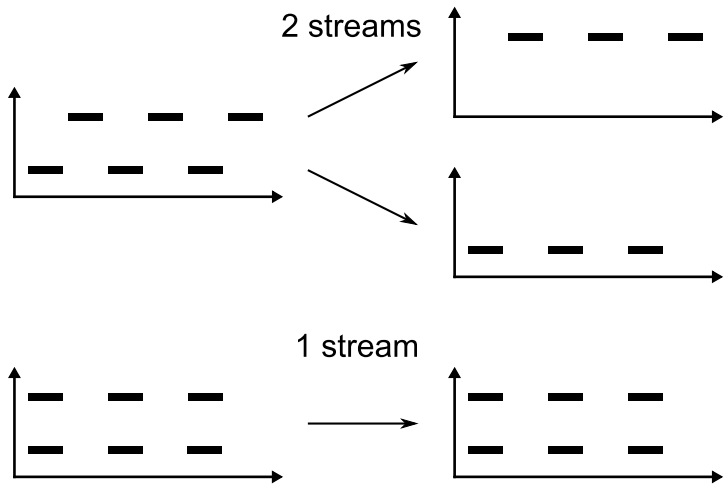
Technical University of Denmark

CONTRIBUTIONS TO
HEARING RESEARCH

Volume 17

Simon Krogholt Christiansen

The role of temporal coherence in auditory stream segregation



Centre for Applied Hearing Research
Department of Electrical Engineering

The role of temporal coherence in auditory stream segregation

PhD thesis by
Simon Krogholt Christiansen

Technical University of Denmark
2014

The thesis was defended on 28 November 2014.

This PhD-dissertation is the result of a research project at the Centre for Applied Hearing Research, Department of Electrical Engineering, Technical University of Denmark (Kgs. Lyngby, Denmark). Part of the project was carried out at the Auditory Perception and Cognition Lab, Department of Psychology, University of Minnesota (Minneapolis, MN, USA).

The project was financed by a stipend from the Oticon Foundation. The external subproject was supported by a grant from the U.S. National Institute of Health (R01 DC007657). The external stay at the University of Minnesota was further supported by travel grants from Knud Højgaard's Fund, Otto Mønsted's Fund and the Augustinus Fund.

Supervisors

Main supervisor

Prof. Torsten Dau
Centre for Applied Hearing Research
Department of Electrical Engineering
Technical University of Denmark
Kgs. Lyngby, Denmark

Co-supervisor

Morten L. Jepsen
Widex A/S
Lynge, Denmark

External advisor

Prof. Andrew J. Oxenham
Auditory Perception and Cognition Lab
Department of Psychology
University of Minnesota
Minneapolis, MN, USA

Abstract

The ability to perceptually segregate concurrent sound sources and focus one's attention on a single source at a time is essential for the ability to use acoustic information. While perceptual experiments have determined a range of acoustic cues that help facilitate auditory stream segregation, it is not clear how the auditory system realizes the task. This thesis presents a study of the mechanisms involved in auditory stream segregation. Through a combination of psychoacoustic experiments, designed to characterize the influence of acoustic cues on auditory stream formation, and computational models of auditory processing, the role of auditory preprocessing and temporal coherence in auditory stream formation was evaluated. The computational model presented in this study assumes that auditory stream segregation occurs when sounds stimulate non-overlapping neural populations in a temporally incoherent manner. In the presented model, a physiologically inspired model of auditory preprocessing and perception was used to transform a sound signal into an auditory representation, and a subsequent temporal coherence analysis grouped frequency-channels of the model together if they were stimulated in a temporally coherent manner. Based on this framework, the model was able to quantitatively predict perceptual experiments on stream segregation based on frequency separation and tone repetition rate, and onset and offset synchrony. Through the model framework, the influence of various processing stages on the stream segregation process was analysed. The model analysis showed that auditory frequency selectivity and physiological forward masking play a significant role in stream segregation based on frequency separation and tone rate. Secondly, the model analysis suggested that neural adaptation, and the resulting enhancement of neural responses to onsets, increases the sensitivity to onset synchrony for auditory stream formation. The effect of sound intensity on auditory stream formation was investigated, under the assumption that the wider auditory filters at high sound pressure levels should lead to a decreased ability to perceptually segregate sounds presented at high intensities. The results of listening experiments confirmed this hypothesis, showing that the minimum frequency separation required for stream segregation increases with increases in sound intensity. The computational model results also showed an increased tendency to group sounds presented at high intensities, but the size of the effect was overestimated relative to the experimental data, suggesting that the computational model does not fully reflect the auditory stream formation process. Lastly, an experimental paradigm designed to measure perceptual organization through an indirect, performance-based measure was investigated. This measure used comodulation masking release (CMR) to assess the conditions under which a loss of temporal coherence across frequency can lead to auditory stream segregation. The study indicated that CMR may be used as an indirect measure of stream segregation, and further supports the hypothesis that temporal coherence acts as a strong grouping cue. Overall, the findings of this thesis suggest that temporal coherence plays a significant role in the grouping of sounds into a single stream, and more generally, that a temporal coherence analysis may provide the framework for determining the perceptual organization of sounds into streams.

Resumé

Evnen til perceptuelt at adskille lydkilder og fokusere på én enkelt lydkilde ad gangen er essentiel for vores evne til at anvende akustisk information. Lytte-forsøg har identificeret en række akustiske parametre der påvirker vores evne til at adskille lydkilder, men det er endnu uvist hvordan vores auditive system rent faktisk udfører opgaven. Denne afhandling er et studie i de mekanismer der er involveret i auditiv lydkildeseparation. Gennem en kombination af lytte-forsøg og matematiske modeller af den menneskelige hørelse er indflydelsen af det perifere auditive system samt tidsmæssig kohærens søgt belyst. Den matematiske model anvendt i denne afhandling antager at auditiv lydkildeseparation opstår når lyde stimulerer separate populationer af neuroner inkohærent. I modellen anvendes en fysiologisk inspireret model af auditiv signalbehandling og lydopfattelse til at transformere et lydsignal til en auditiv repræsentation af lyden, hvorefter en kohærens-analyse anvendes til at gruppere frekvens-kanaler i modellen hvis de er kohærente. Baseret på denne konstruktion er modellen i stand til kvantitativt at forudsige auditiv lydkildeseparation baseret på frekvensforskelle og præsentationsrate af toner, samt på baggrund af start- og stop-synkroni af lyde. En efterfølgende modelanalyse blev brugt til at analysere indflydelsen af forskellige trin i den perifere hørelse på auditiv lydkildeseparation. Modelanalysen viste at auditiv frekvens-selektivitet og fysiologisk tids-maskering spiller en vigtig rolle i lydkildeseparation baseret på frekvensseparation og præsentationsrate af toner. Modelanalysen antyder yderligere at neural adaptation, og den deraf følgende skærpelse af det neurale respons når en lyd starter, øger følsomheden over for synkroni af frekvenskomponenter af lyde hvilket spiller en signifikant rolle i auditiv lydkildeseparation. Indflydelsen af lydtryksniveau på auditiv lydkildeseparation blev også undersøgt, under den antagelse at den reducerede frekvensselektivitet observeret ved høje lydtryksniveauer i mennesker vil resultere i en begrænset evne til at adskille lydkilder ved høje niveauer. Lytte-forsøg bekræftede denne hypotese, og viste at den minimale frekvensseparation hvorved lydkilder kunne blive adskilt perceptuelt steg som funktion af lydtryksniveau. Den matematiske model viste også en øget tendens til at opfatte frekvenskomponenter som en enkelt lydkilde ved høje niveauer, men modellen overvurderede indflydelsen af lydniveau i forhold til resultaterne fra lytte-forsøget. Dette indikerer at den matematiske model ikke fuldt ud afspejler mekanismerne bag auditiv lydkildeseparation. Afslutningsvis blev et nyt eksperimentelt paradigme til at måle lydkildeseparation undersøgt. Dette paradigme måler auditiv lydkildeseparation indirekte ved hjælp af forsøgspersonens præstationsevne i et *"comodulation masking release"* (CMR) eksperiment. CMR beskriver en øget evne til at detektere en svag lyd præsenteret sammen med andre "maskerende" lyde hvis de maskerende lyde er amplitude moduleret med den samme modulator. Forsøget undersøgte indflydelsen af tidsmæssig kohærens på den målte CMR og derigennem på forsøgspersonens evne til at adskille lydkilder. Resultaterne indikerer at CMR kan anvendes som et indirekte mål for lydkildeseparation, og understøtter yderligere hypotesen om at tidsmæssig kohærens kan gruppere frekvenskomponenter til en enkelt lydkilde. Samlet set indikerer resultaterne af denne afhandling at tidsmæssig kohærens spiller en signifikant rolle i grupperingen af frekvenskomponenter til et enkelt lydobjekt, og mere generelt, at en tidsmæssig kohærensanalyse kan danne rammerne for auditiv lydkildeseparation.

Preface

Through my years as a PhD student, many people have interfaced with me and my project, and I would like to take this opportunity to thank some of these.

Torsten Dau, for his contagious enthusiasm, and for his guidance through the last three years. More than once I lost faith in this project, but somehow Torsten was always able to get me back on track.

Morten Løve Jepsen, for introducing me to the world of auditory models, and for answering many a clarifying question - particularly in the beginning of the project.

Andrew J. Oxenham, for making me a part of his research group for three months, for his collaboration, and for providing perspective on how hearing research can be approached.

Christophe Micheyl, for sharing his office with me during my stay in Minneapolis and for many a discussion on auditory streaming, statistics, or just the world status in general. I hope your career in the industry is everything you hoped for, and that sunny California suits you better than Minnesota.

Ewen N. MacDonald, for always taking the time to answer my questions about statistics, English grammar, and the many strange wonders of Canada.

All my previous and current coworkers at CAHR, for many fruitful discussions, many questions asked and answered and for many shared laughs. You have made the last three years a wonderful time, and I sincerely doubt I will ever find another job with such a nice atmosphere.

All researchers and students at the APC lab. You made me feel at home within days of arrival, and I hope I will run into you again someday.

All of the test persons who participated in my experiments. You may have been in it for the money at first, but I am convinced that most of you came back out of pity for me. I thank you for the many hours you have spent listening to beeps and noises. Without you, none of this would have been possible.

Lastly, a special thank goes to Stine Ziska Jensen. Thank you for your patience, for your encouragements, for your love and your support. You are what kept me going through these three years, and I doubt I would have made it through without you.

Simon Krogholt Christiansen, 18 July, 2014.

Related publications

Journal papers

- Christiansen, S. K., and Oxenham, A. J. (2014). Assessing the effects of temporal coherence on auditory stream formation through comodulation masking release. *J. Acoust. Soc. Am.*, 135, 3520-3529.
- Christiansen, S. K., Jepsen, M. L., and Dau, T. (2014) Effects of tonotopicity, adaptation, modulation tuning, and temporal coherence in "primitive" auditory stream segregation. *J. Acoust. Soc. Am.*, 135, 323-333.
- Christiansen, S. K., and Dau, T. (2015). Effects of sound intensity on auditory stream segregation of pure tone sequences. *J. Acoust. Soc. Am.*, submitted

Conference papers

- Hauth, C., Christiansen, S. K., and Dau, T. (2013). Level dependency of auditory stream segregation. *Proceedings of the Deutsche Gesellschaft für Akustik, Joint 39th German and Italian Convention on Acoustics*, Merano, Italy, March 2013
- Christiansen, S. K., Jepsen, M. L., and Dau, T. (2012). A physiologically inspired model of auditory stream segregation based on a temporal coherence analysis. *Proceedings of Meetings on Acoustics 2012*, Hong Kong, May 2012.
- Christiansen, S. K., Jepsen, M. L., and Dau, T. (2012). A computational model of auditory stream segregation based on a temporal coherence analysis. *Proceedings of the Deutsche Gesellschaft für Akustik, 38th German Convention on Acoustics*, Darmstadt, German, March 2012.
- Christiansen, S. K., Jepsen, M. L., and Dau, T. (2011). Modelling auditory grouping based on a temporal coherence analysis. *Proceedings of Forum Acusticum*, Aalborg, Denmark, June 2011.

Contents

1

Introduction

Natural acoustic environments often consist of multiple, simultaneously active sound sources. Therefore, the acoustic signal that reaches a listener's ear is not the simple acoustic signal from a single sound source, but rather a mixture of all the sound sources. Despite the complexity of the resulting acoustical signal, normal-hearing listeners are typically able to focus their attention on a single sound source and, for example, follow a conversation in a crowded room or hear out a single instrument in a piece of music. The problem of perceptually segregating one target source from a mixture of multiple sources is commonly referred to as *auditory stream segregation* (Bregman, 1990). An *auditory stream* refers to the sounds that are perceived as coming from a single sound source, and involves the grouping of spectral components across frequency and time into a single perceptual object.

Auditory stream segregation is similar to the problem of separating a mixture of acoustic signals into its original constituents. For most practical purposes, the acoustic transmission path can be considered as a linear, time-invariant system (Jacobsen and Juhl, 2013), and therefore the superposition principle applies. Thus, if a violin creates the sound signal $x_1(t)$ and a trombone creates the sound signal $x_2(t)$, the mixture of the two sound sources can be described as

$$x_3(t) = g_1 x_1(t) + g_2 x_2(t),$$

where g_1 and g_2 are gain constants for each of the two signals. While the linear combination of acoustic signals is, in theory, reversible, a listener usually only has access to the mixture of the sound signals $x_3(t)$. Therefore, the problem is underdetermined (and increasingly so with more sound sources), and this is the heart of the problem of stream segregation.

1.1 Perceptual auditory stream segregation

A multitude of studies have sought to quantify when and how listeners segregate sound sources. These studies suggest that auditory stream segregation relies on processes which are stimulus-driven and primarily bottom-up based, referred to as *primitive streaming*, as well as processes that rely on prior exposure and are top-down processes, referred to as *schema-based streaming* (Bregman, 1990). While the schema-based processing may allow a listener to focus his or her attention on specific spectral or temporal aspects of a known sound source, the primitive processing directly helps in solving the underdetermined problem of stream segregation by setting up constraints on the solution. Natural sound sources typically consist of vibrating objects, and the constraints realized

by the primitive processes are related to the acoustic characteristics of natural sound sources. These constraints, or "cues", are traditionally classified by whether they group acoustic information across frequency (simultaneous grouping) or across time (sequential grouping).

For simultaneous grouping, the timing of individual frequency components of a sound has been shown to act as a cue for stream formation. The spectral components of a single sound source often start and stop at the same time as the sound generator is being activated or deactivated. This could, for example, be a person speaking a vowel, where the vibration of the vocal chords creates harmonic components that start and end at the same time. The importance of the synchronous onset was demonstrated by Darwin and Sutherland (1984), who showed that starting one harmonic of a synthetic vowel more than 32 ms before the rest of the harmonics altered the overall quality of the vowel, suggesting that the harmonic was perceptually segregated from the rest. Other studies (e.g. Bregman and Pinker, 1978; Elhilali et al., 2009; Micheyl et al., 2013a) have shown that synchronous presentation may cause spectral components to fuse into a single perceptual stream under circumstances where they would not have fused without the synchrony (e.g. for spectral components that are not harmonically related, or with large frequency separations). The streaming due to onset and offset (a)synchrony may be interpreted as a more general mechanism that tends to group frequency components that change "coherently" over time, as spectral components also tend to fuse together into a single stream if they are amplitude modulated by the same modulator (e.g. Bregman et al., 1985; 1990), or frequency modulated by the same modulator (e.g. Summerfield et al., 1992). A second cue that plays a major role in simultaneous grouping is harmonicity. Many natural sound sources vibrate repetitively to generate periodic waveforms whose spectra can be described by harmonic series. For such sounds, the individual frequency components are located at integer multiples of the fundamental frequency (F0) of the harmonic complex. These harmonic complexes occur due to the natural resonance frequencies of objects, as in musical instruments, or through repetitive acoustic events, such as the vocal fold vibration in voiced speech. The auditory system appears to use this harmonic relationship to perceptually group spectral components of a sound into a single auditory stream if they are harmonically related (Broadbent and Ladefoged, 1957). This is particularly useful in situations with e.g. multiple talkers where the differences in the pitch of each talker's voice allow the auditory system to group the spectral components correctly.

While the simultaneous grouping helps to determine which frequency components should be grouped together at a given time instant, sequential grouping groups spectral components over time; ensuring that successive sound events from a single sound source are grouped into the same auditory stream. Sequential streaming plays a major role in how humans perceive sounds, as Plack (2005) eloquently describes: "*Sequential grouping is a precursor to the identification of temporal patterns, such as musical melodies or sentences in speech, and we are usually only interested in the temporal order of sounds from the attended source*". The classification of whether successive acoustic events should be grouped into the same perceptual stream generally depends on whether they represent a "good continuation" of the previous sounds. This means that the auditory system is more likely to group sounds that change slowly over time, and more likely to segregate sounds that vary a lot from one time instant to the next. This was shown to be the case for spectral cues by van

Noorden (1975), who investigated the effect of frequency separation on stream segregation. While he found that tone sequences with a large frequency separation between successive tones were more likely to perceptually segregate than tone sequences with small frequency separations, he also found that the maximum frequency separation where it was possible for the tones to perceptually group into a single stream depended on the rate of the tones, and that fast tones were more likely to split into separate streams than slow tones for a given frequency separation. While frequency separation can lead to stream segregation, it is not a necessary cue, and several studies have shown that stream segregation can occur without any spectral differences, but instead based on e.g. pitch (Vliegen and Oxenham, 1999; Vliegen et al., 1999), waveshape-induced timbre differences (Roberts et al., 2002), amplitude modulation rate (Grimault et al., 2002), spatial location (Middlebrooks and Onsan, 2012) or even simple intensity differences (van Noorden, 1975; 1977). In general, any sufficiently salient perceptual difference may lead to stream segregation (Moore and Gockel, 2002), i.e., if the perceived difference between two successive sounds is large, the sounds may split into separate streams.

The segregation of an acoustic stimulus into separate auditory streams is rarely instantaneous. This process usually takes a short amount of time, where the auditory system "collects evidence" for the perceptual organization (Bregman, 1990). The time required for stream segregation is typically in the order of a couple of seconds, but depends on the exact stimulus conditions (Anstis and Saida, 1985) and the acoustic stimulus is typically perceived as a single fused stream until the segregation takes place.

1.2 Models of auditory stream segregation

Models and theories of the underlying mechanisms behind auditory stream formation have evolved along with the accumulation of evidence from perceptual experiments. One of the earliest theories of how the auditory system might realize the task was suggested by van Noorden (1975). He proposed that stream segregation primarily relied on peripheral filtering or tonotopic separation, and that the observed dependency on the tone rate in his experiments was explained by a concept of "frequency-jump" detectors, which were responsible for directing the listener's attention to the correct peripheral channel during frequency transitions of a stimulus. These "frequency-jump" or "pitch-motion" detectors were associated with a maximum rate of change, and thereby explained why a sequence of rapidly alternating tones would split into two streams at a lower frequency separation than slowly repeating tone sequences. A similar concept was proposed by Anstis and Saida (1985), suggesting that the pitch-motion detectors suffered from fatiguing or adaptation which made them unable to keep up with rapidly fluctuating spectral patterns for longer periods, thereby explaining the time interval required for a stimulus to split up into two streams. This concept was further supported by the findings of Hartmann and Johnson (1991) who defined what is broadly recognized as "the peripheral channeling theory". The peripheral channeling theory proposes that stimuli which stimulate the same or overlapping populations of tonotopically tuned neurons are perceptually grouped, whereas stimuli that activate well separated populations of tonotopically

tuned neurons are perceptually segregated. These concepts have inspired several computational models of stream segregation (e.g. Beauvois and Meddis, 1996; McCabe and Denham, 1997) which are able to account for various perceptual phenomena of stream segregation relying primarily on frequency separation for stream segregation (e.g. Miller and Heise, 1950; van Noorden, 1975; Anstis and Saida, 1985).

The peripheral channeling theory is, however, unable to account for stimuli that produce segregated percepts without tonotopic separation, such as the streaming based on differences in e.g. fundamental frequency (F0) (e.g. Vliegen and Oxenham, 1999), modulation rate (Grimault et al., 2002) or timbre (Roberts et al., 2002). Nonetheless, the principle of neural separation for stream segregation might still hold in populations of neurons that are sensitive to higher-level features, such as fundamental frequency (F0) or pitch (Bendor and Wang, 2005). A model based on such a multi-dimensional representation was suggested by Elhilali and Shamma (2008) which projects a sound stimulus onto a feature space consisting of several perceptually relevant features, such as e.g. audio-frequency, pitch, and timbre. Thereby, sounds which are separated along any of the perceptual dimensions can be perceptually segregated from each other.

While this "generalized channeling theory" might explain most stream segregation phenomena, there are several observations it cannot account for alone. One of these is the perceptual fusion of distant spectral components due to synchrony, as such a stimulus would create neural responses that are well separated in the feature space, but are, in fact, perceived as a single stream. The second problem is linked to the "binding problem", which is a general problem in perception: If the different perceptual features of sounds are encoded by separate neural populations, how can the auditory system combine this information into a single coherent percept? Solutions to this problem generally fall into two categories: Hierarchical coding and temporal correlation (Brown, 2010). The hierarchical coding proposes that a hierarchy of increasingly specialized cells encodes the various combinations of features. While there is some physiological evidence for such hierarchical coding, the number of neurons required to represent every possible combination of features is so large that it is unlikely that hierarchical coding can solve the binding problem alone (Brown, 2010). An alternative solution to the binding problem is based on the temporal correlation of neural responses. The temporal correlation theory proposes that neurons which code different features of the same object are bound together by the synchronization in their temporal fine structure. A model of auditory stream segregation based on neural synchronization was first proposed by von der Malsburg and Schneider (1986), and demonstrated to be able to segregate different acoustic events based on their onset asynchrony.

A recent approach towards explaining stream segregation is the concept of temporal coherence (Elhilali et al., 2009; Shamma et al., 2011). The temporal coherence idea is similar to the correlation theory in that it proposes that neural responses are grouped together into a single percept if they are activated at the same time. However, where the correlation theory proposes that this grouping is based on the temporal fine structure of the neural responses, the temporal coherence theory is based on the correlation of neural responses over relatively long integration windows (50-500 ms), consistent with the slow dynamics of stimulus-induced fluctuations in spike rate at the auditory

cortex (Shamma et al., 2011). It is still unknown whether the temporal coherence theory is able to bridge the gap between models of auditory stream segregation and human performance in similar tasks, but it represents a computationally efficient approach to solving the binding problem.

1.3 Structure of the thesis

Chapter 1 aimed at providing a general introduction to the problem of auditory stream segregation, and a brief overview of the history and advancements of research into auditory stream segregation. The following chapters will focus on reporting and discussing the present work. The aim was to better understand the influence of bottom-up processing of the auditory system on the perceptual organization of acoustic stimuli, using temporal coherence as the organizing principle. This was done through a combination of computational models of auditory processing and through psychoacoustic experiments. Chapters 2 to 4 represent three separate studies:

- In **Chapter 2**, the effect of onset/offset synchrony on auditory grouping is investigated through a listening experiment. Secondly, a computational model of auditory stream segregation is presented, based on an auditory pre-processing front-end combined with a temporal coherence analysis back-end. The computational model is able to account for primitive stream segregation phenomena based on frequency separation, tone rate and onset/offset synchrony, and through the model framework, the effects of tonotopicity, adaptation, modulation tuning and temporal coherence is investigated.

- In **Chapter 3**, the influence of peripheral processing on auditory stream formation is investigated. Under the assumption that the excitation of non-overlapping populations of neurons is the underlying principle of stream segregation, the increasing auditory-filter bandwidth with increasing sound pressure level should lead to more fused percepts at high sound intensities. This hypothesis is investigated through both listening experiments, and through a modified version of the computational model presented in Chapter 2. The experimental data confirm the hypothesis and show a small but significant effect of increasing intensity. The computational model also shows an effect of increasing intensity, but overestimates the size of the effect, suggesting that the computational model does not accurately reflect the stream formation processes of the auditory system.

- In **Chapter 4**, the influence of temporal coherence on auditory stream formation is investigated through an indirect, performance-based measure. This measure is based on comodulation masking release (CMR), and relies on the assumption that the CMR produced by flanking bands remote from the masker and target frequency only occurs if the masking and flanking bands form part of the same perceptual stream. The study evaluates the feasibility of using CMR as a measure of stream segregation, and uses the CMR to assess the conditions under which a loss of temporal coherence across frequency leads to auditory stream segregation. The study suggests that CMR may be used as an indirect measure of

stream formation, but also illustrates that the interaction of streaming cues can make it difficult to investigate a single cue in isolation.

Finally, the main findings of the present work are summarized, and their possible implications are discussed in **Chapter 5**.

2

Effects of tonotopicity, adaptation, modulation tuning, and temporal coherence in "primitive" auditory stream segregation*

The perceptual organization of two-tone sequences into auditory streams was investigated using a modeling framework consisting of an auditory pre-processing front end [Dau et al., J. Acoust. Soc. Am. 102, 2892-2905 (1997)] combined with a temporal coherence-analysis back end [Elhilali et al., Neuron 61, 317-329 (2009)]. Two experimental paradigms were considered: (i) Stream segregation as a function of tone repetition time (TRT) and frequency separation (Δf) and (ii) grouping of distant spectral components based on onset/offset synchrony. The simulated and experimental results of the present study supported the hypothesis that forward masking enhances the ability to perceptually segregate spectrally close tone sequences. Furthermore, the modeling suggested that effects of neural adaptation and processing through modulation-frequency selective filters may enhance the sensitivity to onset asynchrony of spectral components, facilitating the listeners' ability to segregate temporally overlapping sounds into separate auditory objects. Overall, the modeling framework may be useful to study the contributions of bottom-up auditory features on "primitive" grouping, also in more complex acoustic scenarios than those considered here.

2.1 Introduction

One of the most extraordinary features of the human auditory system is its ability to group simultaneous and sequential sensory inputs such that the perceptual representations correspond to the different objects in the environment. In a natural acoustic surrounding, there are often multiple, simultaneously active sound sources that create a mixture of acoustic inputs to the receiver's ears. In hearing, the process of grouping auditory input into distinct percepts is commonly referred to as auditory scene analysis (ASA) or auditory stream segregation (e.g., Bregman, 1990). A distinction between "primitive" and "schema-based" processes has been proposed in ASA (Bregman, 1990). Primitive processes have been associated with data-driven phenomena, consisting of pre-attentive auditory processes that are automatic. Such primitive processes have been assumed to group those sound elements that likely come from a common source into a coherent perceptual representation

* This chapter is based on Christiansen et al. (2014).

based on their common acoustic properties. In contrast, schema-based scene analysis refers to perceptual grouping processes that require high-level cognitive input and are influenced by the listener's attention and prior expectations based on previous learning (Bregman, 1990). The present study is focused on components of the primitive processes involved in ASA.

The phenomenon of ASA has inspired various studies investigating the limits of perceptual grouping or segregation using various experimental paradigms (for a review, see, e.g., Bregman, 1990; Moore and Gockel, 2002; Carlyon and Gockel, 2008). One of the most influential experimental paradigms was provided by van Noorden (1975). By presenting two pure tones, A and B, in an alternating temporal sequence, two different percepts are typically evoked. Either the two tones fuse into one single stream or they split into two separate streams, and the listener's attention will be drawn to either the repeating A- or B-tone. Van Noorden (1975) obtained similar results for a simple alternating ABAB sequence and an ABA-ABA sequence; this gives rise to a characteristic "galloping rhythm," showing that the fusion or segregation of the tones into either one or two streams depends on the frequency separation (Δf) between the tones as well as the tone repetition time (TRT). He observed that the two-tone sequences tend to split into two separate streams for small Δf's when TRT was small, whereas a larger Δf was required for the tone sequences to split when the TRT was large.

