Abstract

Models of perceptual grouping are usually using verbal, poorly accurate predictions. As of late, probabilistic models are being used more and more to create more stringent descriptions of the underlying mechanisms, and quantitative predictions. This thesis presents a nonparametric Bayesian clustering algorithm applied to Auditory Scene Analysis (ASA), along with several validations from the classical literature on the subject, and experiments using a new paradigm in different experimental settings. Grouping/segregation processes in ASA, and therefore in the model, follow similar Gestalt principles as in the more studied visual field: the more tones are similar, the more they tend to be clustered in a single auditory stream, and conversely. The first study focuses on a mathematical description of the clustering algorithm and on its validation on well-known studies from the field. A new paradigm has been used to create situations where 3 simultaneous streams could be reached by increasing the distance in frequencies between rapidly played tones, as predicted by the classical ASA model and our own. Results were in line with the hypotheses. The second study expands on the first one and uses qualitative predictions from the clustering algorithm to observe stream segregations using increasing differences in several dimensions at once in two experiments using the same paradigm: namely, frequency and spatial distance in the first one, frequency and timbre in the second one. Results presented an unexpected pattern, suggesting a stronger influence of attention as initially supposed. The third study explored the influence of attention on the stream formation process in the same paradigm, by adding specific attentional instructions to participants. Results suggest a possible limitation to 2 simultaneous attentional streams: the foreground stream, and the background one where all tones are clustered. Overall, while the model was only used to create qualitative predictions, those were useful enough to guide experiments with impactful results.

A nonparametric Bayesian clustering approach to auditory perception

Nathanael C. Larigaldie

Submitted for Degree of Doctor of Philosophy

Durham University

Department of Psychology

2021

Table of Contents

Acknowledgements

First and foremost, I would like to thank Dr. Ulrik Beierholm for trusting me and giving me the fantastic opportunity to do a Ph.D. under his supervision. His patience and understanding have been limitless, and I will forever be grateful for his support.

Maïtena, you have been there for me for years and I hope I was half as good to you as you were to me. I promise that next time we are going to McDonald's, we won't have to count every penny.

Valerie and Michael, you have helped me become a better researcher and after all these years, I still dearly value the advice you gave me.

Amaury, I simply could not have made it through this last year without you. And probably those before either.

Thank you, Josh, for the unexpected British hugs, the shooters and their Tequila, along with the marvellous nights of discussion. Also, thanks Ali for your friendship and kindness.

Thank you, Delphine, for your love and your never-ending desire for knowledge and self-improvement. Your dedication and bravery truly amaze me.

I would also like to thank the incredibly helpful and friendly staff from the Psychology department at Durham University.

Last but not least, I would like to thank my mother and brother, who never needed to understand anything about whatever I was doing to support and be proud of me.

Chapter submissions

Chapters 2, 3 and 4 are written up for journal submission. Parts of them may have been presented in conferences, and titles used in the thesis may ultimately differ from those used for future publications.

Chapter 2

Using 'Occam's Razor' for causal inference of auditory perception (Nathanael Larigaldie, Tim Yates, Ulrik Beierholm, submitted)

Presented at: Basic Auditory Science conference 2019, CogSci 2017 (slightly different version published as a conference paper), BPS Cognitive Section Conference 2017

Chapter 3

Auditory stream segregation in a complex and controlled environment: how do space and timbre interact with frequency differences? (Nathanael Larigaldie, Ulrik Beierholm, writing in progress)

Chapter 4

Testing the role of attention in Auditory Scene Analysis (Nathanael Larigaldie, Ulrik Beierholm, writing in progress)

Chapter 1

–

General introduction

Introduction

The question of how the human brain processes perceptual information still contains many mysteries in contemporary science. The fields of science that usually focus on this type of scientific question such as psychology and neuroscience, while analysing data with quantitative methods, often formulate underlying models in a verbal, qualitative way. In a very influential paper, Marr & Poggio (1976) suggested that the central nervous system should be studied at three complementary levels of analysis to be properly understood: the computational (what does the system do and why?), algorithmic (how does the system achieve it?) and implementation (how is the system physically constructed?) levels. Traditional approaches in psychology are often stuck somewhere in between computational and algorithmic levels, as qualitative descriptions only provide useful but coarse insights into the way a system achieves its functions.

However, mathematical and computational models are more clearly defined and more detailed than these qualitative counterparts, therefore allowing us to develop a better understanding and create predictions of experimental evidence. Indeed, this approach is still infrequent in psychology even if it is gaining more and more interest through the years, as researchers hope to generate models providing more evidence in the algorithmic and implementation levels of analysis. This research project aims to contribute to the progress of this methodology.

This objective is very timely, thanks to the ever-growing accessibility of extremely powerful hardware during the past few decades. Computational models will most certainly have a larger role in future research, as this new hardware allows the computation of entirely new algorithms or theoretical ones that used to be intractable. For instance, deep neural network-based emulations of human behaviour are now possible to generate both in reasonable time and at a reasonable cost and are among the best candidates for implementation-level simulations. As a result, a growing number of computational models

are now being developed, based both on brain observations and hypotheses, and computer science advancements. This multidisciplinary approach has already been fruitful for research in both fields (Russell & Norvig, 2016).

Until recently, serious limitations inherent to computational complexity and statistical decisions have made it difficult to design models applicable to some aspects of the human mind. Indeed, humans have to deal daily with the uncertainty that can be caused by partial observability or nondeterminism. Frequentist approaches to probability and statistics that are widely used in social and biological sciences provide scientists with an invaluable toolbox to do inferences on populations provided they have access to very large data samples. However, only a limited set of questions can be answered by these methods, and they lack robustness when it comes to making inferences from very small samples. More importantly, as probabilities are interpreted only to be long-term frequencies, they lack the ability to model uncertainty or credence altogether (Hájek, 2012).

As a consequence, this type of approach by itself cannot be used to simulate any kind of algorithm used by our nervous system to deal with new and unique situations, or simply uncertainty and subjective belief. Bayesian statistical approaches, on the other hand, are specifically designed to handle these situations of uncertainty and are therefore strong candidates to an algorithmic-level analysis of the human brain. They also allow for a broader range of statistical questions, to simultaneously estimate model parameters and the uncertainty about them (Turner & Sederberg, 2014), while sometimes granting better control of a model complexity through conditional independence (Russell & Norvig, 2016). These models can be used in a wide range of applications in artificial intelligence and neuroscience.

Bayesian modelling in perception

Imagine a primate being confronted with a trembling bush that he can see right in front of him, and the roaring of a lion that he can hear. From an evolutionary perspective, it makes sense that this observer has to infer if that is the case, that the two cues are coming from the same dangerous source and therefore that there is a predator on the verge of jumping on him - and that he should run for his life in the opposite direction. But the information he gets from his senses is quite imprecise, as he does not directly see the animal, and pinpointing a sound is difficult by nature. However, combining those two cues together can clearly give a more reliable estimation of the actual location of the animal, as the observer now has several pieces of information instead of one about the same phenomenon. If the sound seemed to originate slightly more to the right than the rustling bush, then the predator is most probably located somewhere in-between (see Figure 1.1 for a graphical representation of this cue combination). This evolutionary pressure should be common for any being dealing with multisensory stimuli, and might be among the biggest advantages of developing several senses in the first place. Such a system also has the ability to reduce the total amount of data it has to face (going from estimates of central tendency and uncertainty for each cue to only one once those are combined in the example shown in Figure 1.1), although it may be seen as a side effect rather than the aim.

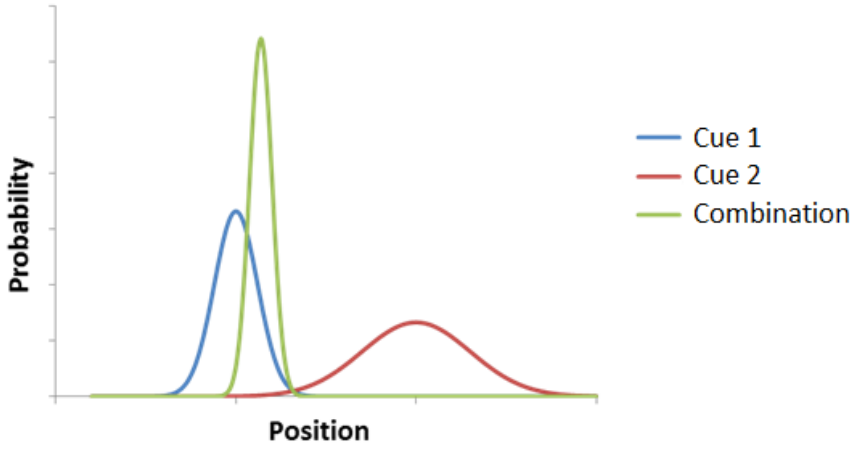


Figure 1.1: graphical representation of the gain in precision thanks to a perceptual cue combination

Thankfully, we have some practical mathematical tools at our disposal to model this situation. Traditionally used in the context of higher cognitive reasoning and decision-making, Bayesian statistics are specifically designed to represent and deal with the same kind of uncertainty that our ancestor is facing. The general idea behind the utilization of Bayesian inference models in perception is that human beings and other animals using sensory organs can only extract noisy and incomplete cues from the world surrounding them. One of the most important tasks of the nervous system is then to combine these cues and extract the relevant information in a way that allows the observer to comprehend and/or act in an optimal way through an unconscious inference (Knill & Richards, 1996)[1].

