004) demonstrated that improvements obtained by training people suffering from dyslexia in "low-level" auditory tasks significantly transferred to language related and phonological skills and to spelling. The auditory tasks used for training involved for example intensity, frequency discrimination, and gap detection. In their framework such tasks were "low-level" per definition and it is not clear at which stages of perception the performance actually improved. It could reasonably be argued that the benefit of training would be assumed biggest at the higher levels of cognition, indicating that for example gap detection puts high demand on the higher levels of cognition. This would be at odds with the way results of gap detection and for example

masking experiments are used to derive properties of a temporal window of auditory perception (as in Oxenham and Moore, 1994, for example).

D. A task probing several levels of perception

The work described in this paper will put different demands on listeners' ability to identify temporal properties via a task which is, in its basic form, similar to the three-tone sequences used by Divenyi and Hirsh (1974). But rather than changing frequency over the three elements of a sequence (as Divenyi and Hirsh), the sequence is composed of three bursts of noise of different (ascending or descending) overall level (Figure 1). In its basic form there is a difference in quality between the ascending and descending patterns. Later the difference in quality will be diminished by the addition of non-informative noise on the sides of the patterns. The listener thus has to identify the relevant pattern embedded in the noise in order to do her/his task, thus putting demands on a different level of perception. The use of level changes, as opposed to changes in frequency, makes the task conceptually more similar to temporal masking and gap detection, possibly providing a bridge between the different classes of experiments. It will be explained and illustrated why a traditional "temporal window" model hardly predicts the difference in performance in the relevant tasks.

According to a model based on "smoothing" by a temporal window it is expected that, in all tasks of the current experiment (different degrees of added noise), the performance of the listeners declines at similar rate as the duration of the stimulus segments is decreased, depending on the temporal extent of the "smoothing" window. If this is not the case, however, that outcome suggests that different types of processing influence performance, possibly operating at different cognitive stages across the different tasks. The outcome will not explain how to formulate a mathematical model of hearing, but it may provide important hints as to which aspects must be included and which concepts are incompatible with the behavior of listeners.

II. METHOD

A. Listeners

Five listeners (3 female, 2 male, including the author (BP in the figures)) participated, and were recruited from students at the university. The mean age of the participants was 24.6 years (range: 19 to 30 years). None of the listeners had a significant hearing loss (>20 dB HL at any of the frequencies 250, 500, 1k, 2k, 4k, 6k, 8k Hz). Listeners BP and MC had previous experience in listening tests, while LP, MK and DP participated in a listening experiment for the first time.

B. Apparatus

Sounds were generated on a computer and converted at a sampling rate of 96 kHz and a resolution of 24 bit using an external digital to analog converter (Edirol UA-5). A headphone amplifier (Tucker-Davis Technologies System 3 HB7) was used for powering the headphones. All sounds were presented diotically via headphones (Sony MDR v6). The listeners were positioned in a sound-insulated booth and gave responses via a custom-made button box with lights providing feedback.

C. Stimuli

The task of the listeners was to discriminate sounds containing a descending pattern in the envelope from sounds containing an ascending pattern. This was done in four different conditions as illustrated by the rows in Figure 1. The central part of the stimulus always contained the 3-segment long ascending/descending pattern, while the number of non-informative noise segments preceding and succeeding the pattern was varied across conditions: In the 3-segment condition no noise segments were added, and in the 5-, 9-, and 21-segment conditions 1, 3, and 9 non-informative noise segments were added on each side of the target pattern. The segment duration was the parameter that was varied by the psychophysical procedure. So for a given segment duration the overall duration of stimuli in different conditions was different because the number of segments per sound was different.

1. Noise carrier

The patterns shown in Figure 1 represent schematic envelopes of modulated noise carriers. The noise carrier was the same type in all conditions, and it had a broadband frequency spectrum, but not the properties of white noise.

Since the noise was used for probing temporal resolution of the sensory system, a special carrier noise was designed to avoid problems of inherent fluctuations of the noise carrier, which could be confused with fluctuations in the modulator amplitude. At short segment durations the amplitude of the modulated noise carrier must be able to change rapidly if it has to closely reflect the envelope of the modulator. The carrier noise was designed fulfill both these requirements. Each sample in the noise had a binary character, having a value of either A or $-A$, where A was given by the amplitude of the modulation envelope. The amount of negative and positive samples was balanced within groups of two samples, which means that all samples of the noise can be arranged in pairs of adjacent samples, always having opposite sign for the amplitude A. This is illustrated in Figure 2. Within each pair of samples the order of A and $-A$ was random. As a consequence the amplitude of

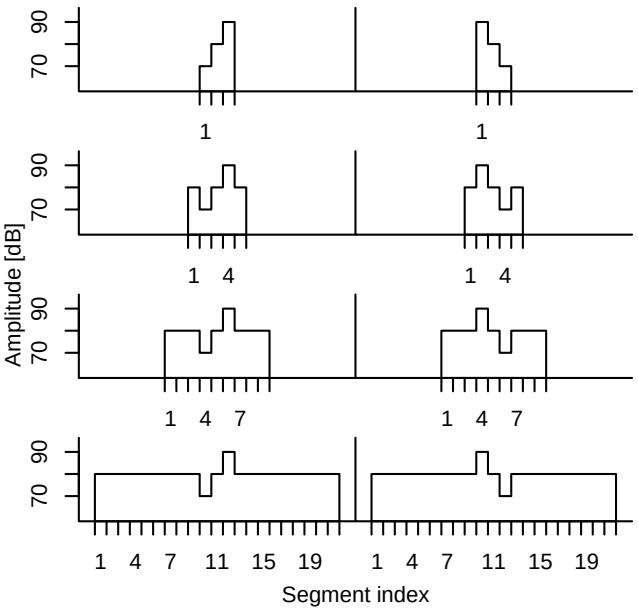


Figure 1. Stimuli used in different conditions: 3-, 5-, 9-, and 21-segment conditions in each row respectively. The y-axis shows sound level for each segment. The x-axis shows the temporal index of each segment, the actual duration of which was varied.

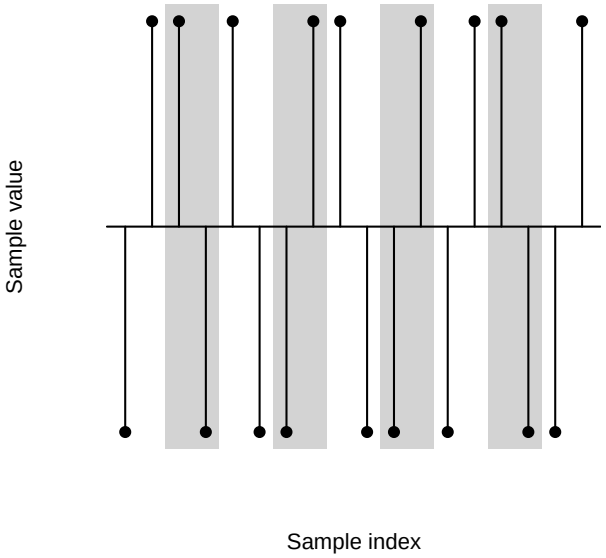


Figure 2. Example of a noise carrier excerpt. Blocks marked by gray and white background contain equally many negative and positive samples in random order.

the modulated noise very closely followed the amplitude of the modulator, because the mean of the squared sample values of an excerpt will be very close to A^2 for any given number of samples included in the excerpt, as long as the excerpt is more than one sample. Thus, for a running mean applied to the modulated noise, the output of the running mean will reflect amplitude with a temporal resolution determined only by the extent of the window used in the running mean.

For one track of trials the noise carrier was "frozen". This was especially important for short segment durations, where the perceived quality of a sound varies more between different noise carriers than between different directions of the target pattern. For short durations of the noise carrier, its long term properties are not valid, meaning that the spectrum can change dramatically across different short excerpts of the noise carrier. For each new track of trials a new "frozen" sequence of samples was used.

A descending sound was generated by time-reversal of the corresponding ascending sound. Therefore the amplitude spectra of the two sounds are identical over the duration of the sounds. It may be argued that the beginning of a sound is different in its fine-structure after it has been time-reversed, which potentially provides a cue for discrimination even if non-informative noise is present at the beginning. Therefore the noise carrier was always mirror-symmetric over the duration of a sound, with the line of symmetry in the middle of the central segment. For this reason non-informative segments in the 5-, 9- and 21-segment conditions are unaffected by time-reversal of the stimuli, which was used to generate the target patterns in opposite directions. Another consequence is that descending and ascending sounds started and ended with the exact same samples within the three conditions. Further, it makes the decomposition of the stimuli, as illustrated in Figure 14 in the appendix, possible. The smallest possible segment duration is equal to the duration of 4 samples. To understand this, consider the central segment of a sound, which must obey the following criteria: The segment has to contain equally many positive and negative samples and at the same time be mirror-symmetric, which cannot be fulfilled with less than four samples. In general the segment duration must always be a multiple of four samples.

A more thorough description of the statistical properties of the stimuli is given in Appendix B.

2. Calibration

As shown in the Figure 1, the calibrated level of the non-informative noise was always 80 dB SPL and the levels of segments of the target pattern were either 70, 80, 90 dB SPL (ascending pattern) or 90, 80, 70 dB SPL (descending pattern). The setup was calibrated relative to a setup with the headphones positioned on a dummy-head (Kemar) where the level of a sine tone with a frequency of 1 kHz was measured in the ear of the dummy-head. Also, the transfer function from the headphones to the ear was recorded. In this paper all sound pressure levels in dB refer to what would have been the level of a 1 kHz sine tone

at the applied gain settings. The actual sound pressure level at the ears will be lower, because the noise contains more power at higher frequencies where the headphone transfer function is highly attenuated. The actual sound level at the ears is estimated in Appendix C, where it is found that the actual level at the ears is approximately 13 dB lower than reflected by the calibrated levels.

D. Experimental procedure

A two-interval, forced-choice paradigm was adopted. In one trial the listener heard two sounds, one of which contained a descending pattern and the other contained an ascending one. The task of the listener was to identify which sound contained the descending pattern. The experiment was divided into tracks in which trials contained only sounds from the same condition (see Figure 1).

The use of an adaptive procedure was avoided to be able to detect non-monotonicities in the psychometric functions, and as to obtain "full" psychometric functions as a function of the segment duration. Also, the perceived quality of sounds of very different segment durations is dramatically different. Therefore it seemed natural to gradually reduce the segment duration throughout one track in order to avoid a drop in performance because the listener had to change his/her decision criterion from trial to trial. However, it cannot be ruled out that effects of this way of varying the segment duration is reflected in the psychometric functions derived. Indeed it can be argued that the performance is worse in the beginning of each track because the listener has to become familiar with the task. Thus the performance, as a function of segment duration, would be underestimated for long segment durations, which always occurred in the beginning of a track. It could also be argued that the listener always looses attention toward the end of a track, leading to an underestimation of the performance as a function of segment duration. Further it could be argued that the performance is better for short segment duration because of long term learning effects. This is however unlikely, because in the time frame of the entire experiment, different segment durations are almost perfectly balanced.

In one track the segment duration was slowly decreased; in the 3- and 5-segment conditions from 10 ms to 0.1 ms and in the 9- and 21-segment conditions from 100 ms to 0.1 ms. The segment duration was reduced with a logarithmic step size with 80 steps per decade. Thus one 3- or 5-segment track contained 161 trials and a 9- or 21-segment track contained 241 trials.

E. Data collection

The presented results are based on 8 tracks per listener per condition. So the total number of trials per condition per listener was $8 \times 161 = 1288$ in the 3- and 5-segment conditions and $8 \times 241 = 1928$ in the 9- and 21-segment conditions. The listeners first went through the 3- and 9-segment conditions where tracks of both conditions were presented in an alternating order. Only thereafter the listener started the 5- and 21-segment conditions, also in an alternating order. All listeners were initially trained in the 3- and 9-segment conditions. The same segment durations were used in the training as in the experiment proper, except for the very first tracks of trials in the 3-segment condition, which started with a segment duration of 100 ms. The amount of training required varied from 1 to 2 hours depending on the listener. It was very different how easily listeners learned the task, especially in the 9-segment condition (the first condition with non-informative segments). No extensive training in the 5- and 21- segment conditions was given, because of the similarity to the 3- and 9-segment conditions. However a few initial tracks of trials in each of the conditions were used for familiarization before the data collection started.

III. RESULTS

A. Procedure for deriving psychometric functions

For each listener, psychometric functions were derived for each of the 6 experimental conditions. These are displayed in Figure 3. Figure 4 offers psychometric functions for pooled results across listeners. The psychometric functions were constructed in the following way: The results of all trials for a given listener and condition, were sorted (according to the segment duration) and arranged in 50 overlapping bins. The bins were equally spaced on a logarithmic time-axis, and the results of 200 trials were included in each bin. This means that the curves are considerably "smoothed" and only approximately 3 estimates per decade are based on completely non-overlapping data. Performance was estimated within each bin by dividing the number of correct responses with the total number of trials in that bin. By considering each trial a Bernoulli trial, 95%-confidence limits were estimated based on the Binomial distribution. The 95%-confidence limits are marked as gray-shaded areas in the figures.

Different number of trials were included in the bins when the results were pooled across listeners (1000 trials per bin).

Generally the psychometric functions show how listeners' performance worsens as the segment duration becomes smaller, which happens at relatively long segment durations when noise segments are added.

B. Procedure for estimating performance level

The segment durations corresponding to a performance level of 0.75 (75% correct responses) were estimated for

Condition	Perf. level	BP	LP	MK	DP	MC	ALL
3–segment	0.75	0.5	0.7	1.0	2.0	2.2	1.1
	0.60	0.4	0.4	0.6	1.0	1.3	0.5
5–segment	0.75	0.3	0.8	0.6	14.4	2.4	1.1
	0.60	0.1	0.2	0.3	1.5	0.8	0.2
9–segment	0.75	4.7	41.3	19.9	21.0	99.4	22.8
	0.60	3.2	17.6	12.3	16.9	35.0	8.3
21–segment	0.75	9.5	22.6	29.8	27.9	–	30.0
	0.60	5.7	13.0	25.4	16.8	–	11.7

Table I. Segment duration in ms at 0.75 and 0.60 performance levels for all listeners in all conditions. The first row in each condition is the 0.75 limit and the second is the 0.60 performance limit. The last column, ALL, is derived from the pooled data for all listeners. The limit of MC in the 21-segment condition was outside the range of segment durations used in the experiment.

all listeners in all conditions. That was done by fitting curves to the obtained data and calculating the segment duration where the fitted curve had an ordinate value of 0.75. A logistic curve was used to model the psychometric functions, taking the following form:

$$\Psi(T) = \frac{1}{2} + \frac{1}{2} \cdot \frac{e^{c(log(T)+k)}}{1 + e^{c(log(T)+k)}} \qquad (1)$$

Ψ models the probability of a correct response as a function of the segment duration T. In the formula k determines the segment duration of the transition point of the psychometric function, and c determines the steepness of the fitted function at the transition. T is the segment duration measured in seconds. The function is fitted on the logarithmic time axis by using $log(T)$ in the equation. As can be seen, when the segment duration approaches 0, Ψ approaches 1/2 and when T approaches ∞, Ψ approaches 1. That means the psychometric function is forced to approach the chance level of 50% correct for small segment durations, and for long segment durations it is forced to approach a performance level of 100% correct. The coefficients c and k were estimated in a least square fit of Ψ to the binary response data. That is, Ψ is *not* directly fitted to the estimated and smoothed performance curves (solid lines) in Figure 3 and 4.

C. Temporal limits for individual listeners

When considering the limits for temporal resolution estimated in the different conditions as summarized in table I and illustrated in Figure 5 and 6, it is evident that there is a difference in performance between the 3- and 9-segment condition, for which data were collected concurrently. When looking at actual values it is observed that the limit in the 9-segment condition is 10 to 50 times higher than the limit in the 3-segment condition. This of course means than the duration of one segment in the 9-segment condition must be 10 to 50 times as long as in the 3-segment condition for the task to be of similar difficulty. A similar difference is observed when comparing the 5- and 21-segment conditions. Here the factor is in the range 2 to 50. If listener DP is excluded due to the problematic estimate in the 5-segment condition (see Figure 5) the factor is in the range 30 to 50. For the pooled results of all listeners there is a difference of a factor 20 in the 3- and 9-segment conditions, and in the 5- vs. 21-segment conditions the factor is approximately 30.

1. Problematic estimates

In general the fitted curves follow the estimated psychometric functions quite well (see Figure 5). In some situations, however, it is problematic to derive the 0.75 performance level: In the 21-segment condition listener MC never performs significantly better than chance, hence a function cannot be fitted and no estimate of the performance limit given. In the 9-segment condition MC's performance may be underestimated as is also the case for listener LP in that condition. In the 5-segment conditions the psychometric functions in general have a shallow slope, which questions the validity of the estimated limit. For listener LP in the 5-segment condition, for example, a reasonable limit seems to be at 0.2 ms segment duration (see Figure 5), however 0.8 ms was estimated based the fitted curve. From inspection of the curves it seems as if the knee-point where the psychometric function reaches the level of chance is less variable across listeners than the point of the psychometric function where performance starts falling from "all correct". Thus it can be argued that a better estimate of the performance limit is obtained at a relatively small performance level; 0.60 for example. For this reason estimated limits at this level are also calculated and shown in table I.

D. Individual psychometric functions

An surprising phenomenon, not accounted for by the fitted curve, can be observed for listener MK in the 9-segment condition (see Figure 3). After the expected decline of the psychometric function it starts rising again, starting from a segment duration of approximately 2 ms and peaking around 1 ms. Even below this segment duration MK's performance stays significantly better than chance. Before any of MK's results were analyzed he noticed this effect himself and for the 9-segment condition expressed that, in the beginning of the track of trials the task was easy, and soon got very difficult, but later became easier again. The 95%-confidence limits in Figure 3 show that this increase in performance at low segment durations is significant. A similar effect, but much weaker, can be observed for BP and LP: BP's performance stays better that chance below 1 ms segment du-

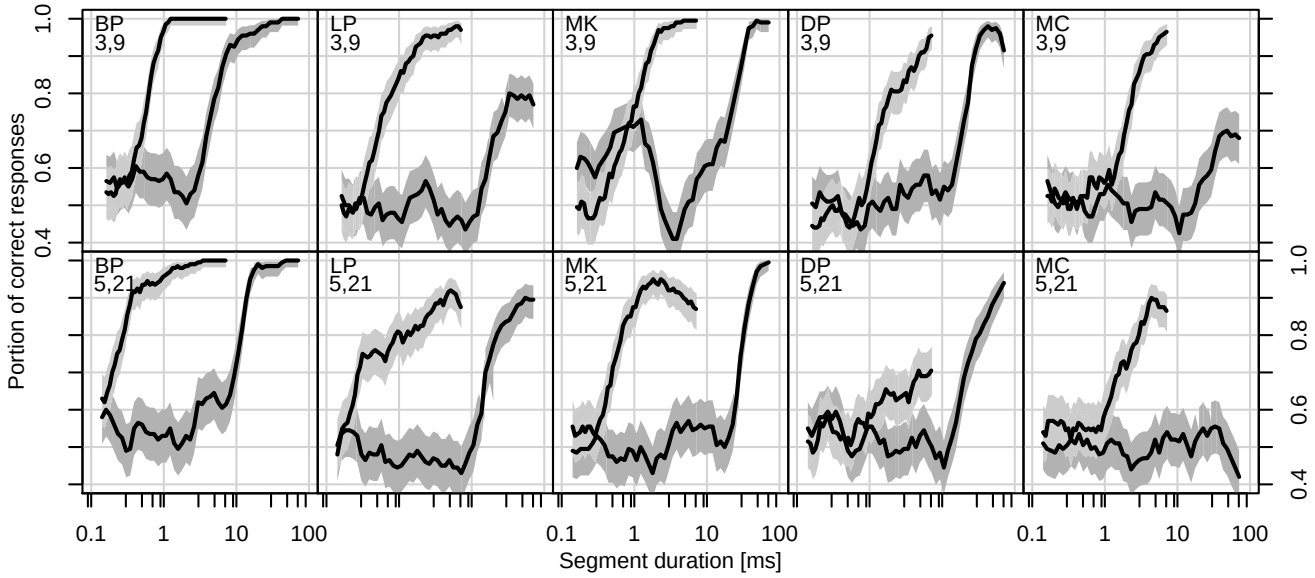


Figure 3. Estimated psychometric functions for all listeners in all conditions. Segment duration on the x-axis and the proportion of correct responses on the y-axis. The upper row shows the results for the 3- and 9-segment conditions with the 3-segment condition more to the left. The lower row shows the 5- and 21-segment conditions where the 5-segment condition is more to the left. The shaded areas illustrate the 95%-confidence limits.

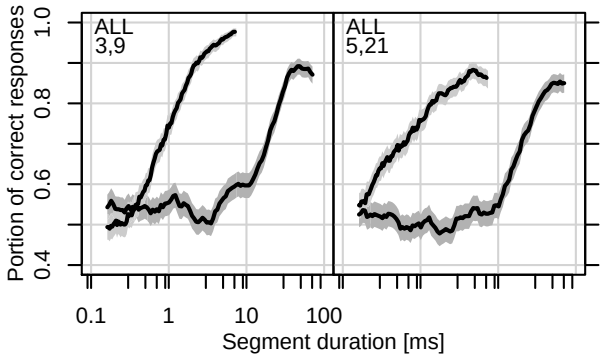


Figure 4. Same as Figure 3, but for the pooled data of all listeners. The results of the 3- and 9-segment conditions are presented in the left panel, and the results of the 5- and 21-segment conditions in the right.

ration, and there is a small increase in LP's performance at the same point where MK's performance started increasing. However, the increase in performance is not significant for LP.

E. Comparing 3- with 5-segment condition and 9- with 21-segment condition

When making these comparisons, caution should be taken, because the 3-segment condition was performed before the 5-segment condition and the 9-segment before the 21-segment condition. In terms of the temporal limit for discrimination, performance in the 5-segment is surprisingly similar to the 3-segment condition for most listeners. There are two possible reasons for this: (1) The 3-segment condition was performed before the 5-segment condition, so training effects could be an explanation and (2) small amounts of added noise do not worsen the performance (some listeners actually performed better in the 5-segment condition). In general the 21-segment condition is harder for the listeners than the 9-segment condition, however, LP performs better in the 21-segment condition, but is the only exception, and further, when the relevant curves for LP are compared it is clear that the evidence for LP performing better in the 21-segment condition is very weak. Indeed, when the very beginning of the curves are not considered, the psychometric functions very closely follows each other in the two conditions.

IV. DISCUSSION

The addition of non-informative noise segments pre- and succeeding a target pattern dramatically impaired listeners' performance in discriminating ascending from descending target patterns. What are the perceptual reasons for this? The influence of the noise could be speculated to be of two different origins: (1) Energetic masking or (2) informational masking (Durlach *et al.*, 2003). Since the current task bears some resemblance with gap detec-

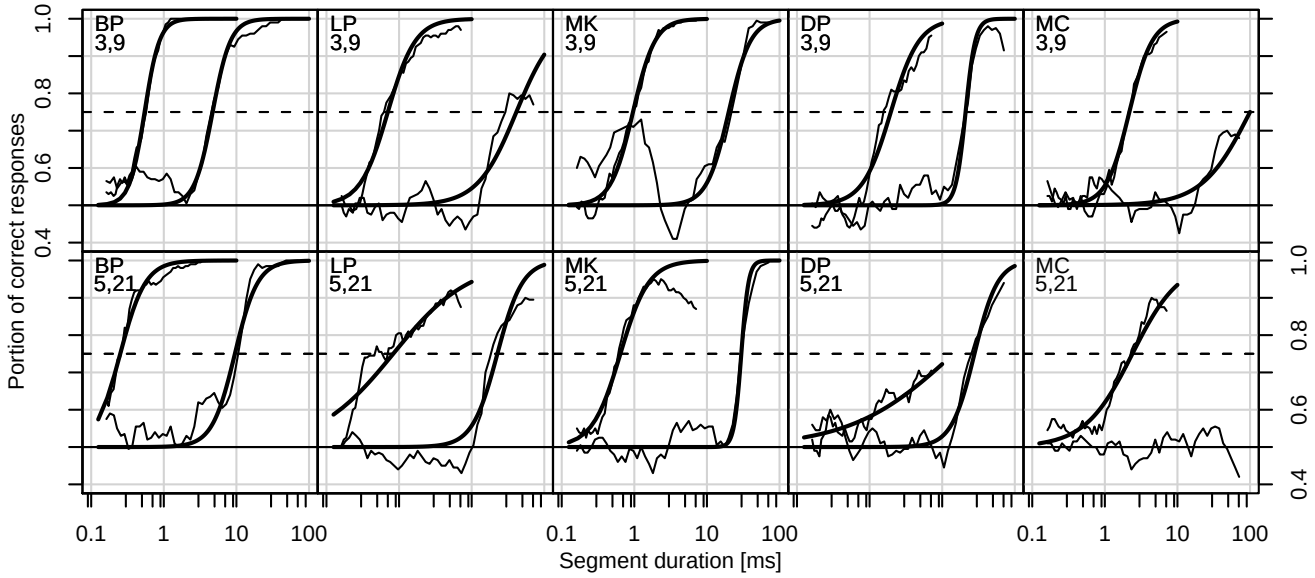


Figure 5. Same data as presented in Figure 3, but also including the fitted curves for estimating the limits of performance (thick lines).

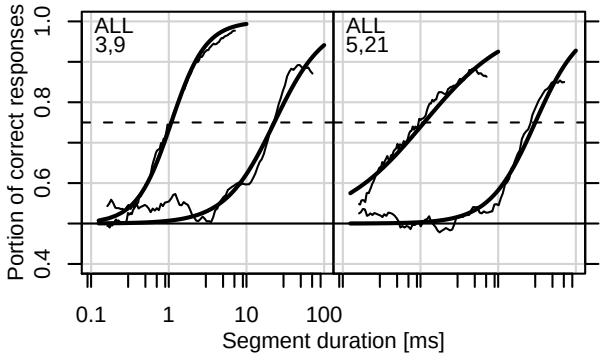


Figure 6. Same as Figure 5, but for the pooled data of all listeners. The results of the 3- and 9-segment conditions are presented in the left panel, and the results of the 5- and 21-segment conditions in the right.

tion, it seems worthwhile to consider the former, which is often used via a temporal window model (Oxenham and Moore, 1994) to explain the performance of listeners in experiments where similar stimuli is used. As it will be shown, a temporal window model does not readily explain the difference in performance in the different conditions, so also the concept of informational masking is worth deeper thoughts. The concept is very broad though, and no formal model is readily available for testing the ideas.

A. Predictions of a temporal window model

The principle working of a temporal window model, is a temporal "smoothing" of the envelope of a given sound. This is achieved by convolving (sliding) a fixed window, whose width determines the temporal acuity, with the envelope of the sound. In Appendix D the procedure is explained more carefully, and Figure 10 to 13 give examples of the output of the sliding window when it is applied to the stimuli of the current experiment. The figures thus give some hints as to which cues can be used for the discrimination task. The top row of a given figure contains descending stimuli and the bottom row ascending. From left to right the segment duration is decreasing, and at the same horizontal position stimuli in the two rows are at the same segment duration. So, in terms of the experimental task, stimuli in the top row had to be discriminated from that in the bottom (at the same horizontal position).

From inspection of the figures is not clear why the discrimination of patterns in for example the 5-segment condition, should be much easier than in the 21-segment condition: Local peaks and dips might be hypothesized to provide important cues, but these are more pronounced in the 21-segment condition (compare for example Figure 10 and 13 at $T/\tau = 1$). Alternatively, skewness (or the slope of the envelope) may be thought to be important for the discrimination, but also in this case there is no indication that the addition of non-informative segment makes the envelope less skewed. What may be observed however, is that potential onset and offset cues are more dominant in the 3- and 5-segment conditions, and almost non-existing in the 21-segment condition. This difference

is most prominent when the segment duration is relatively long compared to the time coefficient of the temporal window. Compare for example the onsets and offsets of the envelopes in Figure 11 and 13 at $T/\tau = 2$. If onset and offset cues are crucial for the discrimination it may be observed that for example the 9-segment condition gets easier when the segment duration gets so short that the "target pattern" begins to interfere with the onset. This explanation is intriguing since it makes it possible to explain why listener MK was able to improve his performance for shorter segment durations in the 9-segment condition. The fact that MK was the only listener for whom this effect was pronounced indicates that he was more sensitive to onset/offset cues. In the 5-segment condition it also appears that the "target pattern" interferes with the onset to a large extent, which would explain why listeners performed surprisingly well in this condition. It should be noted that temporal smoothing is not a prerequisite for the onset/offset hypothesis.