The potential mechanisms underlying ASA have been discussed in various studies, ranging from Gestalt grouping principles like proximity (e.g., Bregman, 1990) toward more physiologically inspired concepts considering effects of peripheral auditory filtering and "frequency jump" or "pitch motion" detectors (van Noorden, 1975; Anstis and Saida, 1985). Recent physiological studies in the songbird forebrain (Bee and Klump, 2004, 2005) and monkey auditory cortex (Fishman et al., 2004) supported the hypothesis that perceptual stream segregation of alternating tone sequences is, to a large extent, a consequence of pre-attentive auditory processes, such as frequency selectivity and physiological forward masking. For fast repeating tone sequences (i.e., a small TRT), physiological forward masking was found to promote the spatial separation of neural responses in the tonotopic space, giving rise to a two-stream percept for a smaller Δf than that required for slowly repeating tones (i.e., a large TRT).

The contribution of tonotopicity in ASA (as in the example of the two alternating tone sequences) is consistent with the principle of streaming that has been referred to as the "channeling hypothesis," implying that streams form when they activate distinct neuronal populations or processing channels (Hartmann and Johnson, 1991). However, this requirement has recently been challenged and shown to be insufficient to account for stream formation. Elhilali et al. (2009) demonstrated that if two tone sequences are made fully correlated by creating synchronous sequences, the channeling hypothesis fails as the two pitch percepts bind together, forming a repeating complex tone as one stream, irrespective of the tone sequences' spectral separation Δf. Thus temporal coherence appears to provide the organizing principle necessary to make the correct perceptual assignments as to which sequences form a stream.

Previous attempts to construct computational models of auditory stream segregation have largely

relied on tonotopic separation for stream segregation (e.g., Beauvois and Meddis, 1996; McCabe and Denham, 1997). These models are able to account for various perceptual phenomena of stream segregation that rely primarily on frequency separation (e.g., Miller and Heise, 1950; van Noorden, 1975; Anstis and Saida, 1985) but fail to predict perceptual grouping of distant spectral components due to, e.g., harmonicity or onset and offset synchrony (Elhilali et al., 2009; Micheyl et al., 2010).

Elhilali et al. (2009) and Shamma et al. (2011) suggested a conceptual model for detecting auditory stream formation based on a temporal coherence analysis, such that, after decomposition of the acoustic stimulus into a set of perceptually relevant features, those features that vary coherently are grouped together. However, no explicit predictions have been provided to evaluate this concept and to explore its capabilities and limitations, e.g., in conditions similar to the ones provided in van Noorden (1975).

In the present study, a computational model of (primitive) stream segregation is proposed to further investigate the role of tonotopicity (i.e., auditory frequency selectivity) and forward masking as well as across-channel coherence on perceptual stream segregation. The model consists of a front end that involves assumptions about the spectro-temporal auditory processing in humans and corresponds to the processing stages earlier proposed in Dau et al. (1997a) and related studies. This model has been shown to account for many detection and masking phenomena in humans, including effects of spectral masking and forward masking (e.g., Derleth and Dau, 2000; Jepsen et al., 2008). The back end, which is an optimal detector in the original models of Dau et al. (1996, 1997a), was replaced in the present study by the coincidence detection stage as proposed by Elhilali et al. (2009).

The computational model framework was evaluated in two experimental conditions where quantitative predictions were compared to corresponding data using the same stimuli. In the first experiment, stream segregation of two-tone sequences was studied as a function of Δf and TRT, as in van Noorden (1975). In the second experiment, stream segregation of distant spectral components was studied as a function of onset and offset asynchrony across frequency. An analysis was undertaken in an attempt to specify which auditory processes, in the framework of the proposed model, contribute to the predicted outcome in the individual tasks, with focus on effects of peripheral frequency selectivity, adaptation, and modulation-frequency selectivity.

2.2 Model description

Figure 2.1 shows the structure of the model proposed in the present study. The "auditory spectrogram" and "temporal integration" stages are based on the preprocessing of the model described in Dau et al. (1997a). The back end is represented by a coherence analysis as suggested in Elhilali et al. (2009).

The front end consists of a fourth-order gammatone filterbank (Patterson et al., 1987) with one equivalent rectangular bandwidth (ERB; Glasberg and Moore, 1990) spacing between each filter to simulate the frequency-selective filtering on the basilar membrane (BM). This is followed by

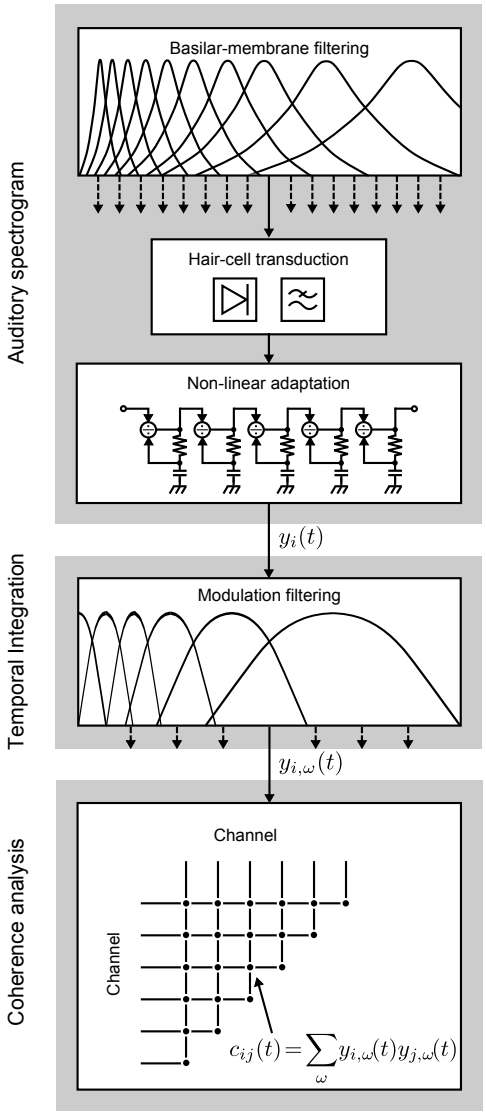


Figure 2.1: Block diagram of the processing model proposed in the present study. The model includes a gammatone filterbank, half-wave rectification, low-pass filtering at 1 kHz, an adaptation stage, and a modulation band-pass filterbank. An across-frequency coherence network is applied to the output of the preprocessing. See main text for further details.

the hair-cell transduction stage that roughly simulates the physical transduction of the mechanical vibration of the BM into receptor potentials in the inner hair cells. It consists of a half-wave rectification stage followed by a low-pass filtering at 1 kHz, realized by a second-order Butterworth filter, preserving phase information at low frequencies and only envelope information at high frequencies. The output serves as the input to the adaptation stage of the model that simulates adaptive properties of the auditory periphery. Adaptation refers to dynamic changes in the gain of the system in response to changes in input level. In the model, the effect of adaptation is realized by a chain of five simple nonlinear circuits, or feedback loops, with different time constants (Püschel, 1988; Dau et al., 1996; Dau et al., 1997a). Each circuit consists of a low-pass filter and a division operation. The low-pass filtered output is fed back to the denominator of the devisor element. For a stationary input signal, each loop realizes a square-root compression. Such a single loop

was first suggested by Siebert (1968) as a phenomenological model of auditory nerve adaptation. The output of the series of five loops approaches a logarithmic compression for stationary input signals, mapping a logarithmically scaled input in dB to a linear internal model unit scale. For input variations that are rapid compared to the time constants of the low-pass filters, the transformation through the adaptation loops is more linear, leading to an enhancement of fast temporal variations or onsets and offsets at the output of the adaptation loops. The time constants, ranging between 5 and 500 ms, were chosen to account for perceptual forward masking data (Dau et al., 1996; Jepsen et al., 2008). In response to signal onsets, the output of the adaptation loops is characterized by a pronounced overshoot.

The adaptation stage is followed by a temporal modulation filterbank. The lowest modulation filter is a second order low-pass filter with a cut-off frequency of 2.5 Hz. The modulation filters tuned to 5 and 10 Hz have a constant bandwidth of 5 Hz. For modulation frequencies at and above 10 Hz, the modulation filters are distributed logarithmically and have a constant Q value of 2. The magnitude transfer functions of the filters overlap at their -3 dB points. The highest modulation filter frequencies in the filterbank are limited to a quarter of the center frequency of the corresponding peripheral auditory filter, and maximally 1 kHz. As in the model of Dau et al. (1997a), the modulation filters are realized by complex frequency-shifted first-order low-pass filters. These filters have complex-valued outputs and either the absolute value or the real part can be considered. For modulation filters centered at or below 10 Hz, the real valued part of the output is considered. For modulation filters above 10 Hz, the absolute value is considered. This is comparable to the Hilbert envelope of the band-pass filtered signal and only conveys information about the presence of modulation energy in the modulation filter (i.e., no phase information). The modulation filters represent a set of integration time constants corresponding to the inverse of their respective bandwidths.

The output of the temporal integration stage is processed by the back end, a coherence analysis, to determine which channels vary coherently over time (Elhilali et al., 2009). The coherence analysis identifies temporally coherent activity across the tonotopic axis by creating a dynamic coherence matrix, $C(t)$, where the value of each element $c_{ij}(t)$ of the matrix is given by Eq. 2.1

$$c_{ij}(t) = \sum_{\omega} y_{i,\omega}(t) y_{j,\omega}(t) \tag{2.1}$$

with i, j indicating gammatone-filter indices and ω representing a given modulation filter. In the present study, as in Elhilali et al. (2009), the dynamic property of the coherence matrix was discarded by integrating the coherence matrix across time. Furthermore, the output of the temporal integration stage was half-wave rectified prior to the calculation of the coherence matrix. This was done to avoid negative values of the coherence matrix. To quantify the coherence matrix, a decomposition into independent eigenvectors was performed. The eigenvectors indicate which peripheral filter outputs are correlated over time and, thus, which filters would be fused into a single stream according to the temporal coherence hypothesis. The eigenvalues indicate the strength of

each of the eigenvectors, and, by calculating the ratio between the second largest (λ_2) and the largest (λ_1) eigenvalue, the relative importance of the second largest to the largest eigenvalue can be determined.

If an input stimulus consists of temporally coherent spectral components, the largest eigenvector alone will be sufficient to describe the coherence matric C. The eigenvalue ratio (λ_2/λ_1) will therefore be close to 0, indicating a fused percept according to the temporal coherence hypothesis. If the input stimulus consists of incoherent spectral components, more than one eigenvector will be required to describe the coherence matrix C, and the eigenvalue ratio (λ_2/λ_1) will be larger than 0, indicating a segregated percept of the stimulus components.

2.3 Method

2.3.1 Experiment I: Stream segregation as a function of frequency separation and tone repetition time

Stimuli and procedure

The data from this experiment were taken from van Noorden (1975). The stimulus consisted of two repeating tones, A and B, presented in an alternating ABABAB pattern, as illustrated in Fig. 2.2. The frequency of the B-tone (f_B) was fixed at 1 kHz, and the frequency of the A-tone (f_A) was swept -15 to +15 semitones relative to the B-tone over a period of 80 s so that the tone interval $\log(f_A/f_B)$ varied linearly with time. The duration of the tones were 40 ms, including linear onset and offset ramps of 5 ms. The experiment was conducted for TRTs between 60 and 150 ms, in steps of 10 ms, and the stimulus was presented at a level of 35 dB sensation level (SL).

The tones were either perceived as a single stream consisting of a tone sequence with alternating pitch or as two separate tone streams, each with its own pitch and repeating at half the rate of the one-stream percept. During the stimulus presentation, the listeners in the study of van Noorden had to indicate, by pushing a button, whether they perceived the stimulus as one stream or as two streams. The entire experiment was run twice, each with different instructions. In the first run, the listeners were instructed to hold on to the alternating rhythm as long as possible, thereby trying to obtain a fused percept. In the second run of the experiment, the listeners were instructed to follow the string of A-tones as long as possible, trying to obtain a segregated percept. The two measurements represented the temporal coherence boundary (TCB) and the fission boundary (FB), respectively. The TCB characterizes the maximum Δf where it is possible to obtain a single, coherent stream; whereas the FB characterizes the lowest Δf where it is possible to selectively attend to a single of the two tones.

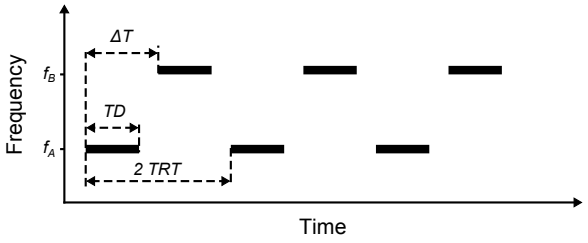


Figure 2.2: Schematic representation of the tone sequences used in the present study, with ΔT representing the onset asynchrony of the A- and B-tones, TRT representing the tone reptition rate, and TD reflecting the tone duration. In Experiment I, following van Noorden (1975), stream segregation was investigated as a function of Δf and TRT. TD was fixed at 40 ms, TRT = ΔT was ranged from 60 to 150 ms in steps of 10 ms, f_A = 1 kHz, and f_B was presented -15 to +15 semitones relative to f_A. In Experiment II, the effect of ΔT on grouping of distant spectral components was considered. Here, f_A = 300 Hz, and f_B = 952 Hz. Three TRTs (75, 100, 125 ms) and two TDs (30, 75 ms) were investigated, while ΔT was varied from -100 to 100 ms.

Simulation parameters

In the proposed model, the stimulus was analyzed in segments of 2 s duration, in which f_A, f_B, and TRT were kept fixed. This differed from the 80 s stimulus duration used in the perceptual experiment where the stimulus parameters varied during the presentation. This change was made to account for the fact that the model provides a single estimate for the entire stimulus analyzed as opposed to the human listeners whose percept changed during the presentation. The frequencies of tones A and B were set to 1 kHz and N semitones above this frequency, respectively. Combinations of 10 TRT values (60-150 ms in steps of 10 ms) and 31 levels of N (0-15 semitones) were considered by the model (310 different conditions in total). A signal level corresponding to 70 dB sound pressure level (SPL) (at earphone) was used and the eigenvalue ratio λ_2/λ_1 was calculated for each condition.

2.3.2 Experiment II: Grouping of distant spectral components due to onset and offset synchrony

Listeners

Four normal-hearing listeners, aged between 24 and 26 yr, participated in the experiment. All listeners had previous experience with psychoacoustic experiments. All listeners received approximately 1 h of listening experience prior to the final data collection. The listeners were compensated monetarily for their participation at an hourly rate. Measurement sessions lasted between 1 and 2 h including breaks.

Stimuli and procedure

As in experiment I, the stimuli consisted of two repeating tones, A and B (see Fig. 2.2). Here, the tones were presented at fixed, distant frequencies f_A = 300 Hz and f_B = 952 Hz (i.e., 20 semitones

above f_A). The A- and B-tones were gated on and off using 10 ms raised cosine ramps. The AB tone pairs were repeated at a rate of $1/(2\mathrm{TRT_A})$ with $\mathrm{TRT_A}$ indicating the repetition rate of tone A. The onset of the B tones occurred ΔT after the onset of the A tones. The overall stimulus duration was 80 s, during which ΔT changed linearly from -100 ms (B leading A) to +100 ms (B lagging A). This linear change was implemented by providing slightly longer TRTs for the B tones than for the A tones. Six experimental sequences were defined with combinations of two tone durations (TD = 30, 75 ms) and three tone repetition times ($\mathrm{TRT_A}$ = 75, 100, 125 ms). The corresponding repetition times for the B-tones were ($\mathrm{TRT_B}$ = 75.2, 100.25, 125.3 ms). Each combination of TRT and TD was presented five times, resulting in 30 measurements per listener, and the stimulus was presented at a level of 70 dB SPL. The entire experiment was repeated with ΔT changing in the opposite direction (from -100 to +100 ms) to avoid a response bias due to the direction of change of ΔT.

The stimulus evoked either two separate streams at the individual tone frequencies, or a single stream consisting of a repeating complex tone. While the stimulus was playing, the listeners had to indicate whether they perceived the stimulus as one stream or as two streams by pushing one of two buttons labeled "1 stream" and "2 streams," respectively. The moment where the percept changed from 2 streams to 1 stream (or vice versa) was recorded, and the asynchrony ΔT at threshold was determined. This method was chosen because it is identical to the procedure used by van Noorden (1975).

Apparatus

All stimuli were generated using MATLAB 2011b (Mathworks). A sampling rate of 44.1 kHz was used, and the signals were presented through a personal computer with a 24-bit soundcard (RME DIGI 96/8 PAD). The stimuli were presented using circumaural headphones (Sennheiser HD580). The listeners were seated in a double-walled, sound insulated booth with a computer monitor that displayed instructions during the experiment.

Simulation parameters

As in experiment I, segments of 2 s duration with fixed f_A, f_B, TRT, and ΔT were considered in the simulation to estimate the perceptual organization in a given stimulus condition. The frequencies of tones A and B were 300 and 952 Hz, as in the measurements. Combinations of three TRT values (75, 100, 125 ms), two different tone duration (30, 75 ms), and 31 ΔT values (-75 to 75 ms in steps of 5 ms) were considered in the model (186 different conditions in total). The signal level of the stimuli in the model corresponded to 70 dB SPL, and the eigenvalue ratio λ_2/λ_1 was calculated for each condition.

2.4 Results

2.4.1 Experiment I: Stream segregation as a function of frequency separation and tone repetition time

Figure 2.3 (left panel) shows a replot of the data from van Noorden (1975), consisting of two data series connected by lines: The TCB and the FB. The TCB shows the largest frequency separation where the alternating rhythm was perceived (i.e., a single stream) when the listener was actively trying to hold on to the rhythm. In the region above the TCB, the A and B-tones always split into two separate streams. The FB represents the smallest frequency separation where the subjects were able to selectively attend to only one of the two tones, forcing a two-stream percept. In the region below the FB, the stimuli were always perceived as a single stream.

The right panel of Fig. 2.3 shows the model predictions obtained with the corresponding stimuli. A bright color represents a low eigenvalue ratio, corresponding to a one-stream percept, and a dark color represents a high eigenvalue ratio, corresponding to a two-stream percept. For illustration, the two solid curves represent iso-eigenvalue-ratio contours assuming $\lambda_2/\lambda_1 = 0.2$ and $\lambda_2/\lambda_1 = 0.05$. The dashed curves represent alternative eigenvalue ratios at 0.3 and 0.1. Some of the characteristics in the data could be described by the model: (i) For small frequency separations Δf, the model predicted a single stream regardless of TRT, consistent with the FB in the data and (ii) tone sequences with a small TRT were more likely to be "perceived" as two streams than tone sequences with large TRTs which can have a larger frequency separation while still producing a fused percept. However, the frequency range over which two streams were produced at large TRTs was clearly smaller in the simulations than in the data. An analysis and discussion of these results is provided further in the following text (Sec. 2.5).

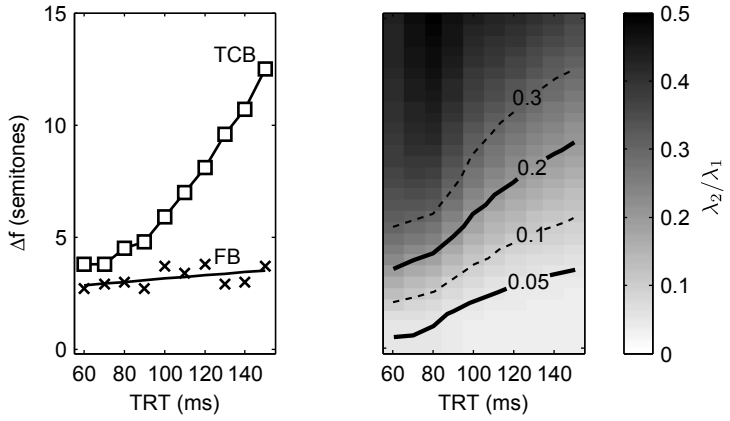


Figure 2.3: Results from experiment I. The left panel shows a replot of the data by van Noorden (1975). The upper curve represents the temporal coherence boundary (TCB) and the bottom curve represents the fission boundary (FB). The right panel shows corresponding simulations obtained with the proposed model. The grayscale intensity indicates the eigenvalue ratio. A bright color represents a small eigenvalue ratio, corresponding to a one-stream percept, and a dark color represents a large ratio, corresponding to a two-stream percept. The curves in the right panel indicate contours with fixed eigenvalue ratios.

2.4.2 Experiment II: Grouping of distant spectral components as a function of their onset and offset asynchrony

The measured data obtained in experiment II are shown in the left panels of Fig. 2.4. The different open symbols represent the measured thresholds for four individual listeners. The asynchrony (ΔT) between the A- and B-tones at which the percept changed between one and two streams is shown for the three TRT values 75, 100, and 125 ms. For ΔT values in the range between the thresholds (i.e., between about -20 to +20 ms), the stimuli were perceived as one stream, whereas for ΔT values above the upper thresholds and below the lower thresholds, the stimuli were perceived as two separate streams. The upper panel indicates the conditions with TD = 30 ms and the lower panel shows the results for TD = 75 ms. The data demonstrate that the tones were fused into a single stream when the asynchrony of the tones was less than roughly 20 ms. A 3-way analysis of variance (ANOVA) was performed to analyze the effect of the three factors TRT, TD, and direction of asynchrony (i.e., the sign of ΔT), as represented in Table 2.1. No significant effect was found for the factors TRT ($p = 0.738$) and TD ($p = 0.329$). However, a significant effect of sign(ΔT) was found ($p = 0.03$), showing that the tones were more likely to fuse into a single stream when the (low-frequency) A-tone was lagging the (higher-frequency) B-tone, as the mean ΔT's for the lagging and the leading A-tones were -20.6 and 17.7ms, respectively. The data were thus not fully symmetrical around ΔT = 0 ms but slightly shifted toward negative ΔT values.

The right panels of Fig. 2.4 show the corresponding simulations, using the same scale for the eigenvalue ratios as in Fig. 2.3. The solid curve indicates the iso-eigenvalue-ratio contour for $\lambda_2/\lambda_1 = 0.2$, while for illustration, other contours are indicated by the dashed curves. The model predicted a one-stream percept (bright color) when the tones were synchronous or close to synchronous. With increasing asynchrony of the tones, the eigenvalue ratio increased (indicated by the darker grayscale). Consistent with the data, the model showed only very little effect of TD and TRT on the eigenvalue ratios. However, in contrast to the data, the simulations showed an asymmetry of the one-stream percept favoring a leading A-tone (a positive ΔT) rather than a (slight) asymmetry toward lagging A-tones, as was observed in the data. The asymmetric effect in the model is a consequence of the frequency-specific (group) delay of the gammatone filters, with a shorter delay of the response to the higher-frequency tones than for the lower-frequency tones. The cochlear travel-time differences across frequencies seem to be largely compensated for by more central auditory processes (not reflected in the current model) that take place before the perceptual grouping of the tone sequences, as further discussed in the following text.

2.5 Model analysis

In the following analysis, the role of auditory preprocessing on the internal representation of the stimuli is discussed in the framework of the proposed model, in an attempt to explore which auditory processes affect auditory stream perception in the stimulus conditions considered in the present study.

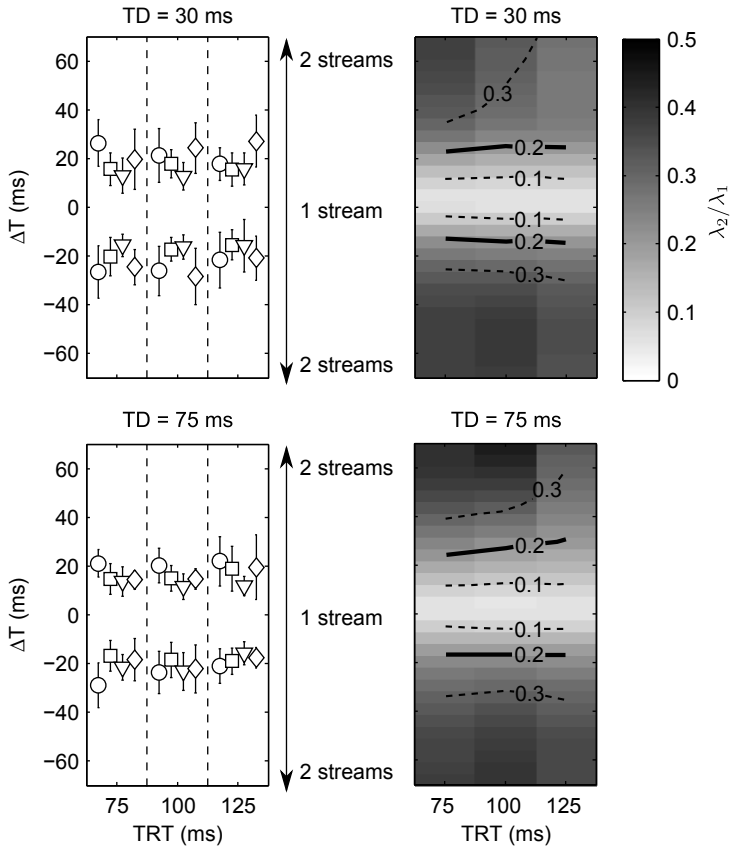


Figure 2.4: Results from experiment II. The left panels show measured ΔT's at the transition between 1 and 2 perceived streams for three different TRTs (75, 100, 125 ms). Data for a tone duration (TD) of 30 ms are shown in the top panel, and data for TD = 75 ms are shown in the bottom panel. The different symbols represent results from the individual listeners. The right panels show the results from the corresponding simulations. The grayscale intensity indicates the eigenvalue ratio, using the same scale as in Fig. 2.3. The solid curves indicate contours with fixed eigenvalue ratios of 0.2.

Table 2.1: Results of a 3 (TRT) × 2 (TD) × 2 (sign(ΔT)) ANOVA comparing the threshold for fusion/segregation (ΔT) between one and two streams.

Source	df.	Mean square	F	p
TRT	2	6.231	0.31	0.738
TD	1	19.943	0.98	0.329
Sign(ΔT)	1	103.782	5.1	0.030
TRT × TD	2	2.079	0.1	0.903
TRT × sign(ΔT)	2	29.051	1.43	0.253
TD × sign(ΔT)	1	15.801	0.78	0.384
TRT × TD × sign(ΔT)	2	1.742	0.09	0.918
Error	36	20.336		

2.5.1 Role of tonotopic separation

According to the coherence hypothesis (e.g., Elhilali et al., 2009), a stimulus must contain spectral components that vary incoherently over time to produce a two-stream (or multiple-stream) percept. However, to split into separate streams, the spectral components need to activate separate peripheral filters. If the frequency separation between the spectral components is too small, the tones will excite the same or overlapping filters as illustrated in the left panel of Fig. 2.5. Here the stimulus from experiment I is shown for a TRT of 140 ms and a Δf equal to three semitones. Because of the small frequency separation between the tones, the simulated neural excitation (with higher excitation indicated by darker areas) produced by these tones overlaps and the peripheral filters tuned to either of the frequencies are excited by both tones (second and third row of the left panel). The output of the peripheral filters tuned to the two frequencies is therefore highly coherent despite the acoustic input consisting of incoherently presented spectral components. Thus, in the model, the two tones are grouped into a single perceptual stream, as also indicated by the corresponding coherence matrix in the bottom row of this panel, consistent with the experimental data from Fig. 2.3 (left panel).