Interestingly, Bayesian modelling has been used in a growing number of studies and has demonstrated a strong capacity to predict human and animal perception. Weiss et al. (2002) were able for instance to make a model be subject to similar velocity illusions as humans by using mathematical descriptions of commonly assumed properties about velocity perception and prior expectations. Other authors could provide evidence that models combining two different cues from the same visual modality could simulate humans' ability to judge a surface slant (Knill & Saunders, 2003). Similar results were

[1] During the course of this thesis, a few assumptions and terminologies about the world and the perceptual system will remain constant. The term *stimulus* will sometimes be used to denote items in the world with objective and non-noisy characteristics. This objective stimulus is then being perceived by an observer who only has access to noisy estimates about these characteristics. Once on the observer's side, this stimulus will often be called a *percept*, and be considered as the atomic unit of perception. At times, the *objective* adjective will also be used to reduce the ambiguity of whether the argument is trying to talk about the nature of the physical world, or the resulting phenomenological state. If we imagine that some source in the physical world is producing two stimuli, for instance reflects light (producing an image) and makes its surrounding environment vibrate (producing a sound), the observer also has two resulting percepts with its own perceived properties. As an example: if produced simultaneously, both stimuli come from the same objective position *X* (which is a real world property), but because of the noise in the physical signal on its course to the observer (e.g. energy loss of sound in the air, light refraction, etc…), along with the imperfect nature of our sensory and information-transmitting organs (unreliable measures, signal compression, etc…), the resulting perceived positions of both percepts, *Y* and *Z*, will probably not be exactly equal to *X*, and there is therefore some uncertainty about the real world. One supposed goal of the perceptual system is to produce an estimation of the true position *X*, as it is what the system is interested in for its survival. Please note that sometimes, *percept* will also refer to the phenomenological result of this estimation process (e.g. the perceived location of *X* once the integration of *Y* and *Z* has been done). Whenever *noise* in perception is mentioned throughout the thesis, we by default integrate all noise sources into a single parameter: the standard deviation of a normal distribution, as represented in Figure 1.1. Using a normal distribution to model the uncertainty is particularly relevant in these conditions considering that it is the result of the sum of noise coming from several sources.

observed in animals such as ferrets, across different sensory modalities (Hollensteiner et al., 2015).

The theory behind it is simple. Each subjective perceptual information (Y) can be regarded as linked to the objective, fixed information (X) that the observer is seeking to know, through a probability distribution ($P(Y|X)$). This so-called likelihood is dependent on the kind of information, and the structure of the sensory organ. Combined with possible prior expectations about this objective information ($P(X)$), Bayes theorem gives us a way to find a posterior probability distribution about it ($P(X|Y)$):

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

The denominator is only a normalizing term, so this equation can be rewritten without losing valuable information as:

$$P(X|Y) \propto P(Y|X)P(X)$$

On top of this, the denominator can be extremely difficult to calculate as the model gains in complexity, even for modern computers. Fortunately, methods called Monte Carlo Markov Chains (MCMC) make it possible to obviate its calculation by sampling from the posterior distribution without explicitly calculating it. Those samples can then be used to do estimates, such as the mean or variance of a variable which is often what is needed. It is therefore possible to do Bayesian inference in rich and complex models (Kruschke, 2014).

If the observer has another perceptual information at hand to use (Z) about the same objective information, he can use both to obtain a more sensible estimation about it:

$$P(X|Y,Z) \propto P(Y|X)P(Z|X)P(X) \qquad \text{(Equation 1)}$$

Equation 1 illustrates how simple it can be to calculate a cue combination and come up with a more reliable estimation of the variable an observer is looking for. In fact, Figure 1.1 was drawn by using this particular equation.

But let's go back to our ancestor for now. Earlier, he assumed that the roaring and the rustling of the bush were both caused by the same event – a dangerous lion. But if he perceives instead that the sound is coming from behind him, he should quickly conclude that the two events are in fact unrelated and that his chances of survival are probably higher if he runs towards the bush he would have run away from in the other scenario. Indeed, it seems very unlikely that a sound originating from behind him has anything to do with a movement from the opposite direction, and therefore the information coming from his visual perception becomes irrelevant for his survival. Combining the two cues together in this particular instance would result in a dangerous inference surely leading to an untimely death. Keeping them separated – even if that means acting on a less accurate estimation than in the first case – is most probably the right way to go. In both cases, this observer did not rely solely on one of the two percepts for his survival, but the inference made about their possible combination gave him a better chance to avoid death. But another question arises: if in some situations a cue combination is beneficial for the estimations made and harmful in others, then how should an observer decide when cues should or should not be combined together?

Fortunately, the Bayesian toolkit easily allows performing model comparisons, which compute relative probabilities of each model given the observed variables. Applied to this situation, it becomes possible to compare if a model combining the different cues seems more or less plausible than another one where the cues are kept separated. In the field of perception, this particular method is called "causal inference", as the decision to combine the cues or keep them separated is, in fact, the same as considering them to be respectively generated by a same underlying cause, or by several. Some statistical

Bayesian model comparison made to simulate a human causal inference ability have proven to be applicable to cues combination in perception (Shams & Beierholm, 2010).

This type of causal inference model has already proven to be a good fit for behavioural observations in humans, especially in multisensory integration situations (Körding et al., 2007; Shams & Beierholm, 2010; Ernst & Di Luca, 2011). It has even been suggested that cue combination and causal inference are not innate properties and are rather the product of a reward dependent training during childhood (Weisswange et al., 2011).

Going back to the theory, Bayesian causal inference models come in handy when an observer cannot know beforehand if it is relevant to treat both percepts as being generated by the same source. Fundamentally, the idea is to obtain a probability for each causal source possibility. For instance, when there are two cues available, there could be one source causing them both (*C=1*) or two different sources (*C=2*), but the theory is applicable to any number of cues. In this simple case, the probability for each causal structure from Equation 1 can be calculated through the following equation:

$$P(C|Y,Z) \propto P(Y,Z|C)P(C) \qquad \text{(Equation 2)}$$

$P(C)$ can be calculated in several ways. It can be an uninformative prior ($P(C = 1) = P(C = 2) = 0.5$), it can be biased towards one of the two possibilities, or be fitted to participants' responses in an experiment. In the case of an individual source causing the two percepts, having one cause is equivalent to saying that there is one fixed piece of information *X* that the observer is trying to infer. It follows that:

$$P(Y,Z|C = 1) \propto P(Y,Z|X)P(X)$$

Since Y and Z are conditionally independent given X:

$$P(Y,Z|X) \propto P(Y|X)P(Z|X)$$

Which are the same terms as in Equation 1. In this situation, Equation 2 therefore becomes:

$$P(C = 1|Y, Z) \propto P(Y|X)P(Z|X)P(X)P(C = 1)$$

Similarly, if we consider the possibility that there are two different causes, there are then two pieces of information X_1 and X_2 for the observer to infer. These can immediately be considered independent, as they are totally separate events. It follows that:

$$P(C = 2|Y, Z) \propto P(Y|X_1)P(Z|X_2)P(X_1)P(X_2)P(C = 2)$$

Armed with this inferential ability, a Bayesian observer could very simply and quickly determine if the perceptual cues he has at hand should be combined or not, and accordingly end up with a stronger estimation of the information he needs. All in all, the Bayesian toolkit seems to not only be a convenient way to mathematically represent reasoning under uncertainty, but it could be that the human brain is actually using similar algorithms to perform its calculations.

Literature about neurological correlates to Bayesian cue combination and causal inference is still scarce, but recent studies suggest that neural activity is indeed similar to what could be expected from a statistical calculator (Fetsch et al., 2011, 2013). Cortical hierarchies could even play similar roles as the different levels of hierarchical Bayesian models do (Rohe & Noppeney, 2015). While these results look promising, it is still too early to conclude that the brain's perceptual processing is indeed acting as a Bayesian observer at a neurological level. Going back to Marr's Tri-Level Hypothesis, these models do not initially have the pretention to go lower than an algorithmic level.

It has already been argued that the ideal Bayesian observer, whose perceptual goal is to obtain an accurate vision of reality and use it for its purposes, may be less susceptible to survival than a Darwinian observer, whose perception is already biased towards survival

and does not need to infer anything about an objective reality at any point (Hoffman et al., 2015). Nevertheless, the two hypotheses have points of convergence, and it could be argued that in many cases, the result is the same. Indeed, the information perceived in an ideal observation probably contains whatever information a positively biased observation could extract from its environment, the main difference being the cognitive resources needed to reach the same useful information. Furthermore, thousands of years of technological advances have shown that human intelligence and perception could go way beyond its sole survival, otherwise, scientific advancements may have been impossible. Finally, an ideal observer can very easily be turned into different kinds of biased observers by manipulating sensory inputs fed to the algorithms. As such, a Bayesian observation can be considered as a good approximation to human perception. Indeed, the Bayesian framework, in particular model comparison and hierarchical modelling, contains statistical tools that allow making very fast decisions about different possible interpretations of the world it is applied to, and can as such be viewed as a mimic of some of the brain's cognitive abilities.

Gestalt psychology and perceptual grouping

Despite the gain in popularity of these mathematical models, perception is one of those fields of study in Psychology in which ideas and theories are often described in a qualitative way. Even if Psychophysics has produced an invaluable amount of quantitative research and tools for the study of perception, some of the oldest and yet most reliable observations continue to evade stringent interpretations through mathematical measures. It is still the case for most of the Gestalt theory despite its century of existence and tremendous efforts spent towards this goal (Jäkel et al., 2016).

The founding principle of the Gestalt theory is generally accepted to be the idea that "the whole is other than the sum of its parts" (Koffka, 1935). Gestalt psychologists refused the idea that perception is only a sum of elementary percepts, but rather thought

that an observer has to actively structure and organise the environment he is confronted with – as opposed to what behaviourism schools of thought theorized at the same period (Wagemans et al., 2012).

As decades went by, Gestalt researchers gathered a great range of observations – especially using visual perception – confirming their original hypotheses. The seeming universality of some of these observations is still striking today, as few psychological experiments result in such consistency across individuals. This compelled early researchers to speculate about general qualitative principles and laws leading to such similar constructs upon perceptions. This was particularly true for perceptual grouping, especially in visual perception. Perceptual grouping is a central aspect of this project, as causal inference can be viewed as a way to create groups of percepts with the assumption from the perceptual system that since they share similarities, they are generated by one or several similar sources.

Perhaps one of the most fundamental principles of Gestalt perceptual grouping is the law of prägnanz, stating that we tend to order our experience in a manner that is regular, orderly, symmetrical, and simple. However, this very general law was too vague to allow for an efficient depiction of perception, which is why several others were devised by Wertheimer (1923) to precise it.