Interestingly, as it is shown in Appendix E, the addition of non-informative noise has no influence on the difference signal obtained when subtracting the descending stimuli from the ascending. This is true both before and after the temporal smoothing, and may be an indication that the temporal window model does not capture the essential cognitive processes, which are heavily influenced by the added noise.

B. Comparison to temporal order judgments of tones

Divenyi and Hirsh (1974) already reported on a similar experiment where listeners had to judge the temporal order of tones. Their study included ascending and descending (in frequency) 3-tone patterns, whose results may be compared to the 3-segment condition of the present experiment. They found that well-trained listeners were able to identify the temporal patterns down to a limit of approximately 2 ms. This compares relatively well to the limits found in this experiment for the 3-segment sequences, though also naive listeners were able to perform the task at such a short segment duration.

C. Quality cues and temporal representation

In their review of neurophysiological experiments Näätänen and Winkler (1999) establish the concepts of pre-representational and representational forms of the percept, where only in the representational form the percept has been mapped to a temporal dimension in perception. It seems worthwhile to interpret the results of the current experiment within this framework.

Auditory "features" of sound, which are already part of the percept at pre-representational stages, are fundamental to their reasoning. They do not mention onset/offset cues as specific features, but since it appears that they may be important concepts in order to understand the results of the present work, it is argued in the following why they may reasonably be interpreted as "features", which emerge already at a low level of perception.

1. Onsets and offsets as perceptual features

First, ecologically it makes good sense that onset and offset transients are important "features" for the understanding of the origin of a sound. Secondly, based on a review of neurophysiological and psychophysical experiments Phillips *et al.* (2002) conclude that the onset of a sound has a particular elaborate representation in the sensory system, especially within the first few ms of the sound. Interestingly this is consistent with MK's increased performance in the 9-segment condition around 1 ms, where the target pattern would enter the critical onset time-frame. Thirdly, further evidence is for example given by Stellmack *et al.* (2005), who show how the auditory system is especially sensitive to level changes at the onset and offset of a sound. They showed this in an experiment where people were asked to detect level increments and decrements of a 5 ms probe at different positions in a sound with an overall duration of 50 ms.

If it is assumed that onsets and offsets are features or "qualities" of a sound, this means that they themselves only indirectly carry information of the fine details of the temporal structure, because only at later perceptual stages are features mapped to a temporal dimension. For the 21-segment condition it was illustrated (see Figure 13) that there is almost no difference in onset and offset cues for ascending vs. descending stimuli, which means that the listener cannot rely on qualitative differences to the same extent as in the other tasks. Rather the listener must single out the "target pattern", which is only possible when the percept has been mapped to the temporal dimension of perception at a relatively high level of cognition. Thus the results in this condition may indicate the precision ("temporal resolution") with which sound is mapped to the temporal dimension. Note that this type of "temporal resolution" is most likely not in the sense of a sliding "temporal window", but may be understood using concepts such as uncertainty, similarity, or attention Durlach *et al.* (2003). If, on the other hand, there is a difference in quality (onsets/offsets) between ascending and descending patterns, then the listener can discriminate based on features alone. If, as it was suggested, this is the case in the 3- and 5-segment conditions, this would explain the good performance of the listeners in these conditions.

V. CONCLUSION

The addition of non-informative noise segments influenced the performance of listeners to an extent not readily predicted by a temporal window model. Rather, the influence could be better understood when different cog-

nitive stages of perception were assumed to influence the performance in different conditions. If onset and offset cues were considered "features" of the percept, then Näätänen and Winkler (1999) provide a good framework for understanding the results. The 21- and 9-segment conditions required awareness of the temporal properties of the patterns to a larger extent as did the 3- and 5-segment conditions where listeners could rely on qualitative onset/offset differences. The relatively poor performance in the 9- and 21-segment conditions may be understood in similar terms as informational masking (Durlach *et al.*, 2003). In experiments involving informational masking large individual differences in performance are often observed (Durlach *et al.*, 2003), which was also the case in the present experiment.

A. Temporal masking and gap detection

Within the explanatory framework outlined in this paper, effects observed in temporal masking and gap detection experiments, reflect the temporal limits of a given "feature"- or "quality"-extraction process. Therefore temporal limits found in such experiments do not reflect "absolute" limits of the auditory processing. Since the number of decoded "qualities" are possible many, it cannot be ruled out that another "feature" may depend on even finer temporal details of the sound. This, of course, can only be detected if a presented sound possesses the relevant "feature". The accompanying paper (Pedersen, 2006) demonstrates that a repeated version of the stimuli used in the present experiment can be discriminated at an extremely fine resolution, indicating the emergence of a new "feature" via the repetition. A temporal window model based on, in this context, relatively long time coefficient is at odds with the very fine resolution with which listeners can perform the discrimination of patterns.

If auditory features play an important role, as suggested by the present results, future research may be concerned with identifying specific features, both via psychoacoustical or physiological experiments. The critical parameters for "feature" extraction need not only be of temporal character, but could also be of spectral or of a more complex nature (e.g. modulation or informational masking). Stimuli may be varied across all of these dimensions as to identify the critical range (temporal or spectral limits for example) for the operation of specific "feature extractors".

VI. ACKNOWLEDGMENTS

The author would like to thank the people of Dr. Neal Viemeister's research group at the Psychoacoustics Laboratory at the Department of Psychology, University of Minnesota. The experiment presented in this paper was performed during a research visit at the laboratory. The author is greatful for their valuable advice and feedback and for providing most of the experimental equipment.

This research was carried out as part of the "Centercontract on Sound Quality" which establishes participation in and funding of the "Sound Quality Research Unit" (SQRU) at Aalborg University. The participating companies are Bang & Olufsen, Brüel & Kjær, and DELTA Acoustics & Vibration. Further financial support comes from the Ministry for Science, Technology, and Development (VTU), and from the Danish Research Council for Technology and Production (FTP).

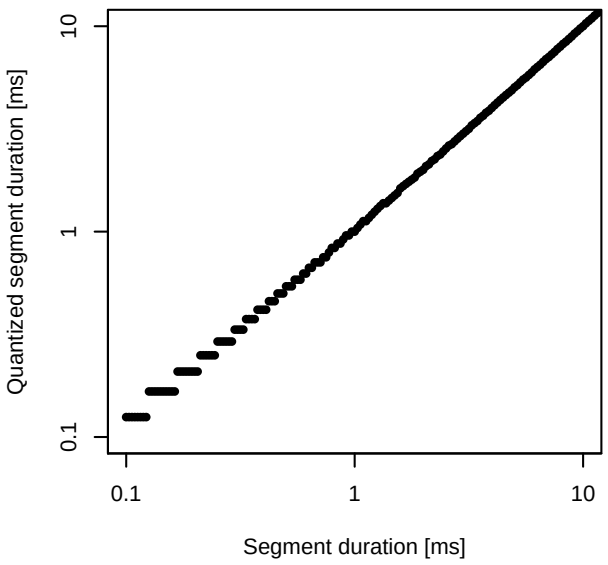


Figure 7. Quantized segment duration (y-axis) plotted against theoretic segment duration used in one track of trials.

Appendix A: QUANTIZATION OF SEGMENT DURATIONS

Because of discrete sampling in the signal generation and playback, actual segment durations of the stimuli played back was also discrete. That is, not all segment durations required in one track of trials could be realized. The procedure used to generate the noise carrier caused the segment duration to be a multiple of four samples. The sampling rate used was 96 kHz, so for the very shortest segment durations the discretization becomes evident. This is illustrated in Figure 7, where the discrete segment duration of all trials in one fixed track is plotted against the segment duration without discretization. All results presented in figures and in the table of this paper are based on the segment durations of the stimuli the listeners heard, that is, the quantized durations.

Appendix B: SPECTRUM OF NOISE CARRIER

The amplitude spectrum of excerpts of the noise carrier, as depicted in Figure 2 for example, is random. However it is possible to derive the long term spectrum in a similar way as it is possible to derive the spectrum of white noise. The power spectrum can be calculated by Fourier transform of the autocorrelation sequence of an infinitely long noise sequence, when the mean of the noise is zero (Oppenheim *et al.*, 1999). So the first step is to derive the autocorrelation sequence, which for a sequence, x, of real numbers, is given by:

$$\phi_{xx}[m] = E\{x_{n+m}x_n\} \tag{B1}$$

Here E denotes the expectation value and ϕ_{xx} the value of the autocorrelation sequence at position m. If now the positive and negative sample values depicted in Figure 2 are taken to be $-A$ and A, then ϕ_{xx} can be calculated for different m. For arbitrary n, and $m = 0$, x_n is either $-A$ or A so:

$$\phi_{xx}[0] = E\{x_n x_n\} = A^2 \tag{B2}$$

For $m = 1$ there are two possibilities: x_{n+1} and x_n are in the same block (see Figure 2). If this is the case their product is always $-A \cdot A = -A^2$. If they are not in the same block, they are uncorrelated and the expectation value of their product is zero, because the product will be $A{\cdot}A$, $-A{\cdot}(-A)$, $A{\cdot}(-A)$ and $-A{\cdot}A$ with equal frequency. In summary the expectation value will be $-A^2$ for half of the sequence and zero for the other half. This gives:

$$\phi_{xx}[1] = E\{x_{n+1}x_n\} = -\frac{1}{2}A^2 \tag{B3}$$

The very same arguments can be used for $m = -1$ and:

$$\phi_{xx}[-1] = E\{x_{n-1}x_n\} = -\frac{1}{2}A^2 \tag{B4}$$

For any other m, x_{n+m} and x_n will be in different blocks and thus uncorrelated. Therefore the expectation value of their product is zero:

$$\phi_{xx}[m] = E\{x_{n-m}x_n\} = 0 \qquad \text{for } m < -1 \text{ or } m > 1 \tag{B5}$$

Now the power spectrum of the sequence x can be found by Fourier transform of ϕ_{xx}:

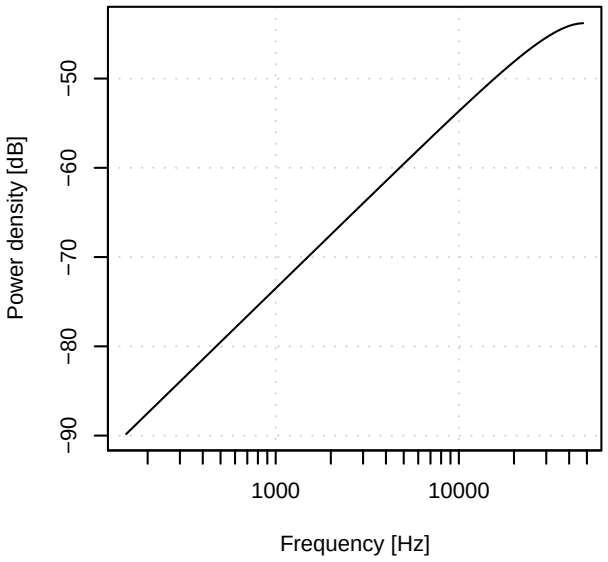


Figure 8. Power density spectrum of the noise carrier (see Figure 2 and equation).

$$\begin{aligned}
P_x(\omega) &\propto \sum_{m=-\infty}^{\infty} \phi_{xx}[m]e^{-j\omega m} \\
&\propto (\phi_{xx}[-1]e^{j\omega} + \phi_{xx}[0] + \phi_{xx}[1]e^{-j\omega}) \\
&\propto -\frac{1}{2}A^2(e^{j\omega} + e^{-j\omega}) + A^2 \\
&\propto -\frac{1}{2} \cdot 2A^2\cos(\omega) + A^2 \\
&\propto A^2(1 - cos(\omega))
\end{aligned}$$

Here P_x is proportional to the power density spectrum, where the normalized frequency, ω, can be related to a frequency in Hz given the sampling rate used in the experiment (f_s =96 kHz), via the Nyquist-Shannon sampling theorem and the relation $\omega = 2\pi f/f_s$. The RMS power of the noise in the full frequency range is 1 for $A = 1$, because the noise samples are always $-A$ or $+A$. The power density should be normalized, so the integral over the full frequency range is the overall RMS value, that is, the overall integral should equal 1 when $A = 1$. This is obtained in the following expression where f has been substituted for ω and the single sided spectrum is obtained by a multiplication by 2:

$$P_x(f) = \frac{2}{f_s}A^2(1 - cos(2\pi f/f_s) \tag{B6}$$

Appendix C: SOUND LEVEL AT EAR

The transfer function from the headphones to the right ear of the dummy-head was measured, and the result is depicted in Figure 9 as the dotted curve. It was calibrated to have unity gain at 1 kHz (0 dB), but has been offset in the figure as to intersect the noise power density spectrum (dashed curve) at 1 kHz. The solid curve is the noise power density spectrum weighted by the transfer function to the ear. The noise spectrum is plotted for the amplitude $A = 1$. Since the samples of the noise carrier are either A or $-A$, the RMS value of the noise sequence is A^2, which for $A = 1$ is 1 (= 0 dB). The RMS value of the noise can also be calculated by integrating the noise power density spectrum over the relevant frequency range. In this way the RMS value of the weighted noise can be estimated by numerical integration. The results of the numerical integration over the range 0.4 to 20 kHz was 0.054 (= -13 dB). Below 0.4 kHz the noise has only insignificant power, and above 20 kHz the power of the weighted noise is also insignificant due to the attenuation of the headphone-to-ear transfer function (hearing is of course not very sensitive to frequencies above 20 kHz either, but this has no influence on the RMS value). In summary, at $A = 1$, the unweighted RMS was 0 dB, and weighting by the headphone-to-ear transfer function reduced this to -13 dB. Consequently, 13 dB should be subtracted from the calibrated levels used throughout this paper to account for the transfer function of the headphones to the ear.

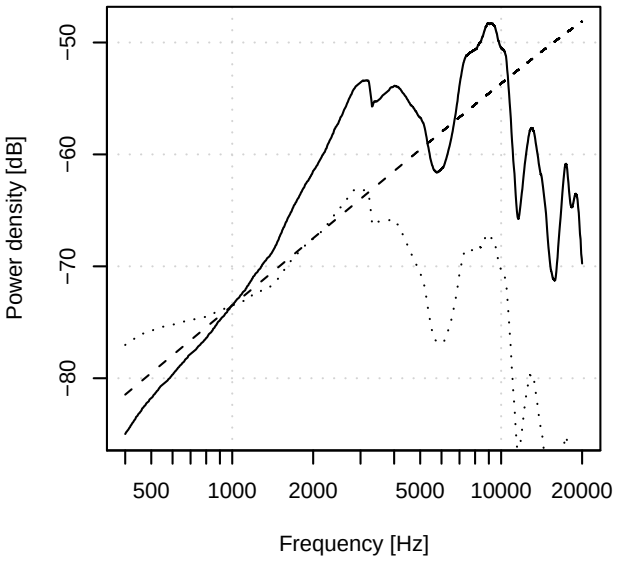


Figure 9. Weighted noise power density spectrum. The unweighted noise spectrum (dotted line, as also depicted in Figure 8) is weighted by the measured transfer function from the headphones to the right ear of the dummy-head (dashed curve), which, in the figure, has been offset from unity gain (0 dB) at 1 kHz, as to intercept the noise spectrum at 1 kHz. The weighted power spectrum is the estimated power density spectrum at the ear (solid curve).

Appendix D: EXAMPLES OF OUTPUT OF SLIDING WINDOW

In Figure 10 to 13 examples are given of the output of a temporal window convolved with the envelope of the stimuli used in different conditions of the experiment and at different values of the segment duration. For a description of the model see for example (Oxenham and Moore, 1994). The temporal window used in this paper is given by:

$$w(t) = \begin{cases} Ae^{t/\tau_b} & \text{t} < 0 \\ Ae^{-t/\tau_f} & \text{t} \geq 0 \end{cases}$$

The symbols τ_f and τ_b are time coefficients for forward and backward masking respectively. A symbolic description of the window convolved with the stimuli of the experiment was derived, and the results are graphically illustrated in the figures. Convolved stimuli is plotted as a function the segment duration divided by τ, where $\tau_f = \tau$ and $\tau_b = 2\tau$ were chosen to illustrate the effect of asymmetry of the temporal window. The envelope of the stimuli before the temporal window is described by the RMS sound pressure levels of the segments. As it is generally true in linear signal analysis the shape of the convolved response reduces to the shape of the impulse response (here the temporal window) when the input signal gets sufficiently short. Different shapes of the temporal window have been proposed by other researchers (see Plack *et al.*, 2002, for example), but since the analysis in this chapter is based on visual inspection it only makes sense to consider different temporal windows to the extent that they look different. If they look very similar also their output will look very similar. The variation in the temporal window proposed by other researchers, will in most cases not give rise to fundamentally different looking graphs of the output of the temporal window. It may be argued that level compression in the sensory system influences the process. The consequence of compression is basically a change in the relative differences between the three segment levels used to describe the envelope patterns. On the output side of the temporal window very much the same compression can be observed, which means that the same fluctuations exists in the output, but settles at different relative levels as compared to the non-compressed case.

Appendix E: IDENTICAL DIFFERENCE BETWEEN ASCENDING AND DESCENDING PATTERNS ACROSS 3-, 5-, 9-, AND 21-SEGMENT CONDITIONS

The convolution with a temporal window has the mathematical property of being distributive, which

means that the difference between ascending and descending stimuli can be calculated in two ways: (1) The two convolved signals for ascending and descending stimuli can be subtracted or, (2) the ascending and descending stimuli can be subtracted prior to convolution and afterward convolved with the temporal window. It is possible to decomposed the used stimuli as illustrated in Figure 14. Using the stimuli components of the figure (x_n and x_t), ascending stimuli ($x_{ascending}$) can be described by:

$$x_{ascending}(t) = x_n(-t) + x_t(t) + x_n(t)$$

And descending:

$$x_{descending}(t) = x_n(-t) + x_t(-t) + x_n(t)$$

Subtracting the ascending and descending signals gives the difference:

$$x_{ascending}(t) - x_{descending}(t) = x_t(t) - x_t(-t)$$

This shows that the difference between ascending and descending stimuli is independent of the added noise (x_n), both prior to or after convolution with a temporal window. The linear analysis is valid for all signals, which can be decomposed as illustrated in Figure 14. This is true both when the envelope or the fine-structure of the signals of the experiment are considered, and further it is also valid in the case where a simple non-linear transform (as a function of signal level only) has been applied earlier in the process as to model compression for example.

Divenyi, P. L. and Hirsh, I. J. (**1974**). "Identification of temporal order in three-tone sequences.", J. Acoust. Soc. Am. **56**, 144–151.

Durlach, N. I., Mason, C. R., Kidd, G., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (**2003**). "Note on informational masking.", J. Acoust. Soc. Am. **113**, 2984–2987.

Farmer, M. E. and Klein, R. M. (**1995**). "The evidence for a temporal processing deficit linked to dyslexia: A review", Psychonomic bulletin & review **2**, 460–493.

Moore, B. C. J. (**2003**a). *An Introduction to the Psychology of Hearing*, 5th edition (Academic Press, San Diego, CA).

Moore, B. C. J. (**2003**b). "Temporal integration and context effects in hearing", Journal of Phonetics **31**, 563–574.

Näätänen, R. and Winkler, I. (**1999**). "The concept of auditory stimulus representation in cognitive neuroscience.", Psychol. Bull. **125**, 826–859.

Oppenheim, A. V., Schafer, R. W., and Buck, J. R. (**1999**). *Discrete-time signal processing*, 2nd edition (Prentice Hall, Englewood Cliffs, NJ).

Oxenham, A. J. and Moore, B. C. (**1994**). "Modeling the additivity of nonsimultaneous masking.", Hear. Res. **80**, 105–118.

Pedersen, B. (**2006**). "Discrimination of temporal patterns on the basis of envelope and fine-structure cues", in *Auditory Temporal Resolution and Integration: Stages of Analyzing Time-Varying Sounds*, 85–96 (Aalborg University).

Phillips, D. P., Hall, S. E., and Boehnke, S. E. (**2002**). "Central auditory onset responses, and temporal asymmetries in auditory perception.", Hear. Res. **167**, 192–205.

Plack, C. J., Oxenham, A. J., and Drga, V. (**2002**). "Linear and nonlinear processes in temporal masking", Acust. Acta Acust. **88**, 348–358.

Schäffler, T., Sonntag, J., Hartnegg, K., and Fischer, B. (**2004**). "The effect of practice on low-level auditory discrimination, phonological skills, and spelling in dyslexia.", Dyslexia **10**, 119–130.

Stellmack, M. A., Viemeister, N. F., and Byrne, A. J. (**2005**). "Comparing monaural and interaural temporal windows: effects of a temporal fringe on sensitivity to intensity differences.", J. Acoust. Soc. Am. **118**, 3218–3228.

Studdert-Kennedy, M. and Mody, M. (**1995**). "Auditory temporal perception deficits in the reading-impaired: A critical review of the evidence", Psychonomic Bulletin & Review **2**, 508–514.

Viemeister, N. F., Rickert, M., Law, M., and Stellmack, M. (**2002**). "Psychophysical and physiological aspects of auditory temporal processing", in *Genetics and the Function of the Auditory System, Proceedings of the 19th Danavox Symposium, edited by L. Tranebjaerg, J. Christensen-Dalsgaard, T. Andersen, and T. Poulsen*, 273–291 (Holmens Trykkeri, Denmark).

Viemeister, N. F. and Wakefield, G. H. (**1991**). "Temporal integration and multiple looks.", J. Acoust. Soc. Am. **90**, 858–865.

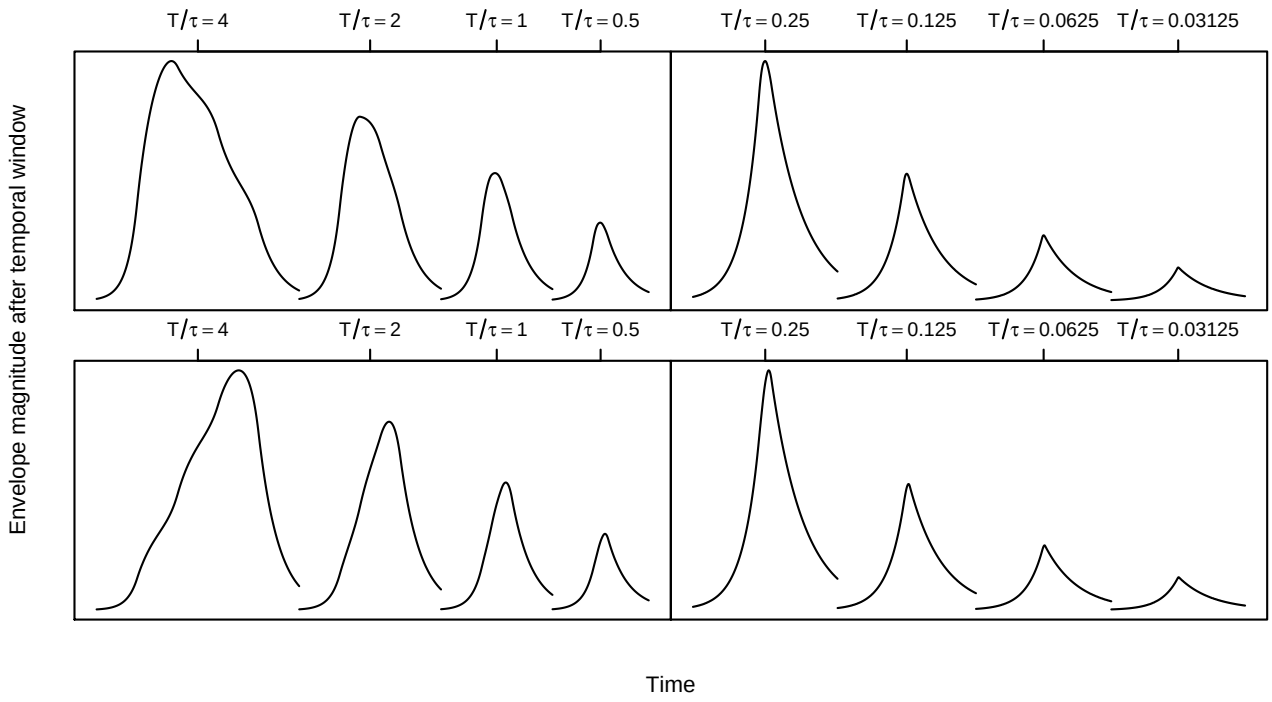


Figure 10. Temporal window convolved with 3-segment stimulus envelope. Descending pattern in top row, and ascending in the bottom row. From left to right the figure shows how decreasing the segment duration changes the output of the temporal window. The ratio of segment duration, T, to the time coefficient of the temporal window, τ, is indicated at the top of each figure. A different scaling of the y-axis is used in the two columns. x-axis shows time, scaled to accommodate all patterns within the same frame.

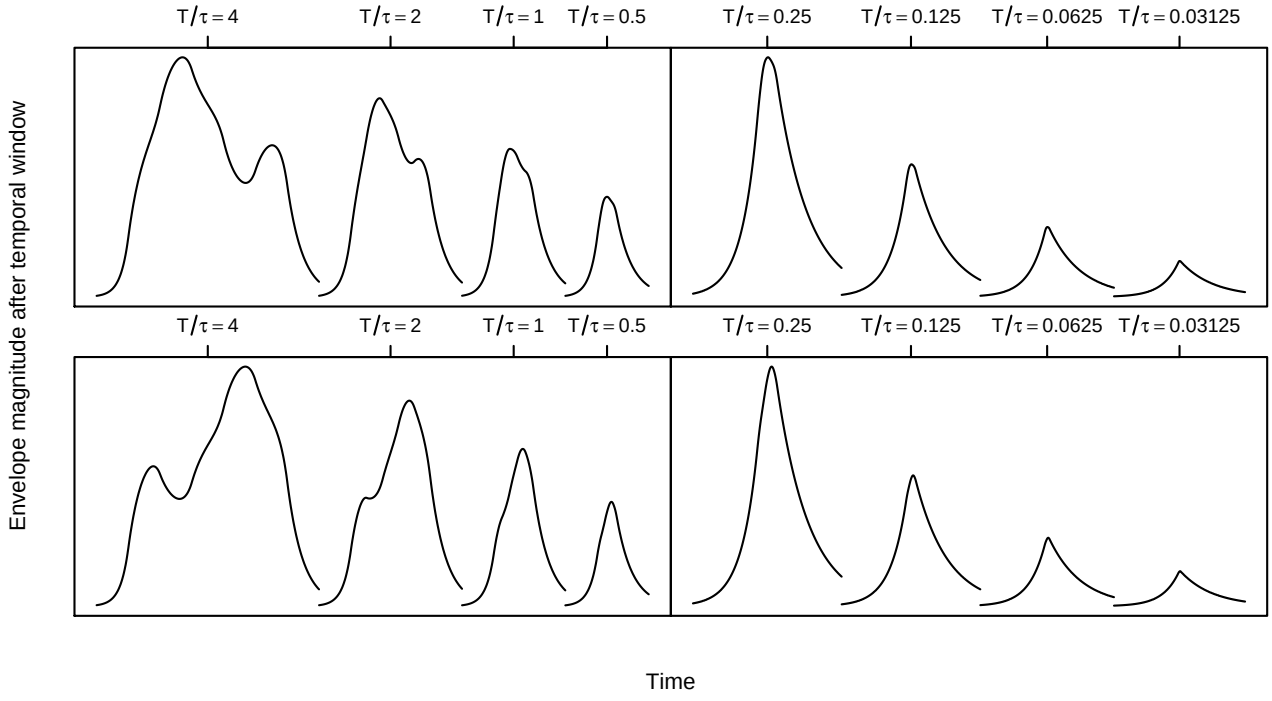


Figure 11. Same as Figure 10, but showing the temporal window convolved with 5-segment stimulus envelope.