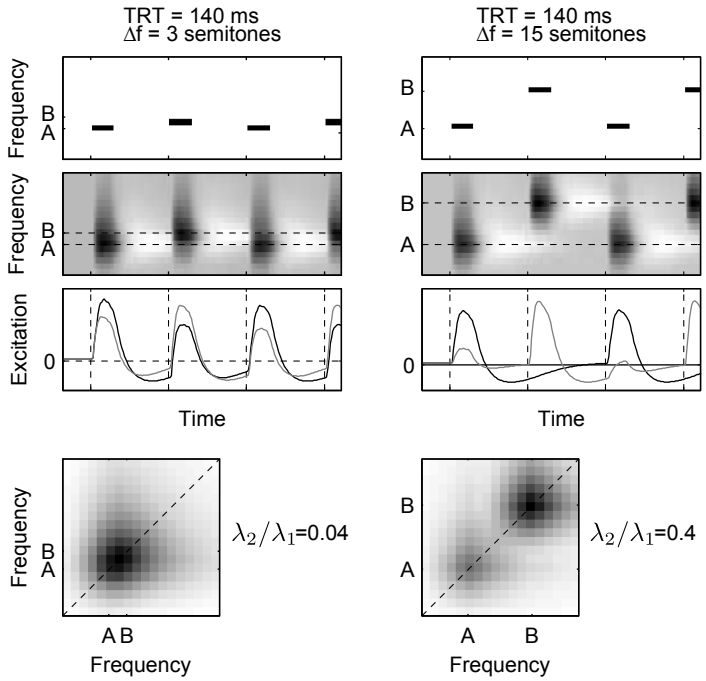


Figure 2.5: Illustration of the role of frequency selectivity in the model in conditions of experiment I. Left: Stimulus conditions with TRT = 140 ms and Δf = 3 semitones. Right: Same TRT but with Δf = 15 semitones. The upper row shows a schematic idealized spectrogram of the stimulus. The second row shows the corresponding auditory spectrogram. The grayscale represents the magnitude of the signals' internal representation in the model with dark representing a large magnitude and white representing negative values (inhibition). The third row indicates the excitation of the two peripheral filters tuned to the A tone (black) and the B tone (gray). The bottom row represents the coherence matrix and the eigenvalue ratios.

The right panel of Fig. 2.5 shows the corresponding results for a stimulus with the same temporal properties but a frequency spacing Δf of 15 semitones. Here Δf is sufficiently large to ensure

tonotopic separation between the spectral components, and the response of the peripheral filters tuned to either of the two tones is dominated by the tone that is closest in frequency. In this condition, the model predicts two separate streams, as visualized in the corresponding coherence matrix at the bottom, consistent with the measured data from Fig. 2.3 (left panel). Thus consistent with the peripheral-channeling hypothesis (Hartmann and Johnson, 1991), the modeling results support the role of frequency selectivity in stream segregation.

2.5.2 Role of forward masking

Bee and Klump (2004, 2005) and Fishman et al. (2004) proposed that the dependency of the TCB on the TRT is mainly a consequence of (physiological) forward masking, i.e., the reduced neural activity after the offset of the tone stimulation. This view is supported by the modeling results from the present study.

Figure 2.6 shows two conditions of experiment I that produced a one-stream percept (left panel) and a two-stream percept (right panel), respectively. The two conditions reflect different TRTs (140 vs 60 ms) for the same Δf of six semitones. In the model, the capacitors of the adaptation loops are charged at the offset of a tone, and the sensitivity of the model in a given peripheral channel is reduced during the period of discharge of the capacitors. For the slowly repeating tones (i.e., the large TRT, left panel), the time interval between the tones is large enough for the sensitivity to return to its state of rest before the onset of the next tone. For the fast repeating tones (i.e., the small TRT, right panel), the capacitors are not fully discharged before the next tone starts. Because the peripheral filter is in a state of reduced sensitivity, the off-frequency tone does not sufficiently activate the (on-frequency) filter, effectively limiting the spread of neural excitation across frequency. Due to forward masking, that is accounted for by the adaptation stage in the model, fast repeating spectral components with only a small frequency separation can thus excite tonotopic regions close to each other without producing overlapping neural excitation, as illustrated in the second and third rows of the right panel of Fig. 2.6. As a consequence, because the outputs of the peripheral filters do not vary coherently, they are not grouped into the same perceptual stream.

These properties in the model lead to the prediction (from Fig. 2.3, right panel) that stimuli with fast repeating tones (i.e., small TRTs) split into two streams at much smaller Δf's than slowly repeating stimuli (i.e., larger TRTs), consistent with the measured TCB described in experiment I (Fig. 2.3). The modeling provided here thus supports the hypothesis that forward masking contributes to perceptual segregation of fast repeating tone sequences. Thus, if the model would not contain effects of adaptation (to account for forward masking), it would fail to predict the data from experiment I despite the coherence stage in the back end.

2.5.3 Role of onset enhancement and multiple time constants

The coherence analysis stage in the computational model estimates the perceptual grouping of distant spectral components across frequency. However, the coherence analysis stage, per se,

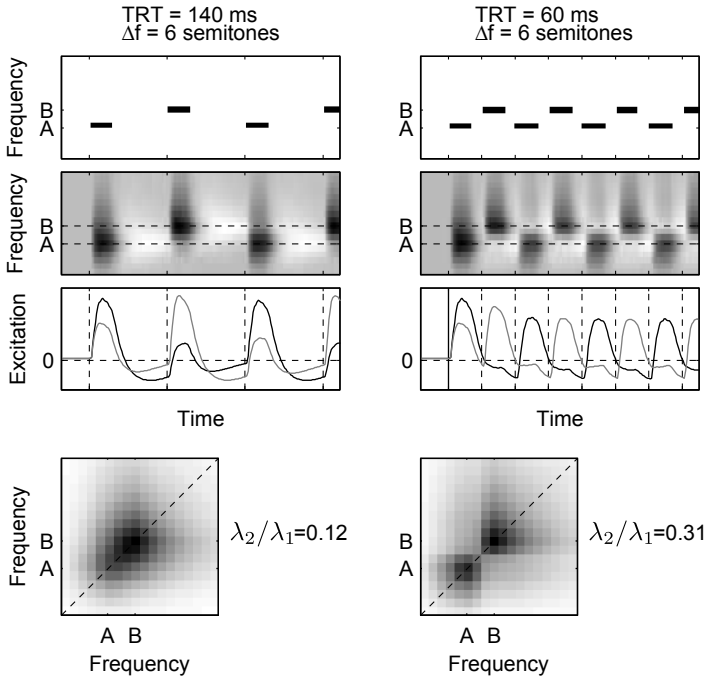


Figure 2.6: Stimuli from experiment I, with TRT = 140 ms (left) and TRT = 60 ms (right), for Δf = 6 semitones in both conditions. The upper row shows a schematic idealized spectrogram of the stimulus. The second row shows the corresponding auditory spectrogram. The grayscale represents the magnitude of the signals' internal representation in the model, with dark representing a large magnitude and white representing negative values (inhibition). The third row shows the excitation of the two peripheral filters tuned to the A tone (black) and the B tone (gray). The bottom row represents the coherence matrix and the eigenvalue ratios.

would not be sufficient to account for the data from experiment II. Assuming that the input to the coherence analysis was an "idealized" spectrogram (where each tone has no spread of excitation to neighboring frequency channels and the temporal envelopes of the tones are perfectly extracted), the corresponding coherence matrix C would have diagonal entries that are proportional to the tone duration (TD) and off-diagonal entries that are proportional to the temporal overlap of the A and B tones (TD-|ΔT|). For such a matrix, the eigenvalue ratio λ_2/λ_1 would correspond to

$$\left|\frac{\lambda_2}{\lambda_1}\right| = \begin{cases} \frac{|\Delta T|}{2TD-|\Delta T|}, & |\Delta T| < TD \\ 1, & |\Delta T| \geq TD \end{cases} \tag{2.2}$$

Equation 2.2 illustrates that the eigenvalue ratio in this case directly depends on the stimulus duration TD. This behavior is also evident from the simulation shown in the left panel of Fig. 2.7, which used such idealized signals as the input to the coherence analysis. The corresponding eigenvalue ratios of the coherence matrix, shown in the lower part of the left panel, strongly depend on the stimulus duration; this is inconsistent with the perceptual data. In the computational model suggested in the present study, the pre-processing stage does not simply extract the original envelope of the input stimulus. Instead due to the adaptation process in the model, the onsets of the tones are enhanced relative to the steady-state parts as illustrated in the top right panel of Fig. 2.7. Due to this onset enhancement, the coherence analysis becomes more sensitive to onset asynchronies

between the A and B tones and less sensitive to the tone duration TD as shown in the bottom right panel of Fig. 2.7.

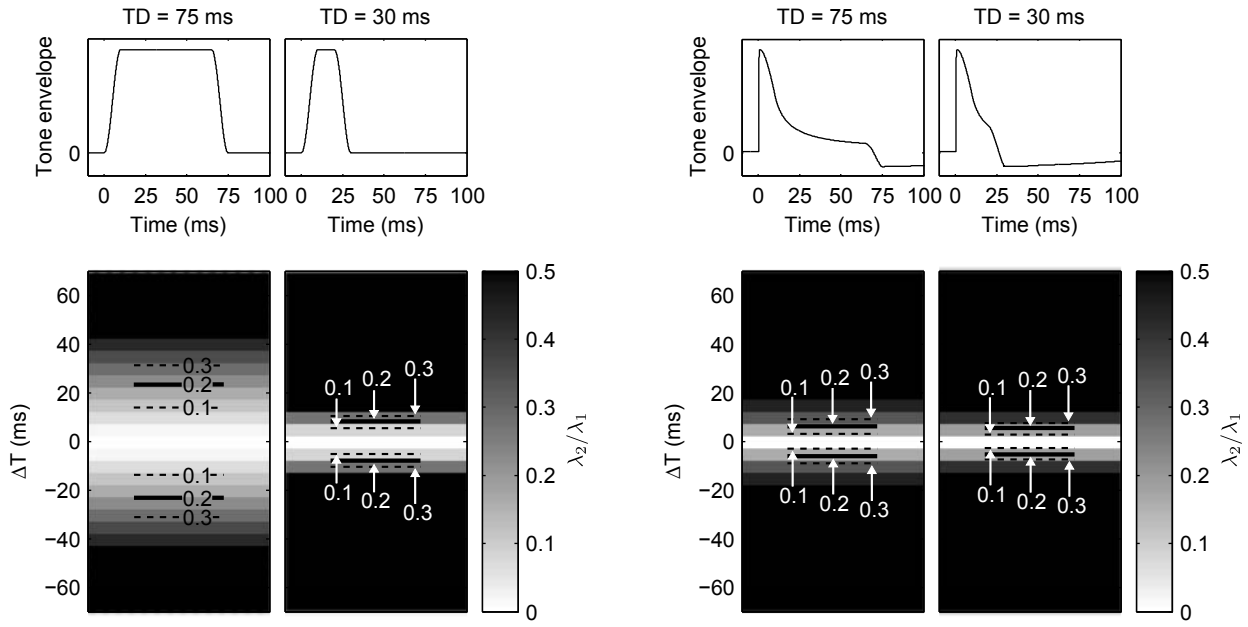


Figure 2.7: Simulation of experiment II with the coherence analysis applied directly to a spectrogram. The top row indicates the temporal envelope of the input stimulus, and the bottom row indicates the eigenvalue ratios. The left panel show the simulation for an "idealized" spectrogram (perfect envelope extraction of the tones), and the right panel shows simulation to an idealized spectrogram processed through the adaptation stage. The grayscale intensity indicates the eigenvalue ratio.

Furthermore, the models suggested in Elhilali et al. (2009) and in the present study include a temporal integration stage prior to the coherence analysis. This stage is realized as a modulation filterbank, reflecting a set of integration time constants corresponding to the bandwidths of the individual modulation filters. An earlier version of the auditory processing model by Dau et al. (1996) suggested a single temporal integrator realized as an 8-Hz low-pass filter (top-left panel of Fig. 2.8), corresponding to an integration time constant of 20 ms. Applying this low-pass filter instead of the modulation filterbank leads to the coherence matrix shown in the bottom left panel of the figure. In this case, the filter response is too slow to follow the rapid onset enhancement resulting from the adaptation stage in the preprocessing. The reduced sensitivity to the tone onset leads to a TD dependency in the model, as in the case described in the preceding text for the idealized constant-amplitude input. In contrast, when applying the modulation filterbank (top right panel of Fig. 2.8), the modulation filters tuned to higher frequencies capture the onset response of the adaptation stage. This leads to predictions with eigenvalue ratios largely independent of TD, consistent with the perceptual data (bottom right panel of Fig. 2.8).

The results thus suggest that the responses of the higher-frequency modulation filters (with center frequencies up to about 50 Hz in the case of the auditory filters considered here) to transients contribute to stream segregation in the framework of the proposed processing model.

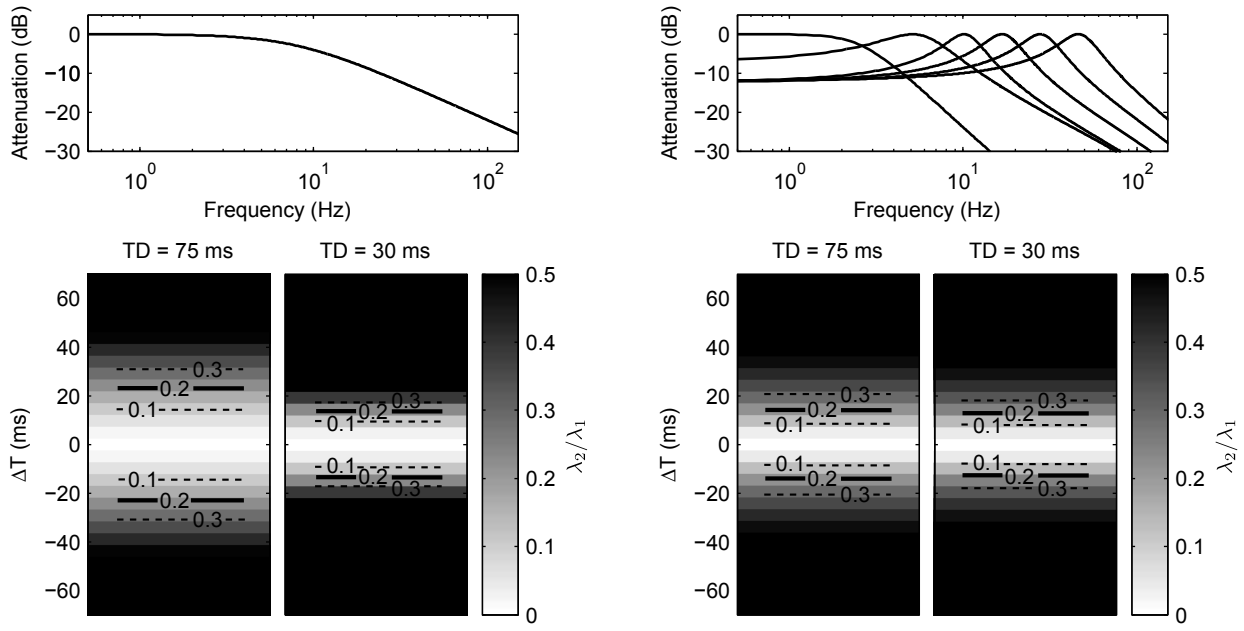


Figure 2.8: Simulation of experiment II with two different temporal integration stages: 8-Hz low-pass filter (left) and modulation bandpass filterbank (right). The input to the temporal integration stage is an idealized spectrogram processed through the adaptation stage. The top panels show the magnitude transfer functions, and the bottom panels show the eigenvalue ratios, indicated by the grayscale intensity.

2.6 Discussion

2.6.1 Stream segregation based on Δf and TRT

van Noorden's (1975) data demonstrated that for a given frequency separation between the A and B tones, fast repeating ABAB sequences are more likely to split into two perceptual streams than slowly repeating sequences. The simulations from the present study showed a pattern consistent with these data with small Δf's and large TRTs promoting a one-stream percept and large Δf's and small TRTs promoting a two stream percept. For an insufficient frequency separation between the tones, the same (or overlapping) peripheral filters were excited such that the activity across peripheral filters becomes more coherent, causing the filters to be perceptually grouped together according to the coherence theory (Elhilali et al., 2009). The assumed peripheral filter bandwidth and spacing in the model determine the amount of overlap of the filters and thus the specific outcome of the predictions. Compared to the model framework provided in Shamma et al. (2013), including a linear inhibitory network to effectively sharpen frequency selectivity, the model proposed here applied comparably wide filters following the ERB scale that has also been used in various previous modeling studies on human masking. The results from the present study suggest that sharper peripheral filters are not required to account for the stream segregation data considered in this study.

Importantly, the data from van Noorden (experiment I of the present study) could not be accounted for by the model solely on the basis of the frequency-selective processing in combination with the coherence analysis because this constellation would not be sensitive to any effect of TRT

on streaming. The predictions suggested that forward masking effectively reduces the amount of spectral overlap, or spread of excitation, for the fast repeating tones compared to the slowly repeating tones. This model-based result is consistent with the results from the study of Bee and Klump (2004) that showed that physiological forward masking (in the starling) increases the spatial separation of neural excitation patterns along the tonotopic axis. Thus a coherence-based analysis per se is not sufficient to account for the data from experiment I, i.e., the model presented in Elhilali et al. (2009) can be expected to be unable to predict a dependency of streaming on TRT.

2.6.2 Grouping of distant spectral components based on onset synchrony

Experiment II further examined the temporal coherence hypothesis. The experimental data supported the hypothesis that temporal synchrony facilitates the fusion of distant spectral components into a single stream despite a large frequency separation. The data furthermore showed that the threshold between a one- and a two-stream percept was essentially unaffected by changes in TRT and TD. This is in contrast to previous hypotheses on the perceptual fusion of spectral components due to onset-/offset-synchrony. Bregman and Pinker (1978) investigated the ability of two pure tones to fuse into a single stream based on the tones' relative synchrony. Their study showed that the tones would segregate into two streams if the temporal overlap was less than 50% and would fuse into a single perceptual stream if the temporal overlap was larger than 88%. This is in contrast to the measured data from the present study, where a ΔT of about 20 ms was observed with no significant effect of either TRT or TD. This implies that the long-duration tones (TD = 75 ms) were fused into one stream at an overlap of 73%, whereas the short-duration tones (TD = 30 ms) were fused into one stream at an overlap of 33%. In the study of Bregman and Pinker (1978), only a single tone duration (TD = 250 ms) was considered. Converting the 88% overlap into an absolute asynchrony yields a value of ΔT = 30 ms, which corresponds fairly well to the data obtained in the present study. The results from the present study suggest that spectral components with an (absolute) asynchrony of less than about 20 ms are fused into a single stream.

The simulations obtained in the conditions of experiment II were in agreement with the perceptual data. The model provided the lowest eigenvalue ratios for low ΔT's, corresponding to a fused percept, whereas for large ΔT's a two-stream percept was predicted. As in the measurements, the simulated thresholds were largely independent of TRT and TD. The model analysis revealed that due to the onset-enhancement in the adaptation stage and the subsequent modulation bandpass filterbank, the coherence analysis becomes sensitive to onset synchrony and less sensitive to the temporal overlap of the stimulus components. Onset synchrony thus appears to be more critical for the fusion of spectral components into a single stream than offset synchrony, which is consistent with previous studies investigating the effect of onset and offset asynchrony on the perception of individual components in complex sounds (e.g., Darwin and Ciocca, 1992; Zera and Green, 1993b; Plack, 2005; Micheyl et al., 2013b).

2.6.3 Effects of cochlear travel times on grouping

The simulations showed an asymmetry in the results from experiment II with respect to ΔT that was not observed in the data. This asymmetry was caused by the bandwidths of the peripheral filters and their associated time constants. The gammatone filters tuned to the frequencies of the A and B tones have time constants of 8.2 and 3.7 ms, respectively. Due to this frequency dependent delay, the maximum coherence at the output of the peripheral stage occurs when tone A leads tone B by 4.5 ms, causing the asymmetry observed in the right panels of Fig. 2.4. The measured data also showed a small (but significant) effect of the sign of ΔT but in the opposite direction.

The discrepancy between the predictions and the data suggests that auditory processes at a level higher than the cochlea, but prior to the mechanisms involved in the grouping phenomena considered here, may roughly compensate for the cochlear delays. This is consistent with earlier results from Uppenkamp et al. (2001) who found a larger perceived "compactness" for click stimuli and downward chirps than for upward chirps that had been designed to produce the largest neural synchronization across frequency at cochlear level. Furthermore, the results are compatible with recent findings from Wojtczak et al. (2012, 2013) where synchrony detection of pure tones with varying frequency separations was investigated. Wojtczak et al. (2012, 2013) found indeed an asymmetry favoring low-frequency components lagging high-frequency components and proposed a higher-level process that slightly overcompensates for the cochlear delay differences across frequency. The perceptual grouping of the stimuli considered in the present study thus seems to take place at a stage subsequent to the compensation process. To account for the "symmetric" ΔT behavior in the measured data from experiment II, the model would need to include such a delay compensation, which is not the case in the present version.

2.6.4 Limitations of the model

The peripheral filtering stage of the model was assumed here to be linear. Effects of a more realistic cochlear processing including, e.g., compressive input-output functions and level-dependent tuning (e.g., Ruggero et al., 1997; Lopez-Poveda and Meddis, 2001; Heinz et al., 2001) were thus not considered. Furthermore, the adaptation stage in the model was the same as the one considered in earlier modeling studies of auditory detection and masking. While this stage has been successful to account for perceptual forward masking and intensity discrimination in human listeners (e.g., Dau et al., 1996; Jepsen et al., 2008), the simulation of adaptation is only phenomenological and not physiologically plausible. Other, more recent, approaches to account for effects of adaptation (e.g., Zilany and Carney, 2010; Wen et al., 2012) might be more reasonable and could lead to conclusions that are different from those drawn in the present study when embedded in a corresponding model of stream segregation.

Furthermore, the parameters of the modulation filterbank were taken from earlier modeling studies focusing on modulation detection and masking; different settings and filter types may lead to different results in the conditions of the present study. Effects of these parameters were

not investigated systematically here and no attempt was undertaken to achieve a best fit to the data. In addition, even though the coherence-analysis back-end used in the presented model as well as in Elhilali et al. (2009) and Shamma et al. (2013) is appealingly straightforward, it represents a pragmatic functional approach and the suggested decomposition of a coherence matrix in Eigenvectors may not be physiologically plausible.

Finally, only two-tone sequences were considered in the present study reflecting simple scenarios of "primitive" grouping, whereas "schema-based" scene analysis commonly associated with high-level cognitive processing was not considered here.

2.6.5 Perspectives

The temporal coherence hypothesis suggests that coherent activity of the neural representation across the "relevant" stimulus dimensions represents the guiding principle for perceptual organization of the acoustic input (Shamma et al., 2011; Shamma et al., 2013). In the present study, only the two dimensions audio frequency and modulation frequency were considered. This approach may also be applicable to conditions of across-frequency processing in comodulation masking release (CMR) where comodulated "off-frequency" flanking noise bands can improve signal detection of a pure tone embedded in an "on-frequency" masking noise band (e.g.; Schooneveldt and Moore, 1987; Hall et al., 1990; Fantini et al., 1993; Piechowiak et al., 2007; Dau et al., 2013). In the model framework presented here, comodulated peripheral filters would be grouped together, and the addition of a pure-tone signal to the on-frequency masker noise band would make that peripheral filter "pop out" and form its own stream, which suggests that temporal coherence may be able to account for the across-frequency processing in CMR.

The more complex task of separating a single talker from a mixture of interfering talkers in a realistic acoustic environment appears too challenging for the current version of the proposed model. The spectral components of natural speech typically involve various frequency-glides and broad-band events, such as fricatives, and the current implementation of the model cannot follow transitions across peripheral filters due to the temporal integration of the coherence matrix over the entire duration of the stimulus. The coherence matrix would have to be analyzed in shorter time windows, in addition to the longer integration times also required for stream segregation.

More generally, however, the concept of grouping based on temporal coherence could be effective across multiple dimensions (Shamma et al., 2011; Shamma et al., 2013). The model could be expanded to other perceptually relevant dimensions, including pitch and timbre, as well as other sensory "channels" such as binaural processors or even other sensory modalities such as vision.

2.7 Summary and conclusion

The proposed model combined a physiologically inspired pre-processing stage with a mathematical coherence analysis to estimate the perceptual grouping obtained with two-tone sequences as a function of spectral separation, tone-repetition rate and across-frequency onset/offset asynchrony.

The model analysis suggested that for fast tone repetition rates, forward masking effectively increases the spatial separation of neural excitation across the tonotopic axis and, thus, effectively facilitates stream segregation according to the channeling hypothesis. This finding supports earlier physiological findings in animals that suggested physiological forward masking to be involved in stream segregation of fast repeating tone sequences.

The results from the present study also supported the hypothesis that onset-synchrony provides an important cue for across-frequency grouping of spectral components. The model analysis suggested that the dependency of streaming on onsets may be caused by effects of neural adaptation (emphasizing stimulus onsets while attenuating steady-state portions of the stimulus) and subsequent processing through modulation frequency specific filters.

The results further suggested that the mechanism for grouping spectral components due to synchrony takes place at a stage above a process that roughly compensates for cochlear delays across frequency.

Overall, the proposed modeling framework may be useful to study the contributions of bottom-up auditory features on perceptual grouping, also in more complex acoustic scenarios than those considered here.

Acknowledgments

This work was supported by the Technical University of Denmark, the Oticon Foundation, and a research consortium with GN ReSound, Widex and Oticon.

3

Effects of sound intensity on auditory stream segregation of pure tone sequences[†]

The perceptual organization of sounds into separate auditory streams has been hypothesized to rely on the activation of non-overlapping neural populations. For example, according to the "peripheral-channeling" theory, spectral differences between sounds facilitate their perceptual segregation due to the tonotopic organization of the auditory system. As a consequence, the level-dependent frequency-selective processing should lead to reduced stream segregation at high sound intensities due to the larger overlap of neural excitation than in the case of lower-intensity sound stimulation. This hypothesis was investigated through listening experiments as well as simulations with a computational model of stream segregation comprising a level-dependent auditory preprocessing followed by a temporal coherence analysis back end. The experimental data obtained with alternating two-tone pulse sequences demonstrated that the stimuli presented at high sound intensities indeed facilitated a fused percept. However, the observed intensity effect was much smaller than expected according to the peripheral-channeling hypothesis and computational modeling. Furthermore, the perceptual data showed a substantial amount of across-listener variability which was partly caused by a response bias with respect to the order of stimulus presentation. Overall, the data from the present study provide strong constraints for future modelling frameworks of auditory stream segregation.

3.1 Introduction

Natural acoustic environments often contain multiple, simultaneously active sound sources. Despite the complexity of the resulting acoustic signal, normal-hearing listeners are usually able to perceptually segregate a single sound source from a mixture of sounds, enabling them to follow a conversation in a crowded room or hear out an instrument from a piece of music. The process of perceptually segregating a single sound source from a mixture is referred to as auditory stream segregation (Bregman, 1990) and relies on a range of acoustic cues, such as spectral content, pitch, or spatial location (for a review see, e.g., Bregman, 1990; Moore and Gockel, 2002; Carlyon and Gockel, 2008).