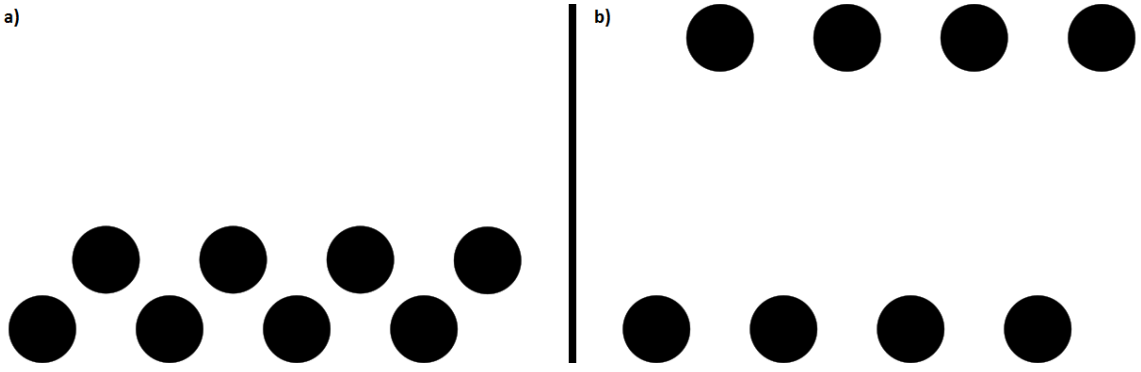


Figure 1.2: Visual example of the law of proximity. Dots in a) all seem part of a unique group, while those in b) seem to form two distinct groups

For instance, the law of proximity states that individuals tend to perceive similar objects as forming different groups when they are closer than they are to others (see Figure 1.2). General similarities of some characteristics of the stimuli, such as colour, size, or orientation, also tend to make the observer group items together, and so does the *common fate*, i.e. the fact that objects are moving in the same way. For example, if in an array of dots half of them are moving upward while the other half is moving downward, an observer would tend to group dots according to the direction of the movement. If every dot is moving in the same direction but at different velocities, then the velocity would become the grouping criteria. A lot of other laws, such as continuity, symmetry, or closure were also devised, but are sometimes less easily applicable to other sensory modalities than vision.

Recent experiments continue to validate the relevance of some Gestalt principles and continue to broaden their range of application. Lee & Blake (1999) showed for instance that temporal synchrony could provide a common fate cue leading to a spatial grouping strong enough to create the sensation of a visual structure. Sekuler & Bennett (2001) demonstrated that in the form of luminance changes, common fate could be a critical clue in the context of object recognition, and to differentiate objects from the background. Palmer (1992) even proposed an entirely new perceptual grouping principle called *common region*, stating that elements being located within a common region of space tend to be so strongly perceived as grouped together that it can overcome the effects of proximity and similarity. Similarly, *element connectedness* is now also considered as an important principle governing perceptual grouping: whenever seemingly distinct elements share a common border, are linked by an object with different properties, or tend to behave like a single object when being manipulated (like the spikes and the handle of a fork), humans tend to perceive them as a single group or entity (Palmer & Rock, 1994).

Traditionally, the vision modality was studied more extensively than other senses in Gestalt perception. However, Gestalt principles were also linked to tactile perception (Gallace & Spence, 2011) and motor action (Klapp & Jagacinski, 2011) among others. The auditory perception was perhaps less investigated than vision because of the difficulty added by the simultaneous and sequential segregation possibilities that are less important in vision, along with the general noise and overlap in the stimuli (Bregman, 1994).

Perhaps one of the most puzzling phenomena in the auditory Gestalt literature is sometimes referred to as the "Cocktail party problem" (Cherry, 1953; Qian et al., 2018): in a cacophonic environment such as a cocktail party, it seems disarmingly easy, fast and automatic for human beings to use two noisy signals (one for each ear) within which all perceptual information is initially fused as input, and yet still extract enough information from them to attend to the one conversation they are interested in and discard the rest as background noise. As speech perception is multimodal, other senses such as vision can participate in refining our ability to select how to reconstruct sensible inner representations through multisensory cue combination. Lipreading has been shown to help with correctly extracting information in the Cocktail party effect (Summerfield, 1992). This can be linked to the McGurk effect, describing how phonemes can be interpreted differently depending on the kind of lips moving video you present while an ambiguous speech sound is being played (Tiippana, 2014). Even if other senses do significantly help, human beings are still able to solve the Cocktail party problem in unimodal situations with great accuracy.

This puts a strong emphasis on at least two, not necessarily independent, questions of interest to study: How are auditory cues extracted and grouped/segregated to form a coherent and usable flow of speech? And how does our selective attentional system select the relevant information in such a situation? These questions can of course be generalized to any auditory context such as music, where one can decide to listen to a specific instrument. Traditionally the Gestalt literature has a stronger focus on extraction and

grouping mechanisms considered separately from attention by hypothesizing that the two processes are independent. In other words, stream formation is often considered a pre-attentive process: our perceptual system first extracts the relevant information from the signal, uses them to form perceptual clusters based on stimuli's characteristics, and only then top-down attention kicks in to select the cluster of interest on which treatments can then be done (Bregman & Rudnicky, 1975).

Albert Bregman and the McGill Auditory Research Laboratory specialized in research on the Auditory Scene Analysis (ASA). One of their aims was to investigate how the auditory system can take a single sound signal and be able to cluster parts of it into different groups they call 'streams' so that the listener does not mix up information from different sources he needs. A lot of studies rose from this initiative, that are most often focusing on high-level mechanisms leading from objective sensory cues (e.g. tone frequencies for each tone in a melody) to streams, rather than low-level mechanisms from a unitary noisy sound signal to a collection of objective sensory cues. We will, throughout this whole research project, also remain on a high level of cognition and therefore consider that our perceptual system has accurate access to objective cues.

One of the first and most fundamental discoveries in this field of research was the fact that when presented with rapid sequences of three or four non-speech sounds lasting 200 ms each, participants were mostly unable to report their order while being able to recall them (Warren et al., 1969). Indeed, previously trained groups of participants who learned names for each of the stimuli could correctly say which ones were present in the recordings they heard, but could not say in which order they were presented better than luck. Later on, Bregman & Campbell (1971) proposed that this inability was due to the fact that the auditory system automatically segregates sounds into several streams based on some characteristics of the signal, such as sound frequencies and the time interval between them. More specifically, they based their idea on former observations from musical theory

where a single instrument could be perceived as being two when alternating very quickly between high and low tones, as if the auditory system made a prior assumption that a single source requires time to adjust the frequency of the sounds it is producing. Their experiment showed a strong tendency to group high tones together and low tones as another stream when participants were presented with sequences of rapidly alternating high and low tones such as HLHLHL. They also proposed that the inability to recall the order of the sounds was due to limitations in attention, which can only be directed towards one stream at a time (see Figure 1.3). Bregman & Achim (1973) later demonstrated that a very similar procedure adapted to the visual modality created the same pattern of results, therefore showing a clear link between visual and auditory perceptual groupings.

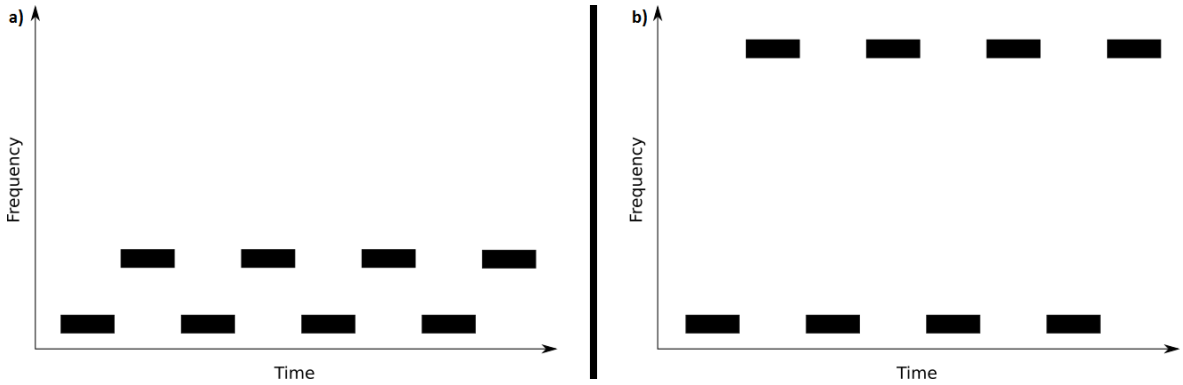


Figure 1.3: Visual representation of auditory stimuli similar to those used in Bregman & Campbell (1971). Order of tones in a) can be easily recalled, while it is harder for those in b). The attentive reader may notice the strong similarity between this figure and Figure 1.2.

The auditory equivalent of *element connectedness* was also uncovered when, in yet again the same experimental settings in which frequency glides were added to link the sounds presented, participants had the tendency not to segregate them into different streams anymore (Bregman & Dannenbring, 1973). Other experiments introducing random noise during these frequency glides, or gaps between different sounds, demonstrated evidence of laws of closure and proximity similar to those seen in the visual modality (Bregman et al., 1999, 2000; Dannenbring, 1976; Dannenbring & Bregman, 1976). This data strongly suggests that common mechanisms are being used across different, if not all, sensory modalities.

Even if these experiments provide some insight on the Cocktail party problem, most of them are limited to the segregation of 2 streams, their application to a higher number of streams being more suggested than verified experimentally. If stream formation is indeed a pre-attentive process, this greatly reduces the ecological validity of the field: in a room full of people talking, our perceptual system should form as many clusters as there are people talking in hearing range (not counting other sound sources). Furthermore, they mostly fall under the same criticisms as those usually applied to the Gestalt theory of perception.

One of the most common of these criticisms is the fact that all of its laws and observations are often qualitatively described and are therefore more descriptive than predictive. Consistently failed attempts to quantify Gestalt laws and principles in a unified way contributed to its decline by the mid-20th century. However, the Gestalt theory of perception can be conceptually linked to Bayesian cue combinations and causal inference: through a mathematical process, individual percepts and their uncertainty can be integrated and give rise to quantitatively and qualitatively different information than their simple sum. As such, the Bayesian framework could very well be the mathematical tool that the Gestalt missed in the past.

Indeed, as was mentioned earlier, causal inference and perceptual grouping are closely related, since percepts should frequently, if not always, be grouped together based on their original source. One of the main objectives of this project is to contribute to perceptual grouping modelling in the context of auditory perception, which could then potentially lead to generalizations in multisensory integration situations.