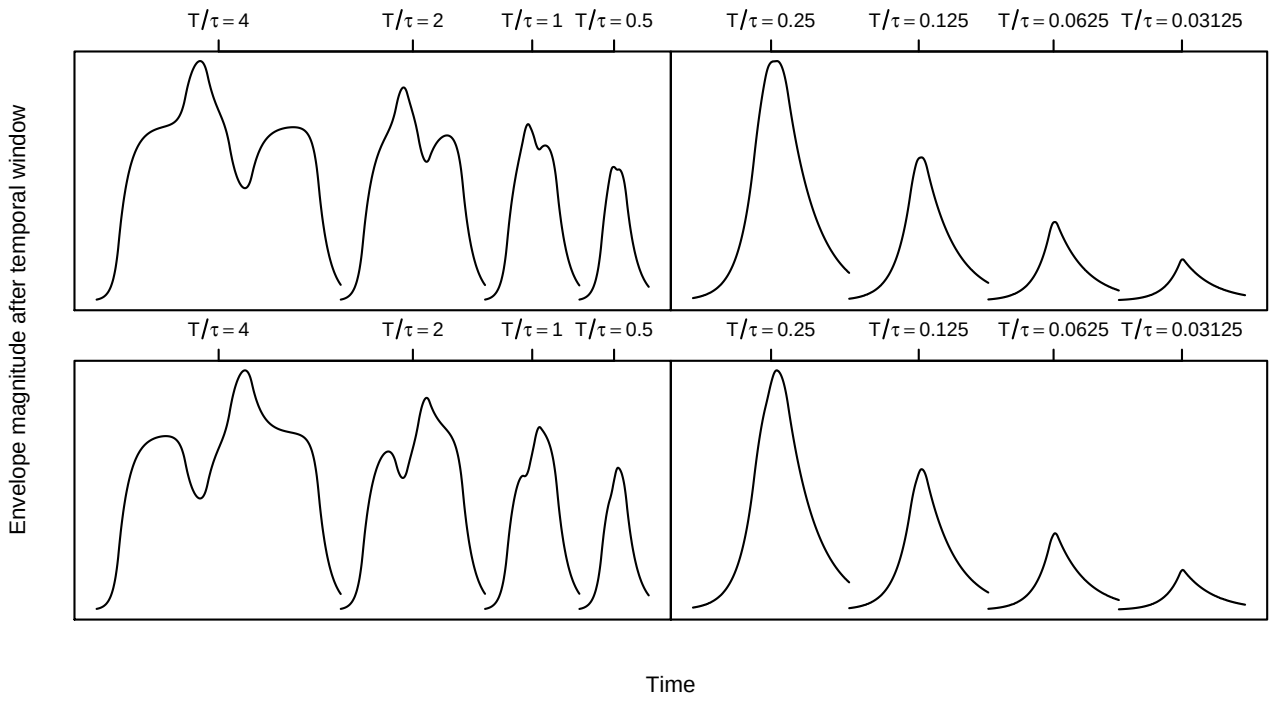


Figure 12. Same as Figure 10, but showing the temporal window convolved with 9-segment stimulus envelope.

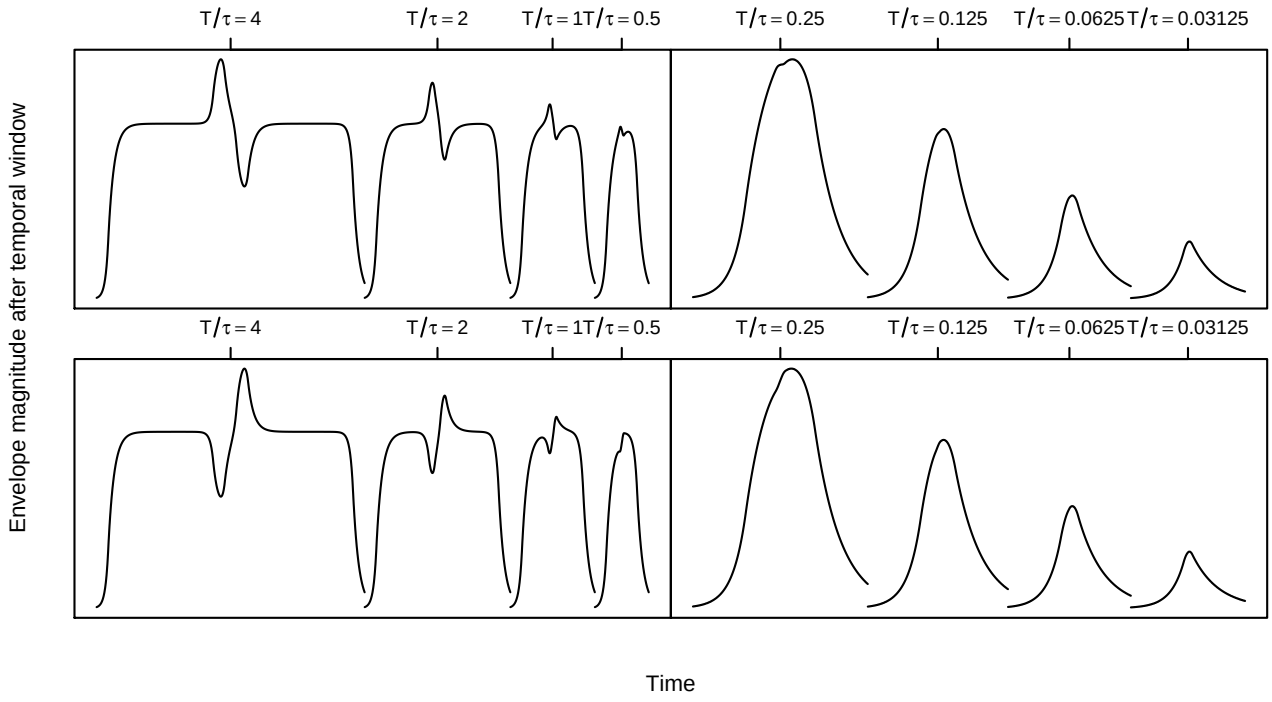


Figure 13. Same as Figure 10, but showing the temporal window convolved with 21-segment stimulus envelope.

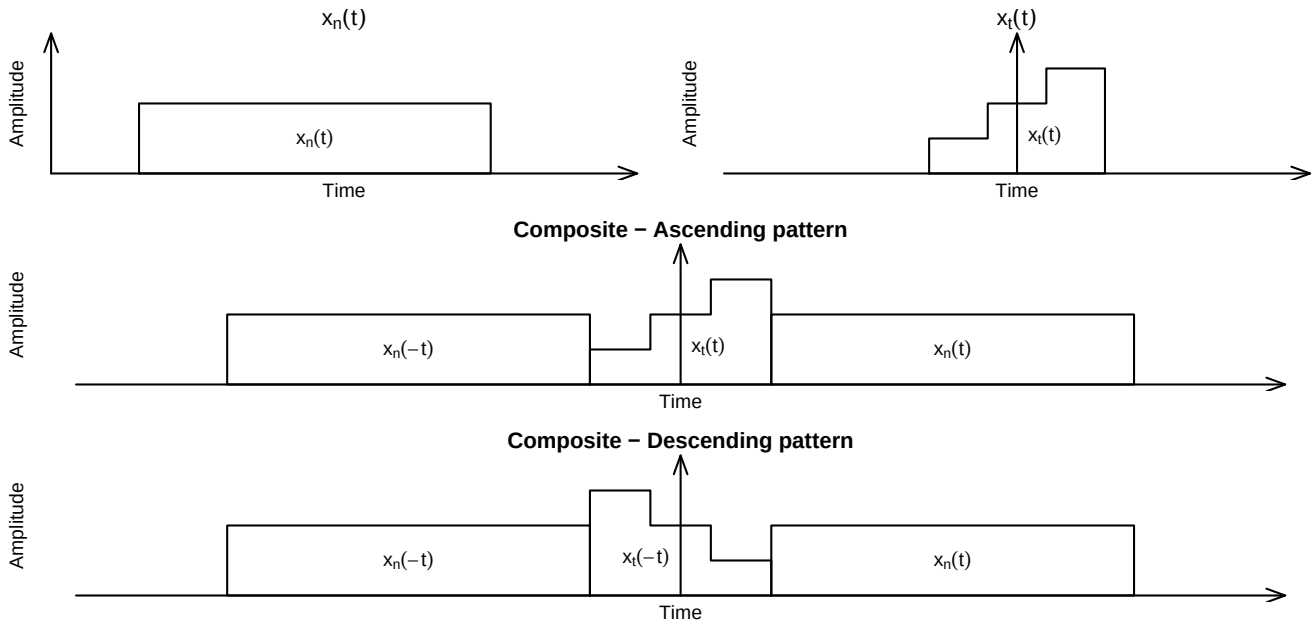


Figure 14. Decomposition of stimuli. All stimuli are based on the signal components x_n and x_t in the top row. Examples of descending and ascending patterns composed from x_n and x_t are shown in the two lower rows. The arrows which mark the y-axes always intersect the time axis at $t = 0$.

Chapter 5

Paper 3:
Discrimination of temporal patterns on the basis of envelope and fine-structure cues

Discrimination of temporal patterns on the basis of envelope and fine-structure cues

Benjamin Pedersen*

Sound Quality Research Unit (SQRU), Department of Acoustics, Aalborg University, Fredrik Bajers Vej 7-B5, 9220 Aalborg Øst, Denmark

(Dated: September 12, 2006)

The importance of envelope vs. fine-structure cues was examined in tasks in which listeners had to discriminate temporally ascending or descending envelope patterns. Decreasing the duration of the patterns made the tasks harder and continuous psychometric functions were obtained as a function of the duration of a single pattern. In two further conditions, a fixed envelope pattern was repeated within a 0.75-s time window, so that, in one condition only the envelope of the pattern was repeated, and in the other, the fine-structure of a single pattern. The temporal limit at which ascending and descending patterns were still discriminable was found to be a duration in the range of 1 ms no matter whether only single patterns were presented or whether the envelope of a pattern was repeated. Fundamentally different performance was observed when the fine-structure was repeated, in which case some listeners were able to discriminate patterns at the finest temporal resolution that could be physically realized (20 μs). Further, noise was added prior to and succeeding the patterns to be identified, which severely degraded performance when the envelope was repeated or if the pattern was not repeated. When fine-structure was repeated with added noise, performance generally increased for the shortest durations of the pattern. The notion of "energetic masking" appears incompatible with this observation, but rather, the fine-structure seems to carry important cues for the discrimination.

I. INTRODUCTION

In an accompanying paper (Pedersen, 2006) it was discussed how different stages of perception may be involved in decoding the temporal information carried in sound. It was demonstrated how the addition of non-informative noise segments influenced listeners' performance in a task where they were asked to identify ascending and descending patterns, to an extent not readily explained by a temporal window model. In accordance with the framework set up by Näätänen and Winkler (1999) it was suggested that pre-representational and representational stages of perception might be the key in understanding the results, the main reasoning being: At an early pre-representational stage the sensory system detects "features" of sounds, which are only later, at the representational stage of perception, mapped to a temporal dimension. Only when the representational stage is reached are the features consciously available. For different stimuli used in the previous experiment it varied to which extent stimuli could be identified by specific features (onset/offset cues) or whether the auditory task required temporal information about the envelope of the stimuli at a conscious level.

The main goal of the experiment described in this paper, is to explore how stimuli with similar temporal properties as that used in the earlier experiment may be manipulated so "static" features, like pitch or timbre, emerge. Such features may depend on temporal details much finer than the temporal limits found in the previous experiment.

*Electronic address: bp@acoustics.aau.dk

A. Repeating pattern - importance of envelope and fine-structure

In the present experiment, the main idea was to let ascending and descending 3- and 21-segment patterns repeat themselves (see Figure 1 versus Figure 2). In two conditions (repetition of 3- and 21-segment patterns) fine-structure was preserved in each repetition and only the envelope of the pattern in two other conditions. This variation was considered especially fit to probe several ideas introduced in the literature to understand perception of time-variance. Viemeister and Wakefield (1991) hypothesized that listeners may integrate their perception over multiple temporal looks. In this framework the proposed stimuli potentially give insight into whether such looks are based on the envelope or the fine-structure of a sound. In the case of only the envelope being important, identical performance would be expected from the listeners for the two types of repetitions (envelope only or fine-structure). Generally, identical performance is predicted by any model, where only the envelope of a sound is assumed to carry temporal information. In contrast, if a difference in performance is observed, it can be concluded that temporal information at a finer level of detail than the envelope is represented in the sensory system. Such an outcome would be in conflict with for example the temporal window model (see for example Moore, 2003).

B. Static features

The difference between the two versions of the repeating stimuli (envelope only or fine-structure), in terms of perception, becomes most evident for high repetition

rates of the pattern, where the repetition of the fine-structure is clearly different in quality (pitch- or timbre-like features) from the repetition of envelope only. This suggests that static features may emerge, making the stimuli discriminable based on pre-representational temporal details.

C. Temporal processing in pitch and timbre perception

In the literature there has been some controversy as to what extent pitch depends on the amplitude spectrum of a sound only, or if temporal coding plays a major role (see for example Moore, 2003, for a review). This is interesting in that it makes several suggestions for temporal processing mechanisms in the sensory system, where both spectral and temporal pattern matching are involved, and is relevant for the interpretation of the results of present experiment. In an experiment Patterson (1994a) presented repeating ramped and damped sinusoids, which bears some resemblance with the condition of the present experiment where the fine-structure of the pattern is repeated. He found that there is a clear difference in the perceived quality depending on the direction of the ramping, the ramped version having a much stronger sinusoidal quality. Since ramped and damped stimuli are spectrally identical, this demonstrates that temporal processing plays an important role. Patterson developed a model of the temporal processing, which has more recently been used by Krumbholz *et al.* (2005) in an attempt to predict listeners' performance in an experiment where they had to detect regularity in a pulse train, and they found that temporal information was combined across frequency bands.

Using wide band-stimuli, the current experiment probed hearing where supposedly different sensory processes are crucial. This was done by varying the temporal extent of patterns over a wide time range. The main goal was to determine whether auditory temporal processing is restricted to analysis of the envelope of sounds only. If this is the case, no difference in the performance of listeners was expected in tasks where fine-structure or envelope was repeated. Contrary to this, the results revealed that when the fine-structure was repeated, the auditory system could rely on extremely fine temporal details, but not when only the envelope was repeated.

II. METHOD

A. Listeners

6 listeners participated in the experiment; 2 males and 4 females with a mean age of 24.9 years (range: 21 to 30). Prior to the experiment the listeners' hearing was screened and nobody had significant hearing loss ($>$ 15 dB HL at more than one frequency at 0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4, 6, and 8 kHz). Except for the author (BP) all listeners were students with little or no experience in listening experiments.

B. Apparatus

A computer program controlling the experiment also generated the stimuli. A sound card (RME HDSP9632) was used for digital to analog conversion at a sample rate of 192 kHz. The high sample rate made short temporal blocks possible. The converter had a resolution of 24 bits though only 7 different sample levels were used in the generated digital signal (see Pedersen, 2006). A power amplifier (Rotel RB-976MkII) modified to have a fixed gain was used. The amplifier's amplitude response covered the range from 2 Hz to 90 kHz. The sounds were presented via headphones (Beyerdynamic DT990Pro). The data sheet of the headphones specified a frequency range of 5 Hz to 35 kHz, so the bandwidth of the headphones determined the bandwidth of the full system.

The listeners were seated in a double-walled listening cabin, and gave their responses via a custom made button box with lights providing feedback.

C. Stimuli

The listening experiment contained 6 different conditions which will be referred to using the following abbreviations, which are more thoroughly described in the next two chapters:

- 3NO: 3-segment pattern with no repetitions
- 21NO: 21-segment pattern with no repetitions
- 3EN: 3-segment pattern with repetitions - the envelope is repeating *not* the fine-structure
- 21EN: 21-segment pattern with repetitions - the envelope is repeating *not* the fine-structure
- 3EX: 3-segment pattern with repetitions - the exact same fine-structure of one pattern is repeating
- 21EX: 21-segment pattern with repetitions - the exact same fine-structure of one pattern is repeating

The task for the listener was always to discriminate ascending from descending patterns. The forms of the basic patterns without repetitions (NO conditions) are illustrated in Figure 1.

1. Noise carrier

A special designed noise carrier was modulated with the envelope patterns of the different conditions. It had a broad-band characteristic, but not the properties of white noise. To allow for rapid level changes, it was designed with several temporal constraints, which are thoroughly

explained by Pedersen (2006). Further, the noise carrier was always mirror-symmetric within a single repetition, and was thus unaffected be time-reversal of the pattern, which was used to generate corresponding ascending and descending patterns.

2. Repetitions of pattern

In the conditions where the pattern was repeating itself, the form of one repetition was the same as in the non-repeating case. In the repeating conditions, the pattern kept repeating within a fixed time frame of 0.75 s. This is illustrated in Figure 2. As illustrated, the pattern was always repeated in a cycle with the duration of 25 segments. This means that the repetition rate changed when the segment duration changed. The 25-segment repetition cycle was used for patterns of both 3 and 21 segments length. Especially in the EX condition the stimuli had a pronounced pitch for short segment durations. The fundamental frequency of the pitch is determined by the segment duration and the repetition cycle, and thus repeated 3- and 21-segment patterns have the same pitch at the same segment duration, because of the fixed 25-segment cycle. The total duration of one sound was not allowed to exceed the 0.75 s duration and no fractional patterns were allowed, so the duration of the sound was sometimes shorter than 0.75 s, most pronounced for the long segment durations. However, for the very longest segment duration starting at 100 ms, a single repetition of 21 segments, exceeded the 0.75 s time-frame. Therefore the duration of a sound was allowed to exceed the 0.75 s limit in the case of only one single repetition of the pattern. Because of these constraints, the 3-segment pattern only started repeating at a segment duration of 27 ms and the 21-segment pattern at 16 ms, so above these segment durations all conditions, repeating and non-repeating, contained identical stimuli.

In the case of non-repeating stimuli it was shown by Pedersen (2006) that across 3- and 21-segment conditions, the difference between ascending and descending stimuli is identical (subtracting a descending 3-segment pattern from a 3-segment ascending gives the same result as subtracting a descending 21-segment pattern from a 21-segment ascending. That is: The noise segments cancels out and have no influence on the difference). When there is no difference for the non-repeated patterns there is of course no difference either when the patterns are repeating. All possible cues for discrimination are contained in the difference signal, so mathematically the possible cues for discrimination are independent of added noise segments.

The difference between repeating the envelope (EN) and repeating fine-structure (EX) is illustrated in Figure 3. The noise carrier has the same properties as in the non-repeating conditions, which means that the carrier is mirror-symmetric within each repetition. Therefore the entire sound is also mirror-symmetric in the EX condition, but *not* in the EN conditions (except when the sound contains only a single repetition). Hence ascending and descending EN patterns do not have the exact same spectrum, only within single repetitions.

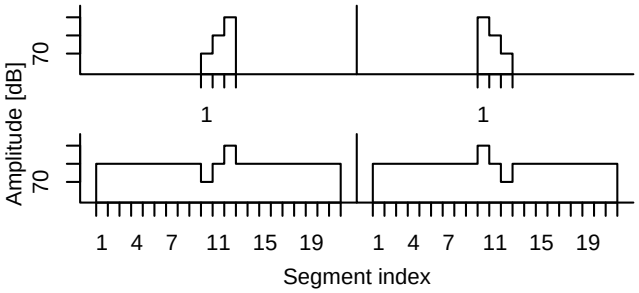


Figure 1. Envelope of stimuli in the conditions without repetitions (NO conditions). Stimuli in the 3-segment condition in the top row and stimuli in the 21-segment condition in the bottom row. The two columns illustrate ascending and descending patterns respectively.

3. Calibration

The setup was calibrated relative to a setup with the headphones positioned on an artificial ear (Brüel & Kjær 4153), with microphone (Brüel & Kjær 4134). The level of a sine-tone with a frequency of 1 kHz was measured in the artificial ear. In this paper all sound pressure levels in dB refer to what would have been the level of a 1 kHz sine tone at the applied gain settings. The actual sound pressure level at the ears will be lower, because the noise contains more power at higher frequencies where the headphone transfer function is highly attenuated. Therefore the (unweighted) level of a continuous noise carrier was also measured, and it was found to be approximately 15 dB lower than the 1 kHz sine-tone at the same gain settings of the system.

As shown in Figure 1 and Figure 2, the calibrated level of the non-informative noise was always 70 dB SPL and the level of segments of the target patter was either 60, 70, 80 dB SPL (ascending pattern) or 80, 70, 60 dB SPL (descending pattern). These levels are lower (by 10 dB) than the levels used in the parallel study described by Pedersen (2006). This was to reduce the annoying sensation introduced by the multiple perceived onsets at longer segment durations in the conditions with repeating pattern.

D. Experimental procedure

In all conditions a three-interval, two-alternative forced choice paradigm was adopted. In one trial the listener heard three sounds, of which either the second or third contained a descending pattern and the two others con-

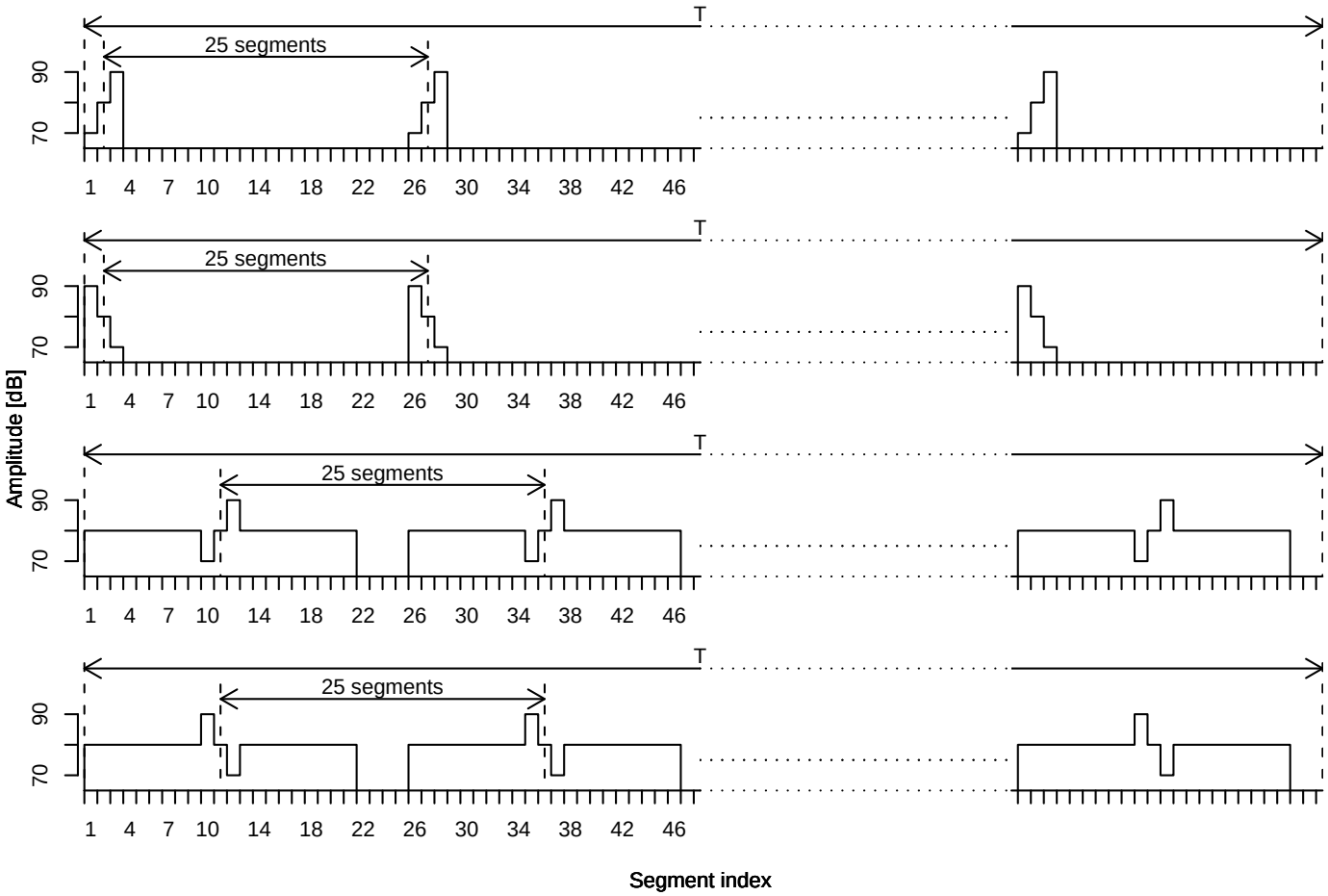


Figure 2. Envelopes of stimuli in the repeating pattern conditions (EN and EX). On the x-axis is the segment index and on the y-axis the sound pressure level. The two top rows show the ascending and descending pattern in the 3-segment conditions, and the two bottom rows in the 21-segment conditions. Each row depicts one single sound where T is the fixed time frame for one sound having a value of 0.75 s. The patterns are always repeated in a 25-segment cycle.

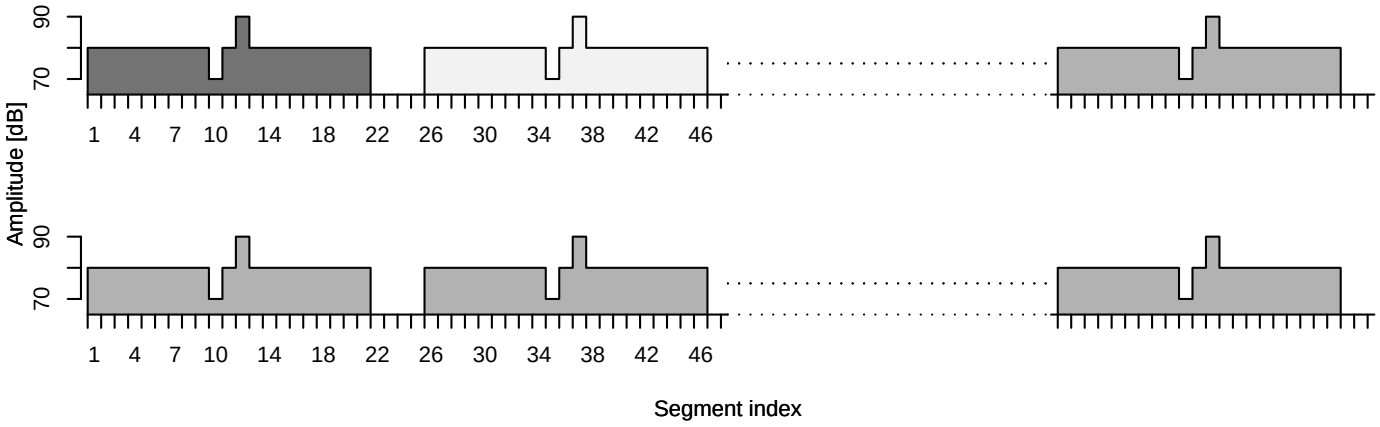


Figure 3. Illustration of repeating envelope (EN) vs. repeating fine-structure (EX) in the top and bottom row respectively. Only ascending 21-segment stimuli are shown. The different shades of gray indicate that a different noise carrier is used for each repetition. In the lower row the same shade of gray illustrates that the exact same noise carrier is used for each repetition. A sound with descending patterns is generated by mirroring each individual repetition, meaning that the order of "colors" from left to right is the same for de- and ascending stimuli.

tained an ascending one. The task of the listener was to identify which sound contained the descending pattern, so the performance at chance level was correct identification in half of the trials.

The use of an adaptive procedure was avoided to be able to detect non-monotonicities in the psychometric functions, and as to obtain "full" psychometric functions as a function of the segment duration. Also, the perceived quality of sounds of very different segment durations is dramatically different. Therefore it seemed natural to gradually reduce the segment duration throughout one track to avoid a drop in performance because the listener had to change his/her decision criterion from trial to trial. However, it cannot be ruled out that an effect of this way of varying the segment duration is reflected in the psychometric functions derived. Indeed it can be argued that the performance is worse in the beginning of each track because the listener has to become familiar with the task. Thus the performance, as a function of segment duration, would be underestimated for long segment durations, which always occurred in the beginning of a track. It could also be argued that the listener always looses attention toward the end of a track, leading to an underestimation of the performance as a function of segment duration. Further it could be argued that the performance is better for short segment duration because of long term learning effects. This is however unlikely, because in the time frame of the entire experiment, different segment durations are almost perfectly balanced.

The experiment was divided into tracks in which trials contained only sounds from the same condition. In one such track the segment duration was slowly decreased, starting from 100 ms and ending at 0.01 ms (conceptually speaking only, since the smallest possible segment duration was 0.02 ms due to the discrete sampling, see Appendix A). The segment duration was reduced in logarithmic steps with 40 steps per decade. Thus one track contained 161 trials.