Early studies of auditory stream segregation suggested that "peripheral channeling", or tonotopic

[†] This chapter is based on Christiansen and Dau (2015).

separation produced initially by cochlear filtering, may provide the physiological underpinnings of the phenomenon of "auditory streaming" (van Noorden, 1975; Hartmann and Johnson, 1991; Beauvois and Meddis, 1996; McCabe and Denham, 1997). According to this framework, sounds that stimulate different populations of tonotopically tuned neurons are segregated into separate streams, whereas sounds that stimulate the same neural population are integrated into a single perceptual stream (Fishman et al., 2004; Micheyl et al., 2005; Bee et al., 2010).

However, several studies showed that auditory stream segregation can also occur with acoustic stimuli that have strongly overlapping, or even essentially identical excitation patterns, relying instead on other stimulus properties or attributes, such as fundamental frequency (F0) differences (Vliegen and Oxenham, 1999, Vliegen et al., 1999; Grimault et al., 2000), wave-shape induced timbre (Roberts et al., 2002), amplitude-modulation rate (Grimault et al., 2002) or differences in intensity (van Noorden, 1975). The idea that stream segregation occurs when sounds activate distinct (or only weakly overlapping) neural populations may still be valid when generalized to other sound attributes than frequency, i.e. by considering populations of neurons selective to these attributes (Itani and Klump, 2009; Gutschalk et al., 2007; Shamma and Micheyl, 2010). However, while it is possible to segregate sounds into separate streams in the absence of differences in excitation pattern, spectral differences still appear to be a dominant factor for stream segregation (Singh, 1987; Vliegen et al., 1999).

An experimental paradigm that has been used to quantify listeners' ability to perceptually segregate sounds based on spectral differences was developed by van Noorden (1975). By presenting a sequence of two pure tones, A and B, in an ABA-ABA pattern, van Noorden investigated the effect of frequency separation and tone repetition time (TRT) on perceptual organization. The results showed that tone sequences tended to perceptually fuse into a single stream when the frequency separation was small (being perceived as a "galloping" rhythm), and to perceptually split into two streams when the frequency separation was large (one fast tone sequence and one slow). By instructing the listeners to focus their attention on the galloping rhythm, van Noorden determined the maximum frequency separation for which it was possible to hear the tones as one stream (the temporal coherence boundary; TCB). Subsequently, he instructed the listeners to try to "hear out" one of the pure tones from the mixture, and thereby determined the minimum frequency separation required to be able to perceptually segregate the tones into separate streams (the fission boundary; FB). The results showed that the FB was largely independent of the TRT, but that the TCB was small for fast tone sequences (small TRTs) and large for slow tone sequences (large TRTs).

The observations by van Noorden can qualitatively be explained by the peripheral channeling theory, as tones with a small spectral separation tend to stimulate the same or overlapping regions of tonotopically tuned neurons, whereas tones with a large frequency separation tend to stimulate separate neural populations. The observation that the TCB depends on TRT and not only spectral separation can furthermore be explained in terms of physiological forward masking. Neurophysiological measurements in birds (Bee and Klump, 2004; 2005) and monkeys (Fishman et al., 2004) indicated that, for fast repeating tone sequences (small TRTs), physiological forward

masking reduces the overlap of neural responses to alternating tones, effectively leading to a non-overlapping neural response for fast tone sequences, thus promoting a segregated percept.

Further support for this hypothesis was recently presented in a study by Christiansen et al. (2014), where a computational model of auditory streaming was presented. This model combines a preprocessing stage that mimics the signal processing in the peripheral human auditory system (Dau et al., 1997a) with a temporal coherence analysis (Elhilali et al., 2009) that groups neural channels together into a single stream if they are activated coherently over time. Christiansen et al. (2014) showed that human auditory frequency selectivity, in combination with properties of forward masking, enabled the model to account for the findings of van Noorden (1975). The study showed that stimuli with small frequency separations tended to stimulate the same or overlapping auditory filters, whereas large frequency separations were more likely to stimulate separate auditory filters. Furthermore, for fast tone sequences, the forward masking in the computational model limited the spread of excitation along the tonotopic axis, making the tone sequences less likely to excite overlapping auditory filters. Thus, the modelling work of Christiansen et al. (2014) suggested that stimulus components are grouped together if they evoke neural synchronization at the level of the internal representation of the stimuli after auditory preprocessing.

Following this idea, wider auditory filters should result in a fused percept for larger frequency separations. As the auditory bandwidth increases with increasing intensity of the stimuli (e.g. Glasberg and Moore, 1990), at least for medium and high frequencies, the TCB and FB defined by van Noorden (1975) should also increase with increasing intensity. The dependency of stream segregation on sound intensity was investigated by Rose and Moore (2000) who measured the effect of stimulus intensity on the FB by presenting ABA-ABA stimuli at levels of 40, 55, 70 and 85 dB sound pressure level (SPL). They found that the increase in intensity from 40 to 85 dB SPL resulted in an increase of the FB consistent with the increase of the equivalent rectangular bandwidth (ERB) of the auditory filter with increasing intensity (Glasberg and Moore, 1990). However, the study by Rose and Moore (2000) only considered a single TRT and only estimated the effect of stimulus intensity on the FB (and not the TCB).

The present study explored the effect of sound intensity on both FB and TCB for several TRTs. The hypothesis was that the increased bandwidth of the auditory filters at high intensities should lead to more fused percepts. The experimental conditions were also analyzed in the framework of the computational model by Christiansen et al. (2014), modified to reflect the non-linear behavior of the auditory periphery with changes in intensity.

3.2 Method

3.2.1 Experimental design

Stimuli

The stimulus consisted of two pure tones, A and B, presented in an ABA-ABA- sequence as illustrated in the top panel of Fig. 3.1. The duration of the tones (TD) was 40 ms including raised cosine onset and offset ramps of 5 ms. The onset-to-onset time between successive tones was controlled by the TRT and the experiment was conducted for TRTs of 60, 80, 100, 120 and 140 ms. The duration of the silent interval between ABA triplets was equal to the TRT. The frequency of the A tone (f_A) was kept fixed at 1 kHz and the frequency of the B tone (f_B) was swept from +20 to 0 to +20 semitones relative to the A tone over 60 seconds. The tones were presented such that the frequency separation between A and B decreased linearly on a semitone scale during the first 30 seconds, and increased accordingly during the second half, as illustrated in the bottom panel of Fig. 3.1. The experiment was presented at three different sound intensities, corresponding to levels of 40, 60 and 80 dB SPL for the steady-state part of the tones.

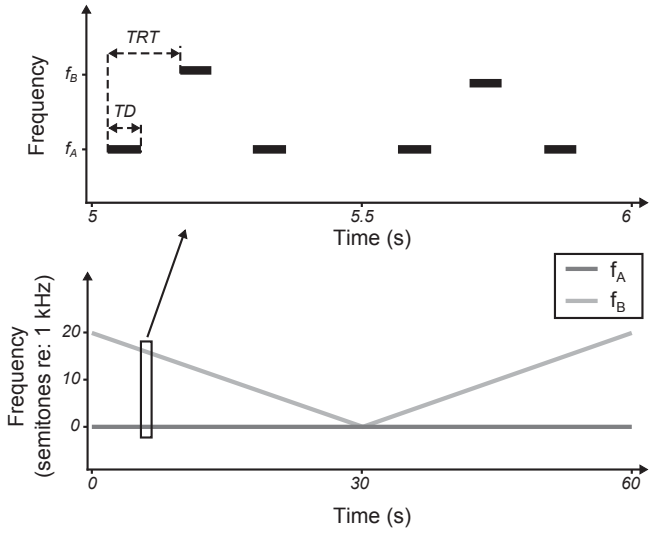


Figure 3.1: Schematic representation of the stimuli. The top panel shows a short segment of the ABA-ABA stimulus, with TD indicating the tone duration, and TRT indicating the tone repetition time. The bottom panel indicates how the frequency of the B tones was swept from +20 to 0 to 20 semitones relative to the A tone over 60 seconds.

Listeners

10 normal-hearing listeners participated in the experiment, including the first author. The group consisted of 4 male and 6 female listeners, aged between 19 and 34 years. The listeners were monetarily compensated for their participation at an hourly rate, and measurement sessions lasted

between 1 and 2 hours including breaks. All listeners received 1 hour of training in the task before measurements began, and 3-4 sessions were required to complete the experiment.

Procedure

During the stimulus presentation, the listeners' task was to indicate whether they heard the stimulus as one stream (galloping rhythm) or two streams (fast tones and slow tones) by pushing one of two buttons labelled "1 stream" or "2 streams", respectively. At the beginning of each stimulus presentation, the "2 streams" button was activated, and the listeners therefore pushed the button labelled "1 stream" when the percept changed to a galloping rhythm, and subsequently pushed the button labelled "2 streams" when the percept changed back into two separate streams.

All five TRTs were tested at all intensities, resulting in 15 conditions. The stimuli in the different conditions were presented in random order. Each condition was measured four times, where the listeners were instructed to focus on the galloping rhythm (thereby measuring the TCB) and another four times where the listeners were instructed to focus on the fast repeating tone (thereby measuring the FB). Thus, overall, 60 stimuli were presented to measure the TCB, and another 60 stimuli were presented to measure the FB. Prior to the data collection, all listeners were trained in the task using 30 stimuli with each set of instructions. Half of the listeners began with a training session measuring the FB, followed by a training session measuring the TCB, while the other half started with the TCB followed by the FB. For the data collection the listeners measured the FB and TCB in the same order as they had trained.

Apparatus

All stimuli were generated and presented using Matlab 2011b (Mathworks). A sampling rate of 44.1 kHz was used, and the signals were presented through a personal computer with a 24-bit soundcard (RME DIGI 96/8 PAD). The stimuli were presented using circumaural headphones (Sennheiser HD580). The listeners were seated in a double-walled, sound insulated booth with a computer monitor that displayed instructions during the experiment.

3.2.2 Simulations

Model structure

Figure 3.2 shows the structure of the computational model used to simulate the data from the different experimental conditions. The model is based on the one presented in Christiansen et al. (2014), but contains a non-linear level-dependent peripheral processing stage instead of the linear gammatone filterbank (Patterson et al., 1987). The "auditory spectrogram" and "temporal integration" stages of the model are the same as in Jepsen et al. (2008) and the "coherence analysis"

stage in the back-end is the same as in Elhilali et al. (2009). The main properties of the model are described in the following.

The model takes a digital signal as input, where a signal with a root mean square (RMS) value of 1 corresponds to a maximum sound pressure level of 100 dB SPL. The digital signal is processed through an outer- and middle-ear stage realized by two linear phase finite impulse response (FIR) filters. The output of the stage is assumed to represent the peak velocity of vibration at the stapes as a function of frequency. The signal is then processed through a dual-resonance non-linear filterbank (DNRL; Meddis et al., 2001), accounting for level-dependent frequency selectivity and compression. The DRNL is followed by a hair-cell transduction stage, roughly simulating the physical transduction of the mechanical vibration of the basilar membrane (BM) into receptor potentials at the inner hair cells. The output of the hair-cell stage is transformed into an intensity-like representation through a squaring expansion, simulating the square-law behavior of rate-versus-level functions of AN fibers near the AN threshold. Lastly, the output of the expansion stage is processed through the adaptation stage of the model, which simulates adaptive properties of the auditory periphery, and effectively simulates forward masking in the model framework. Together, these processing stages transform the digital sound signal into an auditory spectrogram-like representation.

The temporal integration stage performs a short-term integration of the auditory spectrogram produced by the front-end of the model. This is realized by processing the output of the adaptation stage through at temporal modulation filterbank (Dau et al., 1997a), where the modulation filters represent a set of integration time constants corresponding to the inverse of their respective bandwidths. Details of the processing of the auditory spectrogram and the temporal integration stage can be found in Jepsen et al. (2008).

The back-end of the model contains a temporal coherence analysis on the basis of which channels with coherent activity over time are fused into a single stream, and incoherently activated channels are segregated into separate streams. This is done by creating a dynamic coherence matrix, $C(t)$, where each element $c_{ij}(t)$ of the matrix is given by Eq. 3.1

$$c_{ij}(t) = \sum_{\omega} y_{i,\omega}(t) y_{j,\omega}(t) \tag{3.1}$$

with i, j indicating DRNL filter indices, and ω representing a given modulation filter. As in Christiansen et al. (2014), the output of the temporal integration stage was half-wave rectified before the calculation of the coherence matrix to avoid negative values of the coherence matrix, and the dynamic property of the coherence matrix was discarded by integrating the coherence matrix over time.

The coherence matrix was quantified through an eigenvalue decomposition, where the eigenvectors indicate which peripheral channels are coherently activated over time, and thus, which channels should be grouped into the same perceptual stream. The eigenvalues indicate the

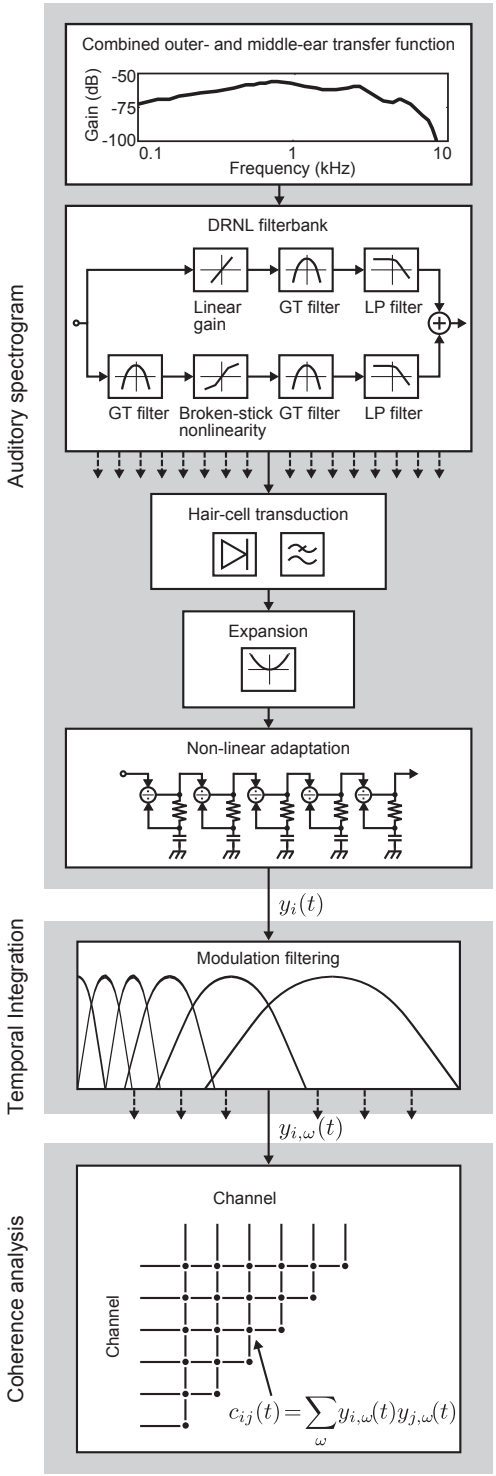


Figure 3.2: Block diagram of the processing model used in the present study. The first stage of the model consists of a finite impulse response (FIR) filter simulating the outer- and middle-ear transfer function, a DRNL filterbank, a simple inner hair-cell model, a square expansion and an adaptation stage. The second stage consists of a modulation band-pass filterbank, integrating the auditory spectrogram over several time scales. The last stage is an across-frequency coherence network which determines the perceptual organization of the stimulus.

strength of the eigenvectors, and by calculating the ratio between the second largest (λ_2) and the largest (λ_1) eigenvalue, the relative importance of the second largest to the largest eigenvector is determined. If the output of the different channels of the preprocessing is temporally coherent, the largest eigenvector is sufficient to describe the coherence matrix C, and the eigenvalue ratio (λ_2/λ_1) will be close to 0, indicating that the input stimulus creates a fused percept. If the channels are temporally incoherent, more than one eigenvector will be required to describe the coherence matrix C, and the eigenvalue ratio (λ_2/λ_1) will be larger than 0, indicating that the stimulus gives rise to a segregated percept.

Simulation parameters

In the simulations, the stimuli were analyzed in segments of 3 seconds duration in which f_B was kept constant, as opposed to the perceptual experiment where f_B was varied continuously over the 60 s duration. This change was made to account for the fact that the model provides a single estimate of the perceptual organization of the given stimulus analyzed, whereas a human listener's percept may change during the stimulus presentation. The model was applied to combinations of 5 TRT values (60 to 140 ms in steps of 20 ms), 3 stimulus intensities (40, 60 and 80 dB SPL), and 81 frequency separations (f_B 0 to 20 semitones above f_A in steps of .25 semitones) for a total 1215 conditions. For each condition, the eigenvalue ratio (λ_2/λ_1) was calculated and used as a measure of the perceptual organization.

3.3 Results

3.3.1 Experimental data

The results of the experiment showed a substantial inter-individual variability. Therefore, both mean data, averaged across listeners, as well as individual data are presented. The data were analyzed through 3-way mixed-model analyses of variance (ANOVA), with the FB or TCB as the dependent variable, intensity and TRT as within-listener factors, and experiment order as a between-listeners factor. For ease of reading, only the main results are presented here, whereas the results from the full statistical analysis can be found in the appendix of this chapter.

The left panel of Fig. 3.3 shows the mean FB (filled squares) and TCB (filled circles) as a function of TRT obtained for the three stimulus intensities 40 (black), 60 (dark gray) and 80 dB SPL (light gray). For a given intensity, the FB was roughly constant with respect to the TRT, and no significant effect of TRT was found [$F(1.56,12.48) = 0.42$, $p = 0.61$]. With increasing intensity, the FB tended to increase, and this increase was significant [$F(2,16) = 24.77$, $p < 0.001$]). Regarding the TCB, an increase was observed with increasing TRT at 40 and 60 dB SPL, whereas at 80 dB SPL the TCB was roughly constant with respect to TRT. This was also found in the statistical analysis that showed a significant effect of TRT [$F(1.88, 15.04) = 6.01$, $p = 0.01$] and a significant interaction

between TRT and intensity [F(3.28, 26.24) = 6.01, p < 0.01]. The statistical analysis also showed that the main effect of intensity was significant for the TCB [F(2, 16) = 16.56, p < 0.001].

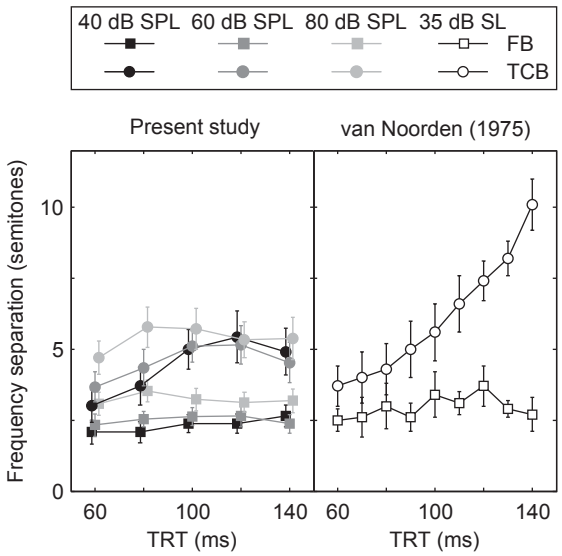


Figure 3.3: Mean results from the experiment plotted as a function of TRT (left panel) together with the results from van Noorden (1975) (right panel) for comparison. The grayscale indicates the stimulus intensity on the left panel, and for both panels the circles and squares represent the TCB and FB, respectively. The error bars represent ±1 standard error of the mean.

For comparison, the right panel of Fig. 3.3 shows the mean results from van Noorden (1975), measured at 35 dB sensation level (SL). The data from van Noorden showed an almost constant FB with respect to TRT, similar to the data from the present study (left panel). For the TCB, the data from van Noorden showed a steep increase with increasing TRT which is different from the results of the present study, where only a moderate increase in TCB with increasing TRT was found, and only at the lowest intensities.

To illustrate the change of FB and TCB with increasing intensity, Fig. 3.4 shows the differences between the TCB and FB at 60 and 80 dB SPL relative to the TCB and FB at 40 dB SPL. Post-hoc paired t-tests, Bonferroni corrected for multiple comparisons (30 comparisons), were applied to test which FBs and TCBs were significantly different, and the results of the t-tests are indicated in the figure by asterisks. The post-hoc tests showed that the TCB was significantly increased at 80 dB SPL relative to both 40 and 60 dB SPL, but only for the two shortest TRTs. For the FB, the 80 dB condition resulted in a significant increase for all TRTs relative to the 40 dB condition and for the two shortest TRTs relative to the 60 dB condition.

Figure 3.5 shows the individual data obtained in the present study, represented in a similar way as in Fig. 3.3. Each panel of Fig. 3.5 shows the results for an individual listener. The listeners who began with the measurement of the TCB (S1-S5) are shown on the left and the listeners who began with the measurement of the FB (S6-S10) are shown on the right. Large inter-individual differences were observed in terms of the dynamic range of the results. The listeners who started with the TCB measurement (left panels) tended to show larger TCBs and FBs than those listeners who started

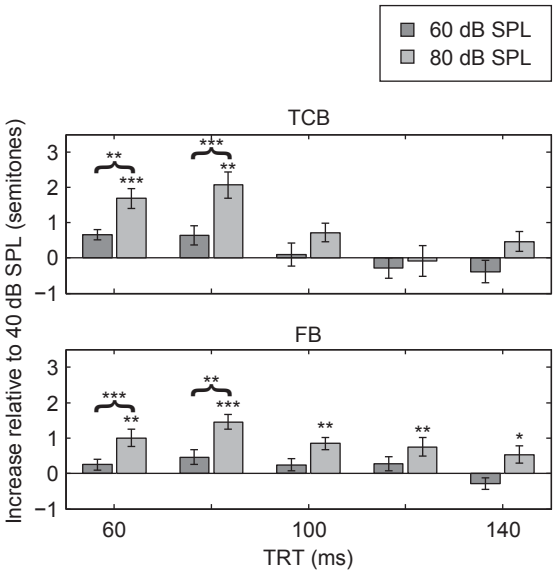


Figure 3.4: Differences in TCB and FB due to stimulus intensity. The panels show the change in TCB (top panel) and FB (bottom panel) of increasing stimulus intensity from 40 to 60 dB SPL (dark gray) and from 40 to 80 dB SPL (light gray), and the errorbars indicate ± 1 standard error. The asterisks show the results of post-hoc paired t-tests, where *, ** and *** indicates $p < 0.05$, $p < 0.01$ and $p < 0.001$, respectively.

with the FB measurement (right panels), but the main effect of experiment order was not significant for either the FB [$F(1,8) = 2.53$, $p = 0.15$] or the TCB [$F(1,8) = 1.80$, $p = 0.22$]. For listeners S1-S5, the TCB tended to increase monotonically with increasing TRT, but for listeners S6-S10, the TCB was highest at medium TRTs and lower for both smaller and larger TRTs. The statistical analysis revealed that this interaction between experiment order and TRT was significant [$F(1.88, 15.04) = 4.65$, $p = 0.03$].

Figure 3.6 shows the FB and TCB obtained at 40 dB SPL (filled black symbols), grouped by experiment order, together with the results from van Noorden (open light gray symbols). The stimulus intensity of 40 dB SPL correspond well to the stimulation intensity of 35 dB SL in the van Noorden study, as the reference hearing threshold for circumaural headphones in the frequency range used in the present study (1-3.2 kHz) was between 2.5 and 6 dB SPL (ISO-389-8, 2004). The data show that the listeners who performed the experiment in the same order as in van Noorden's study (TCB first, FB second; left panel) provided results that are largely consistent with the results from van Noorden, whereas the listeners who performed the experiments in the opposite order (FB first, TCB second; right panel) showed, on average, more "compressed" results, i.e. the boundaries were closer to 0 semitones. This suggests that a substantial amount of the difference observed between the results from the present study and those from van Noorden (1975) may be explained by the experimental procedure. Regarding the influence of sound intensity on the FB and the TCB, the statistical analysis showed no significant interaction between experiment order and intensity (FB: [$F(2,16) = 0.06$, $p = 0.94$], TCB: [$F(2,16) = 1.78$, $p = 0.20$]), indicating that, for the effects of stimulus intensity, all listeners can be analyzed as a single group.

In summary, the data showed that with increasing sound intensity, the FB increased for all TRTs

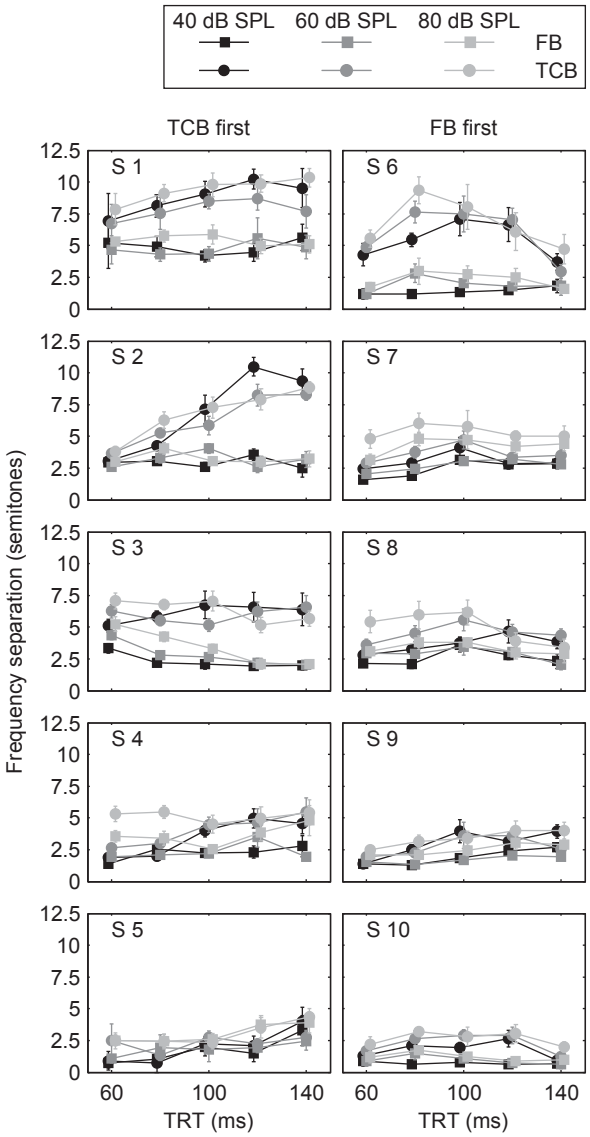


Figure 3.5: Individual results for all 10 listeners participating in the experiment. The grayscale indicates the stimulus intensity and the circles and squares represent the TCB and FB, respectively. The error bars represent ± 1 standard error of the mean. Subjects S1-S5 started by measuring the TCB, and subjects S6-S10 started by measuring the FB.

and the TCB increased for short TRTs. These increases were significant, despite a substantial inter-individual variability which seems to be related to the order of stimulus presentation in the experimental procedure.