Bayesian modelling in perception revisited

A few causal inference model examples applied to sensory integration were presented in the last section, whose main goal was to combine sensory cues together in an

optimal way to extract the most accurate information possible from what is available. Although they were successfully used, common *parametric* Bayesian statistics present some serious limitations in this situation. Among them is the fact that as sensory cues stack up, the number of possible causal inferences grows factorially, and calculation methods used to judge their respective credence require to compute them all to be able to take a decision (as shown by Equation 2). When only two cues are available, it is only necessary to consider 2 possibilities: either both cues were created by 1 common source and one should combine them to extract more information, or they were generated by different sources. With just 3 cues, it is necessary to examine 5 possibilities (see Figure 1.4 for a

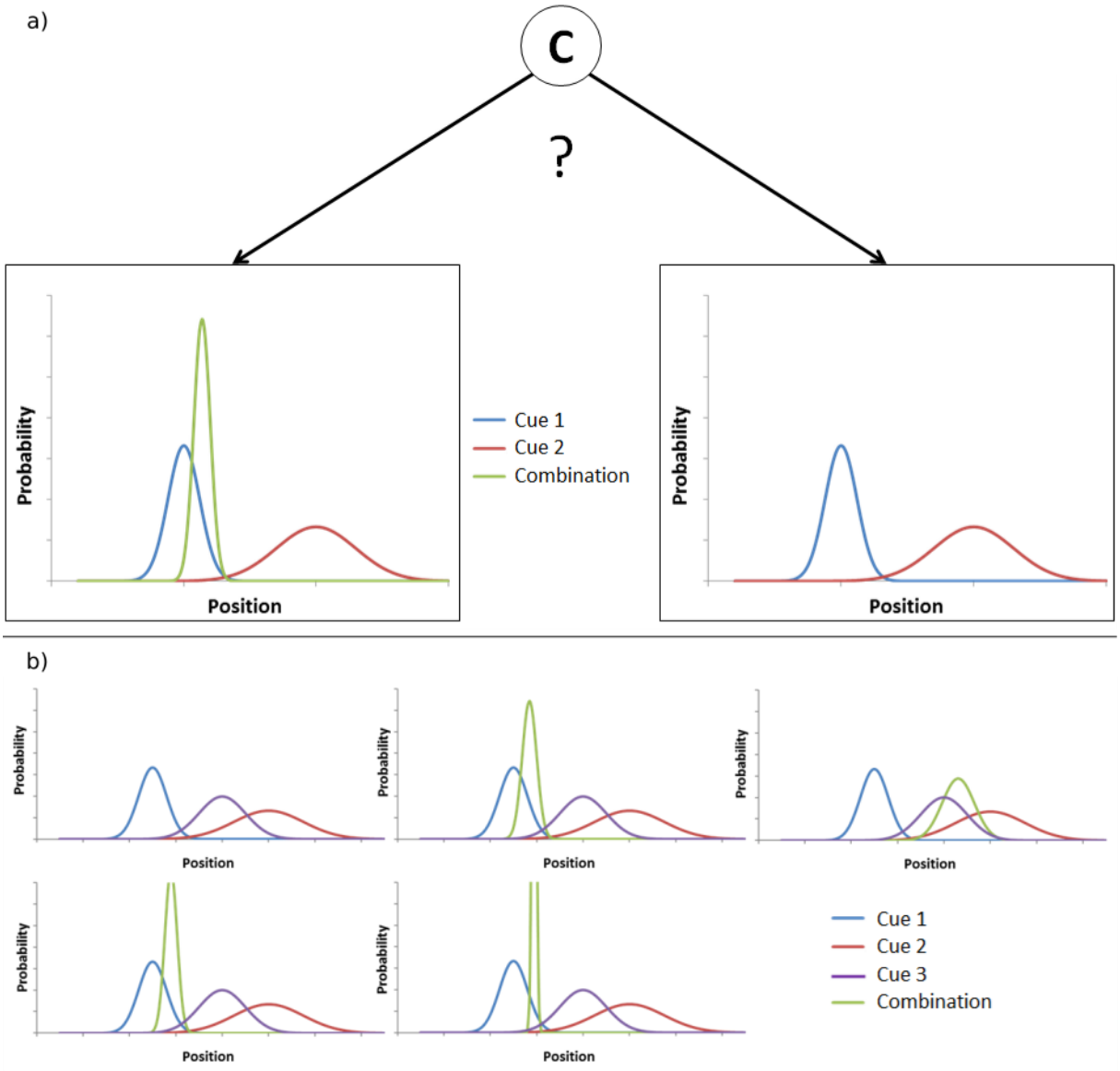


Figure 1.4: Causal inference using parametric Bayesian methods on 1D position cues. a) represents the two possibilities to consider for inference in a double cues situation. b) represents the 5 possibilities to compute when 3 cues are available

visual example). Considering only 6 cues, there are already a total of 858 possible causal combinations to consider. Since human beings are bombarded with hundreds of sensory cues at any moment, the computation of a realistic model becomes virtually impossible – and it seems unlikely that the regular human brain is doing calculations on millions of possible causal combinations every second, effortlessly and nearly instantaneously. On top of this, Equation 2 implies that in order to compute the credence of a particular clusterization, you also have to compute the information gained by combining these cues in the process. It would seem more parsimonious to only compute this information once the cluster has been formed based on heuristics, as opposed to deeming the cluster useful because the information it brings is the best candidate among billions.

Bringing in clustering algorithms as a prior step before the Bayesian parametric combination of cues therefore seems of interest. These come in very different forms but understanding how simple ones work and build up knowledge from there can give good insights as to why they are particularly interesting to the task at hand.

K-means may be the most taught clustering technique thanks to its conceptual simplicity (Jain, 2010). Given a number k of clusters to find in the data on an n-dimensional space, algorithms try to partition the data points in a way that minimizes the total variance within each cluster around its centroid. Intuitively, this means regrouping data points so that they are always the closest possible (in terms of Euclidian distance) to the mean of their own cluster. Once applied to a perceptual context, datapoints would correspond to individual stimuli and each dimension to a sensory cue. A pure tone could for instance be represented as one data point on a 4D space, with perceived X, Y and Z spatial locations and frequency as dimensions. In this example, tones would be assigned to k clusters by proximity, so that tones that were perceived in the same spatial area and a similar frequency are probably created by the same source and should therefore be

clustered together. In the absence of more information, this already feels like a natural decision to make about our surroundings.

However, using Euclidian distances poses several problems, notably ones of scaling and controlling the way distances from centroids are penalized. Should a distance of 1cm in X be penalized the same way as a distance of 1Hz? And more importantly, should a distance of 1 unit on all axis be penalized the same way as a distance of 2 units on any one axis? Probably not, and we would certainly need to be able to control the distance penalization differently for each axis. This *k-means* clustering technique cannot be used for this, but that is one of the reasons why *mixture models* were designed. Mixture models are probabilistic models made to represent subpopulations within a population, each one having their own probability distribution and parameters (McLachlan et al., 2019). In this class of algorithms, data points are partitioned as a function of their likelihood to belong to the same distribution. Using multi-dimensional Gaussian functions, we can very easily model uncertainty in the usual way and choose distance penalization for each dimension by simply choosing (or fitting) the corresponding standard deviation parameter. Even if the computation behind the scenes is very different from *k-means*, the concept is fundamentally equivalent: pure tones on a 4D space would still be clustered together in a way made to minimize a measure of distance from the mean of the *k* clusters (since Gaussians are centred on means). For each dimension, each cluster would have its own standard deviation parameter determining the likelihood that tones belong to this cluster using the z-score. Clusters are chosen to maximise this likelihood on average for every tone (see Figure 1.5 for a graphical representation of a typical Gaussian mixture model clustering).

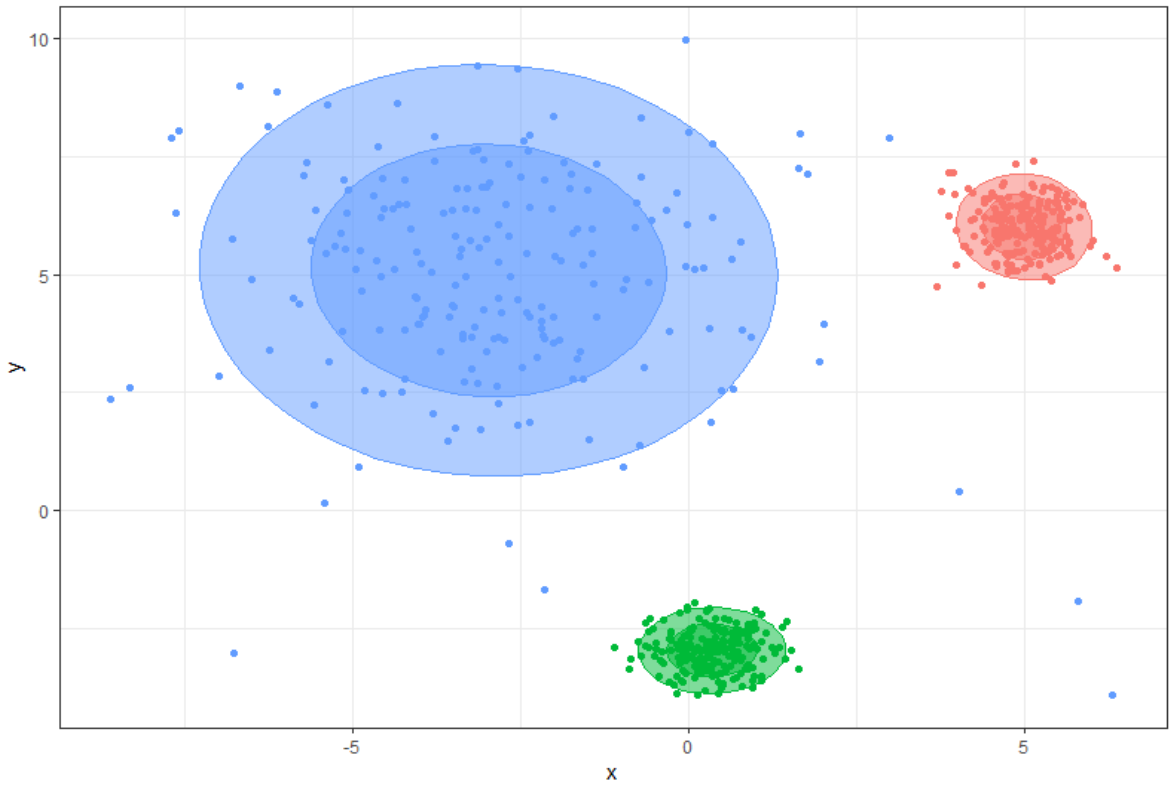


Figure 1.5: Gaussian mixture model clustering example with *k*=3 clusters and *n*=2 dimensions. Outliers belonging to the blue cluster while being closer (in terms of Euclidian distances) to other cluster means illustrate how z-scores now determine the likelihood of belonging to the different clusters.