In the NO conditions there was a silent period of 0.3 s between the three sounds of one trial, and in the repeating conditions the break was 0.1 s. In the NO condition a longer break was used, so the individual sounds were clearly separated also when the segment duration was short and each sound was "click" like. A shorter break was used in the repeating condition, because the impression from informal listening was that the shorter the break the easier the task was. Making the break too short allows for interaction cues of the two sounds, but still, 0.1 s is a long break compared to the temporal resolution eventually achieved. Further, the impression of the author is that the cue which distinguished sounds for short segment durations was a difference in the overall "quality" of the sounds, not a transition cue. However, the contrast of the difference in quality was more evident for shorter break periods.

E. Data collection

Data were collected in two phases. In the first phase 6 tracks of trials were collected per condition, where tracks from different conditions were presented in a different random order for each listener. After the completion of the first phase it was observed that the performance curves were irregular for most listeners for short segment durations in the EX conditions. To examine the listeners' performance in this region, an additional 6 tracks were collected in both conditions. The new data were collected in the second phase in a similar way as in the first, but including only trials for short segment durations and in the EX conditions only. In the second phase a track of trials started at a segment duration of 1 ms, and, in one track, 80 trials were collected per decade.

In the first phase $6 \times 161 = 966$ trials were collected in each condition, which gives 240 trials per decade to estimate each listener's performance. In the second phase also 966 trials were collected in each condition, but the range of the segment durations was halved, giving a total of 480 trials per decade to estimate performance.

III. RESULTS

A. Procedure for deriving psychometric functions

For each listener, psychometric functions were derived for each of the 6 experimental conditions. These are displayed in Figure 4. Figure 5 offers psychometric functions for pooled results across listeners. The psychometric functions were constructed in the following way: The results of all trials for a given listener and condition, were sorted (according to the segment duration) and arranged in 150 overlapping bins. The bins were equally spaced on a logarithmic time-axis, and the results of 100 trials were included in each bin. This means that the curves are considerably "smoothed" and only approximately 2.4 estimates per decade are based on completely non-overlapping data. Performance was estimated within each bin by dividing the number of correct responses with the total number of trials in that bin. By considering each trial a Bernoulli trial, 95%-confidence limits were estimated based on the Binomial distribution. The 95%-confidence limits are marked as gray-shaded areas in the figures.

Different number of trials were included in the bins when the results were pooled across listeners (500 trials per bin), and the number was also different for the curves of results in the second phase (75 for individual data and 500 for pooled data).

B. Procedure for estimating performance level

The segment durations corresponding to a performance level of 0.75 (75% correct responses) were estimated for all listeners in all conditions. That was done by fitting

Condition	OH	OS	BP	HT	TM	AU	ALL
3NO	2.12	1.69	0.47	2.30	2.17	9.06	2.20
21NO	22.33	40.56	3.60	34.49	33.44	92.36	25.92
3EN	0.63	1.03	0.35	1.41	0.89	1.70	0.88
21EN	25.80	35.65	2.62	24.50	36.88	118.71	26.24
3EX	0.25	0.48	0.11	0.31	0.29	1.04	0.33
21EX	6.96	266.80	0.01	31.65	45.69	76.60	43.04

Table I. Segment duration in ms at 0.75 performance level for all listeners in all conditions. The last column, "ALL", was derived from the pooled data of all listeners.

curves to the data obtained in the first phase of the experiment and calculating the segment duration where the fitted curve had an ordinate value of 0.75. A logistic curve was used to model the psychometric functions, taking the following form:

$$\Psi(T) = \frac{1}{2} + \frac{1}{2} \cdot \frac{e^{c(log(T)+k)}}{1 + e^{c(log(T)+k)}} \quad (1)$$

Ψ models the probability of a correct response as a function of the segment duration T. In the formula k determines the segment duration of the transition point of the psychometric function, and c determines the steepness of the fitted function at the transition. T is the segment duration measured in seconds. The function is fitted on the logarithmic time axis by using $log(T)$ in the equation. As can be seen, when the segment duration approaches 0, Ψ approaches 1/2 and when T approaches ∞, Ψ approaches 1. That means the psychometric function is forced to approach the chance level of 50% correct for small segment durations, and for long segment durations it is forced to approach a performance level of 100% correct. The coefficients c and k were estimated in a least square fit of Ψ to the binary response data. That is, Ψ is *not* directly fitted to the estimated and smoothed performance curves (solid lines) in Figure 4 and 5.

C. Results of the first phase

The segment durations corresponding to a 0.75 performance level (75% correct responses) are summarized for all listeners in Table I: When comparing the 3-segment conditions, it is evident that for all listeners the 75%-threshold is lowest in the 3EX (fine-structure repeated) condition followed by the 3EN (envelope repeated) and finally the 3NO (no repetitions). So for the 3-segment conditions there is a unequivocal rank order of the conditions. Such a pattern cannot be identified in the 21-segment conditions, the main reason being that the fitted curves provide a very poor fits to the listeners' actual performance in the 21EX condition, some of the derived limits are almost meaningless. Listener "BP" does in general have a lower limit than other listeners, which may be explained by his greater experience in listening experiments. The difference in performance between "BP" and the other listeners seems to be more pronounced in the 21-segment conditions than in the 3-segment conditions, indicating that experience more heavily influences the performance in the 21-segment conditions.

A closer look at Figure 4 allows for a more detailed comparison of the listeners' performance in different conditions: First it can be observed that performance is almost identical across NO, EN and EX conditions for long segment durations (above 20 ms). This was expected, since the tasks are identical across the mentioned condition before the pattern started repeating (segment duration $\geq$ 27 ms for 21-segment and 16 ms for 3-segment pattern), except for the somewhat shorter period of silence between the three sounds of a single trial in the NO conditions. Considering all psychometric functions, going from long segment durations towards shorter durations, performance starts to get worse at a certain point and continuously becomes worse for a while. In most cases performance continuously declines until it reaches the chance level of 0.5, but this is not generally true. For some listeners in some conditions performance improves again for shorter segment durations. This is more pronounced in the EX conditions (in which the fine-structure repeats), but it may also occur in other conditions, consider for example 21EN for listener "OH" or 3NO for listener "OS" around segment duration of 0.06 ms. The irregularities of the psychometric functions for shorter segment durations were examined in greater depth in the second phase of the experiment, but only for the EX conditions. The effects were considered too small in the other conditions to deserve further study.

D. Results of the second phase

The results of the second phase of the experiment (fine-structure conditions only) are depicted in Figure 6 for each individual listener and in Figure 7 offers pooling across listeners. Only the range of segment durations from 1 ms down to 0.02 ms was covered in the second phase. No psychometric functions were fitted, because of the non-monotonicity of performance as a function of segment duration, which is evident from the figures.

1. Overall results of the second phase

First the overall effects can be considered by comparing Figure 5 and Figure 7. In the 3EX conditions the overall performance of the listeners were almost identical in the two phases, but what was only weakly evident in the first phase is much clearer in the second phase, namely: There are two ranges of the segment duration with "flat" performance (better than chance), the first one starting from a segment duration of 1 ms down to about 0.3 ms where performance is 0.85, and the second

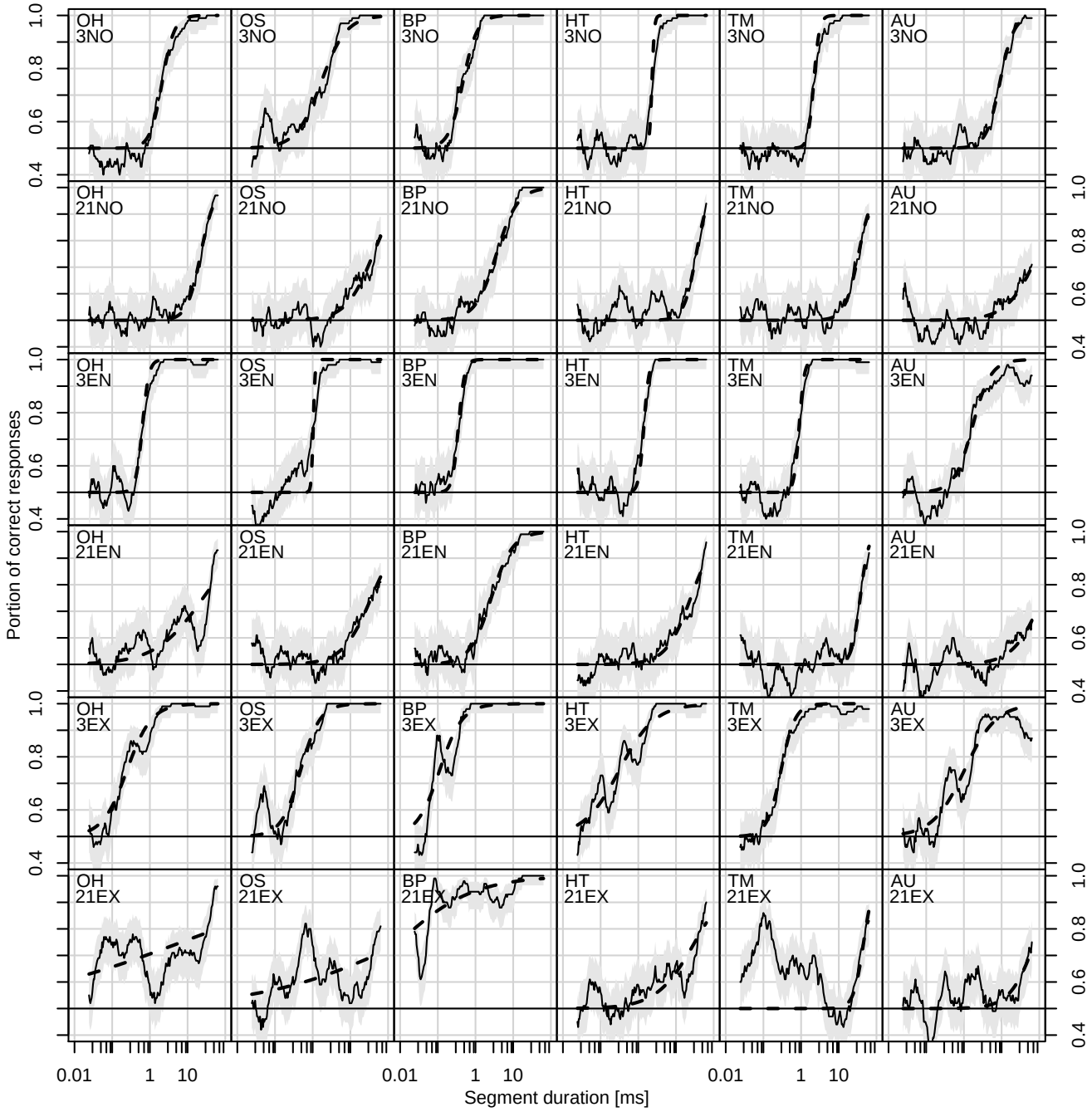


Figure 4. Estimated psychometric functions for all listeners in all conditions. Segment duration on the x-axis and the proportion of correct responses on the y-axis. The results of each listener are depicted in each column. Each row contains the results of each of the conditions: 3NO, 21NO, 3EN, 21EN, 3EX and 21EX from top to bottom respectively. The shaded areas illustrate the 95%-confidence limits. The dashed lines are the fitted curves used to estimate the 0.75 performance limits in Table I.

starting from 0.15 ms and down to 0.07 ms with a performance level of 0.65. Comparing the two phases, but for the 21EX conditions, the overall performance may by slightly higher in the second phase. This indicates that there is only a modest training effect between the two phases. In both phases performance remains almost constant at a level of 0.7 for segment durations smaller than 1 ms. Note that performance at a given segment duration is based on trials with the actual segment durations in a narrower range in the second phase as in the first, which gives twice as high a resolution of the estimated performance curves.

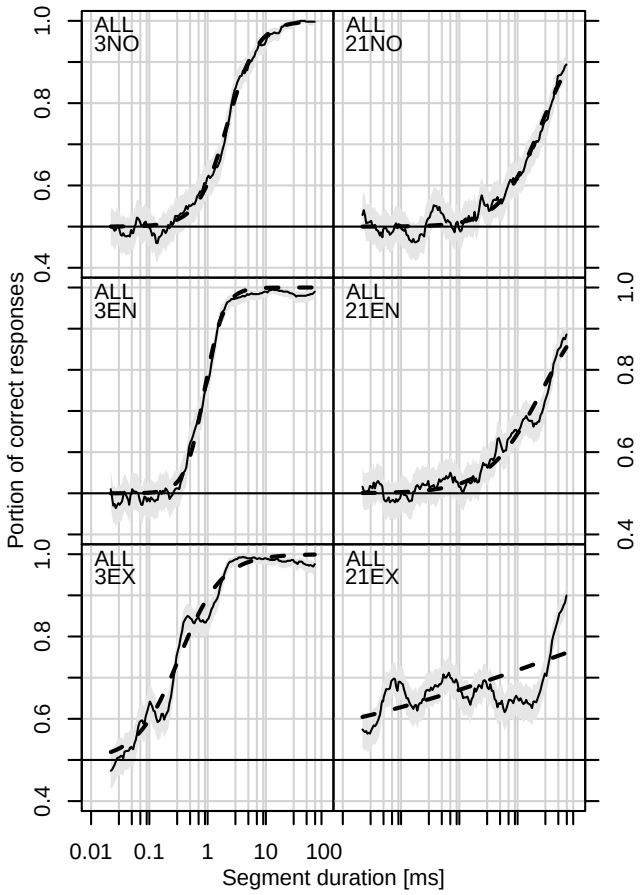


Figure 5. Same as Figure 4, but with pooling of data across listeners.

2. Individual results of the second phase

The individual performance curves (Figure 6) are much harder to discuses in general terms. First, for most listeners, there is no significant deviation in performance in the second phase compared to the first. For some, however, performance is better in certain regions, considering listener "HT" in the 3EX condition, for example, there is a peak in performance at a segment duration of 0.1 ms, which is more pronounced in the second phase of the experiment. There may be two reasons for this: (1) The resolution of the estimated performance curve is higher in the second phase, giving more pronounced peaks and dips, and (2) the listener has improved in performance, which is easily understood in terms of the experience gained throughout the experiment. Performance within each listener seems to decay and increase at the same segment durations across the two phases of the experiment. Some listeners show strikingly similar behavior, compare "HT" and "BP" in the 3EX conditions for example: Both have a "double" peak in performance centered at 0.1 ms segment duration. There is a weak indication that "TM" may also have one of these peaks at 0.08 ms segment duration. All other listeners do not show signs of increasing performance after the first decline in the 3EX condition. A drop in performance between 0.2 and 0.15 ms segment duration seems to be occur for all listeners. In the 21EX condition, most listeners show a peak in performance around a segment duration of 0.1 ms. Before this increase in performance most listeners have a dip, this is however not true for "OH" whose performance continually increases from 1 to 0.1 ms segment duration, and even below 0.1 ms "OH"s performance remains at a high level of approximately 0.8. Comparing 3- and 21-segment EX conditions in which the fine-structure is preserved, the same type of effects can be observed; performance both increases and decreases in various time-ranges: Where exactly this occurs, however, varies across the 3- and 21-segment conditions and also varies across listener. A good example of this is "OH" who, in the 3-segment conditions, performs increasingly worse towards shorter segment durations, while the opposite is the case in the 21-segment condition.

IV. DISCUSSION

The results showed that performance of the listeners was fundamentally different when the fine-structure of a pattern was repeated compared to the case where only the envelope of a single pattern was repeated. When only repeating the envelope, only a modest improvement in performance compared to the non-repeating case could be observed (compare for example the psychometric functions for NO and EN conditions in Figure 5). When repeating the fine-structure, performance was especially different in the 21-segment condition, where performance remained significantly above the chance level over the full range of segment durations used in the experiment (see Figure 5). Further, in the non-repeating and repeating envelope conditions, performance was always worse in the 21-segments condition as compared to the 3-segment conditions, which was not the case when fine-structure was repeated. When analyzing the performance of the listeners in the repeating fine-structure conditions in the second experimental phase, it was clear that individual psychometric functions were strongly non-monotonic as a function of the segment duration (Figure 6).

A. Sensory stages of processing temporal information

To understand the present results, different perceptual processes are proposed as being responsible for the performance at different segment durations:

1. Segment duration above 25 ms: Listeners are aware of the pattern of level-fluctuations.

2. Segment duration above 1 ms: Listeners can discriminate patterns based on onset and offset cues.

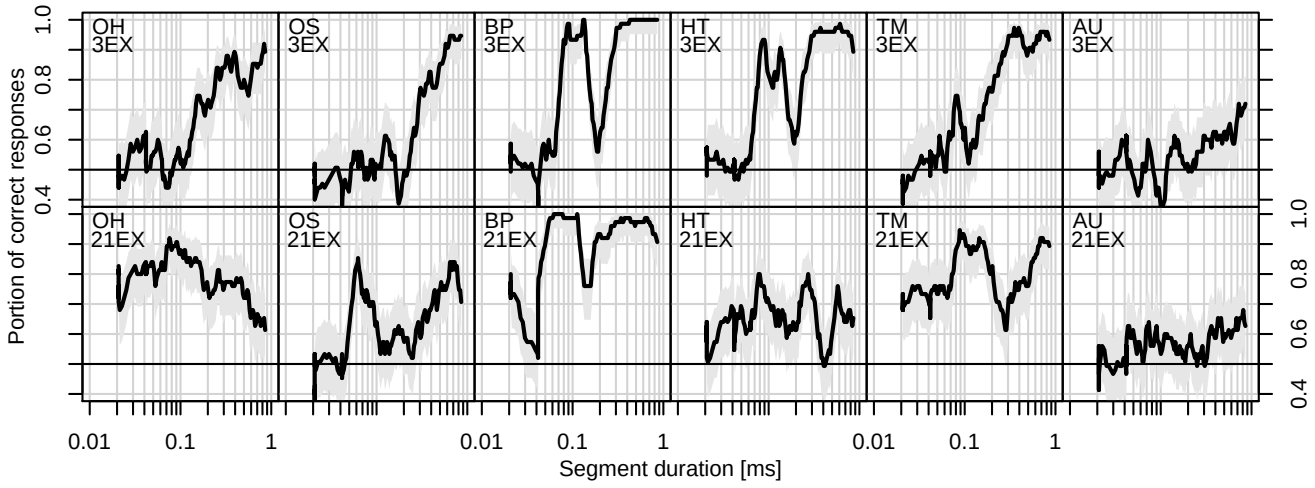


Figure 6. Estimated psychometric functions for all listeners in the two conditions of the second phase of the experiment. Segment duration on the x-axis and the proportion of correct responses on the y-axis. The results of each listener are depicted in each column. Each row contains the results of each of the conditions: 3EX in the top row and 21EX in the bottom row. The shaded areas illustrate the 95%-confidence limits.

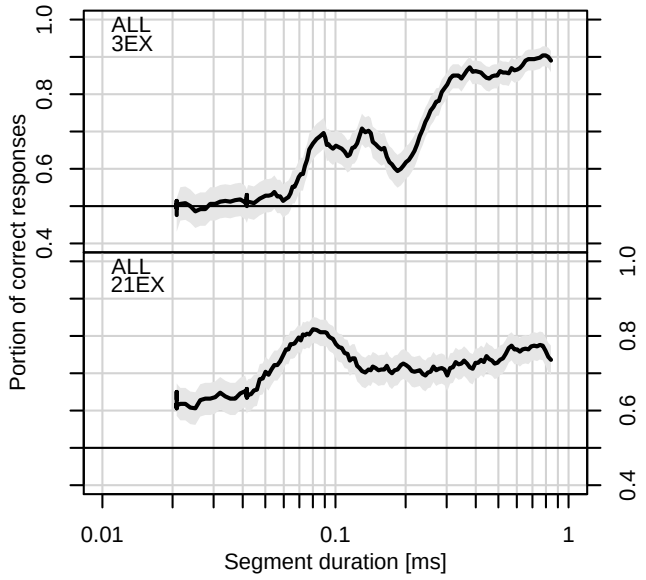


Figure 7. Same as Figure 6, but with pooling of data across listeners.

3. Segment duration below 1 ms: When the fine-structure causes special static features to emerge, listeners can discriminate by perceived differences in quality.

When no flanking noise is present (3-segment conditions), the target patterns occur at the onset/offset of the entire sound. In that case listeners might base their judgments on onset/offset cues (second item on the list). If, however, the onset and offset is dominated by flanking noise (21-segment conditions) the listeners must be able to follow the actual envelope of the sound to identify the target pattern (first item on the list). Further support for this notion is given by Pedersen (2006). The critical numbers for the time ranges come from Table I where 1 ms seems close to an average limit across all listeners in the 3NO and 3EN conditions (onset/offset cues available), whereas in the 21NO and 21EN it is close to 25 ms (listeners must identify envelope fluctuations). Further, these time constant are supported by the results obtained by Pedersen (2006).

The third item on the list is mainly supported by the results of the current experiment's EX-conditions (repeating fine-structure), where it was demonstrated that the discrimination task was not completely impossible at any segment duration (when the overall results are considered). When performing the task for shorter segment durations it was clear that the cue for discriminating the patterns varied, suggesting that different extracted "features" may be responsible for the discrimination as a function of segment duration. This notion is supported by the irregularities observed in the individual performance curves. If the same sensory processing was involved for all segment durations, much smoother curves would be expected.

B. Temporal masking and modeling

Most experiments reported in the literature measuring temporal resolution arrive at a limit in the range of a few ms (Moore, 2003). This means that they fall in the critical region where onset/offset cues are important

according to the reasoning outlined above. A reinterpretation of for example gap detection experiments where no "smoothing" process is assumed, could take the following form: Traditionally, a sliding temporal window is thought to account for the masking, and is for example fitted by Oxenham and Moore (1994) giving different time coefficients for pre and post masking. These time coefficients might be a measure of the width of a time window within which onsets and offsets are decoded, Or, to put it in other words, the auditory system "sees" the temporal variation through a temporally limited window. Relatively long and short post and pre masking respectively is normally found, which translates into a relatively long window for analyzing offsets and short window for onsets. That is, the "temporal window" is not responsible for smoothing, but decisive in determining when an event is considered a new onset. Ecologically this makes good sense since sounds are most often characterized be a relatively abrupt (short window) onset and slow (longer window) decay.

The results of the repeated fine-structure (EX) conditions are clearly at odds with models including a sliding temporal window, supported in two ways: (1) The discrimination task was possible at a resolution far better that what is normally predicted (observed in the order of 60 μs as compared to 2 ms, which is in the typical range of a temporal window), and (2) a temporal window gives a smooth decline in performance (as a function of the extent of temporal details, which in this case is given by the segment duration), which is inconsistent with the observed discontinuities in the performance curves. Surprisingly the 21EX (21-segment, repeated fine-structure) condition was generally easier for very short segment durations than 3EX (3-segment, repeated fine-structure). This argues against any "power" or "energy" masking. Rather, because the difference between ascending and descending patterns is identical across 3- and 21-segment conditions, as argued by Pedersen (2006), relative phase differences may be important properties.

Ideas introduced to understand pitch perception (see for example Moore, 2003) may be helpful to a deeper understanding. Both spectral and temporal pattern matching have been suggested to play a role in pitch perception, and support for both ideas has been found. Sounds in the repeated fine-structure (EX) conditions had a clear pitch quality for short segment durations, but identical spectra. So temporal properties are crucial for the discrimination and may be reflected in pitch, but also different concepts may be of importance, as for example timbre, and cues normally used for spatial localization.

Patterson (1994b) applied a model utilizing temporal properties to explain the difference in perceived quality of ramped and damped sinusoids, but argues that the temporal information may not be preserved with a finer resolution than 1 ms, which is significantly worse that the limit suggested by the result of the current experiment, and thus the model seems incompatible with the stimuli of the present experiment.

C. Monaural and binaural phase sensitivity

Very different measures of the auditory ability to detect phase differences are obtained in studies using binaural or monaural stimuli. In the monaural case, the finest temporal resolution measured is in the range of 0.2 ms as found by Henning and Gaskell (1981), who used Ronken's paradigm (Ronken, 1970), in which the delay between two clicks of different levels is varied. The task of the listener is to detect the order of the clicks (high-low or low-high level). Time-reversal is used to change the order of the levels so no spectral cues are available. The shortest delay between pulses at which the patterns can be discriminated is thus a measure of temporal resolution. If the task of the listener is to discriminate temporal disparities across the ears using binaural stimuli, much shorter temporal limits have been found (in the order of some μs). For example Klumpp and Eady (1956) found that listeners could discriminate stimuli where the interaural time differences was in the order of 9 μs. This vast discrepancy in temporal resolution found in binaural in monaural listening tests has lead to the suggestion that binaural processing (as for example cross-correlation of the stimulation of the two ears) is responsible for the fine temporal resolution in the binaural case. But it could reasonably be argued that the binaural processing has to take place on the basis of the monaural input from each ear. Hence, monaural temporal resolution should be at least as good as that of the binaural system. The discrimination between temporal resolution in binaural and monaural hearing is mostly motivated by the fact that no listening experiment using monaural stimuli has demonstrated temporal resolution in a comparable range to results of binaural experiments. Interestingly the present experiment demonstrated very fine temporal resolution in a monaural task (EX conditions), where some listeners were still able to perform temporal discrimination when the segment duration was as low as 20 μs. This value is comparable to temporal limits found in binaural listening tests. This leads to the suggestion that there is no reason for making a sharp division between the sensory system working in a "binaural" or a "monaural" mode. Thus the different critical time regions found in the present experiment may be compared to effects known to exist in binaural hearing: It was suggested that in the critical time-region above 1 ms, onset and offset played an important role. Interestingly, in term of the critical time range, this coincides with the "precedence effect" in binaural hearing (Blauert, 1999), where 1 ms has been found to be the critical lower limit for the precedence effect to occur. The upper limit is in the range 5 to 50 ms. Ecologically it makes good sense that the precedence effect is tightly connected to the decoding of onsets, and to a smaller extent, offsets. Another effect found in binaural experiments is that listeners' ability to follow spatial movements of a sound source seems to be limited to a certain speed of movement. This effect has been described as "binaural sluggishness", and Blauert (1972)

suggests that fluctuations in location at a rate of approximate 2.5 Hz may be followed well. This coincides, to some extent, with the suggested limit of this paper where listeners must be able to follow the level fluctuations to perform the discrimination task (25 ms). The identification of similar critical time-ranges in both binaural and monaural listening, may lead to the conclusion that they are based on the same processes. Often, models utilizing cross-correlation are used to explain the performance in binaural tasks. This may however seem incompatible with the notion of fundamental "feature extraction" processes at the lowest peripheral levels as explained by Näätänen and Winkler (1999). A full trace of the fluctuations of the sound is required to perform cross-correlation, and a full trace may not be available if the percept has been encoded as different features. Rather it may be hypothesized that the binaural system utilizes the extracted features to determine the spatial location of a given sound source. As was demonstrated by Patterson (1994b), the perceived pitch of a sound depends on temporal details, so for example pitch-like cues might be compared across the two ears and would reflect differences in temporal details of the sound across the ears.

V. CONCLUSION

The presented results suggest that fine-structure play an important role in auditory temporal processing, which is evident from the observation that listeners always performed better in conditions where the fine-structure was repeated compared to the case of only the envelope being repeated. For short segment durations, listeners showed remarkably good performance when non-informative noise segments were added and fine-structure was repeated (21EX condition). No lower limit for the segment duration at which listeners could perform the task was observed in this condition even though patterns were presented at segment durations as low as 20 μs. This is hard to reconcile with the notion of energetic masking, which one would assume to cause performance to be generally worse in conditions where non-informative nose is present. Further, it seems to contradict the notion of "smoothing" of the envelope at peripheral stages of temporal processing in the auditory system. Since a similarly good performance was not observed in the conditions where only the envelope was repeated, this may suggest that fine-structure cues have to be "static" (steady-state) to some degree, as the stimuli of the repeated envelope conditions in principle contain a wealth of fine-structure cues. This supports the notion of the extraction of "static" features relying on very fine temporal cues.

A. Topics for further inquiry

Future experiments may try to more specifically identify "features" and which properties, spectral and temporal, are crucial for their perception. Temporal patterns may also be frequency filtered to examine how temporal information is combined across critical frequency bands, possibly showing different effects in the different critical time regions. Hopefully such experiments would lead to a better understanding of different sensory factors responsible for decoding the temporal properties of sound.