3.3.2 Simulations

Figure 3.7 shows the simulations obtained with the model described in Sec. 3.2.2. Areas with light gray indicate small eigenvalue ratios, corresponding to a one-stream percept, and areas with dark gray represent larger eigenvalue ratios, corresponding to a two-stream percept. The three panels represent the results for the stimulus intensities of 40 (left), 60 (middle) and 80 dB SPL

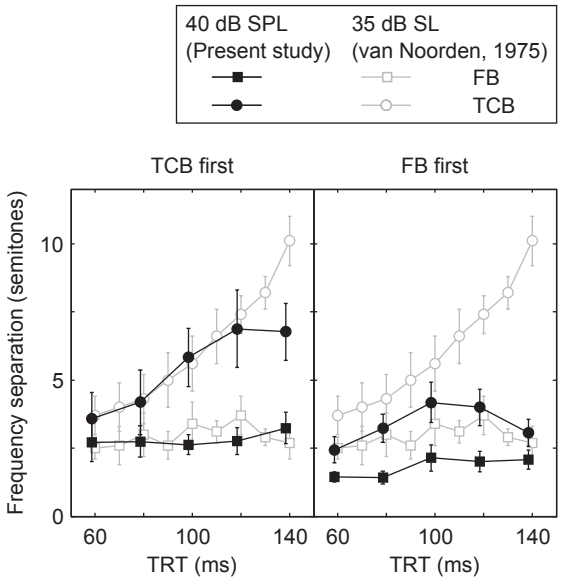


Figure 3.6: Data from the experiment with a sound intensity of 40 dB SPL, grouped based on experiment order. The black symbols in left panel show the results from the listeners who began by measuring the TCB and the black symbols in the right panel show the results from the listeners who began by measuring the FB. The open grey symbols are the results from van Noorden (1975), replotted to ease comparison. For both panels the circles and squares represent the TCB and FB, respectively. The error bars represent ± 1 standard error of the mean.

(right). For illustration, the solid curves in the three panels represent iso-eigenvalue-ratio contours of $\lambda_2/\lambda_1 = 0.026$ and $\lambda_2/\lambda_1 = 0.11$. The dashed curves represent alternative eigenvalue ratios of 0.06 and 0.15.

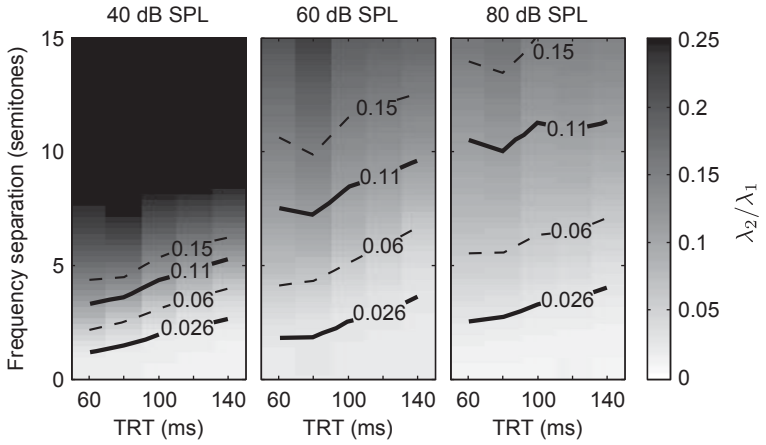


Figure 3.7: Simulation results obtained with the computational model. The greyscale intensity indicates the eigenvalue ratio for a specific combination of TRT and frequency separation, where a small eigenvalue ratio (light gray) corresponds to a one-stream percept, and a large eigenvalue ratio (dark gray) corresponds to a two-stream percept. The three panels show the results for the three sound intensities of 40 (left), 60 (middle) and 80 dB SPL (right). The curves in the three panels indicate contours with fixed eigenvalue ratios.

Some trends in the simulations are similar to those observed in the experimental data: (i) The model predicted a one-stream percept for small frequency separations, and (ii) tone sequences with

a small TRT were more likely to produce a two-stream percept than large TRTs which can have a larger frequency separation and still produce a one-stream percept. However, the simulations showed a strong effect of intensity, as the eigenvalue ratios decreased with increasing intensity for all combinations of TRT and frequency separation. This is reflected in the lighter gray areas in the middle and right panels compared to the left panel, as well as in the position of the iso-eigenvalue-ratio-contours that are shifted towards larger frequency separations with increasing intensity.

To directly compare the simulations to the measured data, the eigenvalue ratios which provided the best fit to the data at 40 dB SPL were selected ($\lambda_2/\lambda_1 = 0.026$ for the FB, $\lambda_2/\lambda_1 = 0.11$ for the TCB) and were also chosen to predict FB and TCB at 60 and 80 dB SPL The resulting FBs and TCBs are shown in Fig. 3.8, using the same scale and symbols as used for the experimental data in Fig. 3.3 (left panel). The simulated FBs (squares) are similar in magnitude to the values observed in the experimental data and the increase of FB with increasing intensity is comparable to the increases observed in the data. However, in contrast to the measured FBs, the simulated FBs show a clear increase with increasing TRT whereas the experimental data showed no effect of TRT. Regarding the TCB (filled circles), the model also predicts an increase with increasing TRT which is roughly consistent with the data. However, the predicted increase with increasing intensity is much larger than that observed in the experimental data, thus representing a major discrepancy between simulations and data, as discussed in more detail further below (Sec. 3.4.4).

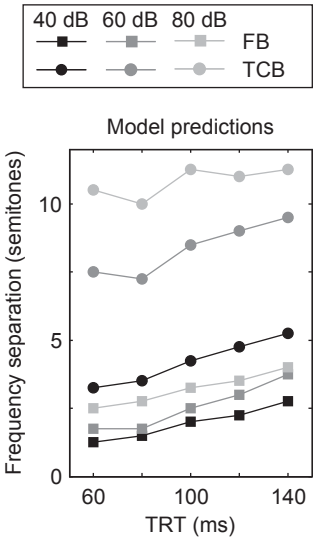


Figure 3.8: Model predictions of FB (squares) and TCB (circles) at sound intensities of 40 (black), 60 (dark gray) and 80 dB SPL (light gray). The model predictions correspond to eigenvalue ratios of $\lambda_2/\lambda_1 = 0.026$ for the FB and $\lambda_2/\lambda_1 = 0.11$ for the TCB.

3.4 Discussion

3.4.1 Inter-individual variation of FB and TCB data

The measured data showed a substantial inter-individual variation of the obtained FB and TCB. The experimental paradigm used the same stimuli for measuring both FB and TCB and the listeners' task was to subjectively evaluate "when a galloping rhythm was heard" and "when the fast and slow tones were heard". This required the listeners to establish internal criteria for the two percepts, and these criteria appeared to be sensitive to listener bias. The inter-individual variation observed in the present study was much larger than that described in the data by van Noorden (1975), and this may have been partly due to the stimulus presentation order. The data suggest that the listeners who began with a stream segregation task (the FB measurement) were biased towards perceiving the stimuli as two streams, leading to smaller frequency separations for both FBs and TCBs, whereas the listeners who began with a stream fusion task (the TCB measurement) were biased towards perceiving the stimuli as one stream, leading to larger frequency separations for both FBs and TCBs. These effects were obtained despite the training before data collection and persisted over time (spanning several days). While the statistical analysis did not reach significance for the main effect of experiment order for either the TCB or the FB, the TCB did show a significant interaction effect of experiment order and TRT, indicating that the increasing TCB with increasing TRT, as reflected in the van Noorden (1975) data, may be sensitive to listener bias. Thus, the dependency on the experimental order, as well as the substantial inter-individual variation, indicate that the paradigm used in this type of auditory streaming experiment needs to be carefully designed. These results seem relevant given the popularity and widespread citation of the original study by van Noorden (1975).

3.4.2 Effect of stimulus intensity on FB

The data showed that the FB increased with increasing stimulus intensity, consistent with the findings of Rose and Moore (2000). The results thus seem to support the hypothesis that the increased auditory filter bandwidth at high sound intensities increases the overlap of the neural responses to the tones, promoting the perceptual fusion of the tone sequences. The similarity in the increase of the FB observed in the data and in the simulations further suggests that the increase may be accounted for primarily on the basis of the changes in auditory filter frequency selectivity with sound intensity.

3.4.3 Effect of stimulus intensity on TCB

For the TCB, the data also showed an increase with increasing intensity, but only for the smallest TRTs. Consequently, the increase of TCB with increasing TRT as observed at the low intensities (40 and 60 dB) was reduced at the highest intensity (80 dB), where the TCB-function flattened out. The dependency of the TCB on TRT at low intensities has been suggested to be a consequence of

forward masking (e.g. Bee and Klump, 2004; Fishman et al., 2004; Christiansen et al., 2014), where the stimulation of a frequency channel with a tone at the corresponding characteristic frequency (CF) limits the ability of a subsequent non-CF tone to excite the CF channel. This mechanism may reduce the overlap of excitation for rapid tone sequences and, thus, may lead to an increased likelihood for a segregation of the tones into separate perceptual streams.

If forward masking causes the sloping TCB-function, the different pattern observed at the high intensity may be due to corresponding changes in the amount of forward masking. Two factors may contribute to this: Firstly, the forward masking effect of a sound decays exponentially from the offset of the sound, and the rate of decay (in dB/ms) at the offset of the sound is higher at high intensities than at low intensities (e.g., Jesteadt et al., 1982; Dau et al., 1996). Secondly, the wider auditory filters at high intensities provide a smaller attenuation of a non-CF tone than the narrower auditory filters at low intensities. This reduces the time-interval where a CF tone is able to mask a subsequent non-CF tone. Both factors suggest that the reduced spread of excitation of fast tone sequences is most pronounced at low intensities whereas, at high intensities, the reduced spread of excitation would only be effective for very fast tone sequences.

Corresponding effects can be observed in the preprocessing of the computational model, as illustrated in Fig. 3.9A and 3.9B. Figure 3.9A shows the processing of a stimulus with a frequency separation of 3 semitones, presented at 40 dB SPL. The stimulus in the top panels has a TRT of 120 ms, and the stimulus in the bottom panels has a TRT of 60 ms. The left panels show the auditory spectrogram, measured at the output of the adaptation stage, with dashed lines indicating the channels tuned to the A- and B-tones. The right panels show the excitation of the channels tuned to f_A (black) and f_B (gray), with a value of 0 representing a level corresponding to the hearing threshold in the model. For the slowly repeating tones (TRT = 120 ms, top right panel), the excitation produced by the A-tones also excites the channel tuned to the B-tones, and vice versa. This is also reflected in the spread of excitation in the auditory spectrogram (left panel). For the fast repeating tones (TRT = 60 ms, bottom panels), the forward masking in the model limits the spread of excitation. Therefore, the channel tuned to f_A only responds to the A-tones (bottom right panel). The channel tuned to f_B primarily responds to the B-tone, but also responds to the first A-tone in each triplet due to the longer time interval between successive B-tones than between successive A-tones in the ABA-ABA stimulus configuration.

Figure 3.9B shows the processing of the same two stimuli presented in the same way, but for an intensity of 80 dB SPL. For the slow tone sequence (top panels), the channels tuned to f_A and f_B respond to both tones, similar to the processing described earlier for the 40 dB stimulus. However, in contrast to the processing at 40 dB SPL, the fast tone sequence also stimulates both channels. This supports the hypothesis that the reduced spread of excitation of fast tone sequences due to forward masking is mainly observed at low intensities. This suggests that the TCB should shift towards lower TRTs with increasing intensity, consistent with the perceptual data.

However, despite the observation that the internal representations of these stimuli in the model seem to be consistent with the stated hypotheses and the experimental findings, the predicted TCB

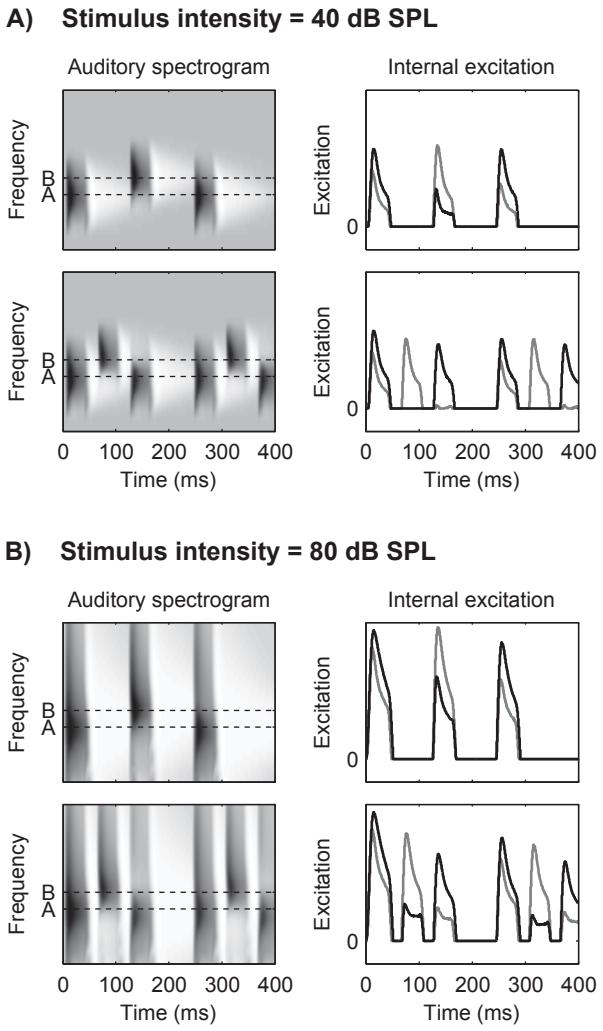


Figure 3.9: Illustration of the role of forward masking for a stimulus presented at 40 dB SPL (A) and at 80 dB SPL (B). The stimuli had a frequency separation of 3 semitones, but for the top panels in both A and B the TRT was 120 ms, whereas the bottom panels had a TRT of 60 ms. The left panels show auditory spectrograms measured at the output of the adaptation stage, where the grayscale represents the magnitude of the signal's internal representation in the model with dark representing a large magnitude and white representing negative values (inhibition). The dashed lines indicate the channels tuned to the A-tones and B-tones. The right panels show the excitation of the two peripheral filters tuned to the A-tone (black) and the B-tone (gray).

largely overestimated the effect of increasing intensity on the magnitude of the TCB, as discussed below.

3.4.4 Discrepancies between simulations and data

The simulation results showed a substantial overestimation of the effect of sound intensity on stream formation in comparison to the measured data, demonstrating a clear limitation of the model in terms of its ability to account for auditory stream formation in human listeners. A step-by-step model analysis revealed that the strong effect of intensity observed in the model is a consequence of the interaction between the level-dependent preprocessing and the coherence analysis. For

a stimulus consisting of temporally non-overlapping spectral components, the across-channel coherence depends entirely on the overlap of the excitation patterns, defined as the output of the auditory filters as a function of their center frequency (Moore and Glasberg, 1984), produced by the two pure-tone sequences. Therefore, the coherence analysis back-end effectively produces a measure of the overlap of excitation across the tonotopic axis. With increasing intensity, the excitation patterns produced by the model broaden, mainly resulting in an increased upward spread of excitation at high intensities.

This is illustrated in the top panels of Fig. 3.10, where the excitation patterns for a stimulus with a frequency separation of 6 semitones is shown for a 40 dB SPL stimulus (left) and a 60 dB SPL stimulus (right). The middle panels show the corresponding auditory spectrograms, where the grayscale indicates the magnitude of the signals' internal representation in the model, with dark gray representing a large magnitude and white representing zero. At both intensities, the auditory spectrograms show that the low-frequency channels primarily respond to the A-tones, whereas the high-frequency channels are activated by both A- and B-tones. The auditory spectrograms also illustrate that the range of frequencies stimulated by both tones is much larger for the stimulus presented at 60 dB than that at 40 dB, resulting in more coherent activation after the preprocessing. The bottom panels show the coherence matrices and eigenvalue ratios for the same stimuli. Due to the increased overall coherence at 60 dB SPL, the resulting eigenvalue ratio is much smaller for the 60 dB stimulus than for the 40 dB stimulus. The eigenvalue ratios suggest that the 40 dB stimulus would be perceived as a two-stream percept, whereas the 60 dB stimulus would be perceived as a one-stream percept (assuming the same thresholds as defined in Sec. 3.3.2). However, the data demonstrated that two streams were perceived at both stimulus intensities. This suggests that the model either overestimates the spread of excitation at high sound intensities, or that the *overall* coherence, as reflected by the eigenvalue ratio, is not always predictive of auditory stream formation.

The model front-end has previously been shown to quantitatively account for various aspects of temporal and spectral masking (e.g., Derleth and Dau, 2000; Jepsen et al., 2008; Jepsen and Dau, 2011). Nonetheless, some other studies suggested a sharper frequency selectivity in the internal representation of the sounds, caused by, e.g., the processing through a lateral inhibitory network (e.g., Shamma, 1985; Chi et al., 2005; Elhilali et al., 2003). Narrower filters at high sound intensities would reduce the predicted increase with increasing intensity in the model framework of the present study. Thus, it may be that the internal representation produced by the model front-end is not appropriate in the context of the coherence analysis proposed here.

Alternatively, the eigenvalue ratio used to estimate the temporal coherence may not be sensitive to the features determining the perceptual organization. The principal idea of the temporal coherence theory is that channels that are activated coherently are grouped together, whereas channels that are activated incoherently are segregated into separate streams. In the auditory spectrograms presented in Fig. 3.10, the frequency channels below 1.1 kHz respond only to the A-tones, whereas the frequency channels above 1.1 kHz respond to both A- and B-tones. Intuitively, there appear to be two incoherently activated frequency regions, which might produce two separate streams

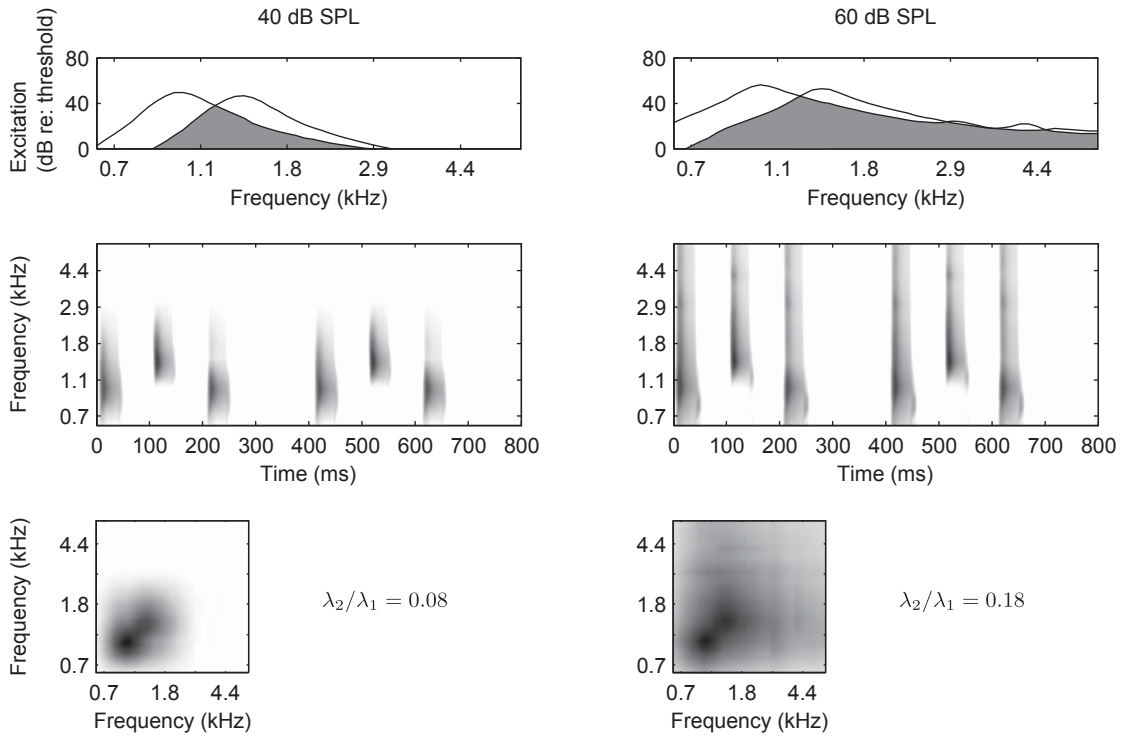


Figure 3.10: Illustration of the role of sound intensity in the model for a stimulus with a frequency separation of 6 semitones and a TRT of 100 ms. The left panels show the processing for a presentation level of 40 dB SPL and the right panels for a presentation level of 60 dB SPL. The top panels show the excitation patterns produced by the two tones, where the dark gray areas indicate the overlap of the excitation patterns. The middle panels show the corresponding auditory spectrograms, where the grayscale indicates the magnitude of the signals' internal representation in the model with dark gray representing a large magnitude and white representing zero. The bottom panels represent the coherence matrices and the corresponding eigenvalue ratios.

according to the temporal coherence theory. However, this is not reflected in the eigenvalue ratio which depends on the ratio of the energy in the two incoherently activated regions. For the 60 dB stimulus, the energy contained in the high-frequency region is much larger than at 40 dB causing the decreased eigenvalue ratio. Thus, the eigenvalue ratio does not directly reflect whether there are channels that are activated incoherently, but rather whether the majority of the energy in the internal representation is coherent. This suggests that the eigenvalue ratio only partially reflects the organizing principle of the temporal coherence concept.

Nonetheless, for a fixed stimulus intensity, the eigenvalue ratio varies monotonically with the overlap of excitation, and consequently also with the amount of channels that are excited by only one of the tones. The eigenvalue ratio can thus be used as a "relative" measure of temporal coherence (and thus, perceptual organization), but the current model cannot offer a general 'threshold' for stream fusion or segregation that successfully account for the data across the different stimulus intensities. Alternative decision metrics may be required to robustly predict the perceptual organization of sounds that vary across intensity and likely also with respect to other stimulus properties.

3.5 Summary and conclusion

The present study investigated the influence of sound intensity on stream segregation as a function of frequency separation and tone repetition rate following the experimental paradigm of van Noorden (1975). This was investigated using psychoacoustic experiments and a computational model of auditory stream segregation. The assumption of the study was that the activation of non-overlapping neural populations in a temporally incoherent manner is the underlying mechanism behind auditory stream segregation and that wider auditory filters at high sound intensities should increase the overlap of the neural responses leading to more fused percepts.

The obtained data supported this hypothesis, showing that the minimum frequency separation required to perceptually segregate two tones increased with increasing stimulus intensity. The study also estimated the maximum frequency separation at which the tones could be perceived as a single stream. The data showed that, at low intensities, fast tone-repetition rates led to low frequency separations relative to slow tone rates, but this effect was not observed at the highest intensity. A model analysis suggested that this interaction between tone rate and intensity may be due to a decrease in effective forward masking with increasing stimulus intensity. The same model also showed an increased tendency to group stimuli presented at high intensities. However, this increase was largely overestimated relative to the experimental data. The analysis revealed that the decision metric of the computational model does not robustly reflect the grouping mechanism of the temporal coherence concept with respect to changes in intensity. This does not reject the temporal coherence theory, but suggests that a different decision metric may be required if a generalizable model of auditory stream formation is desired.

With respect to the dependency of the data on TRT, a substantial inter-individual variation compared to the data from van Noorden (1975) was found. Part of this variation could be accounted for by the stimulus presentation order as those listeners who began with a stream segregation task were biased towards perceiving all stimuli as segregated, whereas the listeners who began with a stream fusion task were biased towards stream fusion. The effect of experiment order indicated that the paradigm used in this type of auditory streaming experiments needs to be carefully designed and that the data showing increasing stream segregation for fast tones vs slow tones may be sensitive to listener bias.

Appendix

Table 3.1: Result of a 3-way, mixed-model ANOVA of the FB .The stimulus intensity and TRT are within-listener factors, and the experiment order is a between-listener factor. Mauchly's test indicated violations of sphericity for the main effect of TRT [$\chi^2(9) = 23.78, p = 0.02$] and for the interaction effect (intensity by TRT) [$\chi^2(35) = 70.46, p = 0.02$] and the degrees of freedom were corrected using Greenhouse-Geisser estimates for sphericity ($\varepsilon = 0.39$ for the main effect of TRT and $\varepsilon = 0.46$ for the interaction effect (intensity by TRT)). The corrected degrees of freedom are indicated in parentheses.

Source	SS	df.	MS	F	p	
Main effects						
Intensity	23.49	2	11.75	24.77	<0.001	***
TRT	1.34	4 (1.56)	0.34	0.42	0.61	
Experiment order	36.60	1	36.60	2.53	0.15	
Two-way interactions						
Intensity × TRT	3.14	8 (3.68)	0.39	2.09	0.11	
Intensity × exp. order	0.06	2	0.03	0.06	0.94	
TRT × exp. order	2.84	4 (1.56)	0.71	0.89	0.41	
Three-way interaction						
Intensity × TRT × exp. order	1.05	8 (3.68)	0.13	0.70	0.59	
Residuals Between listeners	115.62	8	14.45			
Within intensity	7.59	16	0.47			
Within TRT	25.50	32 (12.48)	0.80			
Within intensity × TRT	12.05	64 (29.44)	0.19			

Table 3.2: Result of a 3-way, mixed-model ANOVA of the TCB .The stimulus intensity and TRT are within-listener factors, and the experiment order is a between-listener factor. Mauchly's test indicated violations of sphericity for the main effect of TRT [$\chi^2(9) = 41.49, p < 0.001$] and for the interaction effect (intensity by TRT) [$\chi^2(35) = 71.6, p = 0.02$] and the degrees of freedom were corrected using Greenhouse-Geisser estimates for sphericity ($\varepsilon = 0.47$ for the main effect of TRT and $\varepsilon = 0.41$ for the interaction effect (intensity by TRT)). The corrected degrees of freedom are indicated in parentheses.

Source	SS	df.	MS	F	p	
Main effects						
Intensity	27.28	2	13.64	16.56	<0.001	***
TRT	46.60	4 (1.88)	11.65	6.01	0.01	*
Experiment order	104.82	1	104.82	1.80	0.22	
Two-way interactions						
Intensity × TRT	16.69	8 (3.28)	13.10	6.01	<0.01	**
Intensity × exp. order	2.93	2	1.47	1.78	0.20	
TRT × exp. order	36.04	4 (1.88)	9.01	4.65	0.03	*
Three-way interaction						
Intensity × TRT × exp. order	1.89	8 (3.28)	0.24	0.68	0.58	
Residuals Between listeners	465.68	8	58.13			
Within intensity	13.18	16	0.82			
Within TRT	62.01	32 (15.04)	1.94			
Within intensity × TRT	22.21	64 (26.24)	0.35			

4

Assessing the effects of temporal coherence on auditory stream formation through comodulation masking release[‡]

Recent studies of auditory streaming have suggested that repeated synchronous onsets and offsets over time, referred to as "temporal coherence," provide a strong grouping cue between acoustic components, even when they are spectrally remote. This study uses a measure of auditory stream formation, based on comodulation masking release (CMR), to assess the conditions under which a loss of temporal coherence across frequency can lead to auditory stream segregation. The measure relies on the assumption that the CMR, produced by flanking bands remote from the masker and target frequency, only occurs if the masking and flanking bands form part of the same perceptual stream. The masking and flanking bands consisted of sequences of narrowband noise bursts, and the temporal coherence between the masking and flanking bursts was manipulated in two ways: (a) By introducing a fixed temporal offset between the flanking and masking bands that varied from zero to 60 ms and (b) by presenting the flanking and masking bursts at different temporal rates, so that the asynchronies varied from burst to burst. The results showed reduced CMR in all conditions where the flanking and masking bands were temporally incoherent, in line with expectations of the temporal coherence hypothesis.

4.1 Introduction

An important task of the auditory system is to segregate different sound sources within natural acoustic environments. The ability to perceptually segregate competing sounds and selectively attend to individual sources over time has long been a topic of intense study (for reviews, see Bregman, 1990; Moore and Gockel, 2002; Carlyon and Gockel, 2008). Many experiments have relied on subjective evaluations of perceptual organization, for instance, by asking subjects how many "streams" they perceive. In recent years, an increased emphasis has been placed on more indirect, performance-based measures of auditory stream segregation (e.g., Micheyl and Oxenham, 2010). Measures of performance allow experimenters to eliminate, or at least control for, bias effects and also open up the possibility of studying perceptual organization in non-human species. The aim of the present study was to investigate the effects of temporal coherence on auditory

[‡] This chapter is based on Christiansen and Oxenham (2014).

stream formation using masking and masking release as indirect performance-based measures of perceptual organization.