Although significantly easier to compute than causal inference models presented earlier, these algorithms are still computationally difficult, especially when it comes to finding global optimums on a high number of data points and a high number of dimensions. This is because they usually have to compute a lot (but not all) of possible clustering combinations before converging. However, a deeper reflection into what we are trying to achieve can reveal that this problem is in fact less problematic than can be thought. The world is usually presented to us in a sequential way, especially when it comes to auditory perception. Our memory is limited, and for most estimates, taking the average of dozens of cues observed at different times makes no sense. Let's imagine that two persons are having a discussion. Person A, with eyes closed, is listening to person B, who does not stop talking while walking across a room from point Y to Z. Once B has reached Z, should A cluster every word B has said as being produced by the same source? Obviously, yes. Should A combine the position at which all these words were perceived to infer that person B is now halfway through Y and Z? Obviously not. The only position that

now matters is the last known position of this cluster, and any new position cue should only be compared to the last known positions of every cluster in memory. This, in general, only calls for as many multivariate calculations as there are pre-existing clusters every time a new stimulus is perceived. Hopefully, all these steps seemed logical and justified to the reader, as they are central in the way our model presented in the next chapter is designed.

One last yet major problem remains: the algorithms presented here require providing in advance the *k* number of clusters we want to find in the data. This is where the notion of *non-parametric* Bayesian models come into play. This type of model allows parameters to change with the data, and lets clusters emerge from it (Gershman & Blei, 2012). It assumes an infinite set of latent groups, each one described by a set of parameters, that are slowly uncovered as the sample grows when it is deemed plausible or necessary. In practice, this usually means that not only does the algorithm confront a new stimulus to each pre-existing cluster, but it also has ways to explore the possibility that this new stimulus was created by another, previously unobserved source. There are different methods to achieve this, such as the Chinese restaurant process or the Indian buffet, which should be chosen according to the assumptions made for the model – for instance, if each cue can be included in several clusters at the same time or not (Gelman et al., 2013).

The Chinese restaurant process is often explained through a simple analogy: imagine a restaurant with an infinite number of tables, each having an infinite number of chairs. Each customer represents a datapoint, each (populated) table represents a cluster. The first customer sits at any empty table. The next customer can either sit at the same table or at the next one, and so on, knowing that the probability that the customer chooses to sit at a given table is proportional to the number of people that sat around it: the more people are already around a particular table, the more likely it is that a new customer will pick this one to sit at (a property sometimes called “rich get richer”). This illustrates how a Chinese restaurant process can randomly create clusters without taking into account

attributes of the datapoints. To complete the analogy when considering those attributes, one can imagine that the last person to sit at a table orders food for the next person, based on his own preferences. Every new customer then chooses a table based not only on the number of customers already present on that table, but also based on the proximity of the food that will be served to their own food preference (similar to a Euclidian distance on a plane). The two mechanisms will interact, sometimes in harmony (e.g. a highly populated table with upcoming food that the new customer loves, that is, a cluster with a lot of datapoints whose last datapoint is close to the new datapoint), sometimes in dissonance. The Indian buffet process is quite similar, except that each customer can be present at different tables at once (datapoints can belong to several clusters). This Chinese restaurant process mainly differs from regular centroid clustering using z-transformed scores as presented in Figure 1.5 in that it does not require to set a predetermined number of clusters while its "rich get richer" property ensures that the algorithm will still be parsimonious and not multiply the total number of clusters; and with this particular implementation, the inclusion criterion of a new datapoint is not the proximity to a centroid, but the proximity to the last datapoint (remember from our last example: a new sound should be clustered to others based on the last known position of the source, not the average position of all sounds that the source produced in the past).

We are now armed with the knowledge of clustering algorithm rules that seem to correspond perfectly to the behaviour we wanted it to display: using a Chinese restaurant process, it is possible to cluster a huge number of percepts in a way that is easy to compute when they are presented sequentially, meaningful, and parsimonious.

In fact, researchers have already begun investigating Gestalt perceptual grouping using non-parametric Bayesian statistics. Froyen et al. (2015) used this kind of modelling, which they called Bayesian Hierarchical Grouping, to apply it to investigate the elusive notion of prägnanz in visual perception. Even if their approach does not apply to natural

images yet, their results look really promising as they were able to create very convincing clustering patterns in dot collections.

A quick note on attention

It has been previously stated that Gestalt literature usually considers grouping mechanisms to be a pre-attentive process. It is worth mentioning that, on the other hand, other areas in the literature started with the premise that top-down attention was influencing the stream formation process, or even that this process is purely attentional (Kaya & Elhilali, 2017). Implications differ according to authors, but mostly concern either a simple modulation of the way clusters are formed (Sussman, 2017), or a general constraint on the possible number of streams. Indeed, if streams are dynamically created through attentional processes in order to only perform complex treatments on stimuli when needed, it could be hypothesized that at all times, only a maximum of 2 streams exist: the one we are working on right now, and a *garbage* stream, which is only monitored for salient stimuli to catch the person's attention through bottom-up processes, but within which no grouping happens (Mack et al., 1992). This debate seems to be still going on, with data supporting both views still being gathered. It is therefore important for the sake of ecological validity, to confront major results with various attentional contexts.

Project aims and hypotheses

The Bayesian non-parametric approach still remains to be applied to the problems of multisensory integration. However, doing so immediately would be making the assumption that the auditory system performs inferences the same way as the visual system does. While there are good reasons to think so, this project aims at verifying this assertion by applying non-parametric Bayesian modelling to the auditory system. This could then be used as a first step towards a multisensory (visual-auditory) approach in case of success. Furthermore, while as was mentioned earlier studying auditory perception adds specific challenges to the table, the inherently sequential aspect of audition also presents some

specific advantages. Indeed, it is very difficult in visual perception studies to finely observe possible iterative mechanisms in the grouping process, while auditory perception is a perfect setting to do so thanks to the inherently more sequential nature of the stimuli. This project is therefore also an occasion to dig and hopefully gain a deeper understanding of how perceptual grouping is done across all sensory modalities.

A non-parametric Bayesian model was developed, with the hope of obtaining a convincing model in the future, both in terms of its theoretical justifications and with well-defined and accurate predictions. If such a mathematical model can efficiently predict human responses in perceptual tasks, it becomes possible to reason by analogy and infer that our brains may function in a similar fashion at an algorithmic level. This type of reasoning with meaningful conceptual hypotheses could be a way to bypass limitations in common neuroscientific methodologies (Jonas & Körding, 2017).

The creation of a new predictive model is not the only preoccupation of this project. As was mentioned earlier, Gestalt literature usually considers that stream formation processes are pre-attentive. But former experiments in the field have struggled to design experiments in which the formation of strictly more than 2 simultaneous streams has an important impact on results, or that could not be reinterpreted in terms of only 2 streams. Yet, we have the desire to validate our model in experiments that can be as generalized as possible, while also doing our part in the advancement of the two streams vs. infinite streams debate.

To this end, a new paradigm inspired by Barsz (1988) and Sussman (2017) was created, with melodies comprised of 4 sounds of different frequencies generated specifically to easily favour the emergence of 1, 2, or 3 different perceptual streams, while giving an implicit way to count them. All experiments across all chapters using this new paradigm included a minimum of two conditions in common with the others. Cross-chapters analyses will therefore be conducted to check the validity and consistency of the

paradigm and our results across studies and experiment settings, particularly because some experiments were conducted in a lab while some were online.

The second chapter will give a more mathematically stringent description of our non-parametric Bayesian model, and confront it with well-known results in the Auditory Scene Analysis literature. The model will also be fit to subjects' answers in a recreated classic paradigm. Insights from the model will then be used as predictions to an experiment using the paradigm specifically created for this project. Specifically, this chapter is designed to explore the influence of fast frequency differences on stream segregation.

The third chapter will expand on the first chapter and start generalizing both the model and experimental observations by including several variables at once, and study how their interactions and principal effects are predicted by the model as well as behaviourally observed. The spatial location of tones and timbre will be studied in different experiments, within which frequency differences and a paradigm similar to those present in chapter 2 will still be present.

The fourth chapter will explore the often-overlooked influence of top-down attentional control on stream formation, while still using the same experimental paradigm.

Finally, the last chapter will present the general discussion and conclusions to extract from this research project.

Chapter 2

–

Using 'Occam's Razor' for causal inference of auditory perception

Abstract

Perception relies on being able to segregate stimuli from different objects and causes, in order to perform inference and further processing. For simple binary stimuli we have models of how the human brain can perform such causal inference, but the complexity of the models increases dramatically with more than 2 stimuli. To characterize human perception with more complex stimuli we have developed a Bayesian inference model that allows an unlimited number of stimulus sources to be considered: it is general enough to allow any discrete sequential cues, from any modality. The model uses a non-parametric prior, hence increased complexity of the signal does not necessitate more parameters. The model not only determines the most likely number of sources, but also specifies the source that each signal is associated with. As a test application we show that such a model can explain several phenomena in the auditory stream perception literature, that it gives an excellent fit to experimental data and that it makes novel predictions that we confirm experimentally. These results have implications not just for human auditory temporal perception but for a large range of other perceptual phenomena.

Keywords **–** perception, causal inference, Occam's razor, gestalt psychology, non-parametric, bayesian inference

Introduction

Ambiguity in perceptual systems is a blight for inference. When we hear two sounds sequentially, we may infer that they came from two different sources, e.g. birds A and B, or the same source repeated. A third sound is heard - are the generating sources AAA, AAB, ABA, ABB or ABC (see Figure 2.1 for the possible generative models)? By the time four, five and six sounds are heard the number of possible combinations reaches 15, 52 and 858.