VI. ACKNOWLEDGMENTS

This research was carried out as part of the "Center-contract on Sound Quality" which establishes participation in and funding of the "Sound Quality Research Unit" (SQRU) at Aalborg University. The participating companies are Bang & Olufsen, Brüel & Kjær, and DELTA Acoustics & Vibration. Further financial support comes from the Ministry for Science, Technology, and Development (VTU), and from the Danish Research Council for Technology and Production (FTP).

Appendix A: QUANTIZATION OF SEGMENT DURATIONS

Because of discrete sampling in the generation and playback, actual segment durations of the stimuli played back was also discrete. That is, not all segment durations required in one track can be realized. The procedure used to generate the noise carrier caused the segment duration to be a multiple of four samples. The used sampling rate was 192 kHz, so for the very shortest segment durations the discretization becomes evident. This is illustrated in Figure 8 where discrete segment duration of all trials in one fixed track is plotted against the segment duration with no discretization. All results presented in figures and in table in the paper are based on the segment durations of the stimuli the listeners heard, that is, the quantized durations.

Appendix B: SPECTRUM OF NOISE CARRIER

The amplitude spectrum of the non-modulated and infinitely long noise carrier has broad-band character, but is not similar to white-noise. Rather, it has relatively low power contained at low frequencies and high power at higher frequencies as shown in Figure 9. The noise carrier is described in greater detail by Pedersen (2006), who also describes a procedure for deriving the amplitude spectrum.

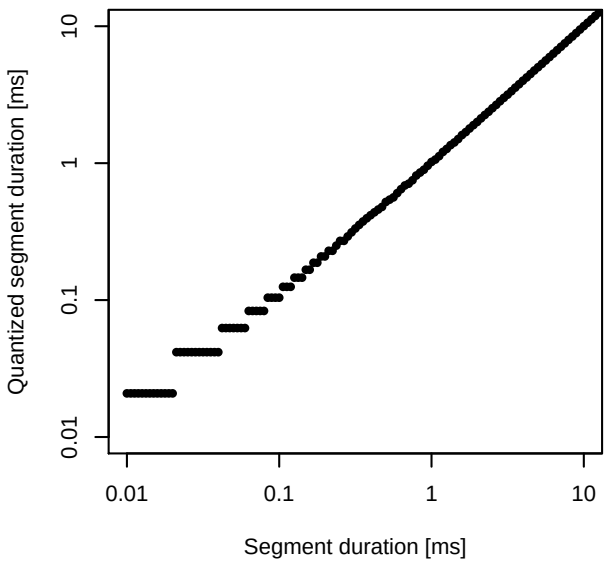


Figure 8. Quantized segment duration (y-axis) plotted against theoretic segment duration used in one track of trials.

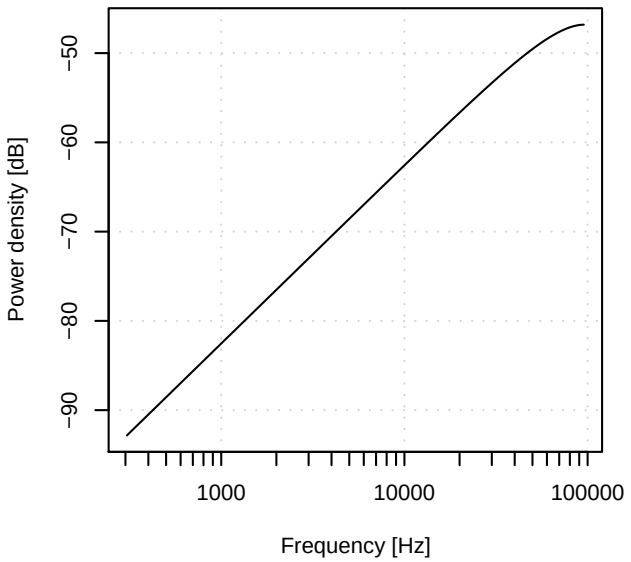


Figure 9. Amplitude spectrum of the noise carrier (see Pedersen, 2006, for its definition).

Blauert, J. (**1972**). "On the lag of lateralization caused by interaural time and intensity differences", Audiology **11**, 265–270.

Blauert, J. (**1999**). *Spatial hearing: The psychophysics of human sound localization*, 2nd edition (Cambridge, Mass, MIT Press).

Henning, G. B. and Gaskell, H. (**1981**). "Monaural phase sensitivity with Ronken's paradigm", J. Acoust. Soc. Am. **70**, 1669–1673.

Klumpp, R. G. and Eady, H. R. (**1956**). "Some measurements of interaural time difference thresholds", J. Acoust. Soc. Am. **28**, 859–860.

Krumbholz, K., Bleeck, S., Patterson, R. D., Senokozlieva, M., Seither-Preisler, A., and Lütkenhöner, B. (**2005**). "The effect of cross-channel synchrony on the perception of temporal regularity.", J. Acoust. Soc. Am. **118**, 946–954.

Moore, B. C. J. (**2003**). *An Introduction to the Psychology of Hearing*, 5th edition (Academic Press, San Diego, CA).

Näätänen, R. and Winkler, I. (**1999**). "The concept of auditory stimulus representation in cognitive neuroscience.", Psychol. Bull. **125**, 826–859.

Oxenham, A. J. and Moore, B. C. (**1994**). "Modeling the additivity of nonsimultaneous masking.", Hear. Res. **80**, 105–118.

Patterson, R. D. (**1994**a). "The sound of a sinusoid: Spectral models", J. Acoust. Soc. Am. **96**, 1409–1418.

Patterson, R. D. (**1994**b). "The sound of a sinusoid: Time-interval models", J. Acoust. Soc. Am. **96**, 1419–1428.

Pedersen, B. (**2006**). "Temporal masking in the auditory identification of envelope patterns", in *Auditory Temporal Resolution and Integration: Stages of Analyzing Time-Varying Sounds*, 67–82 (Aalborg University).

Ronken, D. A. (**1970**). "Monaural detection of a phase difference between clicks", J. Acoust. Soc. Am. **47**, 1091–&.

Viemeister, N. F. and Wakefield, G. H. (**1991**). "Temporal integration and multiple looks.", J. Acoust. Soc. Am. **90**, 858–865.

Appendix A

Loudness models: Specifications and evaluations

A.1 Specification of models for prediction of loudness judgments

This appendix accompanies Section 3.5, where a broad range of models are fitted to listeners' judgments of loudness. The appendix contains the mathematical specification of each of the 15 models applied. Each specification gives the formulation of the model of perceived loudness (L) as a function of the ten segment levels (vector $\mathbf{x}$, and x_i where i is the segment index). In Section 3.5 the actual procedure for fitting the models to the listeners' responses is given.

A short name and number identify each model in its headline, where also the number of parameters is given, which specify the number of parameters which are estimated in the fitting procedure. It may be noted that this number is alway larger than the number of parameters which occurs in the expression for L by one. The "extra" parameter is the threshold criterion, c, as it appears in Equation 3.3.

Further characteristics of each model are given in five fields: "Formula", "model coefficients", "special case of", "extends" and "short description". The "formula" field gives the mathematical equation for L. The "model coefficients" field shortly describes the functioning of the model parameters. The "special case of" field is a list of models, which are more general formulations of the relevant model. That is, by fixing some of the parameters in the more general models, they become identical to the relevant model. "Extends" is a list of models for which the opposite is the case, that is, by fixing parameters of the relevant model it becomes identical to the simpler models. The "short description" field gives a short description of the main ideas and purpose of the model.

If it is not straightforward, it is explained in the "extends" field how the parameters of the model should be fixed to give identical predictions to the listed "extends" models.

M1 *Mean* 2 parameters

Formula:

$$L(\mathbf{x}) = w\text{mean}(\mathbf{x})$$

Model coefficients:

w is a weighting coefficient indicating the sensitivity to changes in the mean of $\mathbf{x}$.

Special case of:

M6, M7, M8, M9, M10, M11, M12, M13, M14, M15

Short description:

Listener judgment is predicted from a simple mean value of the 10 segment levels.

M2 *Maximum* 2 parameters

Formula:

$$L(\mathbf{x}) = w\max(\mathbf{x})$$

Model coefficients:

w is a weighting coefficient indicating the sensitivity to changes in the maximum of $\mathbf{x}$.

Special case of:

M7, M11, M12, M13

Short description:

Listener judgment is predicted from the maximum of the 10 segment levels.

M3 *Minimum* 2 parameters

Formula:

$$L(\mathbf{x}) = w\min(\mathbf{x})$$

Model coefficients:

w is a weighting coefficient indicating the sensitivity to changes in the minimum of $\mathbf{x}$.

Special case of:

M7, M11, M12, M13

Short description:

Listener judgment is predicted from the minimum of the 10 segment levels.

M4 *Variance* 2 parameters

Formula:

$$L(\mathbf{x}) = w\text{var}(\mathbf{x}) = \frac{w}{9}\sum_{i=1}^{10}(x_i - \text{mean}(\mathbf{x}))^2$$

Model coefficients:

w is a weighting coefficient indicating the sensitivity to changes in the variance of $\mathbf{x}$.

Short description:

Listener judgment is predicted from the variance of the 10 segment levels.

M5 *Envelope profile* 3 parameters

Formula:

$$L(\mathbf{x}) = w_{slp}\alpha + w_{int}\beta$$

Model coefficients:

A linear regression fit to the segment levels as a function of segment index determines the slope, α, and intercept, β, of the regression line. The regression coefficients are linearly weighted by w_{slp} and w_{int} respectively.

Extends:

M1

The intercept, β, can be calculated as: $\beta = \text{mean}(\mathbf{x}) - \alpha\text{mean}(\mathbf{segidx})$, where **segidx** is the segment index from 1 to 10, therefore: $\text{mean}(\mathbf{x}) = \beta + \alpha\text{mean}(\mathbf{segidx})$. Thus, if $w_{slp} = 1$ and $w_{int} = \text{mean}(\mathbf{segidx}) = 5.5$, the model gives same predictions as the "mean" model if the two coefficients are further scaled by w from model M1. This proves that M1 is a special case of M5.

Short description:

For a given sound a line is fitted to the levels of the ten segments, with segment index as the x-value and segment level as the y-value. A least square fit gives the parameters α (slope) and β (intercept) for the best fitting line. Negative value for α thus indicates that the sound is descending from the beginning toward the end of the sound and positive value indicates an ascending temporal profile of the sound. The listeners behavior is modeled as the slope weighted by a factor w_{slp} and the intercept by w_{int}.

M6 *Polynomial non-linearity* 3 parameters

Formula:

$$L(\mathbf{x}) = \sum_{i=1}^{10} w_{sq} x_i^2 + w_{lin} x_i$$

Model coefficients:

A quadratic term of the segment levels is scaled by w_{sq}, and linear term by w_{lin}.

Special case of:

M10, M15

Extends:

M1

Short description:

Extends the mean model (M1) and introduces a quadratic term of the segment levels. The sum of the segment levels is scaled by the factor w_{lin} and the sum of the squared levels by w_{sq}.

M7 *Attention weighting* 3 parameters

Formula:

$$L(\mathbf{x}) = w \sum_{i=1}^{10} \frac{x_i^p}{\sum_{j=1}^{10} x_j^p} x_i = w \frac{\sum_{i=1}^{10} x_i^{p+1}}{\sum_{i=1}^{10} x_i^p}$$

Model coefficients:

w is a weighting coefficient indicating the sensitivity to changes in the attention weighted sum of $\mathbf{x}$. p is an exponent determining how strongly attention is guided toward more salient segments in term of their relative level within a sound.

Special case of:

M11, M12, M13

Extends:

M1, M2, M3

With $p = 0$, the attention weight (in front of x_i in the central expression) becomes 0.1, which makes the model identical to M1. With p approaching ∞, the attention weight will approach 1 for the maximum segment, and 0 for the others. With p approaching $-\infty$, the attention weight will approach 1 for the minimum segment, and 0 for the others. Therefore M2 and M3 are also special cases of M7.

Short description:

This model weights each segment level by its own level in relation to all 10 segments of the sound. The relative weight for the i'th segment is given by: $\frac{x_i^p}{\sum_{j=1}^{10} x_j^p}$.

M8 *Moment of inertia* 4 parameters

Formula:

$$L(\mathbf{x}) = w_m m + w_{cm} cm + w_{mi} mi$$

Model coefficients:

"Mass" is weighted by w_m, "center of mass" by w_{cm}, and "moment of inertia" by w_{mi}:

Extends:

M1

Short description:

First a distance vector $\mathbf{r}$ describing the distance for each segment of the sound to the temporal center of the sound is defined:

Distance (from center): $\mathbf{r} = -4.5, -3.5, \ldots, 3.5, 4.5$

The "mass" (m) of a sound is defined as the sum of the levels of the ten segments:

Mass: $m = \sum_{i=1}^{10} x_i$

The "center of mass" (cm) is defined as the sum of the ten segment levels weighted according to their distance to the center:

Center of mass: $cm = \frac{1}{m} \sum_{i=1}^{10} x_i r_i$

The moment of inertia (mi) is defined as the sum of the ten segment levels weighted according to the square of their distance to the "center of mass":

Moment of inertia: $mi = \sum_{i=1}^{10} x_i (r_i - cm)^2$

M9 *Temporal weighting* 11 parameters

Formula:

$$L(\mathbf{x}) = \sum_{i=1}^{10} w_i x_i$$

Model coefficients:

w_i are linear temporal weights as a function of segment index.

Special case of:

M10, M12 M14, M15

Extends:

M1

Short description:

This model linearly weighs each segment level where the weight is a function of segment index.

M10 *Polynomial non-linearity and temporal weighting* 12 parameters

Formula:

$$L(\mathbf{x}) = \sum_{i=1}^{10} w_{sq} x_i^2 + w_{lin_i} x_i$$

Model coefficients:

Same coefficients as M6, but the weight of the linear term (w_{lin_i}) is a function of segment index.

Special case of:

M15

Extends:

M1, M6, M9

Short description:

This is an extension of M6 to further allow for different linear temporal weights.

M11 *Temporal weighting before attention weighting* 12 parameters

Formula:

$$L(\mathbf{x}) = \frac{\sum_{i=1}^{10}(w_i x_i)^{p+1}}{\sum_{i=1}^{10}(w_i x_i)^{p}}$$

Model coefficients:

Same as M7, but $w_i x_i$ has been substituted for x_i.

Extends:

M1, M2, M3, M7

Short description:

This model is an extension of M7 and also includes linear temporal weights as in M9. Temporal weighting is applied before attention weighting.

M12 *Attention weighting before temporal weighting* 12 parameters

Formula:

$$L(\mathbf{x}) = \frac{\sum_{i=1}^{10} w_i x_i^{p+1}}{\sum_{i=1}^{10} x_i^{p}}$$

Model coefficients:

Same as M7, but the attention weighted levels are further weighted by w_i.

Extends:

M1, M2, M3, M7, M9

Short description:

This model is an extension of M7 and also includes linear temporal weights as in M9. Attention weighting is applied before temporal weighting.

M13 *Attention, power as a function of segment index* 12 parameters

Formula:

$$L(\mathbf{x}) = w \sum_{i=1}^{10} \frac{x_i^{p_i}}{\sum_{j=1}^{10} x_j^{p_i}} x_i = w \frac{\sum_{i=1}^{10} x_i^{p_i+1}}{\sum_{i=1}^{10} x_i^{p_i}}$$

Model coefficients:

Same as M7, but p_i is now a function of segment index.

Extends:

M1, M2, M3, M7

Short description:

Extension of the attention model M7, but with different powers (p_i) for the attention as a function of segment index.

M14 *Interaction and temporal weighting* 20 parameters

Formula:

$$L(\mathbf{x}) = \sum_{i=1}^{10} w_i x_i + \sum_{i=1}^{9} wi_i (x_{i+1} - x_i)^2$$

Model coefficients:

Same as M9, but also contains terms of the square of the difference between adjacent segments, which are scaled by wi_i

Extends:

M1, M9

Short description:

The model extends M9 to also include weights for the square of the difference between adjacent segments.

M15 *Polynomial non-linearity as a function of segment index and temporal weighting*

21 parameters

Formula:

$$L(\mathbf{x}) = \sum_{i=1}^{10} w_{sq_i} x_i^2 + w_{lin_i} x_i$$

Model coefficients:

Same as M10, but the factor of the quadratic term (w_{sq_i}) is now a function of segment index.

Extends:

M1, M6, M9, M10

Short description:

This is an extension of M10 to allow for different weighting of the squared segment level as a function of segment index.

A.2 Non-linearity in model predictions

The following figures are similar to Figure 3.13 and 3.14, but for the cases where the models are fitted to the pooled data of all listeners (Figure A.1) and for each individual listener (Figure A.2).

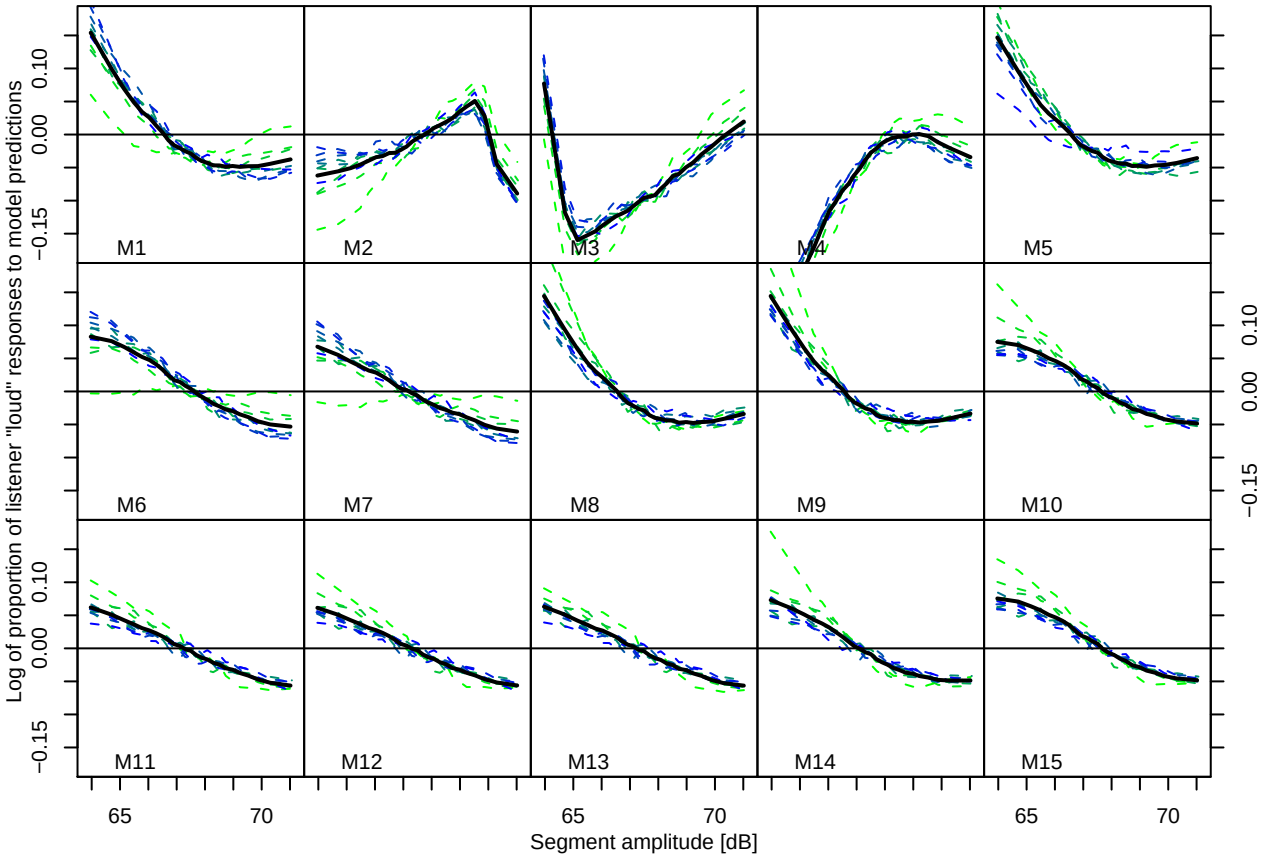


Figure A.1: Same as Figure 3.13, but for the pooled data of all listeners.

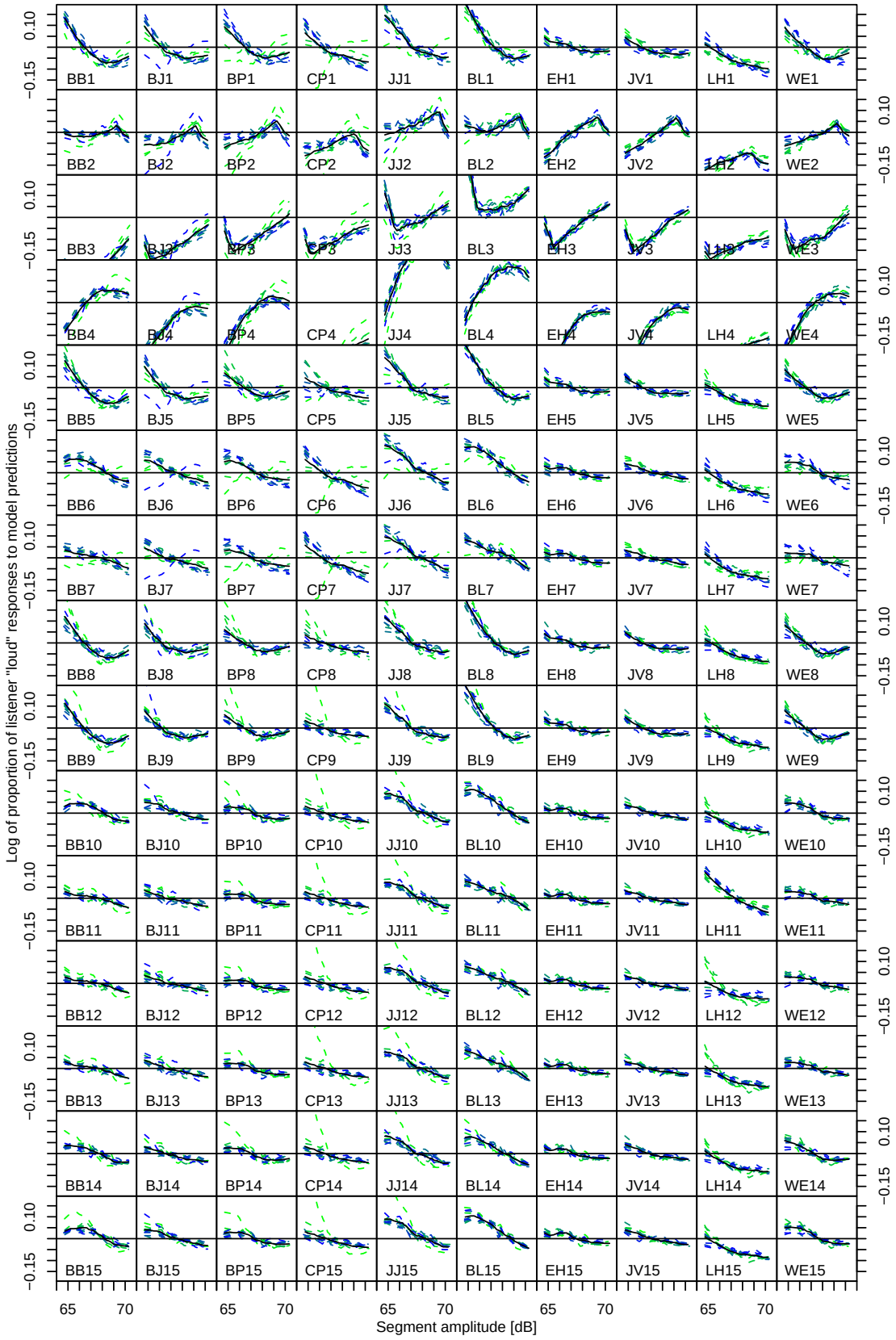


Figure A.2: Same as Figure 3.13, but for each individual listener, identified in the lower left corner of each panel, where the number succeeding the listener ID is the number of the relevant model.

Appendix B

Spectra of repeating patterns

B.1 Spectrum of repeating stimuli

In this appendix spectra for stimuli used in Chapter 5 (Pedersen, 2006b) are shown. There are two reasons why this is of interest: (1) For different segment durations there are different spectral cues and they vary across conditions and (2) it is of interest to validate that de- and ascending patterns have the same spectra as they should according to linear theory as described in the chapter. It should be noted that figures in this section are based on the recording of single sounds only, and not averaged over several sounds. The fine details of the spectra would vary in the experiment, because of the random generation of the noise carrier. However a good overall impression of the spectra can be obtained from the figures, and the harmonic structure, which is most easily observed in the EX conditions and caused by the repetition of the fine-structure, is similar to a large degree across different random noise carriers.

B.1.1 Procedure for recording

Stimuli were recorded for 12 different segment durations, selected on the basis of interesting points on the psychometric functions shown in Chapter 5 (Figure 4, 5, 6, and 7). The same hardware and software were used to generate the stimuli as in the listening experiment described in the chapter. Using the multichannel recording software KRISTAL Audio Engine, both digital and analog signals were recorded in synchrony. The sound card had both digital and analog outputs and inputs making this possible. The purely digital signal was recorded before being converted to an analog signal. The analog recording was obtained with the headphones positioned on a Brüel & Kjær dummy head (BK4128C). All figures presented here were obtained from the right ear. The levels of the 3 segments of the "target" pattern were always in the descending order. Exceptions from this are the graphs where the spectrum of ascending and descending stimuli are compared. In these cases the amplitude spectrum of the descending pattern is subtracted (in dB) from the ascending pattern where all other parameters are similar (segment duration and condition (NO, EN, or EX) and the total number of segments (3 or 21)).

Unfortunately, the equipment used for analog recording had a pronounced noise component at approximately 800 Hz, most easily observed in the lower right panel of Figure B.14 as what may appear to be the fundamental frequency of the harmonic pattern.

B.1.2 Impulse response of headphones

Before analyzing the relevant stimuli, it is of interest to know a little bit more about the physical system. Therefore an impulse played back via the headphones positioned on the dummy head was recorded. In Figure B.1 the recorded sound pressure in Pa is shown as a function of time, and in Figure B.2 the amplitude spectrum of the recorded impulse is shown. Interestingly the duration of the

impulse is relatively long compared to segment durations in some of the tasks in the experiment where the listeners were well beyond the chance level of performance. From the amplitude spectrum it is seen that even though high frequencies exist in the theoretical signal (beyond 25kHz), theses are not transferred to the listeners' ears.

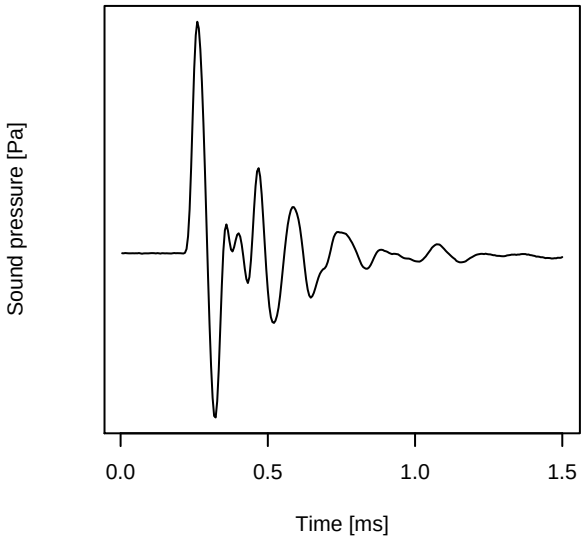


Figure B.1: Recorded impulse response of headphones (DT990) positioned on Brüel & Kjær dummy head.

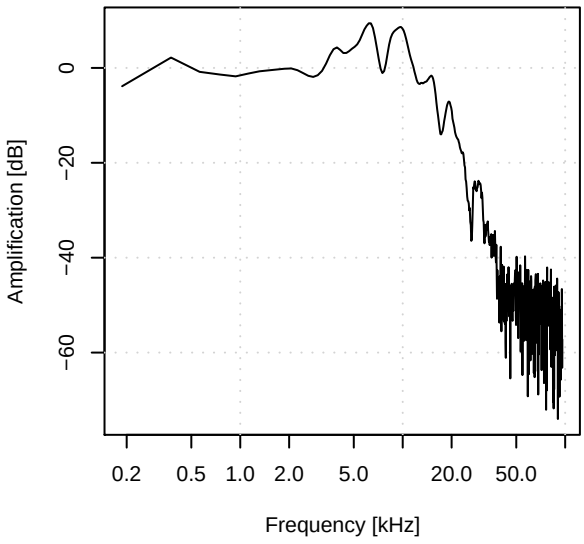


Figure B.2: Amplitude response of recorded impulse response (Figure B.1). The y-axis is scaled to give 0 dB amplification at the "flat" region of the response.