Early studies suggested that "peripheral channeling," or tonotopic separation produced initially by cochlear filtering, may provide the physiological underpinnings of the phenomenon known as "auditory streaming" (van Noorden, 1975; Hartmann and Johnson, 1991; Beauvois and Meddis, 1996; McCabe and Denham, 1997). According to this framework, sounds that stimulate different populations of tonotopically tuned neurons are segregated into different streams, whereas sounds that stimulate the same neural population are integrated within a single perceptual stream (Fishman et al., 2004; Micheyl et al., 2005; Bee et al., 2010). It has, however, been shown that dimensions other than tonotopic separation, such as fundamental-frequency (F0) differences (Vliegen and Oxenham, 1999; Vliegen et al., 1999; Grimault et al., 2000) or waveshape-induced timbre differences (Roberts et al., 2002) can also induce streaming. Nonetheless, the principle of neural separation may still hold in populations of neurons that are sensitive to higher-level features, such as F0 or pitch (Bendor and Wang, 2005).

More recently, emphasis has been placed not only on the *spatial* separation of neural responses to sounds in a sequence, but also on the *temporal* relationships between them (e.g., Elhilali et al., 2009; Shamma et al., 2011; Micheyl et al., 2013a; Micheyl et al., 2013b). The finding that sounds repeatedly presented synchronously tend to form a single perceptual stream has been referred to as the principle of "temporal coherence" (e.g., Elhilali et al., 2009). Although not explicitly accounted for in earlier neural models of streaming (e.g., Fishman et al., 2004; Micheyl et al., 2005), temporal coherence has been reported to be a relatively strong auditory grouping cue, which can bind together components even when they are relatively widely spaced in frequency.

In the study by Elhilali et al. (2009), the role of temporal coherence in grouping was assessed by measuring listeners' ability to detect a small temporal asynchrony between two spectrally distant target tones that were preceded by a series of repeating tones at the same two frequencies as the target tones. Previous studies have shown that listeners are able to detect asynchronies of just a few milliseconds between spectral components of a complex tone that is perceived as a single auditory object when all components are synchronous (e.g., Zera and Green, 1993a,b, 1995), but that they cannot accurately judge the relative timing (Bregman and Campbell, 1971; Broadbent and Ladefoged, 1959; Neff et al., 1982; Roberts et al., 2002) or synchrony (Micheyl et al., 2010) of sounds that fall into separate auditory streams. Thus Elhilali et al. (2009) used the thresholds from their asynchrony detection task as an indirect measure of perceptual grouping.

Elhilali et al. (2009) showed that when the pairs of preceding tones were presented synchronously (temporally coherent condition), the threshold for detecting asynchrony between the final two target tones was around 2-4 ms, whereas when the preceding tones were presented asynchronously (temporally incoherent condition), the threshold for detecting asynchrony was nearly an order of magnitude larger. This outcome is consistent with the idea that temporal coherence leads to perceptual grouping, even when the target tones are separated by a large frequency difference (15 semitones in this case).

In the same paper, Elhilali et al. (2009) reported physiological results obtained from the primary auditory cortex (AI) of awake but passive ferrets. The cortical units that responded to one or the other of the target tones did not show sensitivity to temporal coherence between the two tones, so that the human behavioral data could not be predicted from the ferret neural data without postulating an additional stage of neural processing that included the computation of temporal correlations of the activity from the AI units.

The discrepancy between the human behavioral and ferret neural data may be due to the presence of additional processing in non-primary cortical networks, as hypothesized by Elhilali et al. (2009). However, other alternatives exist. One possibility is that neural differences are observed at the level of AI only in situations where the subject is awake and attending to the stimuli; in the Elhilali et al. (2009) the ferrets were exposed passively and had no incentive to attend to the stimuli. A second possibility is that humans and ferrets perceive the stimuli differently, although behavioral studies to date suggest generally similar patterns of performance (Ma et al., 2010). A third possibility is that the thresholds in the human behavioral task used by Elhilali et al. (2009) do not accurately reflect the perceptual organization of the stimuli. The task involved asynchrony detection in two conditions: In the temporally coherent condition, all the preceding tone pairs were synchronously gated, so the presence of the target resulted in the only stimulus asynchrony, whereas in the temporally incoherent condition, all the tone pairs were asynchronous, so that the target did not introduce the new "feature" of asynchrony to the stimulus. Thus it is possible that the behavioral results of Elhilali et al. (2009) were determined by the number of "distracting" asynchronies rather than by the perceptual organization of the target stimuli.

To test an alternative measure of streaming that does not suffer from the potential confounds associated with the asynchrony detection task used in Elhilali et al. (2009), we used comodulation masking release (CMR; Hall et al., 1984). The term CMR refers to the finding that the detection of a target (usually a tone) in the presence of a masker with slow amplitude fluctuations can be improved when masker energy at remote frequencies has amplitude fluctuations that are coherent with those of the on-frequency masker. Different processing strategies have been proposed as underlying mechanisms for CMR. In general, two different categories exist: The first category involves within-channel processes, whereby changes in the amplitude envelope of a single masker band (or a broadband masker after filtering through a single auditory filter) can be used to explain CMR; the second category involves across-channel processes, where it is necessary to compare the amplitude envelopes at the outputs of multiple auditory filters to account for CMR. Many aspects of within-channel CMR can be explained to some extent by relatively peripheral auditory processes, such as suppression (e.g., Ernst and Verhey, 2008; Ernst et al., 2010), or by modulation processing with a modulation filterbank following each auditory filter (Verhey et al., 1999). In contrast, across-channel CMR requires processes such as across-channel envelope correlation (Richards, 1987), equalization-cancelation (Buus, 1985), dip-listening (Buus, 1985; Buss et al., 2009), or an across-frequency comparison and integration of modulation information (Eddins and Wright, 1994; van de Par and Kohlrausch, 1998; Piechowiak et al., 2007) to account for the data. An empirical feature that seems to distinguish within-channel from across-channel CMR is that

across-channel CMR is affected by stimulus manipulations that are known to influence perceptual grouping (Grose and Hall, 1993; Dau et al., 2005, 2009; Grose et al., 2005; Grose et al. 2009; Verhey et al., 2012). In particular, the coherent amplitude modulation in the remote frequency (flanking) maskers seems only to aid signal detection in across-channel CMR when the flankers are thought to be perceptually grouped with the on-frequency masker. Other studies (Ernst and Verhey, 2008) have shown that CMR can also be observed with remote flanking bands (separated by as much as three octaves) in conditions that traditionally lead to a segregated percept; however, in the study by Ernst and Verhey (2008), the flanking bands were presented at a much higher intensity than the on-frequency band, and the observed CMR could have been the result of more peripheral effects, such as suppression, rather than across-channel processing (Ernst and Verhey, 2008; Ernst et al., 2010).

In the present study, CMR was used as a measure of the perceptual organization of sounds under the assumption that the CMR produced by maskers remote in frequency from the target and on-frequency masker and with similar intensity will only occur if the masking and flanking bands form part of the same perceptual stream (e.g., Dau et al., 2005, 2009). Perceptual streams were manipulated by embedding the masker and flankers that were synchronous with the target within a context of preceding and following maskers and flankers that were designed to lead to either perceptual integration or segregation of the masking and the flanking noise bursts.

Through this measure, temporal coherence was investigated as a grouping cue, eliminating the potentially confounding influence of asynchronous stimuli within a task of asynchrony detection, such as that used by Elhilali et al. (2009). Temporal incoherence was introduced by manipulating the "gating envelope" in two ways, either through a constant asynchrony between the (on-frequency) masking and (off-frequency) flanking bursts or through different repetition rates for the masking and flanking bursts, leading to constantly varying asynchronies between the masking and flanking band onsets and offsets. In addition, the influence of the temporal coherence of the inherent fluctuations of the narrowband noises ("ongoing envelope") was studied by using ongoing masker and flanker envelopes in the preceding and following bursts that were either correlated or uncorrelated.

4.2 Experiment 1: Effects of temporal incoherence and gating asynchrony on CMR

4.2.1 Rationale

Across-channel CMR has been shown to depend on the perception of the masking and flanking bands within the same perceptual stream. Therefore the temporal coherence hypothesis predicts that CMR will be present when the maskers and flankers are preceded by coherent (i.e., synchronous) masker and flanker bursts and that CMR will be reduced or absent when the preceding masker and flanker bursts are presented incoherently or asynchronously, such that they form separate perceptual streams.

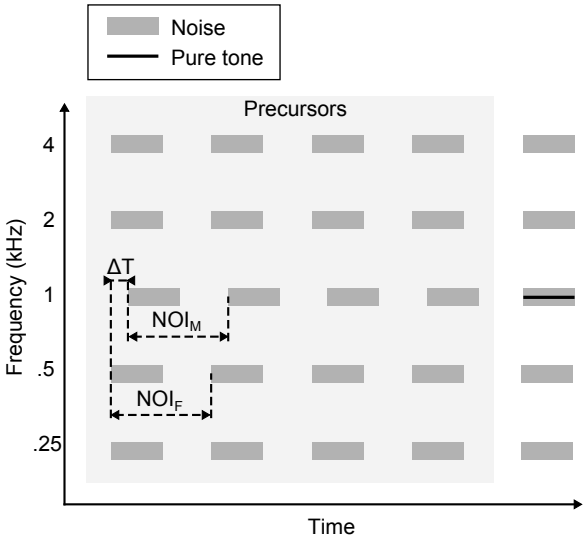


Figure 4.1: Schematic representation of the stimuli used in experiment 1. Five narrow-band noises were presented at octave frequencies between 0.25 and 4 kHz. Each noise burst repeated five times. During the final noise burst a target signal (1-kHz pure tone) was embedded in the central noise band. The final noise bursts were always presented synchronously, but the temporal relationship between the central masking noise-band (MN) and the flanking noise-band (FN) precursors (highlighted in light gray) could be varied. The NOI_M and NOI_F represent the onset-to-onset time between successive MN and FN bursts, respectively. The value of ΔT represents the constant onset asynchrony between the MN and FN precursors in conditions where the NOI_M and NOI_F were equal. In the sketched example $NOI_M = NOI_F$ and $\Delta T \neq 0$ ms.

4.2.2 Method

Stimuli

Figure 4.1 shows a schematic representation of the stimuli used in experiment 1. The target signal (a 1-kHz pure tone) was embedded within a synchronously gated narrow-band (20-Hz-wide) masking noise (MN) centered at 1 kHz. Four flanking noises (FN) were presented synchronously with the target and MN, separated from the MN by ± 1 or 2 octaves (i.e., centered at 0.25, 0.5, 2, and 4 kHz). All FNs were also 20 Hz wide, and the ongoing envelope of each FN was either random or comodulated with that of the MN. Comodulation was achieved by generating the 20-Hz-wide Gaussian noise bands in the spectral domain and using the same amplitudes and phases at the different center frequencies. For the "random" configuration, each noise band was produced with independent randomly generated amplitudes and phases. The target and the noise bursts all had a duration of 187.5 ms, including 20 ms raised-cosine onset and offset ramps.

Prior to the presentation of the target tone and concurrent MN and FN bursts, a series of four precursors was presented (highlighted in light gray). These precursors consisted of noise bursts with the same average spectral and temporal properties as the FNs and MN. The number of precursors was chosen to correspond to the study by Dau et al. (2005). The time intervals between the onsets of successive noise bursts (noise onset interval, NOI) were termed NOI_M and NOI_F for MN and FN

bursts, respectively. All FNs were gated on and off simultaneously across frequency, but the timing of the MN precursors could vary independently from that of the FNs. Asynchronous gating of the MN, relative to the FNs, occurred in conditions where the MN precursors were delayed by ΔT relative to the FNs, or in conditions where $NOI_M \neq NOI_F$. Regardless of the temporal relationship between FN and MN precursors, the final FNs and MN, which were presented together with the target, were always gated on and off synchronously. For conditions with asynchronous precursors ($\Delta T \neq 0$), the NOI between the last precursor and the target interval deviated from the NOI between precursors to enable synchronized target noise bursts. This deviation was implemented by decreasing NOI_M by ΔT/2 and increasing NOI_F by ΔT/2 in the interval between the last precursor and the target. In conditions with either $\Delta T \neq 0$, or $NOI_M \neq NOI_F$, temporally overlapping portions of FN and MN had comodulated ongoing envelopes in the comodulated condition. This was realized by generating long-duration noises with comodulated ongoing envelopes and subsequently applying temporal windows to the noises to obtain the desired temporal gating properties of the precursors. The level of each narrow-band noise was set to 60 dB sound pressure level (SPL), and the level of the target was adaptively varied, as described in the following text, but was initially set to 75 dB SPL to ensure relatively easy detection of the target at the beginning of each adaptive track.

A total of seven different conditions were tested. In condition 1 (baseline), all precursors were synchronized and presented at a NOI_M and NOI_F of 250 ms with a ΔT of 0 ms. Conditions 2-4 were again presented with NOI_M and NOI_F of 250 ms but with values of ΔT of 20, 40, and 60 ms, respectively. Conditions 5 and 6 kept the NOI_M constant at 250 ms, while the NOI_F was either 200 or 300 ms, respectively. Condition 7 had all precursors synchronized with the average NOI_M and NOI_F of 250 ms but with the NOI between each successive burst jittered by $\pm$30 ms with uniform distribution. Condition 7 thus had synchronized on- and offsets across frequency but not the temporal regularity of the baseline condition. All seven conditions were tested in both the random and comodulated configurations.

Procedure

An adaptive, three-interval, three-alternative forced choice procedure was used together with a one-up two-down tracking rule to estimate the 70.7% correct point on the psychometric function (Levitt, 1971). The intervals were marked on a computer monitor, and feedback was provided after each trial. Listeners responded via computer keyboard or mouse. The initial step size of the target level was 8 dB, which was reduced to 4 and 2 dB after the second and fourth reversals, respectively. The adaptive run then continued for an additional six reversals at the final step size, and threshold was defined as the mean of the levels at those last six reversals. Four threshold estimates were obtained and averaged from each listener in each condition.

Listeners

Eight normal-hearing listeners, including the first author, participated in this experiment. The group consisted of four female and four male listeners, aged between 18 and 27 yr. The listeners (except the first author) were compensated monetarily for their participation at an hourly rate, and measurement sessions lasted 1-2 h including breaks. All listeners received 2-3 h of training in the same task before data collection began, and three to four sessions were required to complete the experiment. Data from one of the subjects were excluded from further analysis as the obtained thresholds did not stabilize through either initial training or through the data collection (intra-individual standard deviations remained around 7-10 dB). Thus the reported results are from the remaining seven subjects. The protocol was approved by the University of Minnesota's Institutional Review Board and the listeners provided written informed consent.

Apparatus

All stimuli were generated and presented through MATLAB (Mathworks, Natick, MA), using the AFC toolbox (Ewert, 2013). A sampling rate of 44.1 kHz was used, and the signals were presented through a personal computer with a 24-bit Lynx22 sound card (LynxStudio, Costa Mesa, CA). The stimuli were presented diotically through HD650 circumaural headphones (Sennheiser, Old Lyme, CT). The listeners were seated in a double-walled, sound-attenuating booth with a computer monitor that displayed instructions and feedback throughout the experiment.

4.2.3 Results and discussion

Threshold measurements were generally reliable within the seven subjects whose data were analyzed further; intraindividual standard deviations were typically between 0.5 and 2.5 dB and never exceeded 4 dB across the four estimates. In addition, the pattern of results was very similar across subjects, so only the mean data are reported here and are shown in Fig. 4.2. The top panels show the measured target thresholds in noise bands with random (squares) and comodulated (circles) ongoing envelopes, and the bottom panels show the amount of CMR, defined as the difference between thresholds in the random and comodulated configurations. The left column of Fig. 4.2 shows the results from conditions with a fixed NOI but varying degrees of asynchrony, ΔT; the middle column shows the results from conditions with varying NOI_F; and the right column shows the results from the condition with jittered NOIs. Results from the baseline condition ($\Delta T = 0$ ms, $NOI_M = NOI_F$) are shown in all three columns for comparison (hatched symbols and bars). The thresholds obtained in the baseline condition are consistent with results from Dau et al. (2005, 2009) using no precursor for both random and comodulated noise bands (hatched symbols). A one-way within-subjects (repeated-measures) analysis of variance (ANOVA) with threshold as the dependent variable and condition as the independent variable revealed no significant effect of precursor condition with the random-masker configuration (squares; upper panels) [$F(6,36) = 1.00$,

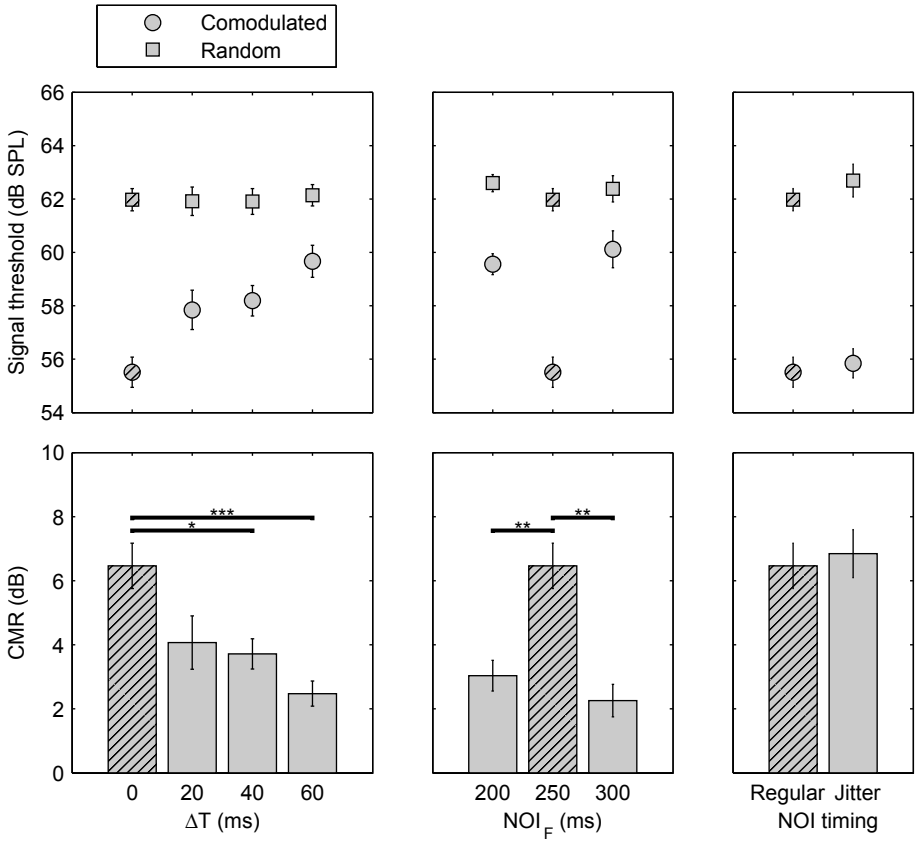


Figure 4.2: Mean results from the seven subjects tested in experiment 1. The upper panels show the detection thresholds for the comodulated (circles) and random (squares) ongoing envelopes of the noise bursts. The bars in the lower panels show the amount of CMR, defined as the difference between the thresholds in the comodulated and random ongoing envelope conditions. The error bars represent $\pm$ 1 standard error of the mean across subjects. The baseline condition (ΔT = 0 ms; $NOI_M = NOI_F$ = 250 ms, no jitter) is indicated by hatching and is shown in all panels. The asterisks in the lower panels show the results of post hoc (Tukey's test) comparisons of CMR between conditions, where *, **, and *** indicate $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively. The leftmost panels show conditions with increasing ΔT but identical NOIs. The middle panels show the effect of varying NOI_F while keeping NOI_M = 250 ms. The rightmost panel show the effect of synchronized, but temporally irregular noise bursts.

p = 0.44]. In contrast, the conditions with the comodulated masker and flankers (circles; upper panels) showed a significant effect of precursor condition [$F(6,36) = 14.7$, $p < 0.001$].

The amount of CMR was treated as the dependent variable in another one-way repeated-measures ANOVA with condition as the factor. A main effect of precursor condition was found [$F(6,36) = 9.97$, $p < 0.001$]. Post-hoc analyses of the CMRs within the three groupings illustrated in the three lower panels of Fig. 4.2 showed a significant reduction in CMR for ΔT of 40 and 60ms, relative to the synchronized precursors (lower left panel), a significant reduction in CMR for conditions with $NOI_M \neq NOI_F$ relative to the $NOI_M = NOI_F$ (lower middle panel), and no significant effect of jittered NOIs (jitter) relative to the baseline condition (regular) (lower right panel).

The results indicate that increasing asynchrony, ΔT, leads to decreasing CMR, as would be expected if the asynchrony led to increased perceptual segregation between the MN and FNs. Previous studies have shown that onset/offset asynchronies larger than 20-40 ms lead to increased

stream segregation (e.g., Turgeon et al., 2002; Turgeon et al., 2005; Bregman and Pinker, 1978; Micheyl et al., 2013b; Christiansen et al., 2014), in good agreement with the data from this study.

Similarly, the middle panels of Fig. 4.2 show that presenting the flanking precursors at a different rate from that of the masking precursors also leads to a significant reduction of CMR, again in line with predictions based on segregation based on temporal incoherence (Elhilali et al., 2009). Last, the rightmost panels of Fig. 4.2 show that temporal irregularities in the form of jittered NOIs do not affect the amount of CMR relative to the baseline condition, indicating that the reduced CMR in conditions with onset/offset asynchronies or different MN and FN rates is not caused simply by the reduced temporal regularity of the stimuli.

For all seven precursor conditions, the threshold for the comodulated configuration was significantly lower than that for the random configuration [seven paired t-tests; $t(6) > 4.46$, $p < 0.003$ in all cases; significant after Bonferroni correction], indicating that none of the precursor conditions led to a complete elimination of CMR as might be expected if perceptual segregation of the masking and flanking bands was complete. This outcome differs from the results of Dau et al. (2005, 2009) and Verhey et al. (2012), who reported a complete elimination of CMR in conditions where the MN was perceptually segregated from the FN using repeated FN bursts as pre- or post-cursors. However, unlike the studies by Dau et al. (2005, 2009) and Verhey et al. (2012), experiment 1 had conflicting streaming cues as the temporal incoherence of the gating envelopes should facilitate a two-stream percept, while the coherent ongoing envelopes of the precursors during portions of temporal overlap may have promoted a one-stream percept (e.g., Hall and Grose, 1990). These conflicting streaming cues may have led to an incomplete perceptual segregation of FNs and MN, resulting in some residual CMR. Experiment 2 was designed to test this potential conflict between ongoing envelope cues and temporal onset and offset gating cues.

4.3 Experiment 2: Influence of ongoing envelope comodulation versus gating synchrony

4.3.1 Rationale

In experiment 1, the potential conflict of incoherent gating combined with coherent ongoing envelopes between the masking and flanking bands may have resulted in incomplete perceptual segregation of the masker and flankers; this in turn may have resulted in residual CMR. To test this hypothesis, two of the precursor conditions from experiment 1 were retested, but with random ongoing envelopes on all (FN and MN) precursors, to eliminate any potential fusion due to ongoing comodulation within the precursors. The ongoing envelopes of the final MN and FNs (i.e., those presented simultaneously with the target tone) remained either comodulated or random, as in experiment 1. If comodulation of the ongoing envelopes within the precursors induced some perceptual fusion, then the use of random ongoing envelopes should result in a further reduction or elimination of CMR in conditions with asynchronously gated precursors.

4.3.2 Method

Stimuli and procedure

The stimuli were identical to those used in experiment 1 except that all precursors always had random ongoing envelopes, regardless of whether the temporal envelopes of the MN and FN presented simultaneously with the target signal were comodulated or random. Only two precursor configurations were tested: $\Delta T = 0$ (baseline condition) and $\Delta T = 60$ ms (maximum onset/offset asynchrony from experiment 1). In both conditions, the NOI_M and NOI_F were 250 ms. The procedure and equipment were identical to those of experiment 1.

Listeners

Ten normal-hearing listeners participated in this experiment. Five of the listeners had also participated in experiment 1 (including the first author). The group consisted of five female and five male listeners, aged between 19 and 34 yr. The listeners were compensated monetarily for their participation at an hourly rate, and measurement sessions lasted between 1 and 2 h including breaks. All listeners received at least 1 h of training in the same task before data collection began, and one to two sessions were required to complete the experiment.

4.3.3 Results and discussion

The results from experiment 2 are shown in Fig. 4.3 together with the results from experiment 1 for the corresponding precursor configurations. As in experiment 1, the intra-individual standard deviations were relatively small (0.5-2 dB, rarely exceeding 4 dB), and all subjects showed similar patterns of results, so only the mean data are shown here. The left panel shows the measured target detection thresholds for random (squares) and comodulated (circles) noises, and the right panel shows the CMR (the difference between random and comodulated thresholds). In both panels, the gray and open symbols indicate results from experiments 1 and 2, respectively. Note that there is some, but not complete, overlap between the subjects in the two experiments.

Mixed-model ANOVAs were carried out separately for the random and comodulated configurations, with threshold as the dependent variable, experiment as a between-subjects factor, and ΔT as a within-subjects factor.[1] The results of the ANOVA for the random configuration showed no significant effect of experiment (1 or 2) [$F(1,15) = 0.39$, $p = 0.54$] or ΔT [$F(1,15) = 0.01$, $p = 0.91$] and no interaction [$F(1,15) = 0.36$, $p = 0.55$]. The absence of an effect of experiment was expected as the stimulus properties were identical across the two experiments. For the comodulated configuration, significant main effects were found for both experiment [$F(1,15) = 6.41$, $p = 0.02$]

[1] Although some subjects participated in both experiments, they were treated as independent for the purposes of this analysis to avoid problems of missing values. Treating the subjects as independent across experiment likely results in a loss of statistical power, making the current analysis a relatively conservative test of significance.

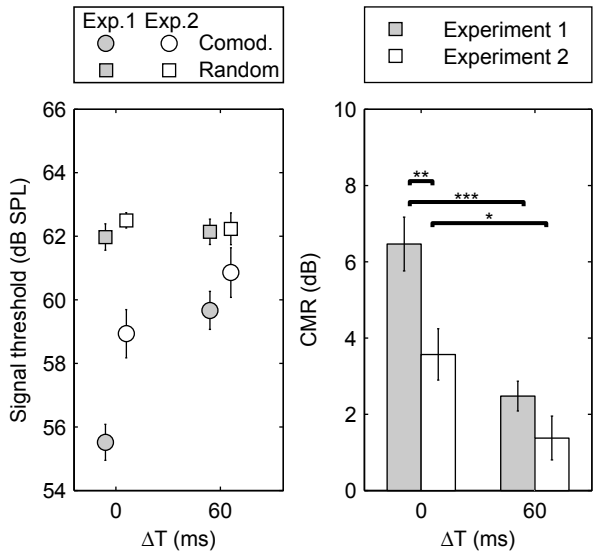


Figure 4.3: Results from experiment 2 (open symbols), with data from experiment 1 replotted for comparison (gray symbols) for conditions with ΔT of 0 and 60 ms. The left panel shows the detection thresholds for the comodulated (circles) and random (squares) ongoing envelopes of the final noise bursts. The bars in the right panel show the CMR, and the error bars indicate ±1 standard error of the mean between subjects. The asterisks in the right panel show the results of post hoc (Tukey-Kramer) comparisons of CMR, where *, **, and *** indicate $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively. In experiment 1, both the precursors and the noise bursts concurrent with the target signal had either random or comodulated ongoing envelopes. In experiment 2, the precursors always had random ongoing envelopes, and only the noise bursts concurrent with target signal were random or comodulated.

and ΔT [$F(1,15) = 34.95$, $p < 0.001$] along with a significant interaction [$F(1,15) = 4.70$, $p = 0.047$]. Post-hoc t-tests indicated that the threshold increased significantly in experiment 2 relative to experiment 1 for ΔT = 0 ms [t-test; $t(15) = 3.52$, $p < 0.01$], but the small increase for ΔT = 60 ms was not significant [$t(15) = 1.19$, $p = 0.14$].