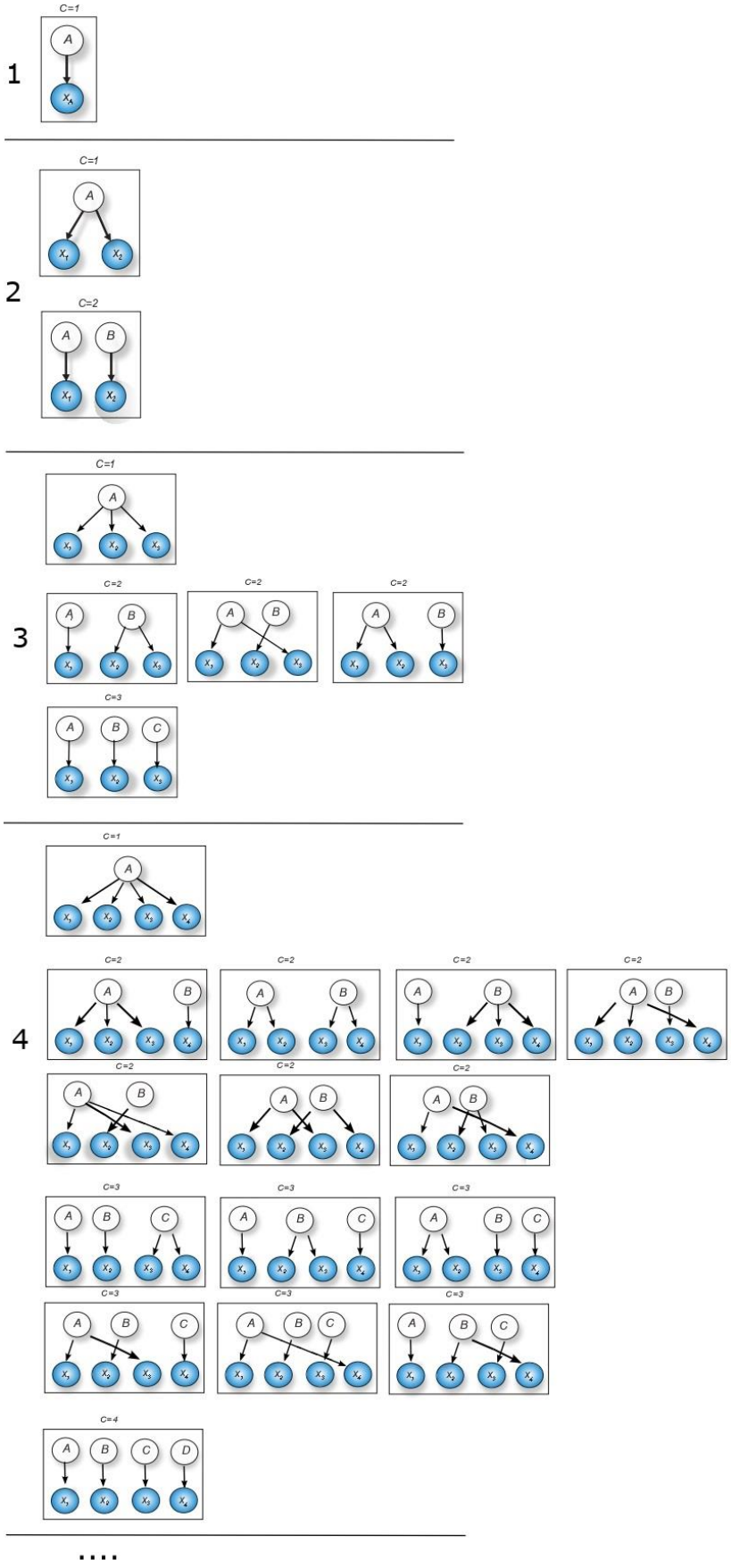


Figure 2.1: As the number of stimuli increases ($X_1, X_2, X_3, X_4, ...$) the number of potential causes(C) increases at the same rate, while the number of combinations of causes that could have generated the stimuli increases according to the number of ways to partition a set of n objects into k non-empty subsets. It is easy to differentiate between the two potential generative structures when there are only two stimuli, but much harder when four stimuli can be created from fifteen different generative structures.

The ambiguity breeds to generate a combinatorial explosion, and yet the human auditory system is able to reliably allocate multiple sources of sound in complex, real world situations. Features of the signal are consistently associated with different sources, allowing us to keep track of a speaker's voice and the wail of an ambulance siren, separate from the noise of background traffic and falling rain.

This is a general problem faced by the perceptual system, inferring the generative model that caused the observed stimuli. To perform such a task for simple stimuli the brain relies on causal inference (Körding et al., 2007; Shams & Beierholm, 2010), probabilistically estimating the most likely cause of the stimuli in the environment. This has been shown to be a good model of perceptual inference for ambiguous stimuli, when they are in small numbers, i.e. two. However, the increase in number of stimuli causes the complexity of possible generative structures to rapidly increase (Figure 2.1), rendering a causal inference strategy that relies on enumerating all possible structures impossible.

An important realisation is however that given a specific set of a large number of stimuli, this process is essentially one of *clustering*, combining together stimuli that are similar while keeping separate from those dissimilar. This idea, that perception involves clustering, has a long history in the Gestalt psychology literature, although not always expressed in those terms (Wagemans et al., 2012). However, the brain would need to choose the right number of clusters, and have a way to specify the prior expectations over clusters, which is hard before even knowing the number of stimuli.

The key proposal in this paper is that the brain can perform this clustering process using a modern version of 'Occam's Razor', *non-parametric Bayesian clustering*. With this approach the numbers of clusters do not have to be pre-specified, instead the algorithm adapts the number of clusters to the data. Likewise, there is no need for a large number of

parameters to specify the prior expectations, instead a single-meta parameter specifies the degree of clustering.

Individual data points (stimuli) are assigned to different groups (or clusters) based on the other existing data points. This type of algorithm is renowned for allowing the complexity of the model to grow with the data set (Aldous, 1985; Ghahramani, 2013) as larger number of clusters can emerge as the number of data points increase.

Non-parametric Bayesian inference has previously been used in cognitive and perceptual studies (Froyen et al., 2015b; Gershman & Niv, 2013), but not to study the segmentation of perceptual cues.

To exemplify how human perception of large number of sources can be modeled by non-parametric Bayesian inference we will present modeling and experimental results on auditory stream segregation.

Auditory stream segregation

For several decades, the human ability to segregate sequential sounds into streams corresponding to sources has been investigated using simple sequences of either pure tones or more complex sounds (review in Moore & Gockel, 2012). The time interval between tones, their pitch difference and the duration of a sequence are among the factors that play an important role (Anstis & Saida, 1985; Bregman & Campbell, 1971; van Noorden, 1975): explanations of how the factors are used based on principles such as Gestalt laws and Occam's razor have been incorporated into the conceptual model of Bregman (Bregman, 1994).

Descriptive models based on peripheral excitation (Beauvois & Meddis, 1997), coherence of coupled oscillators (Wang, 1996) and cortical streaming modules (McCabe & Denham, 1997) provide mechanisms to estimate the number of streams, but do not specify which sound is associated with which source. While some of the models are expandable to

allow more sources to be inferred, it is not known if they would cope with the combinatorial explosion. Furthermore, Moore and Gockel (2012) conclude from an extensive review of the literature that any sufficiently salient factor can induce stream segregation. This indicates that a more general model of inference is needed, that can incorporate any auditory perceptual cue and multiple sounds with different sources.

If ambiguity is a blight for inference, regularities in natural signals are the cure. Not all combinations of signal sources are equally likely – when perceptual systems generate a model of the world, we assume that they infer the most likely interpretation because the perceptual systems are optimized to the statistics of natural signals (Barlow, 1961) (McDermott & Simoncelli, 2011). Bayesian inference has had considerable success in modeling many visual and multi-sensory percepts as a generative, probabilistic process (Beierholm, 2013; Shams et al., 2005; Weiss et al., 2002). Despite these successes, we still have no general, principled model of how the auditory system solves the source inference problem.

A Bayesian approach to auditory stream segregation (based on sequential sampling) has been used to model the dynamics of perceptual bistability (Lee & Habibi, 2009) but assumes that only two percepts are possible. Turner (2010) has developed methods of analyzing statistics of sounds based on Bayesian inference, and constructed a model to synthesize realistic auditory textures. While inference in the model can qualitatively replicate many known auditory grouping rules, the expected number of sources in the environment has to be specified.

In our model the probability of many alternative stream configurations (given the input signal) are calculated and the percept generated corresponds to the most probable configuration. The probabilities are calculated using Bayes' rule to combine the likelihood of generating a signal given a postulated stream configuration, with the prior probability of

sounds being associated with different sources. The likelihood and prior probability distributions are iteratively updated in a principled manner as information accumulates.

The forms of the distributions are presumably optimized to natural signal statistics: the likelihood distribution we use is based on considerations of the physical limitations of oscillators. However, the framework of the model allows formulations of multiple explanatory factors, such as those determined by Bregman (1994) from psychophysics experiments, to be simply incorporated in the distributions. Furthermore, while the current study uses simple pure tones (replicating work by Bregman), the framework allows more complex cues from audition and other modalities to be used as long as their perceptual difference can be quantified.

Model

Pure tones are the indivisible atoms of input to the model – each being assigned to just one sound source, or stream. Inspired by work done on non-parametric priors (Froyen et al., 2015b; Orbanz & Teh, 2010; Wood et al., 2006) we assume the existence of an infinite number of potential sources, leading to a sequence of tones with pitch $f_1, f_2, \ldots$,

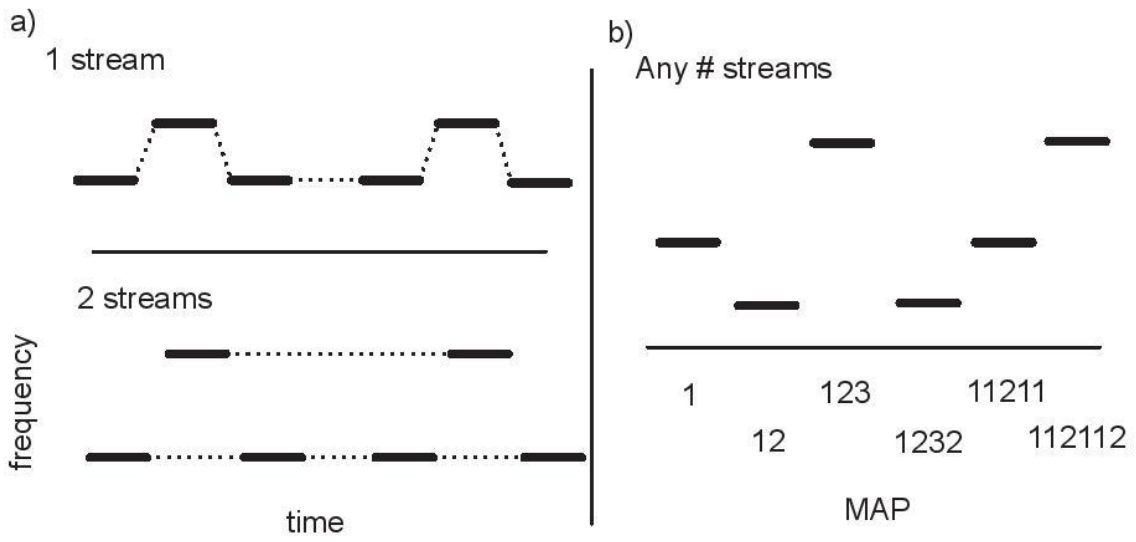


Figure 2.2: Example of stimuli being segregated into one or two streams, using 'galloping' stimuli similar to van Noorden (1975). b) Example of a series of potential stimuli with a representative model Maximum a Posteriori (MAP) assignment of tones to streams below. As each tone is presented the model reassigns the entire set of tones to streams (1 → 12 → 123 etc...).

onset time $t_1^{on}, t_2^{on}, \ldots$ and an offset time, $t_1^{off}, t_2^{off}, \ldots$ and the sound sources/streams that generated the tones are denoted by positive integers $S_1, S_2, \ldots$ We rename the sources when necessary so that the first tone heard will always be generated by source 1 (i.e. $S_1 = 1$), and a subsequent tone, S_n can be associated with source 1: $max(S_1 \ldots S_{n-1}) + 1$.