B.1.3 Computing the amplitude spectrum

Before analysis of the recorded stimuli was possible, relevant regions of the recorded stimuli had to be extracted. Since the digital and analog recording were synchronized, the recording of the digital

signal could be used for precise identification of the onset and the offset of sounds. Additionally 2000 samples before and after the onset and offset were included to allow for delay in the analog signal and the time it takes for the impulse response of the analog transfer function (see Figure B.1) to reach its zero level. Of the extra 4000 samples the first 1000 and last 1000 were scaled by a Hanning window. Additional samples of value zero were added so the total duration of the analyzed sound always was 5 s. The spectrum was calculated using a Fast Fourier Transform, which, because off the addition of the extra zeros, had the same frequency resolution independent off condition and segment duration.

Amplitude spectra of digital stimuli

The figures in this section show the amplitude spectra of the digital version of the recorded stimuli. Because it was recorded before digital to analog conversion there is no noise in the recordings. The figures can be compared to their counterparts in section B.1.3 where the resulting spectrum of the sound pressure is plotted. The figures are directly comparable since they were synchronized in the recording.

Patterns of harmonic frequencies appear in the repeating conditions (EX and EN), but more pronounced in EX conditions. The fundamental frequency of the harmonics can be calculated as:

$$fundamental\_frequency = \frac{1}{25 \times segment\_duration}$$

Where 25 is the number of segments per repetition of the pattern. This fundamental frequency is the same for 21- and 3-segment conditions, and is important for the perceived pitch. The fundamental frequency is the same across the conditions at the same segment duration.

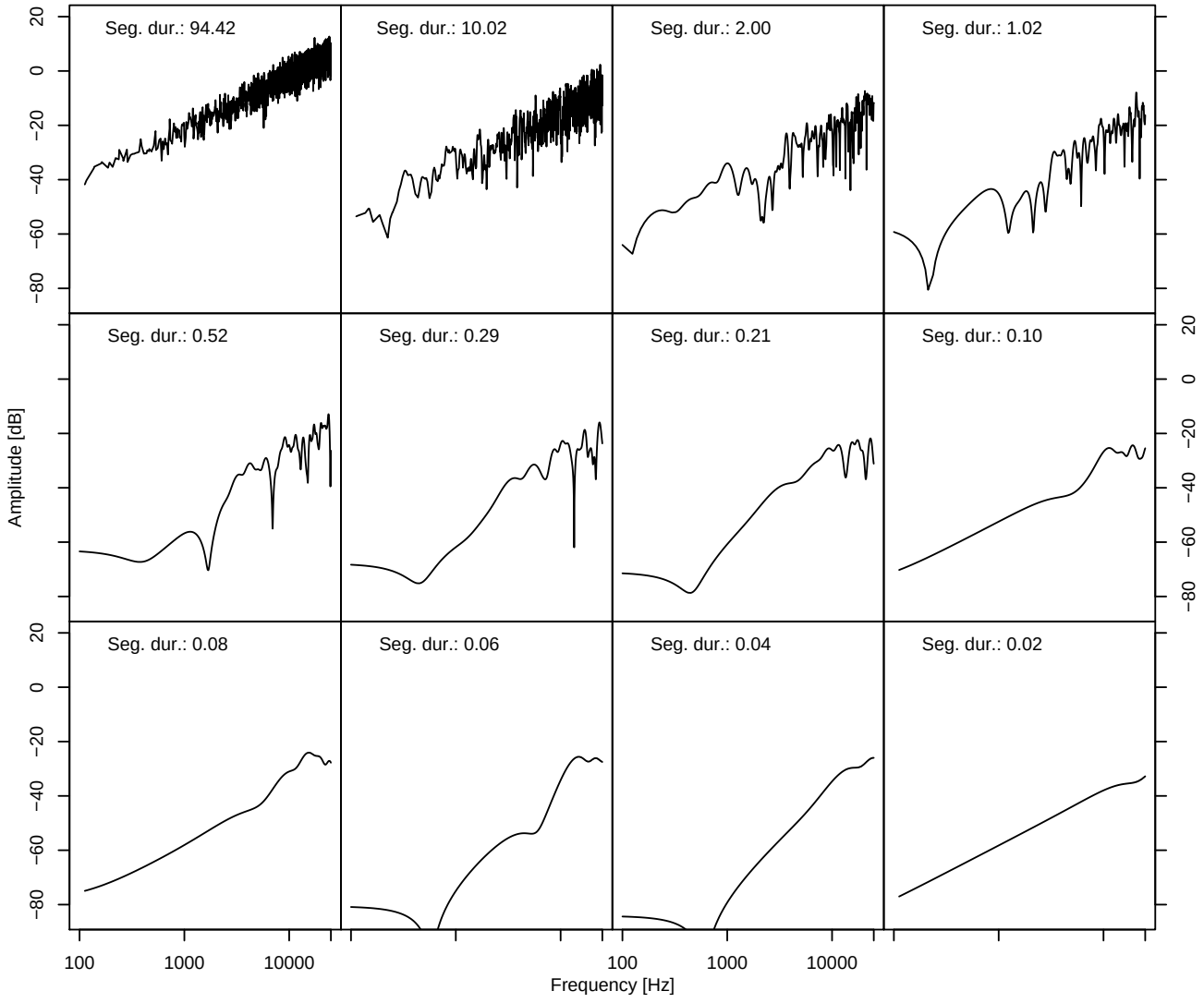


Figure B.3: Spectrum of digital stimuli from the 3NO condition for selected segment durations.

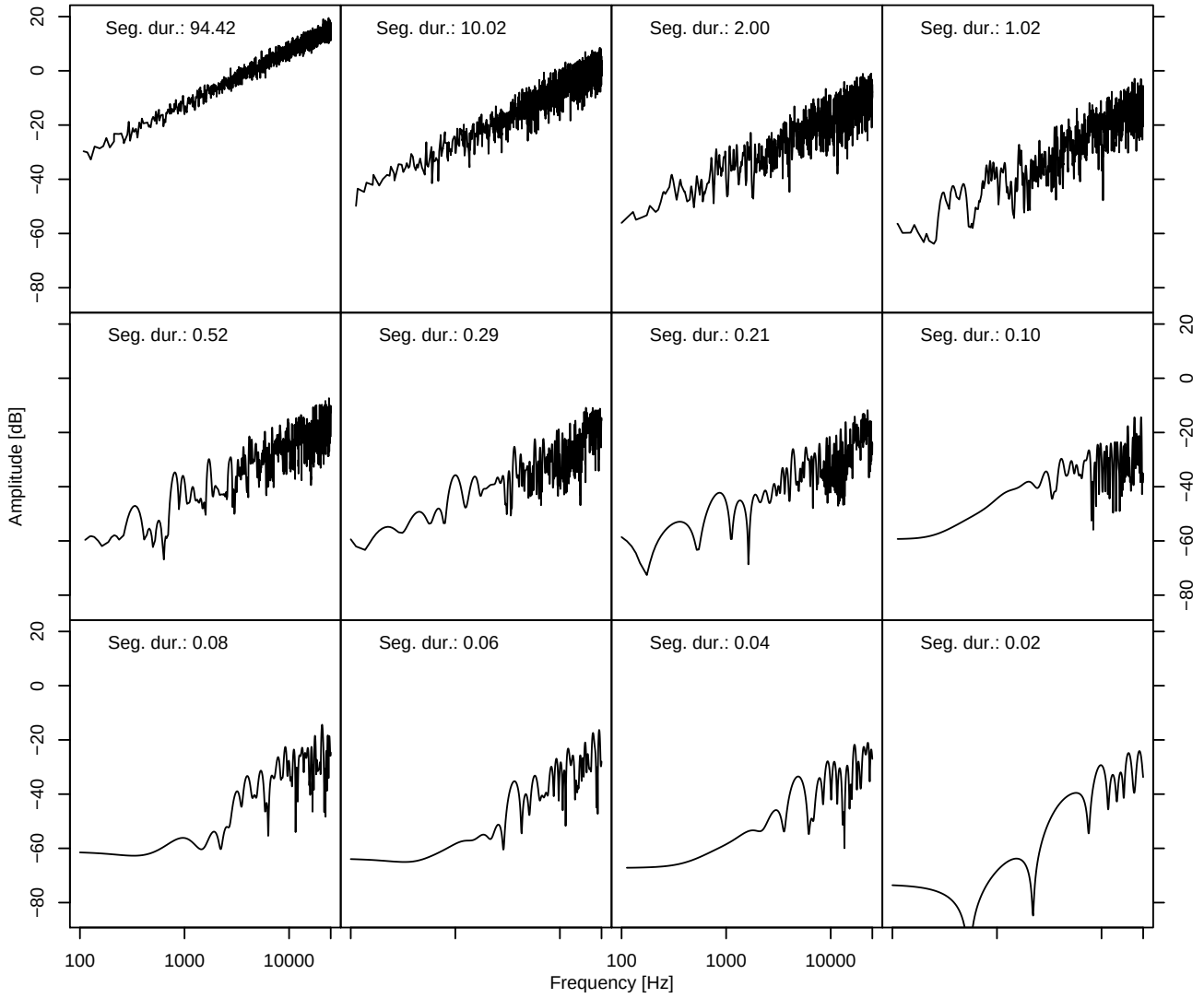


Figure B.4: Spectrum of digital stimuli from the 21NO condition for selected segment durations.

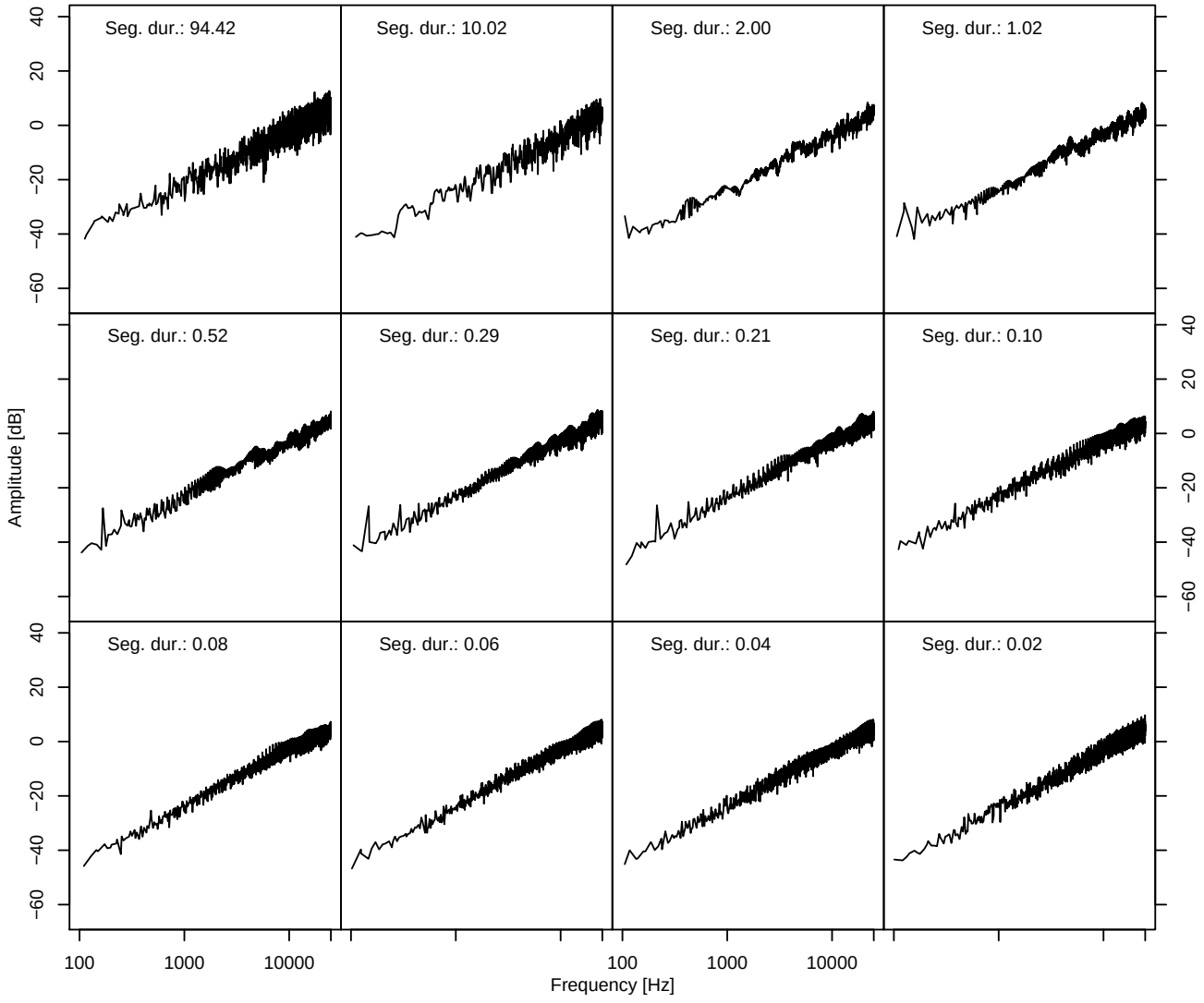


Figure B.5: Spectrum of digital stimuli from the 3EN condition for selected segment durations.

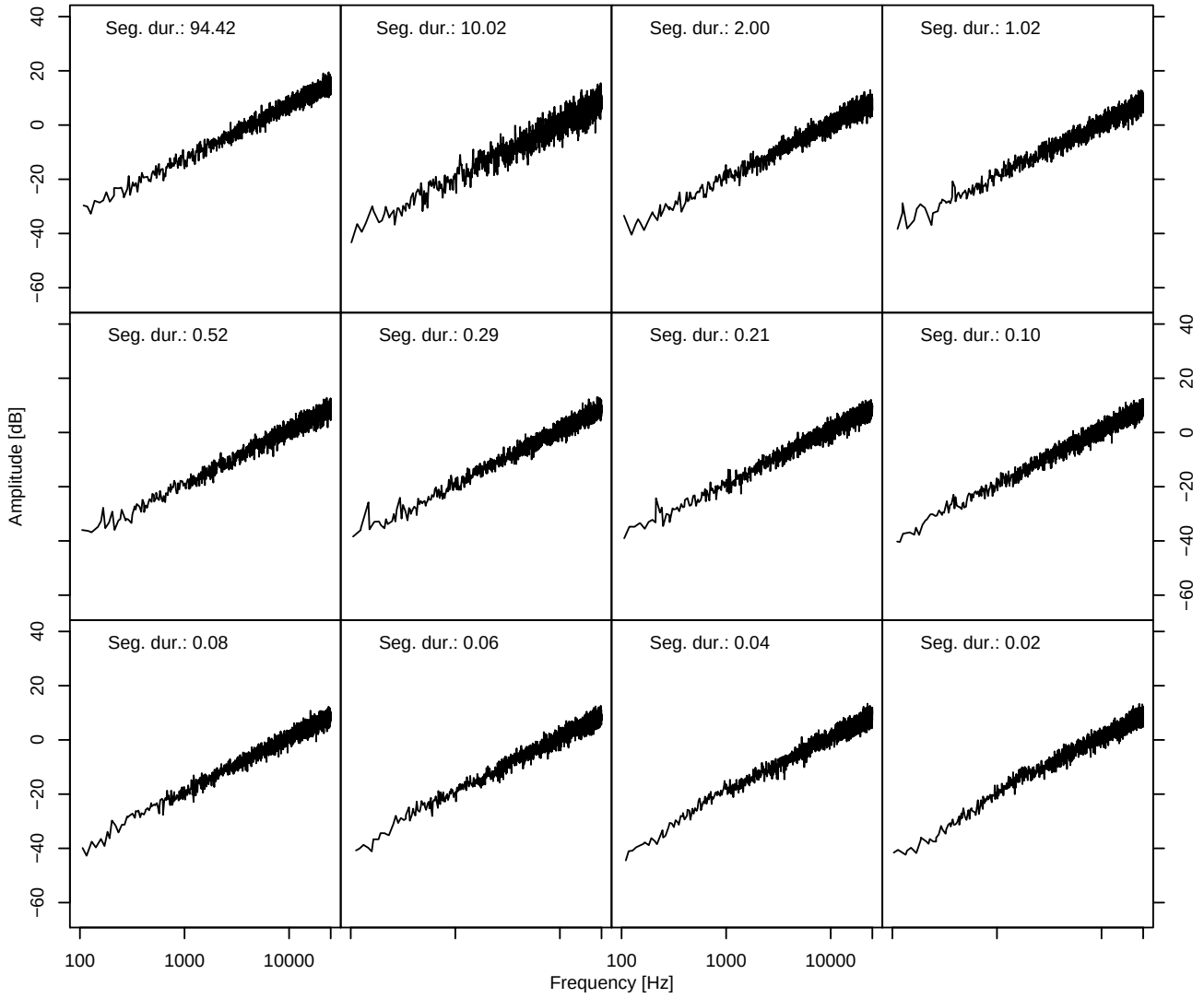


Figure B.6: Spectrum of digital stimuli from the 21EN condition for selected segment durations.

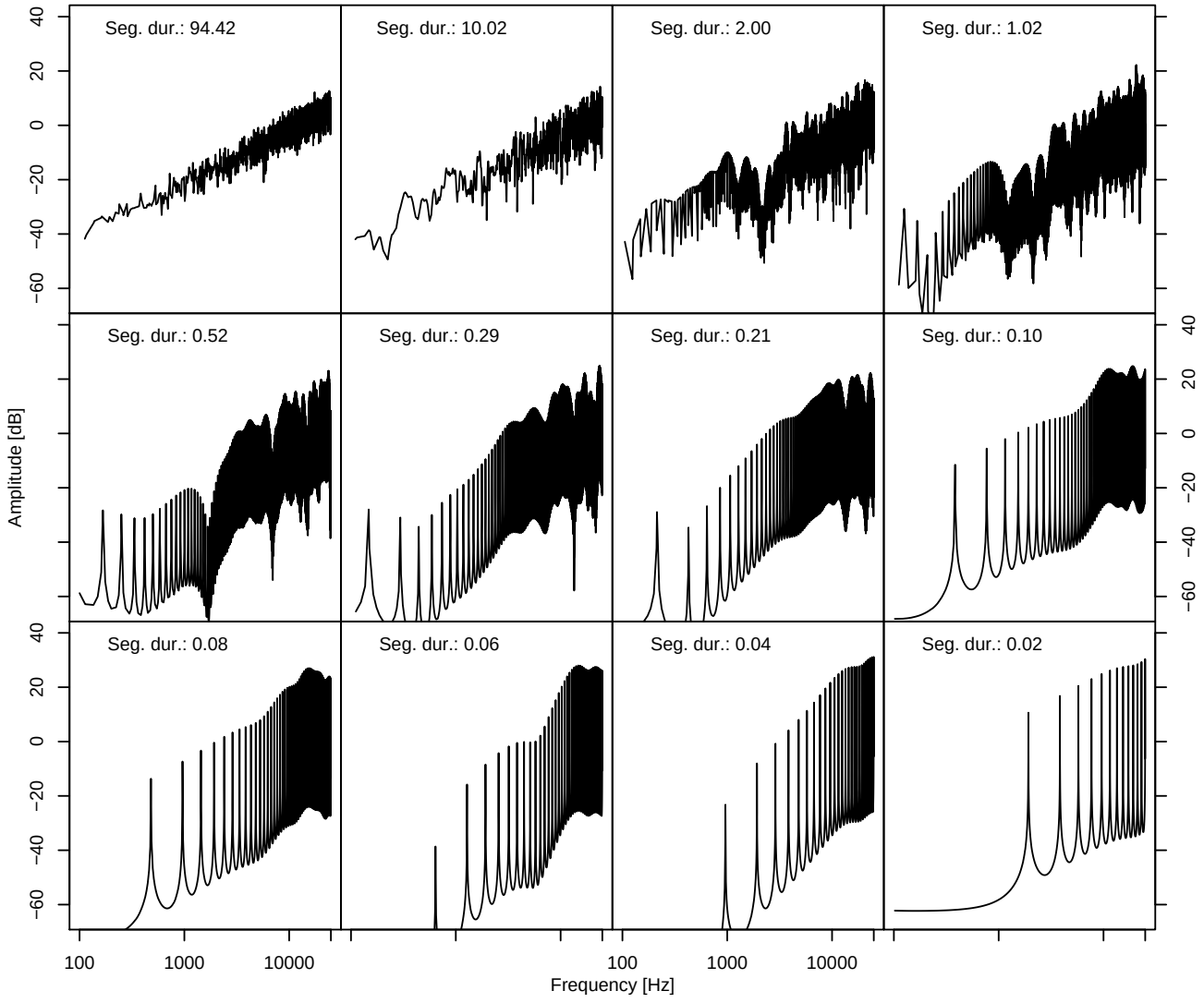


Figure B.7: Spectrum of digital stimuli from the 3EX condition for selected segment durations.

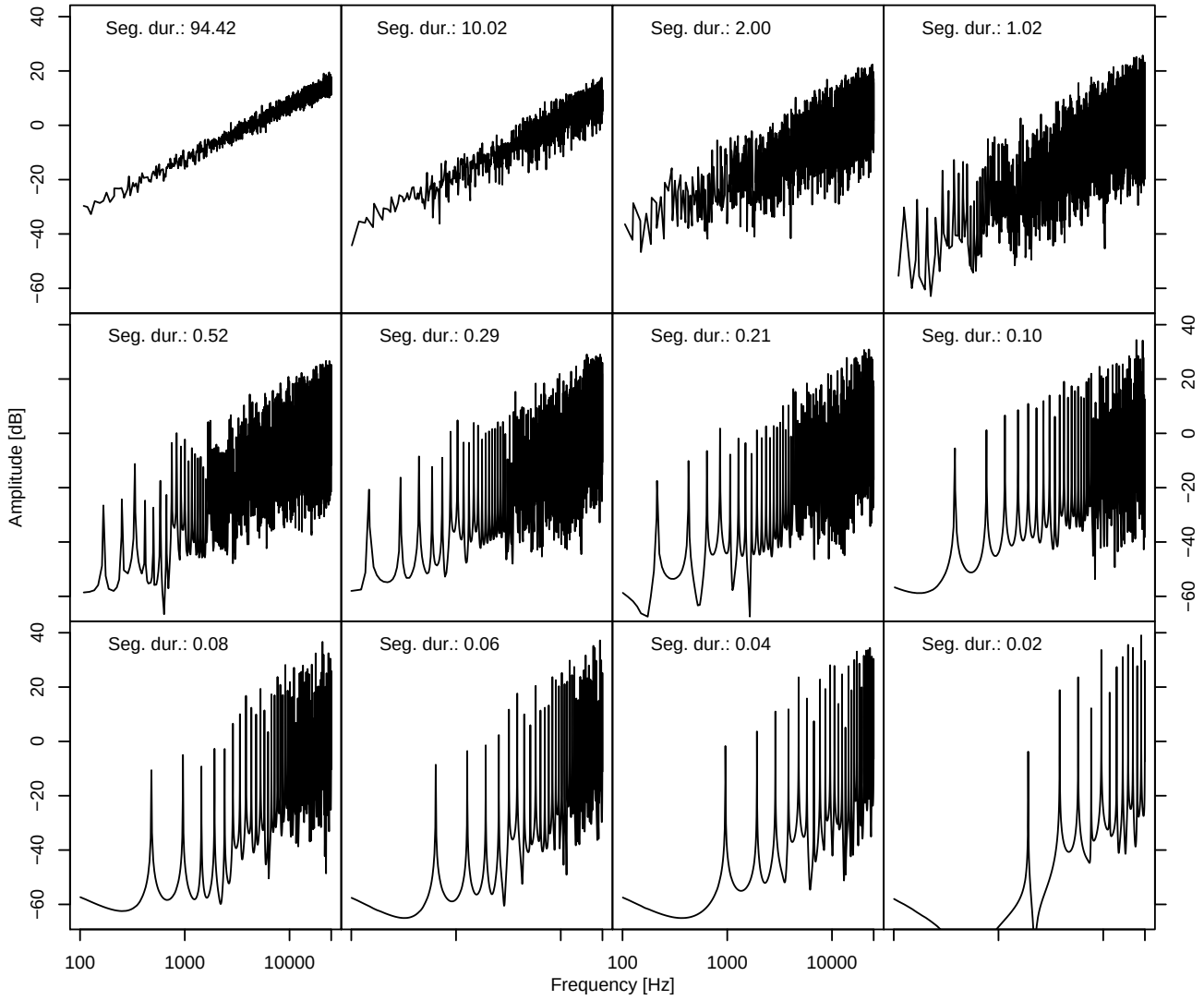


Figure B.8: Spectrum of digital stimuli from the 21EX condition for selected segment durations.

Amplitude spectra of acoustic stimuli

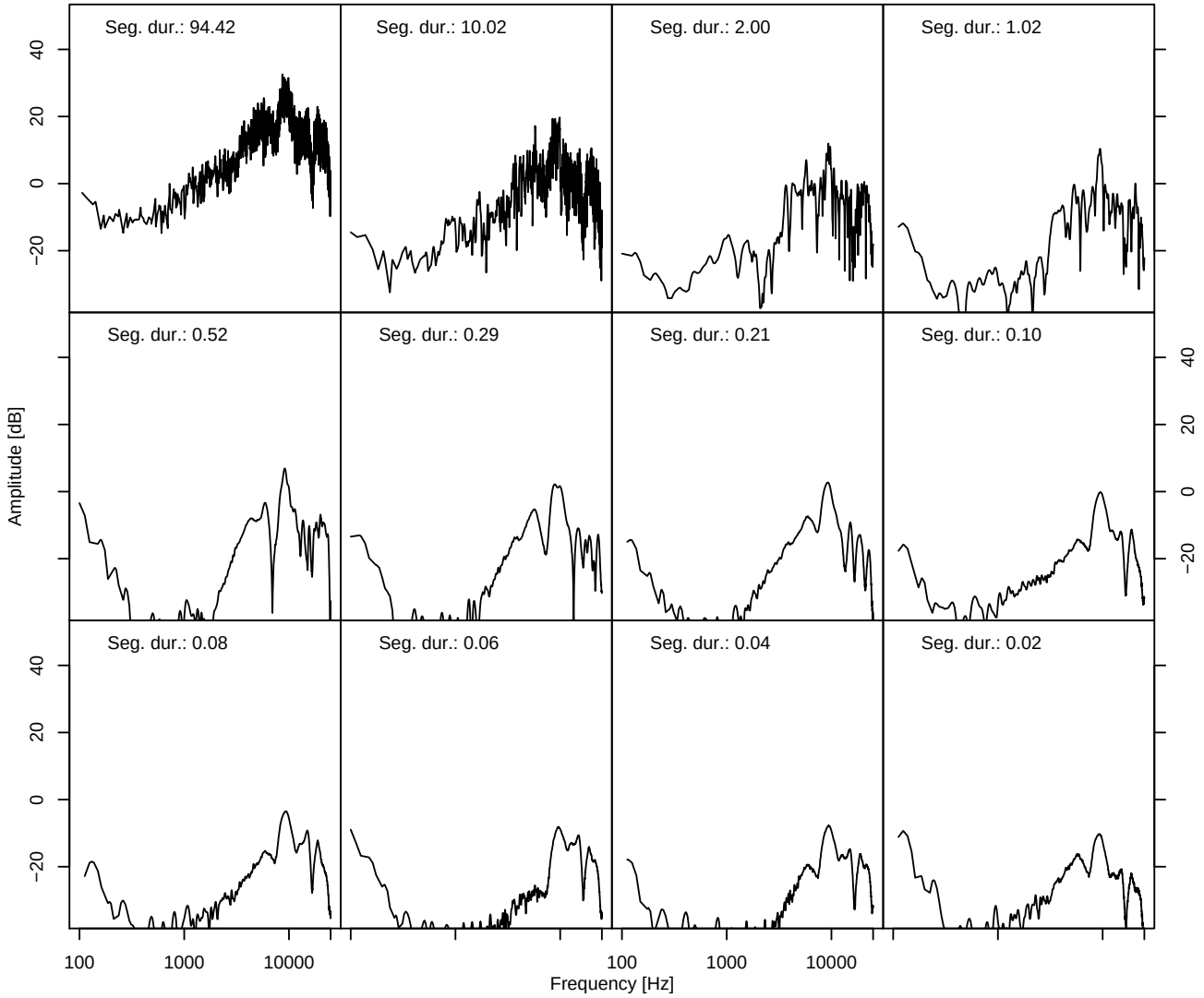


Figure B.9: Amplitude spectrum of acoustic stimuli from the 3NO condition for selected segment durations.