A similar mixed-model ANOVA of the CMRs revealed a significant effect of ΔT [$F(1,15) = 26.21$, $p < 0.001$], a significant effect of experiment [$F(1,15) = 9.363$, $p < 0.01$], but no significant interaction effect [$F(1,15) = 2.215$, $p = 0.157$]. The results of post hoc comparisons are indicated in the right panel of Fig. 4.3. The pair-wise comparisons showed first that regardless of whether the ongoing envelopes of the precursors were comodulated or not (experiment 1 vs experiment 2), the onset asynchrony ΔT = 60 ms significantly reduced the amount of CMR. Second, the reduction in CMR between experiments 1 and 2 was significant for ΔT = 0 ms but not for ΔT = 60 ms. Under the assumption that the amount of CMR reflects the strength of perceptual fusion between the masking and flanking bands, this result indicates that random ongoing envelope fluctuations of the precursors reduce the fusion between MN and FNs, even though they have synchronized on- and offsets. The difference in CMR between ΔT = 0 ms and ΔT = 60 ms in experiment 2 shows that onset synchrony still provides a strong grouping cue when the precursors are not comodulated. Even though the comodulation of the ongoing envelopes of the precursors was removed in experiment 2, the signal thresholds were still significantly lower for the comodulated configuration than for the random configuration for both ΔT = 0 ms [paired t-test; $t(9) = 5.30$, $p < 0.001$] and ΔT = 60 ms [paired t-test;

t(9) = 2.39, p = 0.02], suggesting that the fusion of MN and FNs was not eliminated. The results of experiment 2 thereby support the hypothesis that the comodulation of the precursors in experiment 1 provided a grouping cue that limited the streaming effects of the incoherently gated precursors. However, the results also show that the comodulation of the ongoing envelope of the precursors cannot explain why the CMR effect persisted for precursor conditions that were predicted to lead to a segregated percept.

4.4 Experiment 3: Effect on streaming of embedding the target between pre- and post-cursors

4.4.1 Rationale

Experiment 1 showed that the presence of asynchronously gated precursors reduced the amount of CMR but did not fully eliminate it. Experiment 2 investigated the contribution of comodulation between the ongoing envelopes of overlapping portions of the precursor flanking and masking bands. Although precursors with random temporal envelopes produced less CMR than precursors with comodulated envelopes when gated synchronously, even the precursors with random ongoing envelopes did not completely eliminate CMR when they were gated asynchronously.

In all the experimental conditions tested so far, the target was always presented in the final masker burst. It may be that listeners were able to develop a strategy of "ignoring" the precursors and focusing primarily on the final noise burst. It is known that switching attention can lead to breakdown of auditory stream segregation and lead to more fused percepts in alternating tone sequences (e.g., Carlyon et al., 2001; Cusack et al., 2004). Therefore switching attention to the sounds directly prior to the final noise burst might be an advantageous strategy, as it would lessen the possibility that the masking and flanking bands form separate perceptual streams. In this final experiment, the target was presented in an unpredictable location within a longer series of noise bursts. Because of that listeners were no longer able to ignore the precursors and had to monitor the repeated bursts to detect the presence of the signal. The hypothesis was that forcing listeners to attend to the longer sequence would lead to greater buildup of stream segregation and may thus lead to the elimination of CMR in cases where the flanking and masking bands were gated on and off incoherently.

4.4.2 Method

Stimuli and procedure

Figure 4.4 shows a schematic representation of the stimuli presented in each interval of a trial. The general structure of the stimuli was the same as that used in experiment 1 except that instead of 5 repeating noise bursts there were now 10 repetitions. In addition, the target was no longer embedded in the final noise burst but was instead presented synchronously with the fifth, sixth, or

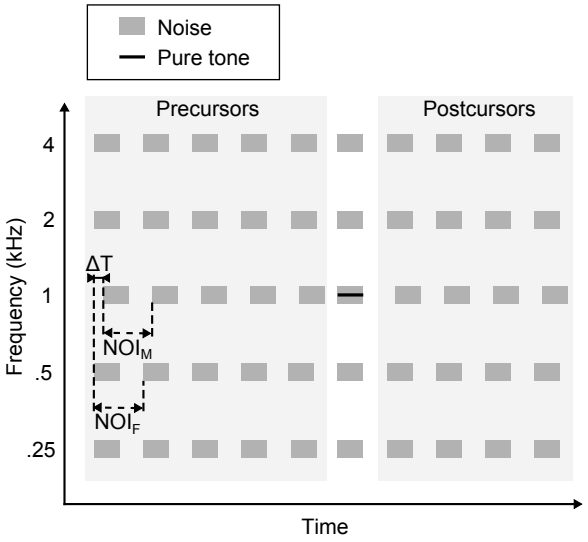


Figure 4.4: Schematic representation of the stimuli used in experiment 3. Five narrow-band noise bursts were presented at octave frequencies between 0.25 and 4 kHz. Each noise burst was repeated 10 times, and the 1-kHz target was presented simultaneously with the 5th, 6th, or 7th noise burst, chosen at random in each trial. The noise bursts presented together with the target were always presented synchronously, but the onset asynchrony (ΔT) between the masking noise band and the flanking noise bands' pre- and post-cursors was either 0 or 60 ms. In the sketched example, the target signal is presented in the 6th interval.

seventh noise burst, selected at random on each trial. The MN and FNs were always presented synchronously during the noise burst containing the target, and the non-target intervals in each trial had the MN and FNs synchronized on the same noise burst (fifth, sixth, or seventh) as in the target interval within a given trial. The MN and FN preceding and following the target (pre- and post-cursors, highlighted in light gray) were either synchronized (ΔT = 0 ms) or had an asynchrony of ΔT = 60 ms. In both conditions, the NOI_M and NOI_F were both set to 250 ms. The procedure and the set up were identical to that used in experiment 1.

Listeners

Eight normal-hearing listeners participated in this experiment. Five of the listeners had also participated in both experiments 1 and 2 (including the first author). The group consisted of four female and four male listeners, aged between 19 and 31 yr. The listeners were compensated monetarily for their participation at an hourly rate, and measurement sessions lasted between 1 and 2 h, including breaks. All listeners received at least 1 h of training in the same task before data collection began, and one to two sessions were required to complete the experiment.

4.4.3 Results

The individual results showed within-subject standard deviations that were typically around 0.5-2 dB and never exceeded 4 dB. In addition, all subjects showed a similar pattern of results across

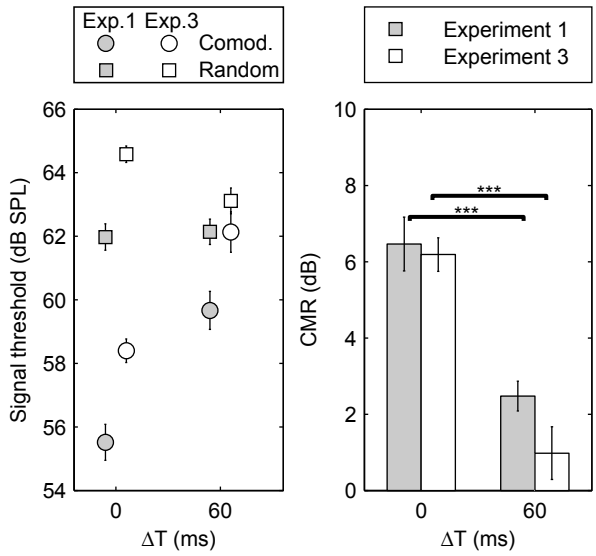


Figure 4.5: Results from experiment 3 (open symbols), with data from experiment 1 replotted for comparison (gray symbols) for conditions with ΔT of 0 and 60 ms. The left panel shows the detection thresholds for the comodulated (circles) and random (squares) ongoing envelopes of the noise bursts. The bars in the right panel show the CMR, and the error bars indicate ±1 standard error of the mean between subjects. The asterisks in the right panel show the results of post hoc (Tukey-Kramer) comparisons of CMR, where *** indicates $p < 0.001$. In experiment 1, the stimuli only contain pre-cursor noise bursts and target interval as depicted in Fig. 4.1. In experiment 3 the stimuli contain both pre- and post-cursors before and after the target interval, as depicted in Fig. 4.4.

the different conditions, and so only the mean data are reported here. The mean data are shown in Fig. 4.5 together with the results replotted from experiment 1 for the corresponding precursor configurations. The left panel shows the detection thresholds from the random (squares) and comodulated (circles) configurations, and the right panel shows the CMR (difference between the random and comodulated thresholds). In both panels, the gray and open symbols indicate results from experiments 1 and 3, respectively.

A mixed-model ANOVA on thresholds in the random configuration showed a significant effect of experiment [$F(1,13) = 14.57$, $p < 0.01$], indicating that the addition of post-cursors and/or the randomized location of the target affected performance in experiment 3, relative to that in experiment 1. The analysis also showed a significant effect of ΔT [$F(1,13) = 7.59$, $p < 0.02$] and interaction (experiment by ΔT) [$F(1,13) = 11.97$, $p < 0.01$]. Post-hoc analyses (paired t-tests) showed that detection thresholds with the random maskers were significantly poorer in the synchronized condition (ΔT = 0 ms) than in the asynchronous condition (ΔT = 60 ms) in experiment 3 [$t(7) = 3.28$, $p < 0.01$]. A mixed-model ANOVA on thresholds in the comodulated configuration revealed a significant effect of experiment [$F(1,13) = 23.47$, $p < 0.001$] and ΔT [$F(1,13) = 51.50$, $p < 0.001$] but no interaction [$F(1,13) = 0.14$, $p = 0.71$], indicating that the addition of post-cursors and/or the randomizing of the target location resulted in a similar increase in signal threshold for both the synchronous and asynchronous conditions.

The elevated thresholds observed in experiment 3 relative to experiment 1 may be due to an

increased signal uncertainty in time. Green and Weber (1980) and Bonino and Leibold (2008) investigated the effect of temporal uncertainty in a detection task involving a 1 kHz pure tone target in a noise masker. Both studies found increased detection thresholds of 2-3 dB when going from a temporally certain to an uncertain position of the target; this is consistent with the observed increase in detection threshold for $\Delta T = 0$ ms for both the comodulated and random configurations, and for the $\Delta T = 60$ ms in the comodulated configuration. However, for the $\Delta T = 60$ ms in the random configuration, the increase is only about 1 dB. One possible explanation for the lower threshold for $\Delta T = 60$ ms relative to $\Delta T = 0$ ms in the random configuration may be that the synchrony of the noise bursts at the target location helped listeners identify where the target is likely to appear, effectively reducing the temporal uncertainty. If the temporal uncertainty is reduced for the $\Delta T = 60$ ms, a similar benefit would be expected for the comodulated configuration. The 3 dB increase for the $\Delta T = 60$ ms comodulated condition may therefore be a combination of a reduced ability to use across-channel information due to a perceptual segregation as well as an increased temporal uncertainty.

A further mixed-model ANOVA performed on the CMR values revealed a significant effect of ΔT [$F(1,13) = 70.6$, $p < 0.001$] but no significant effect of experiment [$F(1,13) = 2.14$, $p = 0.17$] or interaction (experiment by ΔT) [$F(1,13) = 1.24$, $p = 0.29$], indicating that the addition of post-cursors did not significantly affect the amount of CMR.

The original hypothesis was that stream segregation might be increased (and CMR reduced) by embedding the target within a longer stream of noise bursts when the MN and FN bursts were asynchronous. The lack of an effect of (or interaction with) experiment on CMR is not consistent with this hypothesis. On the other hand, when considering just the results from experiment 3, the amount of CMR was significantly greater than zero for the synchronous condition ($\Delta T = 0$ ms) [paired t-test; $t(7) = 14.1$, $p < 0.001$], whereas no significant CMR was found in the asynchronous condition ($\Delta T = 60$ ms) [paired t-test; $t(7) = 1.42$, $p = 0.10$]. Thus in contrast to experiments 1 and 2, and consistent with the original hypothesis, no significant CMR was observed in the condition where the pre- and post-cursors were not temporally coherent, suggesting that the MN and FN bursts were sufficiently segregated to eliminate measurable CMR.

4.5 General discussion

4.5.1 Summary of results

Detection of a 1-kHz tone was measured in narrow bands of noise that were spaced at octave frequencies from 250 Hz to 4 kHz. The bands of noise were either comodulated (shared the same ongoing envelope) or independently generated. The noise bursts that were gated synchronously with the target tone were preceded or temporally surrounded by a series of noise bursts, intended to influence the perceptual organization of the sequence. The first experiment showed that temporally coherent on-frequency masker and off-frequency flanker noise bursts produced CMR and that

reducing the coherence through a fixed asynchrony or through different presentation rates in the preceding noise bursts led to a reduction of CMR. The second experiment explored the effect of comodulation within the precursor bursts and found that random (independent) ongoing envelopes in the precursors reduced CMR when the precursors were synchronously gated, suggesting that incoherence in the ongoing portions of the precursor temporal envelopes can reduce fusion, even when the bursts are synchronously gated. The third experiment added "postcursors" that followed the target burst, in addition to the precursors, and randomized the position of the target, so that subjects were obliged to attend to more of the sequence. In general, thresholds were somewhat higher in experiment 3, and the amount of CMR was not significantly greater than zero when the precursors were asynchronously gated, suggesting that the MN and FN bursts were perceptually segregated from each other.

The results of all three experiments support the hypothesis that temporal coherence between noise bursts widely separated in frequency leads to the formation of a single perceptual stream, as evidenced by the finding of significant CMR in conditions where the masking and flanking bands were presented synchronously across bursts. Also consistent with the hypothesis was the finding that CMR was reduced or absent in conditions where the on-frequency and flanking pre- and post-cursors were not temporally coherent, either through a fixed asynchrony or through different presentation rates. The outcomes do not depend on the temporal regularity of the precursors, as temporally jittered (but synchronously gated) precursors produced as much CMR as the regular sequence of precursors.

Overall, the results provide support for the hypothesis of Elhilali et al. (2009) that temporal coherence plays an important role in the auditory streaming of widely separated frequency components, using a paradigm (CMR) that does not suffer from the potential confound of the asynchrony-detection task used by Elhilali et al. (2009). In addition, the results provide further support for the idea that across-channel CMR provides a viable indirect measure for investigating the perceptual organization of sounds.

4.5.2 Relation to previous studies and interpretations of perceptual segregation

Grose et al. (2005) measured CMR with maskers and flankers that were comodulated for the duration of the 400-ms target but otherwise had random temporal envelopes. They found that the introduction of these random temporal "fringes" significantly reduced CMR even though the maskers and flankers that were simultaneously present with the target tone were unchanged. They argued that the ongoing random noise may put *"the system in a state where comodulation is not expected, and therefore potential cueing mechanisms (perhaps based on grouping by common modulation) are not activated."* Our observation in experiment 2 that random ongoing envelopes in the flankers, even when gated synchronously with the precursor masker bands, led to reduced CMR is consistent with the findings and conclusions of Grose et al. (2005). In fact, it is possible to consider the inherent fluctuations of the 20-Hz-wide noise bands and their gating on and off as coherent or incoherent modulation in two modulation-frequency regions: The gating involved a

period of 250 ms (or 4 Hz), whereas 20-Hz-wide noise bands have modulation energy out to 20 Hz. Thus both the ongoing (inherent) modulation and the gating can be considered cases of temporal coherence. Within this framework, the temporal coherence at the level of the gating modulation, as well as the temporal coherence at the level of the inherent noise fluctuations, influences the perceptual organization of the sound. Thus both our results and those of Grose et al. (2005) can be understood as special cases of the general principle that streams fuse if they are temporally coherent but tend to segregate if they are incoherent at one or more levels of modulation analysis.

A somewhat surprising finding was that it was not possible to completely eliminate the CMR in most of the experimental conditions, given that Dau et al. (2005, 2009) and Verhey et al. (2012) found no across-channel CMR in conditions with pre- or post-cursors. One major difference between the current study and the studies by Dau et al. and Verhey et al. was that the earlier studies only presented FN pre- or post-cursors and no MN pre- or post-cursors. Therefore their stimuli did not have any ambiguous streaming cues, such as overlapping (and sometimes comodulated) portions of the pre- and post-cursor noise bursts. Removing comodulation within the precursors in experiment 2 decreased the CMR but did not eliminate it completely. Even in experiment 3, where the target was embedded between pre- and post-cursors and its position randomized, although the remaining CMR was not significantly different from zero, it was also not significantly different from that found in experiment 1, leaving the result somewhat ambiguous.

It may be that even precursors with random ongoing envelopes, and onset asynchronies of up to 60 ms, are not enough to completely eliminate perceptual fusion and that the remaining temporal overlap of the MN and FN precursors (for 68% of their duration) acted as a grouping cue. Another possibility is that the stimuli were too short to allow a sufficient build-up of stream segregation. Although several studies (e.g., Anstis and Saida, 1985; Bee et al., 2010) suggest that the "build-up" of auditory streams takes place on a timescale of several seconds, the study of Dau et al. (2005) showed a complete elimination of CMR using a relatively short build-up period of only four precursors (for a total duration of 1 s). According to Moore and Gockel (2002), the extent to which sequential stream segregation occurs is directly related to the degree of perceptual difference between successive sounds: In the original study by Dau et al. (2005), the perceptual difference between MN and FNs was likely larger due to the absence of MN precursors. This may have led to faster stream segregation in Dau et al. (2005) than in the present study, and it remains possible that a complete elimination of CMR would have been observed in the present study if more precursors had been used.

Acknowledgments

This work was supported by the Oticon Foundation and by NIH Grant No. R01 DC007657.

5

General discussion

5.1 Summary of main results

In this thesis, auditory stream segregation was investigated through behavioral listening experiments in combination with computational models of auditory processing. The listening experiments were conducted to characterize the influence of acoustic cues on auditory stream formation, and the modelling allowed the evaluation of hypotheses about the mechanisms underlying auditory stream segregation. Specifically, the influence of temporal coherence on auditory stream formation was investigated, as well as the influence of the processing through the peripheral auditory system on the ability to perceptually segregate sound sources.

In **Chapter 2**, a listening experiment demonstrated the effect of onset and offset synchrony on across-frequency grouping of spectral sound components. The experimental data showed that onset asynchronies of less than approximately 20 ms led to the perceptual fusion of spectral components with a large frequency separation, whereas larger asynchronies tended to lead to stream segregation. These results supported the hypothesis that temporal coherence provides an important cue for stream fusion. Chapter 2 also presented a computational model of auditory stream formation that combines a physiologically inspired model of auditory preprocessing and perception (Dau et al., 1997a) with a temporal coherence analysis (Elhilali et al., 2009). The auditory preprocessing transformed a digital sound signal into an auditory representation of the sound, and the coherence analysis estimated the perceptual organization into one or two streams based on the coherence of the preprocessed output. The model was able to reproduce the main trends from perceptual experiments on stream segregation based on frequency separation and tone repetition rate (van Noorden, 1975), and to account for the streaming effects due to onset and offset synchrony measured in the listening experiments of Chapter 2.

The influence of various processing stages of the auditory periphery on auditory stream formation were analyzed in this model framework, and the model results showed that forward masking played a critical role in the stream segregation of fast tone sequences: At the offset of a tone, the reduced sensitivity of the peripheral channel tuned to the tone limited the spread of excitation from subsequent tones, effectively reducing the coherence across channels and making the model more likely to predict two streams than one. This observation supported findings earlier reported in physiological studies in animals (Bee and Klump, 2004; 2005; Fishman et al., 2004), suggesting that physiological forward masking is the underlying mechanism behind the increased tendency to perceptually segregate fast tone sequences into separate streams relative to slow tone sequences.

The model analysis also suggested that the dependency of streaming on onset synchrony may be caused by effects of neural adaptation (emphasizing stimulus onsets while attenuating steady-state portions of the stimulus), and the subsequent processing through modulation specific filters. This may explain why onset-synchrony appears to be more critical for the fusion of spectral components into a single stream than offset-synchrony (e.g. Darwin and Ciocca, 1992). Lastly, the model showed an asymmetry in the synchrony experiment, causing the model to predict the maximum across-frequency synchrony when the low-frequency tone was leading the high-frequency tone by 4.5 ms, caused by the frequency specific (group) delay of the gammatone filters in the model preprocessing. Similar effects are observed in the frequency-to-place transformation in the cochlea with a higher latency for low frequencies than for high frequencies (e.g. Harte et al., 2009), but the effect of such cochlear latencies were not observed in the experimental data of Chapter 2. Recent studies have suggested that auditory processes at a level higher than the cochlea may roughly compensate for the cochlear delays (Uppenkamp et al., 2001; Wojtczak et al., 2012; 2013), and the results of Chapter 2 indicate that a grouping mechanism based on temporal coherence must take place at a stage subsequent to this compensation process.

In **Chapter 3**, the assumptions underlying the computational model were tested by investigating the effect of sound intensity on auditory stream formation. Under the assumption that the stimulation of non-overlapping neural populations is a prerequisite for stream segregation, the wider auditory filters at high intensities should lead to a reduced stream segregation. This hypothesis was investigated through listening experiments, as well as through a modified version of the computational model presented in Chapter 2. The model was modified to account for the level-dependent processing of the peripheral auditory system, by replacing the linear gammatone filterbank in the preprocessing by Dau et al. (1997a) with the non-linear dual resonance filterbank (DRNL) in the preprocessing by Jepsen et al. (2008). Consistent with the hypothesis, the experimental data showed that larger frequency separations were required for stream segregation at high sound intensities than at low.

The study also estimated the maximum frequency separation where it was possible to perceive the tone sequences as a single coherent stream, which is typically small for fast tone sequences and large for slow tone sequences (van Noorden, 1975). Similar results were found in the present study at the lowest intensities (40 and 60 dB SPL), where the experimental data showed a smaller frequency separation for fast tones than for slow tones. With increasing intensity, the maximum frequency separation where it was possible to hear the tones as a single stream increased for the fast tone sequences, but not for the slow tone sequences, and at the highest intensity (80 dB SPL) the experimental data showed similar thresholds for both fast and slow tone rates. Assuming that the increased stream segregation of fast tone sequences is a consequence of forward masking, the interaction between tone rate and intensity may be a consequence of a reduced forward masking at high intensities. This hypothesis was supported by analysis of the stimuli through the preprocessing of the computational model.

The data from Chapter 3 showed a substantial inter-individual variation. An analysis of the individual data revealed that part of the variation was due to an experiment-order bias. Half of the

listeners began with a stream segregation task, whereas the other half of the listeners started with a stream fusion task. The data suggested that the listeners who began with a stream segregation task were biased towards stream segregation in both tasks, and the listeners who began with the stream fusion task were biased towards stream fusion in both tasks. This finding demonstrated the uncertainty inherent in subjective experiments, and showed that auditory streaming experiments based on the paradigm by van Noorden (1975) need to be carefully designed.

The computational model used to analyze the experimental conditions also showed an increased tendency to group stimuli presented at high sound intensities relative to the low intensities. The model predictions for the minimum frequency separation for stream segregation with increasing sound intensity were comparable with the experimental data, but the model clearly overestimated the maximum frequency separation where it was possible to perceive the tones as a single stream. This suggests that some part of the model framework does not accurately reflect the auditory stream segregation processes, which could be because the model front-end overestimates the increase in spread of excitation at high intensities, or that the overall coherence is not indicative of stream fusion.

Chapter 4 considered subjective experiments for measuring perceptual organization by investigating an indirect, performance based measure of auditory stream segregation. The proposed method used comodulation masking release (CMR) to measure perceptual organization, inspired by the observations that across-channel CMR appears to be affected by streaming cues, and, in particular, that across-channel CMR only occurs when the flanking noises are perceptually grouped into the same perceptual stream as the central masking band (Dau et al., 2005, 2009). The experimental setup used a series of "pre-cursor" noise-bursts presented before the traditional CMR stimulus to either group or segregate the flanking and masking noises in the CMR stimulus into either one or two streams. Through this setup, the influence of temporal coherence on auditory stream formation was investigated. The temporal coherence was manipulated by either introducing a constant asynchrony between onsets and offsets of the masker and flanker pre-cursors, or by presenting the masker and pre-cursors at different rates, leading to constantly varying asynchronies. The results of the experiment showed that the temporally coherent conditions resulted in a CMR of approximately 6 dB which is consistent with the CMR measured without pre-cursors in other studies (e.g. Dau et al., 2005, 2009). The experiment also showed that the CMR decreased with increasing temporal incoherence of the precursors, supporting the hypothesis that temporal coherence facilitates the perceptual grouping of spectral components across frequency. While temporally incoherent conditions led to a reduction of CMR, the CMR was not fully eliminated in any of the conditions tested.

A follow-up experiment investigated whether the residual CMR for temporally incoherent conditions were caused by competing streaming cues. In the first experiment, the "ongoing envelope" of the narrowband noises was comodulated when masker and flanker bands were overlapping in time, regardless of whether the on- and off-gating of the precursors was synchronous or asynchronous. This common amplitude modulation may have acted as a grouping cue (e.g. Hall and Grose, 1990), counteracting the asynchronous on- and off-gating of the precursors. The follow-

up experiment therefore tested whether the CMR was affected by having random ongoing envelopes in the precursors, and only having comodulated ongoing envelopes during the presentation of the target signal, using either synchronized pre-cursors, or the maximum onset/offset asynchrony of the first experiment (60 ms). The results of this experiment showed reduced CMR for both the synchronous and asynchronous precursors relative to the first experiment, indicating that perceptual organization of the masker and flanking bands was not entirely controlled by the on- and off-gating of the precursors, but also by the comodulation of the ongoing envelopes in the precursors. The results still showed a significantly higher CMR for conditions with synchronous gating of the precursors than the asynchronous condition, indicating that the onset/offset synchrony facilitated grouping across frequency regardless of the ongoing envelope. For the asynchronous gating there was, however, still a significant residual CMR.

A second follow-up experiment investigated whether the residual CMR might be caused by the listeners employing a strategy of ignoring the precursors and switching their attention to the target signal during the signal presentation, as switching attention can lead to a break-down of auditory stream segregation and lead to more fused percepts in streaming experiments (e.g. Carlyon et al., 2001; Cusack et al., 2004). Therefore, a third experiment was designed where the standard CMR stimulus was embedded between a semi-random number of pre- and post-cursors, forcing the listeners to attend to the stimulus through its entirety. The experiment employed the same two precursor conditions used in the first follow-up experiment; either fully synchronized precursors, or precursors with an on/offset asynchrony of 60 ms. The results of the experiment showed decreased detection thresholds for all conditions, which may be explained by an increased signal uncertainty in the new experimental paradigm. Furthermore, the results showed that for the temporally coherent condition, the CMR was 6 dB, and for the temporally incoherent condition there was no significant CMR.