Generative model

Given a source S_i we assume that the frequency of tone i is governed by physical constraints and statistical regularities of the source. If two sequential sounds with frequencies f_1 and f_2 are produced by the same source, the pitch cannot change at an infinitely fast rate: to make an oscillator change its frequency discontinuously would require an infinite impulse of energy. We assume that, all things being equal, a pure tone sound source is most likely to continue oscillating at the same frequency as it has in the past, and the probability decreases with $\Delta f = f_1 - f_{t-1}$ but increases with $\Delta t = (t_1^{on} - t_{t-1}^{off})$. More specifically we assume a normal probability distribution:

$$p(f_i, t_i^{on} | S_i = S_{i-1}, f_{i-1}, f_{i-1}^{off}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{\left(\frac{\Delta f}{\Delta t}\right)^2}{2\sigma^2}}$$

where σ is a constant. We here assume that the observer has a perfect noise free access to the generated fundamental auditory frequencies. Harmonics in non-pure tones would be considered a unique separate cue.

For successive sources, we assume that sources that have been active previously are more likely to be active again, but do not provide a limit to the number of sources that N tones can be generated from. Concretely we assign the probability of a source i generating the N^{th} tone according to a Chinese restaurant process (CRP; Aldous, 1985), which can be considered as an extension of Occam's rule:

$$p(S_N = i | S_1 \ldots S_{N-1}) = \frac{n_i}{N - 1 + \alpha}$$

If $n_i = \sum_{1:N-1} \delta(S_N - S_i) > 0$, but

$$p(S_N = i | S_1 \dots S_{N-1}) = \frac{\alpha}{N - 1 + \alpha}$$

If $\sum_{1:N-1} \delta(S_N - S_i) = 0$

and where δ is the discrete Kronecker delta function ($\delta(0) = 1$, but 0 elsewhere). α is a parameter that influences the probability of a new source: the lower it is, the lower it is that the new tone comes from a new source.

Inference

The task of the observer is to infer the sources generating each of the tones, i.e. to find the $S_1 S_2 S_3$ ... that maximize $p(S_1 S_2 S_3 | f_1 f_2 f_3 \dots, t_1^{on} t_2^{on} t_3^{on} \dots, t_1^{off} t_2^{off} t_3^{off} \dots)$, as illustrated in Figure 2.2. For simplicity of writing we will refer to the properties of tone i as x_i in place of the set $(f_i, t_i^{on}, t_i^{off})$.

As an example, we use a sequence of three tones x_1, x_2, x_3, for which the observer wishes to infer the likely sources S_1, S_2, S_3. Thus the probability $p(S_1, S_2, S_3 | x_1, x_2, x_3)$ that a sequence of three tones was generated by sources S_1, S_2, S_3, has to be calculated over the five possible combinations: $[S_1 = 1, S_2 = 1, S_3 = 1]$, $[S_1 = 1, S_2 = 1, S_3 = 2]$, $[S_1 = 1, S_2 = 2, S_3 = 1]$, $[S_1 = 1, S_2 = 2, S_3 = 2]$, $[S_1 = 1, S_2 = 2, S_3 = 3]$ corresponding to the five unique configurations of sources generating three sounds. Note that the first source is always assigned the value 1, the next different source is assigned 2, etc.

Bayes' rule relates each conditional probability (the posterior distribution) to the likelihood $p(x_1, x_2, x_3 | S_1, S_2, S_3)$ of each configuration of sound sources generating the sequence of tones, by

$$p(S_1, S_2, S_3|x_1, x_2, x_3) = \frac{p(x_1, x_2, x_3|S_1, S_2, S_3)p(S_1, S_2, S_3)}{Z}$$

where Z is a normalization constant, and $p(S_1, S_2, S_3)$ is the prior probability of the particular configuration of sound sources, regardless of the frequency, etc… of the tones.

Assuming conditional independence of the tones and tone-source causality, this can be rewritten as $p(S_1, S_2, S_3|x_1, x_2, x_3) = \frac{p(x_3|S_1, S_2, S_3)}{p(x_3)} p(S_1, S_2, S_3|x_1 x_2)$

$$= \frac{p(x_3|S_1, S_2, S_3)}{p(x_3)} p(S_3|S_1, S_2) p(S_1, S_2|x_1 x_2)$$

The final term is the posterior generated from the first two tones. The latter two terms can be considered together as the prior for the third source, allowing us to use an iterative approach to the inference. After each tone we grow the tree of possible source sequence (e.g. $11 \rightarrow 111$ and 112), by multiplying the previous posterior $p(S_1, S_2|x_1, x_2)$ with two terms; the likelihood $p(x_3|S_1, S_2, S_3)$ and a prior for how likely the next 'branch' is, $p(S_3|S_1, S_2)$.

We now consider how to determine the likelihood and prior probabilities. The first source can only be associated with one source, so $p(S_1 = 1) = 1$. The principle of Occam's razor would suggest that $p(S_1 = 1, S_2 = 1) > p(S_1 = 1, S_2 = 2)$, i.e., if we have not heard any of the sounds, the most probable acoustic scene is the simplest one: all sounds come from the same source. The value of $p(S_1 = 1, S_2 = 1)$ for an individual can be determined from fitting their data, and the value $p(S_1 = 1, S_2 = 2)$ is simply $1 - p(S_1 = 1, S_2 = 1)$. The values may depend on factors such as the environment, which are not considered in the model: natural signal statistics may provide guidance for how the prior probabilities are assigned.

Regarding the likelihood function, the observer assumes the generative probability $p(x_n|S_n, x_{n-i}, S_n = S_{n-i})$, where tone $n - i$ was the latest tone inside stream S_n, and i is

therefore the number of tones that have been played since the latest tone inside stream S_n. Note that this applies even when the sounds generated by the same source are separated by one or more sounds associated with different sources (e.g. $(S_1 = 1, S_2 = 2, S_3 = 1)$). The only transition that matters is that between the most recent tone and the last tone in the same stream, so if three tones x_1, x_2 and x_3 had all been associated with the same stream (e.g. $(S_1 = 1, S_2 = 1, S_3 = 1)$), we would only consider the transition from x_2 to x_3, whereas if x_2 was associated with a different stream (e.g. $(S_1 = 1, S_2 = 2, S_3 = 1)$), we would only consider the transition from x_1 to x_3.

If a sound comes from a new source, then we assume that the likelihood is independent of previous tones:

$$p(f_n | S_1 \neq S_n, S_2 \neq S_n, \dots S_{n-1} \neq S_n) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{(f_n - f_0)^2}{2\sigma^2}}$$

where f_0 is the midpoint of the range of auditory frequencies presented for the trial.

The resulting model has two parameters, α and σ, plus a parameter for the steepness of the response variability (given by softmax function) for each subject β. α is a parameter that determines how likely it is *a priori* that a newly heard sound is coming from an as-of-yet unobserved source (i.e. creating a new cluster). It is sometimes referred to as a concentration parameter. σ is the dispersion parameter of the likelihood of a newly heard tone of being in each cluster, centred around the frequency of the last tone in the said cluster (or the midpoint of the range of auditory frequencies presented for the trial in the case of a new cluster). Both α and σ are fitted to each subject individually.

For details of implementation of the model see Methods.

Results

In order to evaluate the performance of the model we made qualitative comparisons to studies in the literature and quantitative comparison with experimental data. We

furthermore tested a qualitative prediction based on the experimenters' knowledge of the model, using a novel experimental paradigm.

Modeling example - Time

A well-known basic stream segregation phenomenon (e.g. Bregman & Campbell, 1971) shows that increasing the speed at which auditory tones increases the probability that tones are perceived as coming from separate streams. To examine this in the model we recreate the second experiment of Bregman & Campbell (1971), showing in Figure 2.3 that while a sequence of six slowly presented tones is assigned a low posterior probability of originating from different sources (and should therefore be assigned the same stream), as