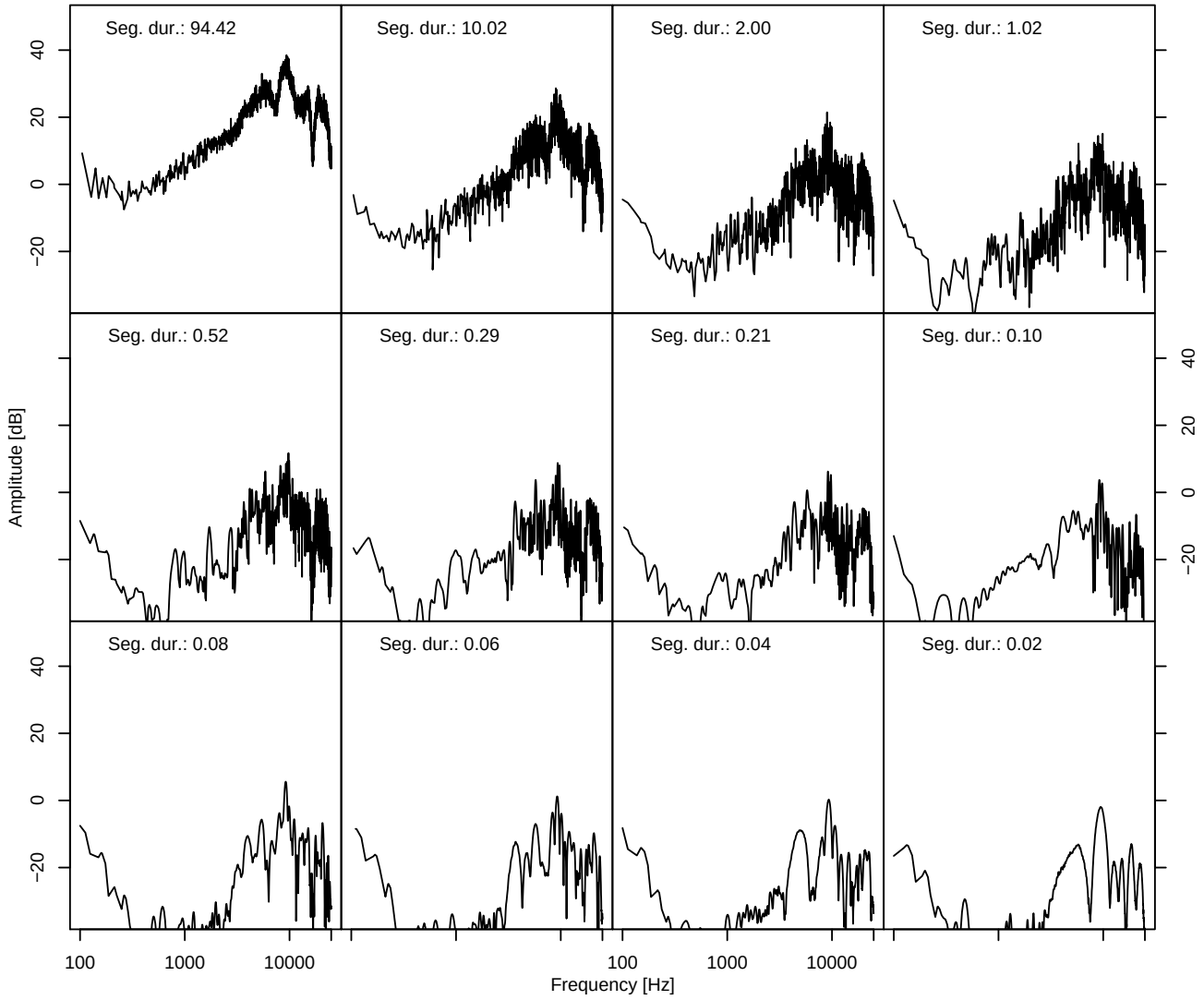


Figure B.10: Amplitude spectrum of acoustic stimuli from the 21NO condition for selected segment durations.

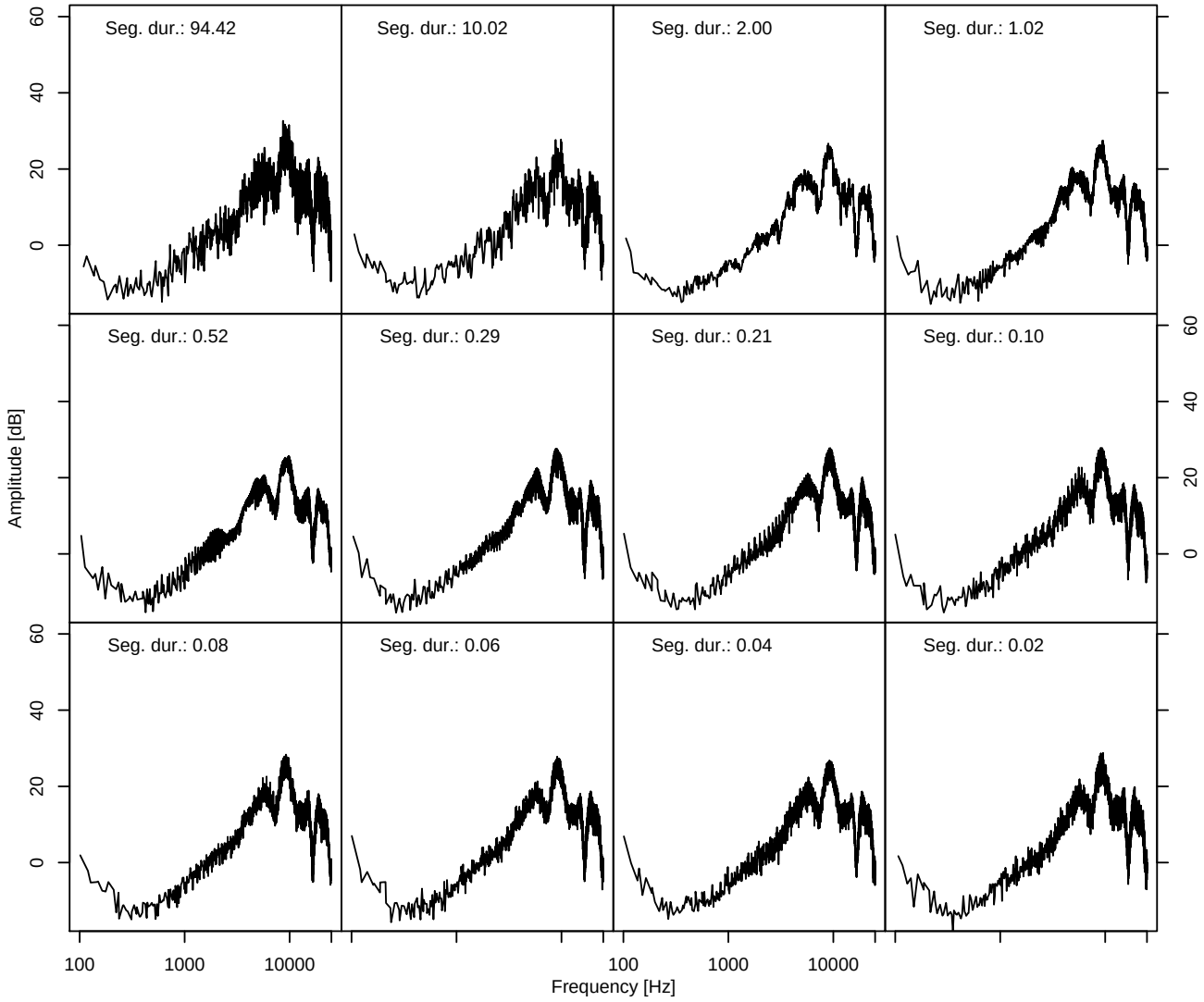


Figure B.11: Amplitude spectrum of acoustic stimuli from the 3EN condition for selected segment durations.

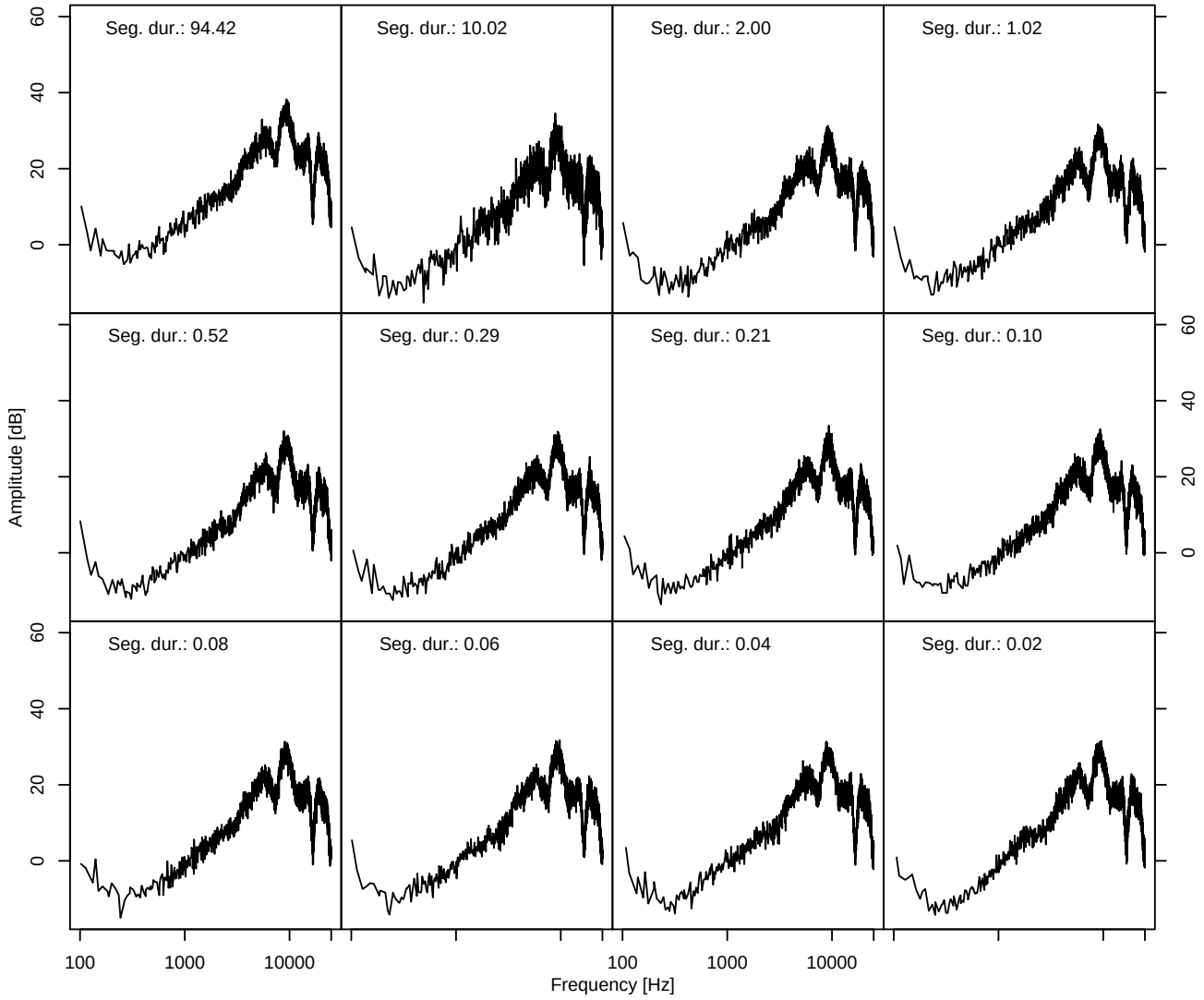


Figure B.12: Amplitude spectrum of acoustic stimuli from the 21EN condition for selected segment durations.

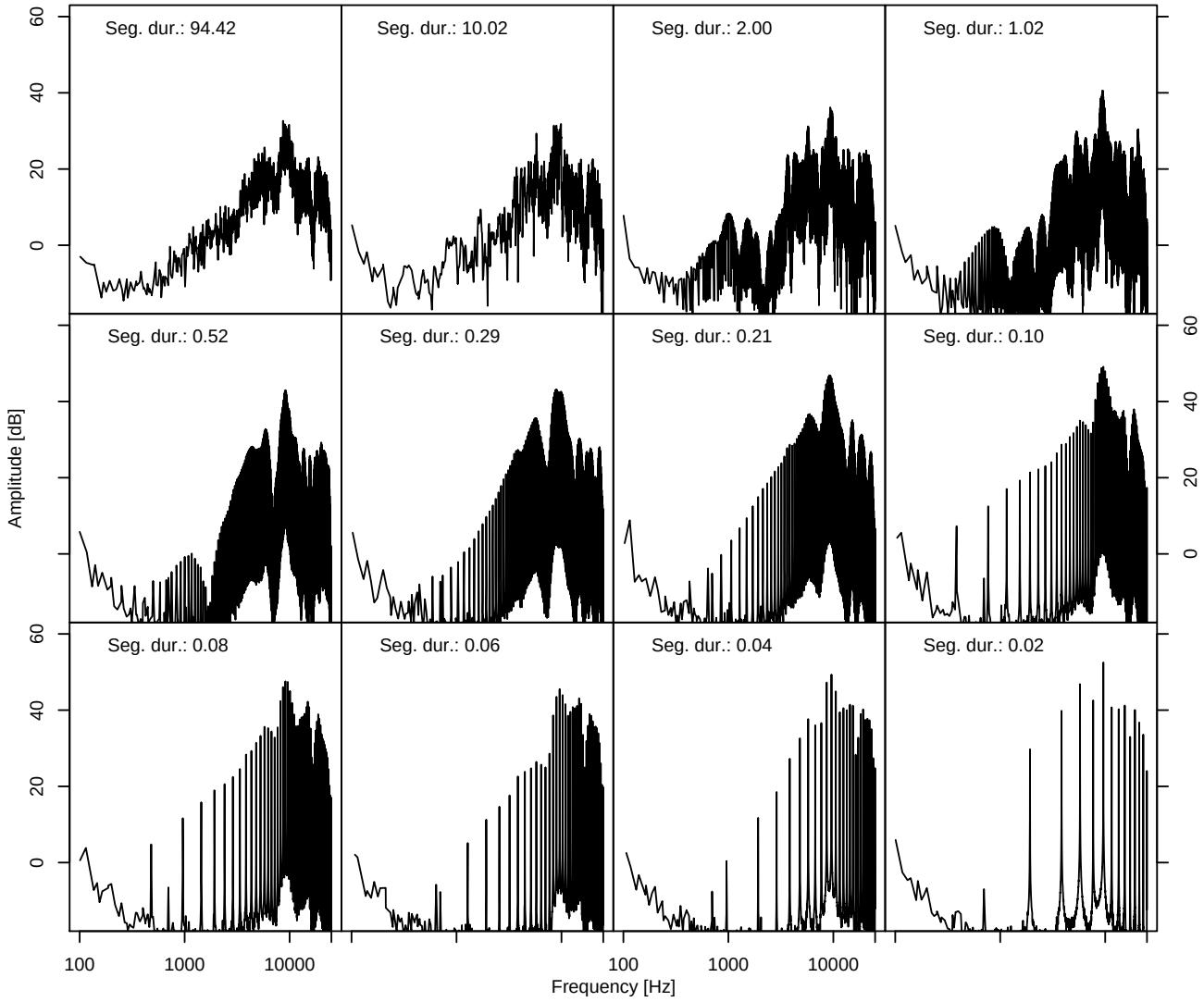


Figure B.13: Amplitude spectrum of acoustic stimuli from the 3EX condition for selected segment durations.

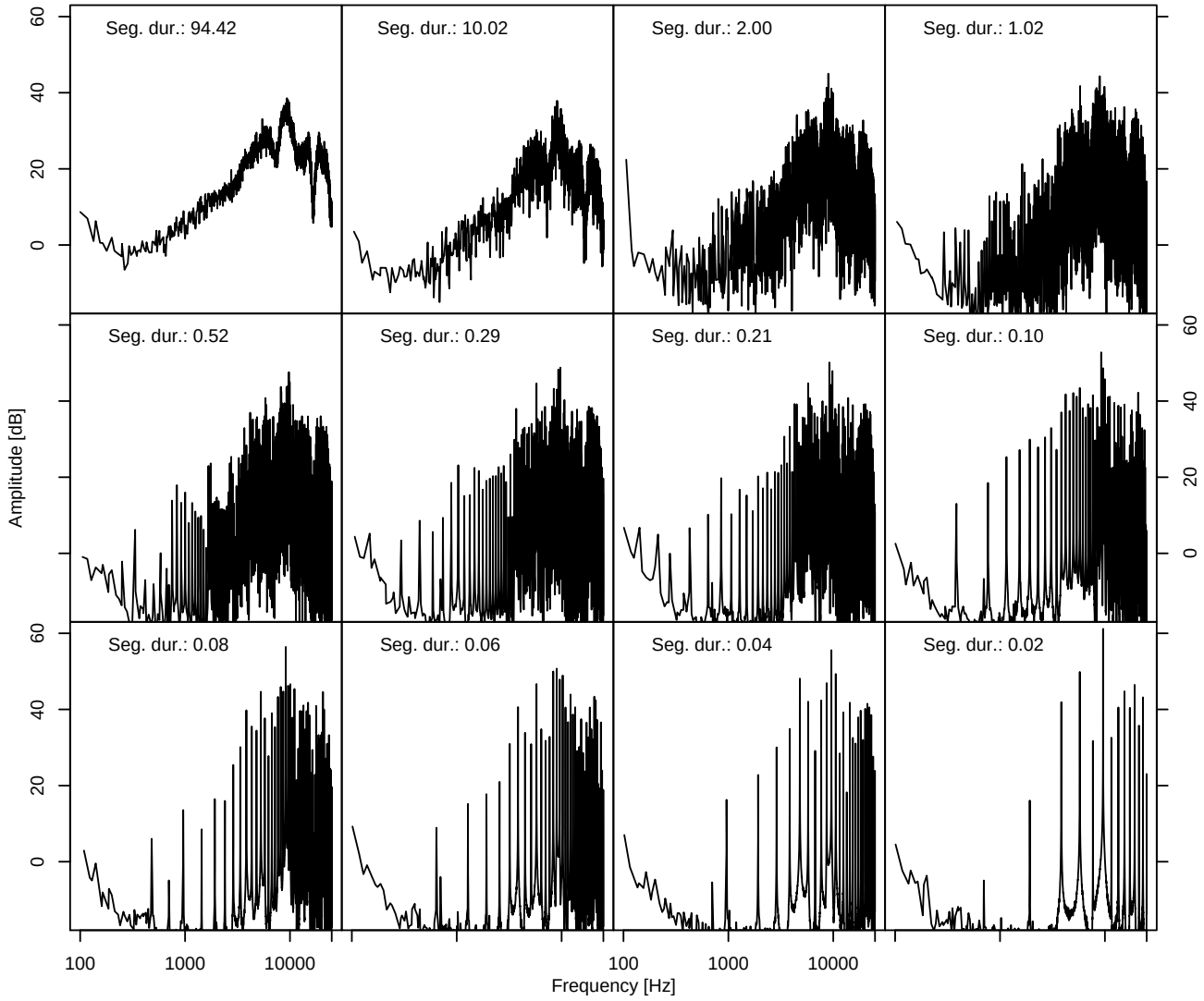


Figure B.14: Amplitude spectrum of acoustic stimuli from the 21EX condition for selected segment durations.

Difference in amplitude spectra of acoustic stimuli where the target is descending and ascending

As already mentioned, one crucial assumption is that within one trial of the experiment, the sounds, descending or ascending, have the same long-term spectrum. When the extent of the time-window of the frequency analysis becomes sufficiently short (comparable to the segment duration) there will of course be differences in the spectrum of descending and ascending patterns. As described earlier, only the spectrum of entire sounds is considered in this chapter. Still it is of interest to verify that no difference exists between the spectra due to for example distortion introduced by the physical system.

Especially the figures which represent the stimuli of the conditions where the listeners showed remarkable good performance is of interest, that is, the EX conditions at segment durations smaller than 1 ms.

When interpreting the graphs it is important to realize that the subtraction of the two log-transformed spectra is identical to a division of the non-transformed spectra. This means, that when there is no power in the input spectrum, the fraction of the two spectra will be governed by noise in the recordings. The reason being that the denominator in that case is close to zero, and only small amount of noise in the spectrum of the numerator changes the fraction dramatically. For this reason, the figures should be considered with the noise-free input spectrum of the stimuli as shown in chapter B.1.3 on the side, for identifying regions where noise is dominant.

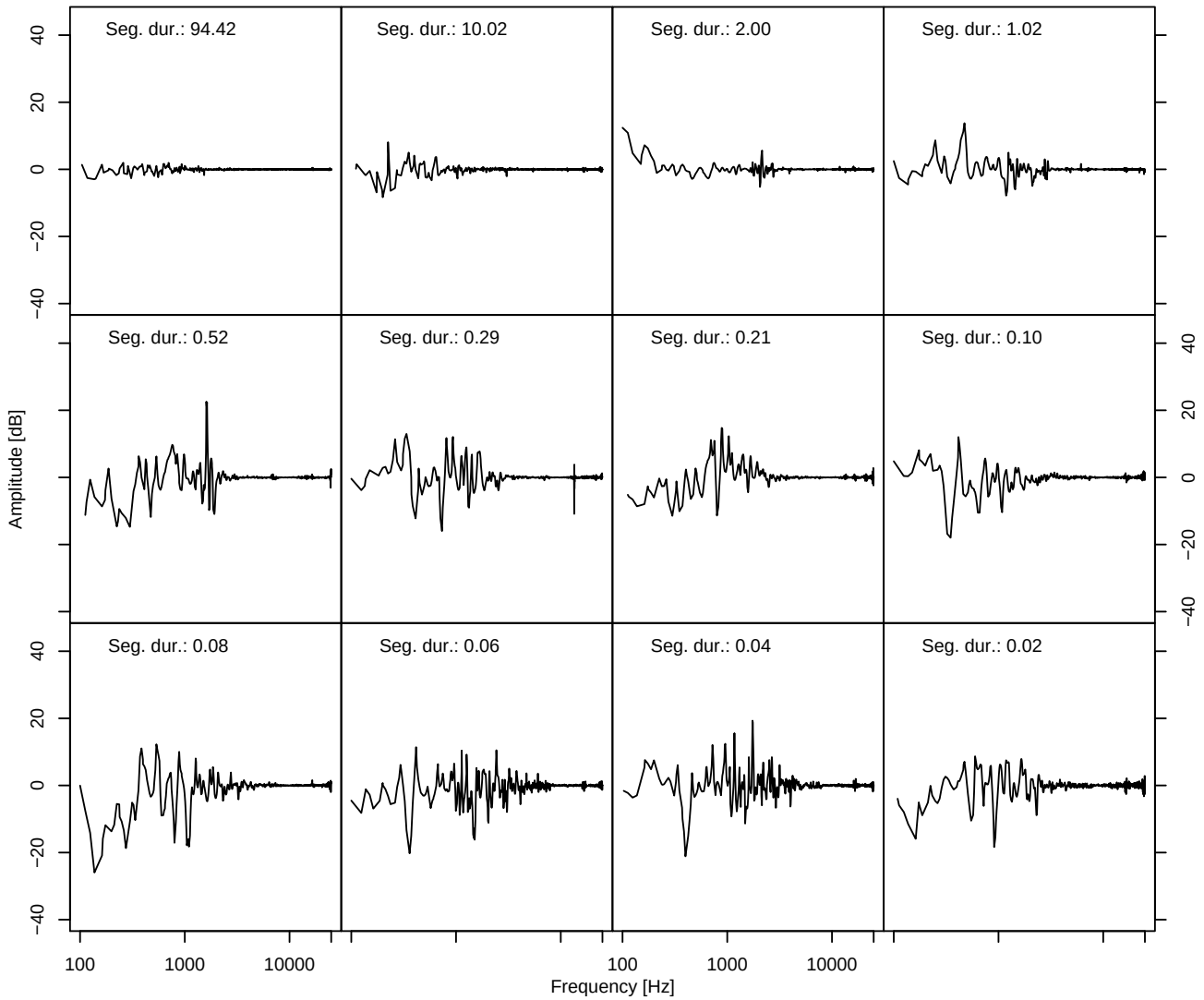


Figure B.15: Difference in the amplitude spectrum of acoustic stimuli from the 3NO condition for selected segment durations. Amplitude spectrum of descending pattern is subtracted from the spectrum of ascending pattern.

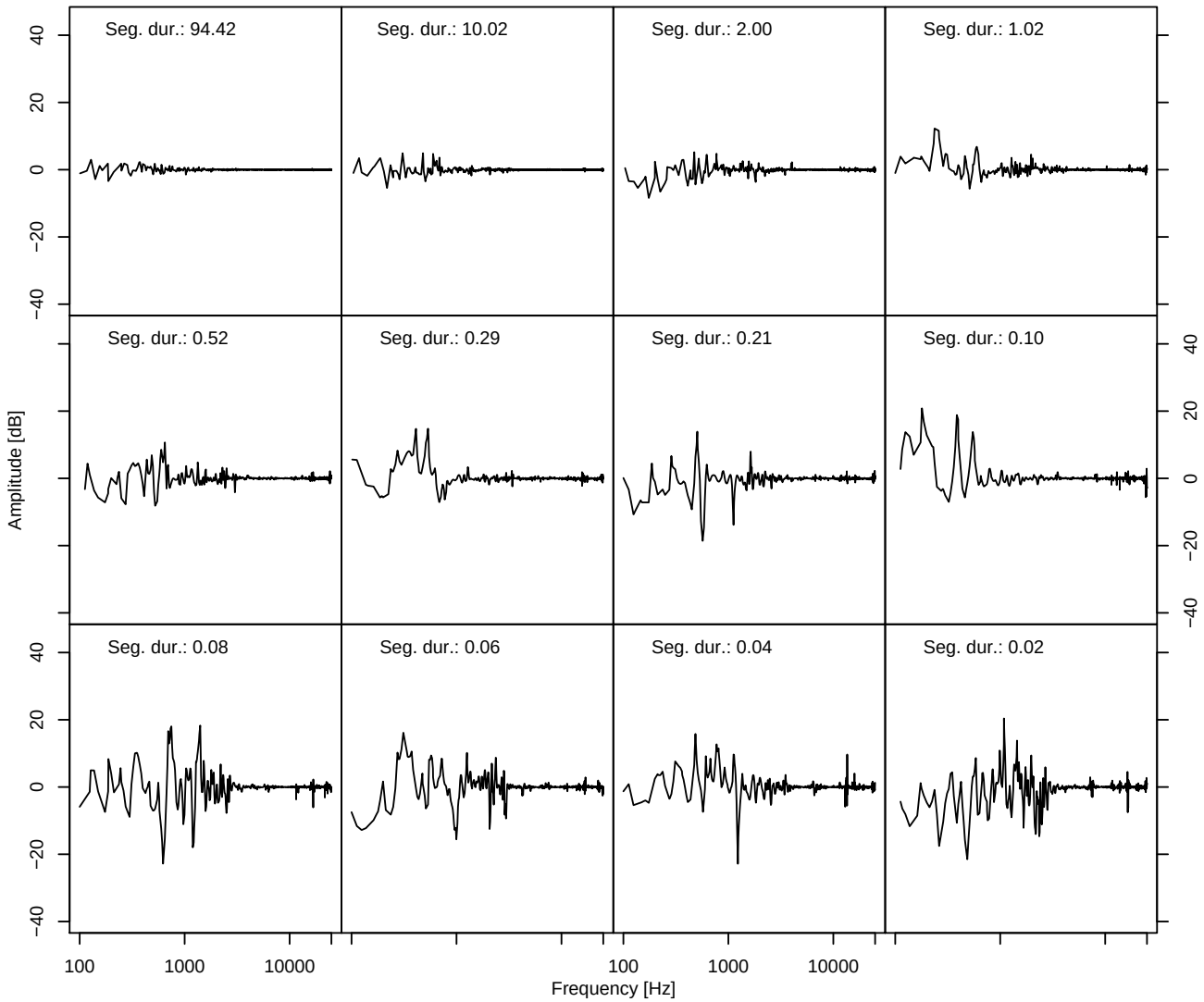


Figure B.16: Difference in the amplitude spectrum of acoustic stimuli from the 21NO condition for selected segment durations. Amplitude spectrum of descending pattern is subtracted from the spectrum of ascending pattern.