Overall, the study suggested that CMR may be used as an indirect measure of perceptual organization into auditory streams. The study also provided support for the hypothesis that temporal coherence facilitates perceptual grouping across frequency. The use of performance-based measures may provide a tool to eliminate listener-bias from psychoacoustic measures of stream segregation, and secondly, may provide an estimate of the "strength" of stream segregation, which is not offered through the binary response-options of most of the subjective streaming experiments, where the listeners must classify a stimulus as either one or two streams. However, using an indirect measure of perceptual organization, such as CMR, inherently introduces confounding factors which makes the analysis of the results less transparent than the direct, subjective measures where a listener reports whether he or she perceives a stimulus as one or two streams.

5.2 Limitations of the modeling framework

The modeling framework presented in this thesis is limited to segregating sounds based on tonotopic separation. Therefore, it cannot account for stream segregation in the absence of spectral cues,

produced by, e.g., pitch, temporal modulations or spatial location. While the model does include a temporal modulation filterbank, which has previously been used to account for perceptual data on modulation detection and masking (e.g., Dau et al., 1997a, b; Ewert and Dau, 2000), it is only used as a means for short-term integration with different time constants in the present study, and the modulation frequency selectivity was discarded by the summation across modulation frequencies for the generation of the coherence matrix C. To allow the model to account for stream segregation based on other cues than tonotopic separation, model stages extracting these cues would have to be added to the pre-processing of the model.

A second limitation of the presented model is that it cannot follow transitions across peripheral filters, as the coherence matrix is integrated over the entire duration of the analyzed stimulus. Therefore, the model is limited to analyzing spectrally stable stimuli and cannot predict a fused percept of, e.g., speech signals which typically involve various frequency glides and broad-band events, such as fricatives. In order for the model to follow such dynamic stimuli, the coherence analysis would have to be applied in shorter time windows. Such a short-term model must ensure that the spectral components of successive time instants are grouped into the same streams. This coherence over time may be realized through the temporal integration stage in the model, where the auditory representation of a stimulus is analyzed over a range of time scales. It is possible that this multi time scale analysis would allow the model to be sensitive to rapid events (e.g. onset asynchrony) while maintaining continuity of successive sounds.

Thirdly, the eigenvalue ratio used as the decision metric in the model was shown to only produce a relative measure of temporal coherence and, thus, perceptual organization. Therefore, the model framework cannot simply be generalized to arbitrary stimulus configurations presented at an arbitrary intensity. While this limits the predictive powers of the model, it is still able to qualitatively evaluate whether one sound stimulus is more or less likely to be produce a two-stream percept than another sound stimulus.

5.3 The role of temporal coherence in auditory stream segregation

In this thesis, the role of temporal coherence in auditory stream formation was investigated. In all listening experiments, temporal coherence, here interpreted as "*sounds that are repeatedly presented synchronously*", led to an increased perceptual fusion relative to stimuli that were presented incoherently. This is consistent with the temporal coherence theory and supports the hypothesis that temporal coherence acts as a strong grouping cue. The experimental data suggested that the term "synchronously" should not be interpreted in the strict physical sense such that deviations from perfect synchrony in the order of tens of milliseconds can still lead to perceptual grouping. This flexibility may be a consequence of the fact that perfect synchrony rarely occurs in nature and that, for a single natural sound source, such as a musical instrument, the onset asynchrony between individual harmonic components can be tens of millisecond (e.g. Risset and Mathews, 1969; Beauchamp, 1975). A second factor may be that natural acoustic environments are rarely

anechoic, and a significant part of the acoustic energy transmitted from the source to the receiver is carried by reflections. The arrival time of reflections will inevitably be delayed relative to the direct sound wave, and studies on the influence of early and late reflections on speech intelligibility indicate that early reflections (within approximately 50 ms) tend to improve intelligibility, while late reflections (later than 50 ms) tend to impair intelligibility (Arweiler and Buchholz, 2011). Early reflections are typically perceptually fused with the direct sound, and contribute to the acoustic energy defining the "target signal", while late reflections are perceived as a different source and therefore act as an interferer. The perceptual fusion of sounds despite the deviations from perfect synchrony may be an evolutionary consequence of living in reverberant environments.

Temporal coherence was also investigated as the potential organizing principle behind auditory stream formation in the framework of the computational model. The simulations demonstrated that the application of a temporal coherence analysis, applied to the internal representation of a stimulus after auditory pre-processing, could quantitatively account for several perceptual streaming phenomena. The success of the model framework was a consequence of the pre-processing in the model, which limited the spectral and temporal information of the sound signals in a manner that is consistent with perceptual detection experiments. This suggests that the temporal coherence of the "correct" internal representation of a stimulus after auditory pre-processing determines how the stimulus is perceived by the receiver.

Overall, the findings of this thesis suggest that temporal coherence plays a significant role for the perceptual grouping of sounds into a single stream and, more generally, that a temporal coherence analysis might help determine the perceptual organization of sounds into streams.

References

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

Arweiler, I., and Buchholz, J. M. (2011). "The influence of spectral characteristics of early reflections on speech intelligibility," J. Acoust. Soc. Am. 130, 996-1005.

Beauchamp, J. W. (1975). "Analysis and synthesis of cornet tones using nonlinear interharmonic relationships," J. Audio Eng. Soc. 23, 778-795.

Beauvois, M. W., and Meddis, R. (1996). "Computer simulation of auditory stream segregation in alternating tone sequences," J. Acoust. Soc. Am. 99, 2270-2280.

Bee, M. A., and Klump, G. M. (2004). "Primitive auditory stream segregation: A neurophysiological study in the songbird forebrain," J. Neurophysiol. 99, 2270-2280.

Bee, M. A., and Klump, G. M. (2005). "Auditory stream segregation in the songbird forebrain: Effects of time intervals on responses to interleaved tone sequences," Brain Behav. Evol. 66, 197-214.

Bee, M. A., Micheyl, C., Oxenham, A. J., and Klump, G. M. (2010). "Neural adaptation to tone sequences in the songbird forebrain: Patterns, determinants, and relation to the build-up of auditory streaming," J. Comp. Physiol. A 196, 543-557.

Bendor, D., and Wang, X. (2005). "The neuronal representation of pitch in primate auditory cortex," Nature 436, 1161-1165.

Bonino, A., and Leibold, L. J. (2008). "The effect of signal-temporal uncertainty on detection in bursts of noise or a random-frequency complex," J. Acoust. Soc. Am. 124, EL321-EL327.

Bregman, A. S. (1990). Auditory Scene Analysis: The Perceptual Organization of Sound (MIT Press, Cambridge, MA), pp. 1-736.

Bregman, A. S., Abramson, J., Doehring, P., and Darwin, C. J. (1985). "Spectral integration based on common amplitude modulation," Percept. Psychophys. 37, 483-493.

Bregman, A. S., and Campbell, J. (1971). "Primary auditory stream segregation and perception of order in rapid sequences of tones," J. Exp. Psychol. 89, 244-249.

Bregman, A. S., Levitan, R., and Liao, C. (1990). "Fusion of auditory components: Effects of the frequency of amplitude modulation," Perception and Psychophysics, 47, 68-73.

Bregman, A. S., and Pinker, S. (1978). "Auditory streaming and the building of timbre," Can. J. Psychol. 32, 19-31.

Broadbent, D. E., and Ladefoged, P. (1957). "On the fusion of sounds reaching different sense organs," J. Acoust. Soc. Am. 29, 708-710.

Broadbent, D. E., and Ladefoged, P. (1959). "Auditory perception of temporal order," J. Acoust. Soc. Am. 31, 1539-1540.

Brown, G. J. (2010). "Physiological models of auditory scene analysis," in Computational Models of the Auditory System, edited by R. Meddis, E. A. Lopez-Poveda, A. N. Popper and R. R. Fay (Springer, New York)

Buss, E., Grose, J. H., and Hall, J. W. (2009). "Features of across-frequency envelope coherence critical for comodulation masking release," J. Acoust. Soc. Am. 126, 2455-2466.

Buus, S. (1985). "Release of masking caused by envelope fluctuation," J. Acoust. Soc. Am. 78, 1958-1965.

Carlyon, R. P., Cusack, R., and Foxton, J. M. (2001). "Effects of attention and unilateral neglect on auditory stream segregation," J. Exp. Psychol. Hum. Percept. Perform. 27, 115-127.

Carlyon, R. P., and Gockel. H. E. (2008). "Effects of harmonicity and regularity on the perception of sound sources" in Auditory Perception of Sound Sources, edited byW. A. Yost (Springer, New York), Chap. 7, pp. 191-213.

Chi, T., Ru, P., and Shamma, S. A. (2005). "Multiresolution spectrotemporal analysis of complex sounds," J. Acoust. Soc. Am. 118, 887-906.

Christiansen, S. K., and Dau, T. (2015). "Effects of sound intensity on auditory stream segregation of pure tone sequences," J. Acoust. Soc. Am. (submitted).

Christiansen, S. K., Jepsen, M. L., and Dau, T. (2014). "Effects of tonotopicity, adaptation, modulation tuning, and temporal coherence in ‘primitive‘ auditory stream segregation," J. Acoust. Soc. Am. 135, 323-333.

Christiansen, S. K., and Oxenham, A. J. (2014). "Assessing the effects of temporal coherence on auditory stream formation through comodulation masking release," J. Acoust. Soc. Am. 135, 3520-3529.

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

Darwin, C. J., and Ciocca, V. (1992). "Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component," J. Acoust. Soc. Am. 91, 3381-3390.

Darwin, C. J., and Sutherland, N. S. (1984). "Grouping frequency components of vowels: When is a harmonic not a harmonic?" Q. J. Exp. Psychol. 36A, 193-208.

Dau, T., Ewert, S., and Oxenham, A. J. (2005). "Effects of concurrent and sequential streaming in comodulation masking release," in Auditory Signal Processing: Physiology, Psychoacoustics, and Models, edited by D. Pressnitzer, A. de Cheveigne, S. McAdams, and L. Collet (Springer-Verlag, Berlin).

Dau, T., Ewert, S., and Oxenham, A. J. (2009). "Auditory stream formation affects comodulation masking release retroactively," J. Acoust. Soc. Am. 125, 2182-2188.

Dau, T., Kollmeier, B., and Kohlrausch, A. (1997a). "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," J. Acoust. Soc. Am. 102, 2892-2905.

Dau, T., Kollmeier, B., and Kohlrausch, A. (1997b). "Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration," J. Acoust. Soc. Am. 102, 2906-2919.

Dau, T., Piechowiak, T., and Ewert, S. D. (2013). "Modeling within- and across-channel processes in comodulation masking release," J. Acoust. Soc. Am. 133, 350-364.

Dau, T., Püschel, D., and Kohlrausch, A. (1996). "A quantitative model of the effective signal processing in the auditory system. I. Model structure," J. Acoust. Soc. Am. 99, 3615-3622.

Derleth, R.-P., and Dau, T. (2000). "On the role of envelope fluctuation processing in spectral masking," J. Acoust. Soc. Am. 108, 285-296.

Eddins, D. A., and Wright, B. A. (1994). "Comodulation masking release for single and multiple rates of envelope fluctuation," J. Acoust. Soc. Am. 96, 3432-3442.

Elhilali, M., Chi, T., and Shamma, S. A. (2003). "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," Speech Commun. 41, 331-348.

Elhilali, M., Ling, C., Micheyl, C., Oxenham, A. J., and Shamma, S. A. (2009). "Temporal coherence in the perceptual organization and cortical representation of auditory scenes," Neuron 61, 317-329.

Elhilali, M., and Shamma, S. A. (2008). "A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation," J. Acoust. Soc. Am. 124, 3751-3771.

Ernst, S. M. A., Rennies, J., Kollmeier, B., and Verhey, J. L. (2010). "Suppression and comodulation masking release in normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. 128, 300-309.

Ernst, S. M. A., and Verhey, J. L. (2008). "Peripheral and central aspects of auditory across-frequency processing," Brain Res. 1220, 246-255.

Ewert, S. D., and Dau, T. (2000). "Characterizing frequency selectivity for envelope fluctuations," J. Acoust. Soc. Am., 108, 1181-1196.

Ewert, S. D. (2013). "AFC - A modular framework for running psychoacoustic experiments and computational perception models," in Proceedings of the International Conference on Acoustics AIADAGA2013, Merano, Italy, pp. 1326-1329.

Fantini, D. A., Moore, B. C. J., and Schooneveldt, G. P. (1993). "Comodulation masking release as a function of type of signal, gated or continuous masking, monaural or dichotic presentation of flanking bands and center frequency," J. Acoust. Soc. Am. 93, 2106-2115.

Fishman, Y., Arezzo, J. C., and Steinschneider, M. (2004). "Auditory stream segregation in monkey auditory cortex: Effects of frequency separation, presentation rate, and tone duration," J. Acoust. Soc. Am. 116, 1656-1670.

Glasberg, B. R., and Moore, B. C. J. (1990). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. 47, 103-138.

Green, D. M., and Weber, D. L. (1980). "Detection of temporally uncertain signals," J. Acoust. Soc. Am. 67, 1304-1311.

Grimault, N., Bacon, S. P., and Micheyl, C. (2002). "Auditory stream segregation on the basis of amplitude-modulation rate," J. Acoust. Soc. Am. 111, 1340-1348.

Grimault, N., Micheyl, C., Carlyon, R. P., Arthaud, P., and Collet, L. (2000). "Influence of peripheral resolvability on the perceptual segregation of harmonic complex tones differing in fundamental frequency," J. Acoust. Soc. Am. 108, 263-271.

Grose, J. H., Buss, E., and Hall, J. W. (2009). "Within- and across-channel factors in the multiband comodulation masking release paradigm," J. Acoust. Soc. Am. 125, 282-293.

Grose, J. H., and Hall, J. W. (1993). "Comodulation masking release: Is comodulation sufficient?," J. Acoust. Soc. Am. 93, 2896-2902.

Grose, J. H., Hall, J. W., Buss, E., and Hatch, D. R. (2005). "Detection of spectrally complex signals in comodulated maskers: Effect of temporal fringe," J. Acoust. Soc. Am. 118, 3774-3782.

Gutschalk, A., Oxenham, A. J., Micheyl, C., Wilson, E. C., and Melcher, J. R. (2007). "Human cortical activity during streaming without spectral cues suggests a general neural substrate for auditory stream segregation," J. Neurosci. 27, 13074-13081.

Hall, J. W., and Grose, J. H. (1990). "Comodulation masking release and auditory grouping," J. Acoust. Soc. Am. 88, 119-125.

Hall, J. W., Grose, J. H., and Haggard, M. P. (1990). "Effects of FB proximity, number, and modulation pattern on comodulation masking release," J. Acoust. Soc. Am. 87, 269-283.

Hall, J.W., Haggard, M. P., and Fernandes, M. A. (1984). "Detection in noise by spectro-temporal pattern analysis," J. Acoust. Soc. Am. 76, 50-56.

Harte et al., Pigasse, G., and Dau, T. (2009). "Comparison of cochlear delay estimates using otoacoustic emissions and auditory brainstem responses," J. Acoust. Soc. Am. 126, 1291-1301.

Hartmann, W. M., and Johnson, D. (1991). "Stream segregation and peripheral channeling," Music Percept. 9, 155-184.

Heinz, M. G., Colburn, H. S., and Carney, L. H. (2001). "Evaluating auditory performance limits. I. One-parameter discrimination using a computational model for the auditory nerve," Neural Comput. 13, 2273-2316.

ISO 389-8 (2004). Acoustics - Reference zero for calibration of audiometric equipment - Part 8: Reference equivalent threshold sound pressure levels for pure tones and circumaural headphones (International Organization for Standardization, Geneva, Schwitzerland).

Itani, N., and Klump, G. M. (2011). "Neural correlates of auditory streaming of harmonic complex sounds with different phase relations in the songbird forebrain," J. Neurophysiol. 105, 188-199.

Jacobsen, F., and Juhl, P. M. (2013). Fundamentals of General Linear Acoustics (Wiley).

Jepsen, M. L., Ewert, S. D., and Dau, T. (2008). "A computational model of human auditory signal processing and perception," J. Acoust. Soc. Am. 124, 422-438.

Jepsen, M. L., and Dau, T. (2011). "Characterizing auditory processing and perception in individual listeners with sensorineural hearing loss," J. Acoust. Soc. Am. 129, 262-281.

Jesteadt, W., Bacon, S. P., and Lehmann, J. R. (1982). "Forward masking as a function of frequency, masker level, and signal delay," J. Acoust. Soc. Am. 71, 950-962.

Levitt, H. (1971). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. 49, 467-477.

Lopez-Poveda, E. A., and Meddis, R. (2001). "A human nonlinear cochlear filterbank," J. Acoust. Soc. Am. 110, 3107-3118.

Ma, L., Micheyl, C., Yin, P., Oxenham, A. J., and Shamma, S. (2010). "Behavioral measures of auditory streaming in ferrets (Mustela putorius)," J. Comp. Psychol. 124, 317-330.

McCabe, S., and Denham, M. J. (1997). "A model of auditory streaming," J. Acoust. Soc. Am. 101, 1611-1621.

Meddis, R., O'Mard, L. P., and Lopez-Poveda, E. A. (2001). "A computational algorithm for computing nonlinear auditory frequency selectivity," J. Acoust. Soc. Am. 109, 2852-2861.

Micheyl, C., Hanson, C., Demany, L., Shamma, S., and Oxenham, A. J. (2013a). "Auditory stream segregation for alternating and synchronous tones," J. Exp. Psychol. Hum. Percept. Perform. 39, 1568-1580.

Micheyl, C., Hunter., C., and Oxenham, A. J. (2010). "Auditory stream segregation and the perception of across-frequency synchrony," J. Exp. Psychol. Hum. Percept. Perform. 36, 1029-1039.

Micheyl, C., Kreft, H., Shamma, S., and Oxenham, A. J. (2013b). "Temporal coherence versus harmonicity in auditory stream formation," J. Acoust. Soc. Am. 133, EL188.

Micheyl, C., and Oxenham, A. J. (2010). "Objective and subjective psychophysical measures of auditory stream integration and segregation," J. Assoc. Res. Otolaryngol. 11, 709-724.

Micheyl, C., Tian, B., Carlyon, R. P., and Raunschecker, J. P. (2005). "Perceptual organization of tone sequences in the auditory cortex of awake macaques," Neuron 48, 139-148.

Middlebrooks, J. C., and Onsan, Z. A. (2012). "Stream segregation with high spatial acuity," J. Acoust. Soc. Am. 132, 3896-3911.

Miller, G. M., and Heise, G. A. (1950). "The trill threshold," J. Acoust. Soc. Am.22, 637-638.

Moore, B. C. J., and Glasberg, B. R. (1983). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. 74, 750-753.

Moore, B. C. J., and Gockel, H. E. (2002). "Factors influencing sequential stream segregation," Acta. Acust. Acust. 88, 320-332.

Neff, D. L., Jesteadt, W., and Brown, E. L. (1982). "The relation between gap discrimination and auditory stream segregation," Percept. Psychophys. 31, 493-501.

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (1987) "An efficient auditory filterbank based on the gammatone function," in paper presented at a meeting at the IOC Speech Group on Auditory Modelling at RSRE, December 14-15.

Piechowiak, T., Ewert, S. D., and Dau, T. (2007). "Modeling comodulation masking release using an equalization-cancellation mechanism," J. Acoust. Soc. Am. 121, 2111-2126.

Plack, C. J. (2005). "The auditory scene," in The Sense of Hearing (Erlbaum Associates, New York), Chap. 10, pp. 193-214.

Püschel, D. (1988). "Prinzipen der zeitlichen Analyse beim Hören" ("Principles of temporal processing in hearing"), Ph.D. thesis, University of Göttingen, Germany.

Rose, M. M., and Moore, B. C. J. (2000). "Effects of frequency and level on auditory stream segregation," J. Acoust. Soc. Am. 108, 1209-1214.

Richards, V. M. (1987). "Monaural envelope correlation perception," J. Acoust. Soc. Am. 82, 1621-1630.

Risset, J.-C., and Mathews, M. V. (1969). "Analysis of musical instrument tones," Physics Today, 22, 23-30.

Roberts, B., Glasberg, B. R., and Moore, B. C. J. (2002). "Primitive stream segregation of tone sequences without differences in fundamental frequency or passband," J. Acoust. Soc. Am. 112, 2074-2085.

Ruggero, M. A., Rich, N. C., Recio, A., Narayan, S. S., and Robles, L. (1997). "Basilar-membrane responses to tones at the base of the chinchilla cochlea," J. Acoust. Soc. Am. 101, 2151-2163.

Schooneveldt, G. P., and Moore, B. C. J. (1987). "Comodulation masking release (CMR): Effects of signal frequency, flanking band frequency, masker bandwidth, flanking-band level, and monotic versus dichotic presentation of the flanking band," J. Acoust. Soc. Am. 82, 1944-1956.

Shamma, S. A. (1985). "Speech processing in the auditory system. II. Lateral inhibition and the central processing of speech evoked activity in the auditory nerve," J. Acoust. Soc. Am. 78, 1622-1632.

Shamma, S. A., Elhilali, M., and Micheyl, C. (2011). "Temporal coherence and attention in auditory scene analysis," Trends Neurosci. 34, 114-123.

Shamma, S., Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., Pressnitzer, D., Yin, P., and Xu, Y. (2013). "Temporal coherence and the streaming of complex sounds," in Basic Aspects of Hearing, Advances in Experimental Medicine and Biology edited by B. C. J. Moore, R. D. Patterson, I. M. Winter, R. P. Carlyon, and H. E. Gockel (Springer Science Ã¿ Business Media, New York), Chap. 59, pp. 535-543.

Shamma, S., and Micheyl, C. (2010). "Behind the scenes of auditory perception," Curr. Opin. Neurobiol. 20, 361-366

Siebert, W. M. (1968). Quarterly Report No. 88, MIT Research Laboratory of Electronics.

Singh, P. G. (1987). "Perceptual organization of complex-tone sequences: a tradeoff between pitch and timbre?" J. Acoust. Soc. Am. 82, 886-899.

Summerfield, Q., Culling, J. F., and Fourcin, A. J. (1992). "Auditory segregation of competing voices: Absence of effects of FM or AM coherence [and discussion]," Phil. Trans. R. Soc. Lond. B. 336, 63-72.

Turgeon, M., Bregman, A. S., and Ahad, P. A. (2002). "Rhythmic masking release: Contribution of cues for perceptual organization to the crossspectral fusion of concurrent narrow-band noises," J. Acoust. Soc. Am. 111, 1819-1831.

Turgeon, M., Bregman, A. S., and Roberts, B. (2005). "Rhythmic masking release: Effects of asynchrony, temporal overlap, harmonic relations, and source separation on cross-spectral grouping," J. Exp. Psychol. Hum. Percept. Perform. 31, 939-953.

Uppenkamp, S., Fobel, S., and Patterson, R. D. (2001). "The effects of temporal asymmetry on the detection and perception of short chirps," Hear. Res. 158, 71-83.

van de Par, S., and Kohlrausch, A. (1998). "Analytical expressions for the envelope correlation of narrow-band stimuli used in CMR and BMLD research," J. Acoust. Soc. Am. 103, 3605-3620.

van Noorden, L. P. A. S. (1975). "Temporal coherence in the perception of tone sequences," Ph.D. dissertation, Institute for Perception Research, Eindhoven, The Netherlands.

van Noorden, L. P. A. S. (1977). "Minimum differences of level and frequency for perceptual fission of tone sequences ABAB," J. Acoust. Soc. Am. 61, 1041-1045.

Verhey, J. L., Dau, T., and Kollmeier, B. (1999). "Within-channel cues in comodulation masking release (CMR): Experiments and model prediction using a modulation filter bank model," J. Acoust. Soc. Am. 106, 2733-2745.

Verhey, J. L., Ernst, S. M. A., and Yasin, I. (2012). "Effects of sequential streaming on auditory masking using psychoacoustics and auditory evoked potentials," Hear. Res. 285, 77-85.

Vliegen, J., Moore, B. C. J., and Oxenham, A. J. (1999). "The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task," J. Acoust. Soc. Am. 106, 938-945.

Vliegen, J., and Oxenham, A. J. (1999). "Sequential stream segregation in the absence of spectral cues," J. Acoust. Soc. Am. 105, 339-346.

Von der Malsburg, C., and Schneider, W. (1986). "A neural cocktail-party processor," Biol. Cybern. 54, 29-40.

Wen, B., Wang, G. I., Dean, I., and Delgutte, B. (2012). "Time course of dynamic range adaptation in the auditory nerve," J. Neurophysiol. 108, 69-82.

Wojtczak, M., Beim, J. A., Micheyl, C., and Oxenham, A. (2012). "Perception of across-frequency asynchrony and the role of cochlear delays," J. Acoust. Soc. Am. 131, 363-377.

Wojtczak, M., Beim, J. A., Micheyl, C., and Oxenham, A. (2013). "Effects of temporal stimulus properties on the perception of across-frequency asynchrony," J. Acoust. Soc. Am. 133, 982-997.

Zera, J., and Green D. M. (1993a). "Detecting temporal asynchrony with asynchronous standards," J. Acoust. Soc. Am. 93, 1571-1579.

Zera, J., and Green, D. M. (1993b). "Detecting temporal onset and offset asynchrony in multicomponent complexes," J. Acoust. Soc. Am. 93, 1038-1052.

Zera, J., and Green, D. M. (1995). "Effect of signal component phase on asynchrony discrimination," J. Acoust. Soc. Am. 98, 817-827.

Zilany, M. S. A., and Carney, L. H. (2010). "Power-law dynamics in an auditory-nerve model can account for neural adaptation to sound-level statistics," J. Neurosci. 30(31), 10380-10390.

Contributions to Hearing Research

Vol. 1: *Gilles Pigasse*, Deriving cochlear delays in humans using otoacoustic emissions and auditory evoked potentials, 2008.

Vol. 2: *Olaf Strelcyk*, Peripheral auditory processing and speech reception in impaired hearing, 2009.

Vol. 3: *Eric R. Thompson*, Characterizing binaural processing of amplitude-modul-ated sounds, 2009.

Vol. 4: *Tobias Piechowiak*, Spectro-temporal analysis of complex sounds in the human auditory system, 2009.

Vol. 5: *Jens Bo Nielsen*, Assessment of speech intelligibility in background noise and reverberation, 2009.

Vol. 6: *Helen Connor*, Hearing aid amplification at soft input levels, 2010.

Vol. 7: *Morten Løve Jepsen*, Modeling auditory processing and speech perception in hearing-impaired listeners, 2010.

Vol. 8: *Sarah Verhulst*, Characterizing and modeling dynamic processes in the cochlea using otoacoustic emissions, 2010.

Vol. 9: *Sylvain Favrot*, A loudspeaker-based room auralization system for auditory research, 2010.

Vol. 10: *Sébastien Santurette*, Neural coding and perception of pitch in the normal and impaired human auditory system, 2011.

Vol. 11: *Iris Arweiler*, Processing of spatial sounds in the impaired auditory system, 2011.

Vol. 12: *Filip Munch Rønne*, Modeling auditory evoked potentials to complex stimuli, 2012.

Vol. 13: *Claus Forup Corlin Jespersgaard*, Listening in adverse conditions: Masking release and effects of hearing loss, 2012.

Vol. 14: *Rémi Decorsière*, Spectrogram inversion and potential applications for hearing research, 2013.

Vol. 15: *Søren Jørgensen*, Modeling speech intelligibility based on the signal-to-noise envelope power ratio, 2014.

Vol. 16: *Kasper Eskelund*, Electrophysiological assessment of audiovisual integration in speech perception, 2014.

Vol. 17: *Simon Krogholt Christiansen*, The role of temporal coherence in auditory stream segregation, 2014.