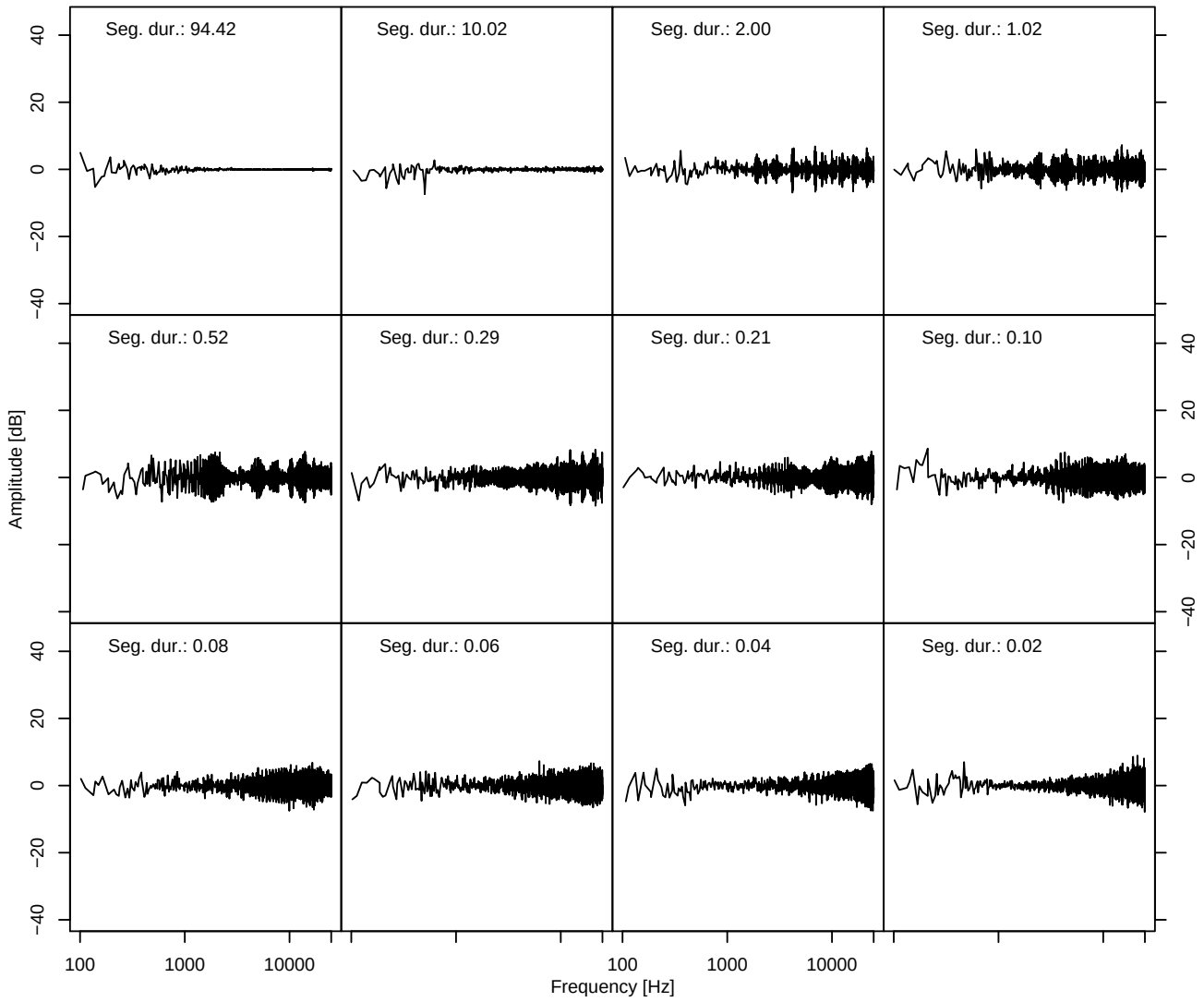


Figure B.17: Difference in the amplitude spectrum of acoustic stimuli from the 3EN condition for selected segment durations. Amplitude spectrum of descending pattern is subtracted from the spectrum of ascending pattern.

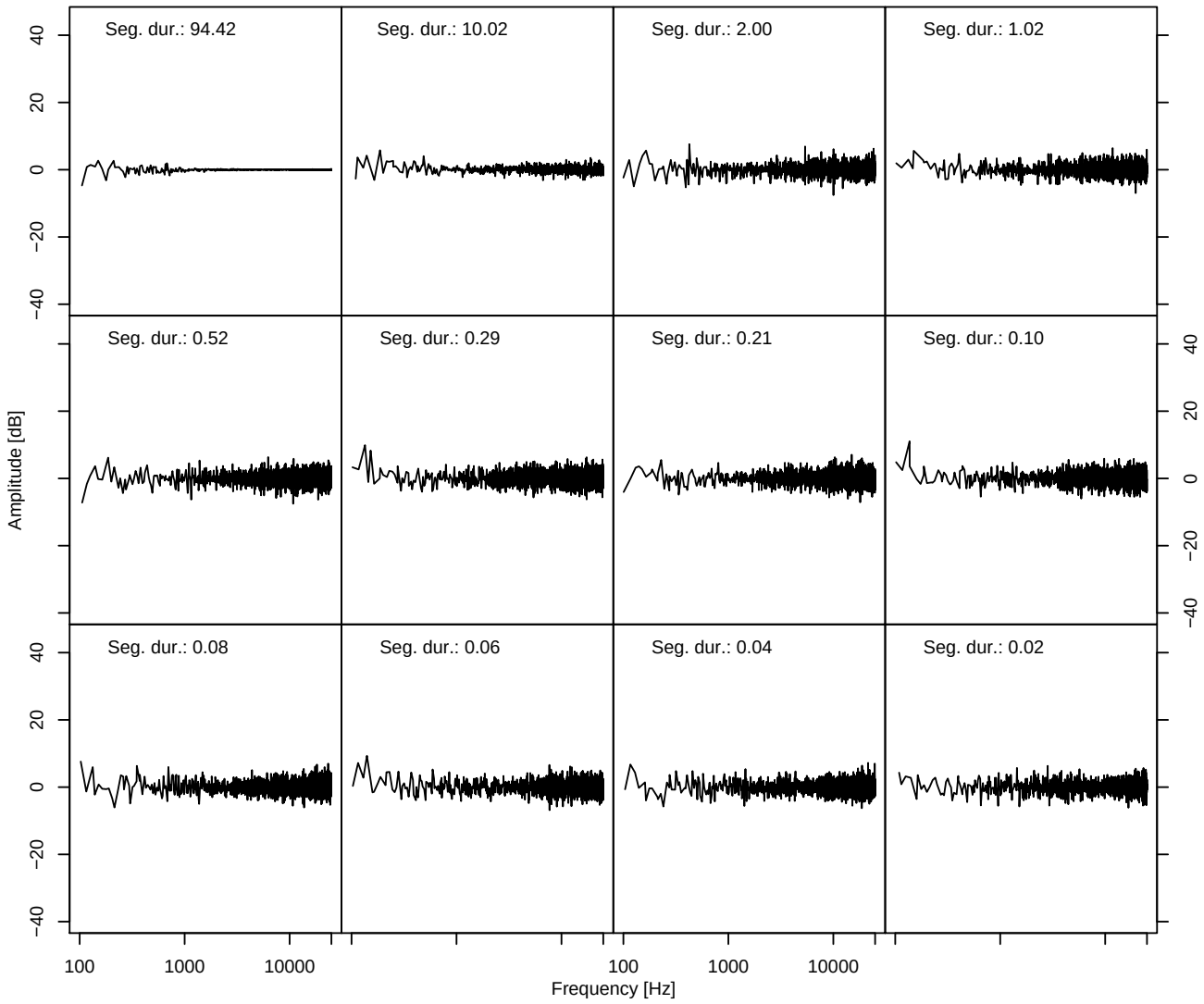


Figure B.18: Difference in the amplitude spectrum of acoustic stimuli from the 21EN condition for selected segment durations. Amplitude spectrum of descending pattern is subtracted from the spectrum of ascending pattern.

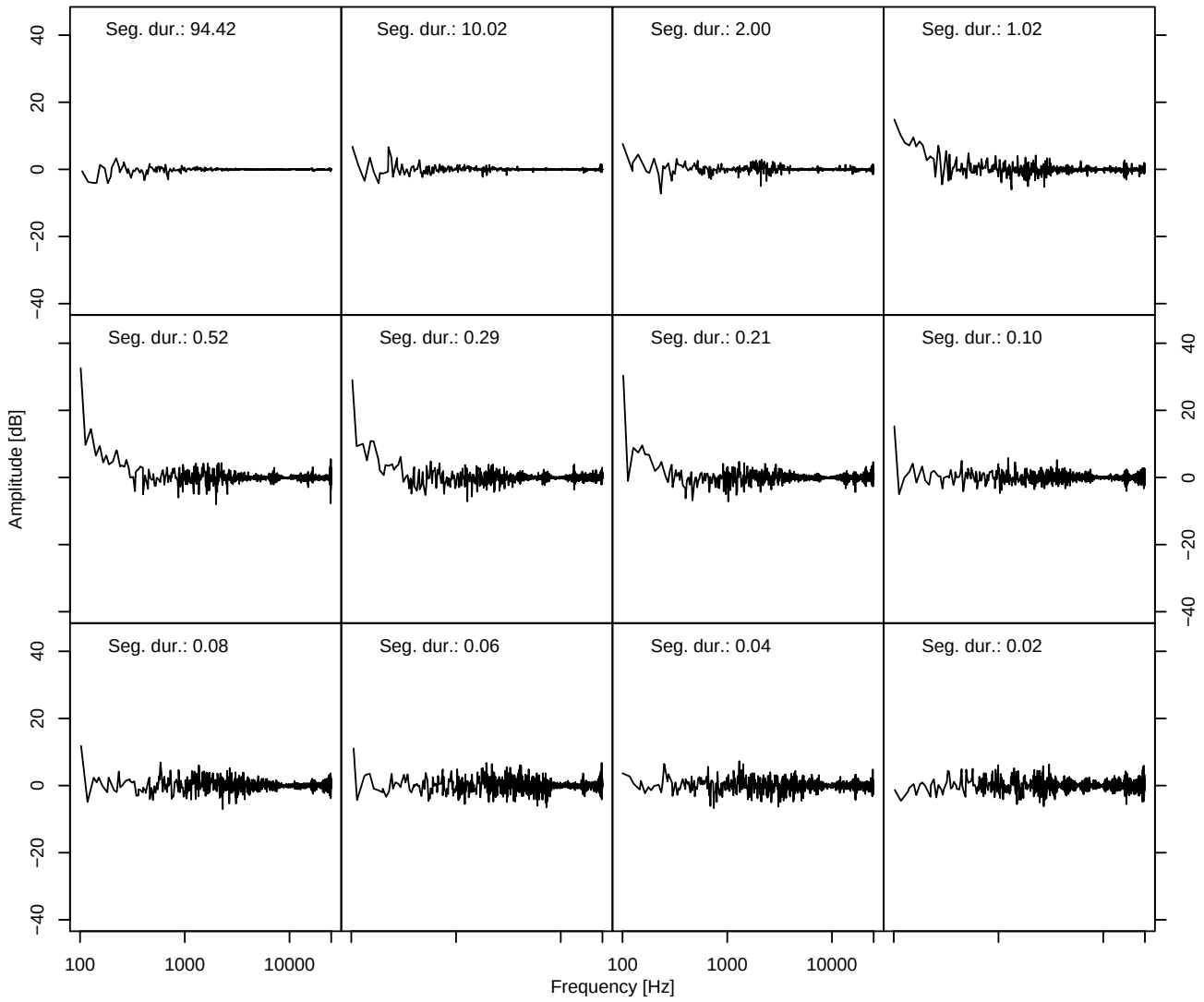


Figure B.19: Difference in the amplitude spectrum of acoustic stimuli from the 3EX condition for selected segment durations. Amplitude spectrum of descending pattern is subtracted from the spectrum of ascending pattern.

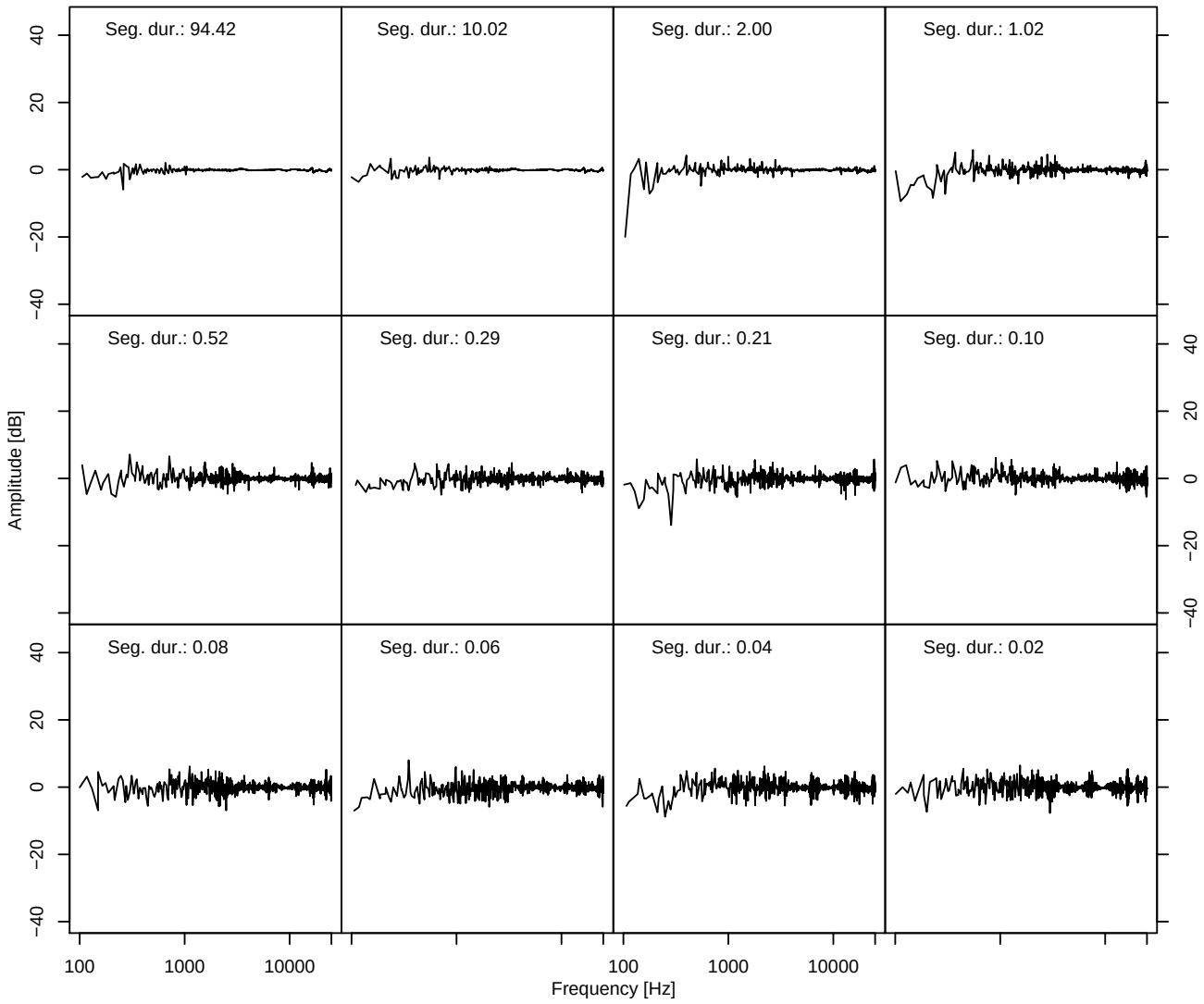


Figure B.20: Difference in the amplitude spectrum of acoustic stimuli from the 21EX condition for selected segment durations. Amplitude spectrum of descending pattern is subtracted from the spectrum of ascending pattern.

Bibliography

ANSI S3.4 (**2005**). *American national standard procedure for the computation of loudness of steady sounds*, American National Standards Institute, New York.

Berg, B. G. (**1989**). "Analysis of weights in multiple observation tasks", J. Acoust. Soc. Am. **86**, 1743–1746.

Blauert, J. (**1972**). "On the lag of lateralization caused by interaural time and intensity differences", Audiology **11**, 265–270.

Blauert, J. (**1999**). *Spatial hearing: The psychophysics of human sound localization*, 2nd edition (Cambridge, Mass, MIT Press).

Buus, S. (**1999**). "Temporal integration and multiple looks, revisited: Weights as a function of time.", J. Acoust. Soc. Am. **105**, 2466–2475.

Buus, S., Florentine, M., and Poulsen, T. (**1997**). "Temporal integration of loudness, loudness discrimination, and the form of the loudness function.", J. Acoust. Soc. Am. **101**, 669–680.

Cohen, J. (**2003**). *Applied multiple regression/correlation analysis for the behavioral sciences*, 3rd edition (Lawrence Erlbaum, Mahwah, NJ).

DeCarlo, L. T. and Cross, D. V. (**1990**). "Sequential effects in magnitude scaling: Models and theory", Journal of Experimental Psychology **119**, 375–396.

Divenyi, P. L. and Hirsh, I. J. (**1974**). "Identification of temporal order in three-tone sequences.", J. Acoust. Soc. Am. **56**, 144–151.

Doherty, K. A. and Lutfi, R. A. (**1996**). "Spectral weights for overall level discrimination in listeners with sensorineural hearing loss.", J. Acoust. Soc. Am. **99**, 1053–1058.

Doherty, K. A. and Lutfi, R. A. (**1999**). "Level discrimination of single tones in a multitone complex by normal-hearing and hearing-impaired listeners.", J. Acoust. Soc. Am. **105**, 1831–1840.

Durlach, N. I., Mason, C. R., Kidd, G., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (**2003**). "Note on informational masking.", J. Acoust. Soc. Am. **113**, 2984–2987.

Ellermeier, W. and Schrödl, S. (**2000**). "Temporal weights in loudness summation", in *C. Bonnet (Ed.), Fechner Day 2000. Proceedings of the 16th annual meeting of the International Society for Psychophysics*, 169–173 (Université Louis Pasteur, Strasbourg).

Farmer, M. E. and Klein, R. M. (**1995**). "The evidence for a temporal processing deficit linked to dyslexia: A review", Psychonomic bulletin & review **2**, 460–493.

Gerken, G. M., Bhat, V. K., and Hutchison-Clutter, M. (**1990**). "Auditory temporal integration and the power function model.", J. Acoust. Soc. Am. **88**, 767–778.

Glasberg, B. R. and Moore, B. C. J. (**2002**). "A model of loudness applicable to time-varying sounds", J. Audio Eng. Soc. **50**, 331–342.

Grimm, G., Hohmann, V., and Verhey, J. L. (**2002**). "Loudness of fluctuating sounds", Acust. Acta Acust. **88**, 359–368.

Hellman, W. S. and Hellman, R. P. (**2001**). "Revisiting relations between loudness and intensity discrimination.", J. Acoust. Soc. Am. **109**, 2098–2102.

Henning, G. B. and Gaskell, H. (**1981**). "Monaural phase sensitivity with Ronken's paradigm", J. Acoust. Soc. Am. **70**, 1669–1673.

Insightful (**2005**). *S-PLUS 7, Guide to statistics, Volume I* (Insightful Corporation, Seattle, Washington).

Kinsler, L. E., ed. (**2000**). *Fundamentals of acoustics*, 4th edition (Wiley, New York).

Klumpp, R. G. and Eady, H. R. (**1956**). "Some measurements of interaural time difference thresholds", J. Acoust. Soc. Am. **28**, 859–860.

Kohfeld, D. L., Santee, J. L., and Wallace, N. D. (**1981**). "Loudness and reaction time: I.", Percept. Psychophys. **29**, 535–549.

Kortekaas, R., Buus, S., and Florentine, M. (**2003**). "Perceptual weights in auditory level discrimination.", J. Acoust. Soc. Am. **113**, 3306–3322.

Krumbholz, K., Bleeck, S., Patterson, R. D., Senokozlieva, M., Seither-Preisler, A., and Lütkenhöner, B. (**2005**). "The effect of cross-channel synchrony on the perception of temporal regularity.", J. Acoust. Soc. Am. **118**, 946–954.

Lutfi, R. A. (**1990**). "Informational processing of complex sound. II. Cross-dimensional analysis.", J. Acoust. Soc. Am. **87**, 2141–2148.

Lutfi, R. A. (**1995**). "Correlation-coefficients and correlation ratios as estimates of observer weights in multiple-observation tasks", J. Acoust. Soc. Am. **97**, 1333–1334.

Lutfi, R. A. and Jesteadt, W. (**2005**). "Test of the linearity assumption underlying coss estimates of listener decision weights", in *Abstracts of the twenty-eighth annual midwinter research meeting of the Association for Research in Otolaryngology*, 360 (New Orleans, Louisiana).

McFarland, D. J. and Cacace, A. T. (**1992**). "Aspects of short-term acoustic recognition memory: modality and serial position effects.", Audiology **31**, 342–352.

Moore, B. C. J. (**2003**a). *An Introduction to the Psychology of Hearing*, 5th edition (Academic Press, San Diego, CA).

Moore, B. C. J. (**2003**b). "Temporal integration and context effects in hearing", Journal of Phonetics **31**, 563–574.

Moore, B. C. J., Glasberg, B. R., and Stone, M. A. (**2003**). "Why are commercials so loud? perception and modeling of the loudness of amplitude-compressed speech", J. Audio Eng. Soc. **51**, 1123–1132.

Neath, I. (**1993**). "Distinctiveness and serial position effects in recognition.", Mem. Cognit. **21**, 689–698.

Neath, I., Brown, G. D. A., McCormack, T., Chater, N., and Freeman, R. (**2006**). "Distinctiveness models of memory and absolute identification: evidence for local, not global, effects.", Q. J. Exp. Psychol. **59**, 121–135.

Neubauer, H. and Heil, P. (**2004**). "Towards a unifying basis of auditory thresholds: the effects of hearing loss on temporal integration reconsidered.", J. Assoc. Res. Otolaryngol. **5**, 436–458.

Näätänen, R. and Winkler, I. (**1999**). "The concept of auditory stimulus representation in cognitive neuroscience.", Psychol. Bull. **125**, 826–859.

Oberfeld, D. and Plank, T. (**2005**). "Temporal weighting of loudness: Effects of fade in", in *Fortschritte der Akustik - DAGA 2005*.

Oppenheim, A. V., Schafer, R. W., and Buck, J. R. (**1999**). *Discrete-time signal processing*, 2nd edition (Prentice Hall, Englewood Cliffs, NJ).

Oxenham, A. J. and Moore, B. C. (**1994**). "Modeling the additivity of nonsimultaneous masking.", Hear. Res. **80**, 105–118.

Patterson, R. D. (**1994**a). "The sound of a sinusoid: Spectral models", J. Acoust. Soc. Am. **96**, 1409–1418.

Patterson, R. D. (**1994**b). "The sound of a sinusoid: Time-interval models", J. Acoust. Soc. Am. **96**, 1419–1428.

Pedersen, B. (**2006**a). "Temporal masking in the auditory identification of envelope patterns", in *Auditory Temporal Resolution and Integration: Stages of Analyzing Time-Varying Sounds*, 67–82 (Aalborg University).

Pedersen, B. (**2006**b). "Discrimination of temporal patterns on the basis of envelope and fine-structure cues", in *Auditory Temporal Resolution and Integration: Stages of Analyzing Time-Varying Sounds*, 85–96 (Aalborg University).

Pedersen, B. and Ellermeier, W. (**2006**). "Temporal weighting in loudness judgments of level-fluctuating sounds", in *Auditory Temporal Resolution and Integration: Stages of Analyzing Time-Varying Sounds*, 13–23 (Aalborg University).

Pfingst, B. E., Hienz, R., Kimm, J., and Miller, J. (**1975**). "Reaction-time procedure for measurement of hearing. I. Suprathreshold functions", J. Acoust. Soc. Am. **57**, 421–430.

Phillips, D. P., Hall, S. E., and Boehnke, S. E. (**2002**). "Central auditory onset responses, and temporal asymmetries in auditory perception.", Hear. Res. **167**, 192–205.

Plack, C. J., Oxenham, A. J., and Drga, V. (**2002**). "Linear and nonlinear processes in temporal masking", Acust. Acta Acust. **88**, 348–358.

Ronken, D. A. (**1970**). "Monaural detection of a phase difference between clicks", J. Acoust. Soc. Am. **47**, 1091–&.

Saberi, K. (**1996**). "Observer weighting of interaural delays in filtered impulses.", Percept. Psychophys. **58**, 1037–1046.

Schäffler, T., Sonntag, J., Hartnegg, K., and Fischer, B. (**2004**). "The effect of practice on low-level auditory discrimination, phonological skills, and spelling in dyslexia.", Dyslexia **10**, 119–130.

Southworth, C. and Berg, B. G. (**1995**). "Multiple cues for the discrimination of narrow-band sounds", J. Acoust. Soc. Am. **98**, 2486–2492.

Stecker, G. C. and Hafter, E. R. (**2000**). "An effect of temporal asymmetry on loudness.", J. Acoust. Soc. Am. **107**, 3358–3368.

Stecker, G. C. and Hafter, E. R. (**2002**). "Temporal weighting in sound localization.", J. Acoust. Soc. Am. **112**, 1046–1057.

Stellmack, M. A., Viemeister, N. F., and Byrne, A. J. (**2005**). "Comparing monaural and interaural temporal windows: effects of a temporal fringe on sensitivity to intensity differences.", J. Acoust. Soc. Am. **118**, 3218–3228.

Studdert-Kennedy, M. and Mody, M. (**1995**). "Auditory temporal perception deficits in the reading-impaired: A critical review of the evidence", Psychonomic Bulletin & Review **2**, 508–514.

Surprenant, A. M. (**2001**). "Distinctiveness and serial position effects in tonal sequences.", Percept. Psychophys. **63**, 737–745.

Tang, Z., Richards, V. M., and Shih, A. (**2005**). "Comparing linear regression models applied to psychophysical data", J. Acoust. Soc. Am. **117**, 2597.

Venables, W. and Ripley, B. (**2002**). *Modern applied statistics with S*, 4th edition (Springer, New York).

Viemeister, N. F., Rickert, M., Law, M., and Stellmack, M. (**2002**). "Psychophysical and physiological aspects of auditory temporal processing", in *Genetics and the Function of the Auditory System, Proceedings of the 19th Danavox Symposium, edited by L. Tranebjaerg, J. Christensen-Dalsgaard, T. Andersen, and T. Poulsen*, 273–291 (Holmens Trykkeri, Denmark).

Viemeister, N. F. and Wakefield, G. H. (**1991**). "Temporal integration and multiple looks.", J. Acoust. Soc. Am. **90**, 858–865.

Vogel, A. (**1975**). "Common model for loudness and roughness", Biological Cybernetics **18**, 31–40.

Wagner, E., Florentine, M., Buus, S., and McCormack, J. (**2004**). "Spectral loudness summation and simple reaction time.", J. Acoust. Soc. Am. **116**, 1681–1686.

Widmann, U., Lippold, R., and Fastl, H. (**1998**). "A computer program simulating post-masking for applications in sound analysis systems", in *Proceedings of NOISE-CON' 98*, 451–456 (Ypsilanti, Michigan).

Willihnganz, M. S., Stellmack, M. A., Lutfi, R. A., and Wightman, F. L. (**1997**). "Spectral weights in level discrimination by preschool children: synthetic listening conditions.", J. Acoust. Soc. Am. **101**, 2803–2810.

Zwicker, E. (**1977**). "Procedure for calculating loudnesss of temporally variable sounds.", J. Acoust. Soc. Am. **62**, 675–682.

Zwicker, E. and Fastl, H. (**1999**). *Psychoacoustics: Facts and models* (Springer, Berlin).

Zwislocki, J. (**1960**). "Theory of temporal auditory summation", J. Acoust. Soc. Am. **32**, 1046–1060.