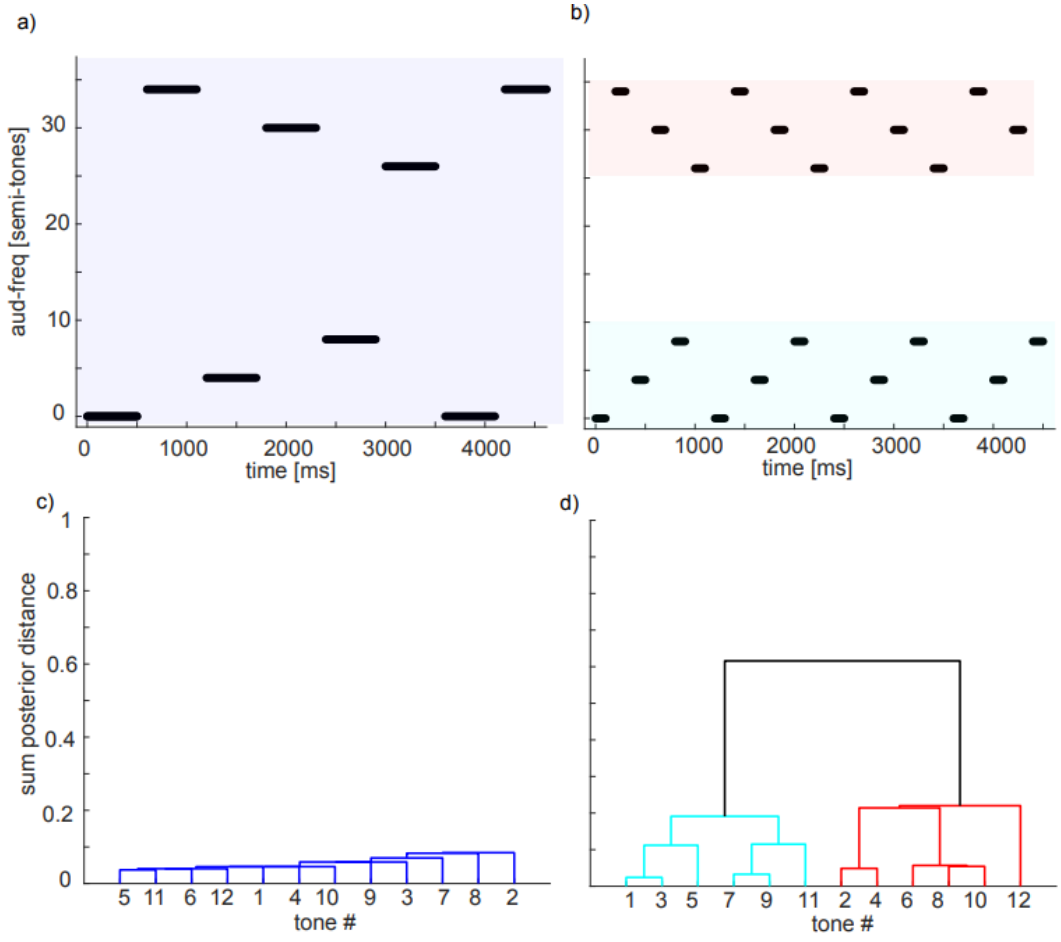


Figure 2.3: Stimuli used in experiments from Bregman & Campbell (1971, second experiment), highlighting how the speed of presentation affects perception of streams of tones. Stimuli are shown at the top, bottom are dendrogram tree-plots based on the posterior distribution over clustering. A unique colour is assigned to clusters with more than 50 percent distance from other clusters. a) Slow sequence, ISI 100 ms, tone duration 500 ms, pitch difference [0 4 8 26 30 34] semi-tones, tone sequence repeated twice. The posterior mode (the sequence combination with the highest posterior probability) was 111111, i.e. all tones assigned to the same stream. b) Fast sequence, ISI 100ms, tone duration 100ms (posterior mode 121212). c-d) show the modelled sum posterior of the two sequences. Parameters for this figure (and subsequent figures) were $\alpha = 1.44$, $\sigma = 40$.

the speed of presentation is increased the probability of originating from two sources increases drastically (implying subjects should segregate the streams).

Modeling example - Galloping

Several studies have shown how the effect of frequency and time can interact. Van Norden (1975) found that in a repeating Low-High-Low sequence of tones the subjects would report one or two streams as a function of both the difference in auditory frequency and the speed of presentation (interstimulus interval). In our computational model the likelihood term directly depends on both of these factors, and the prior probability again constrains the observer from segregating tones into more streams. We replicate the results of this study in Experiment 1 below.

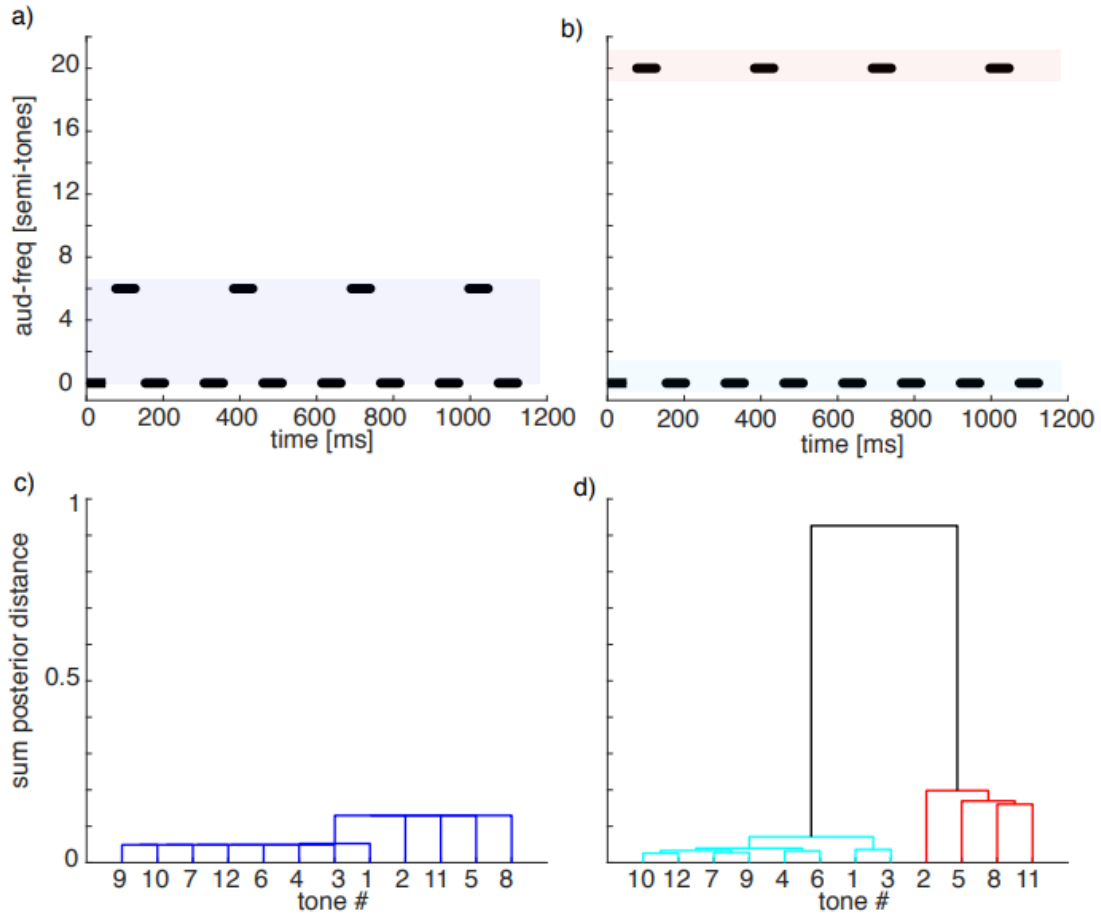


Figure 2.4: Example of a galloping stream, from van Noorden (1975), a) ISI 26.6ms, pitch difference 6 semi-tones (posterior mode 111) b) ISI 26.6ms, pitch difference 6 semi-tones (posterior mode 121). c-d) show the modelled sum posterior distance of the two streams. Modelling parameters were the same as in Figure 2.3.

Modeling example - Cumulative

The galloping sequence Low-High-Low has also been used to highlight the effect of the accumulation of information. Bregman (1978) showed that a short sequence of tones tends to lead to the percept of a single stream, whereas the accumulation of information causes the tones to segregate into two streams. For the model this effect is due to the non-parametric prior initially assigning a low probability to the possibility of two streams, before more information is gathered.

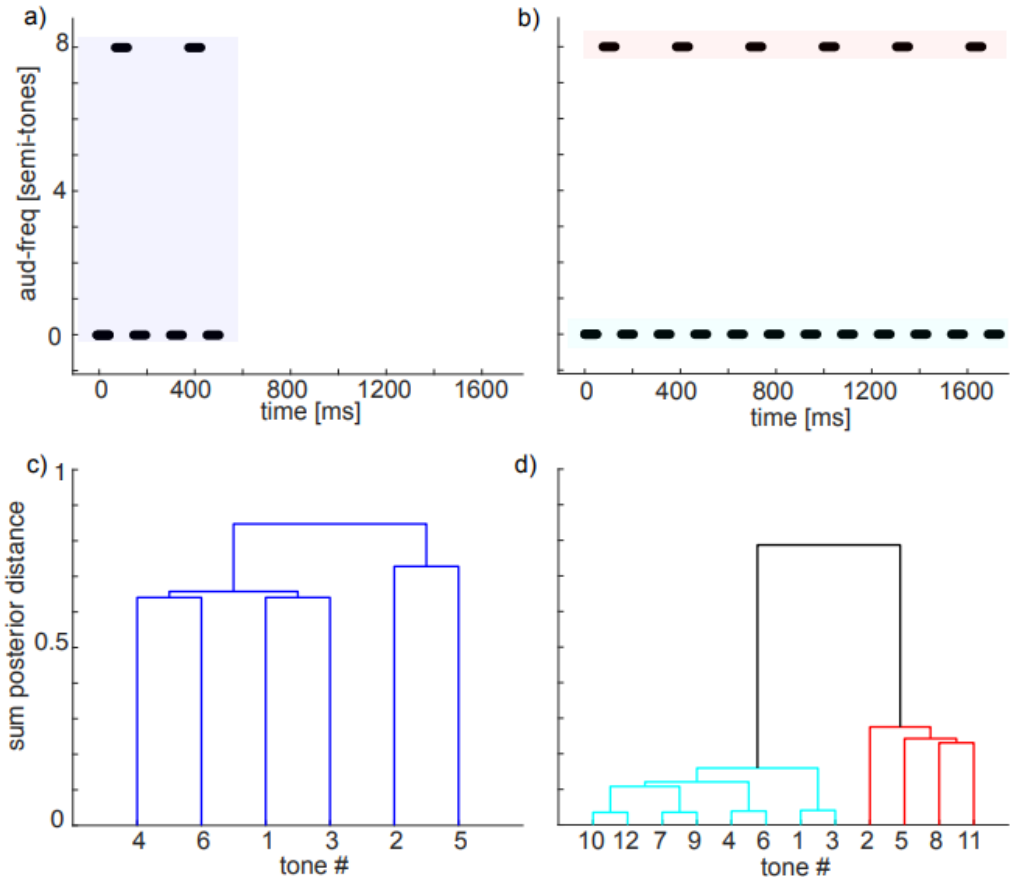


Figure 2.5: Stimuli used in experiments from Bregman (1978), highlighting the cumulative effect of tones. Stimuli are shown at the top, bottom is dendrogram tree-plots based on the posterior distribution over clustering. A unique colour is assigned to clusters with more than 50 percent distance from other clusters. a) Short sequence ISI 26.6ms, pitch difference 7 semi-tones, tone sequence repeated twice (posterior mode 111). b) Long sequence ISI 26.6ms, pitch difference 7 semi-tones, tone sequence repeated eight times (posterior mode 121). c-d) show the modelled sum posterior of the two sequences. Modeling parameters were the same as in Figure 2.3.

Modeling example - Context

An aspect of auditory perception that especially received attention from the Gestalt movement, was the role of context in auditory clustering. Experiments done by Bregman (1978) showed that modifying the context in which tones were presented modified the segregation of unmodified tones. Figure 2.6 shows an example, based on Bregman (1978), where two low tones will be clustered together while two distractor tones are far off in frequency, but will be clustered separately as the distractor tones are placed around them. The model replicates this phenomenon, showing a separation of the first two tones in Figure 2.6b.

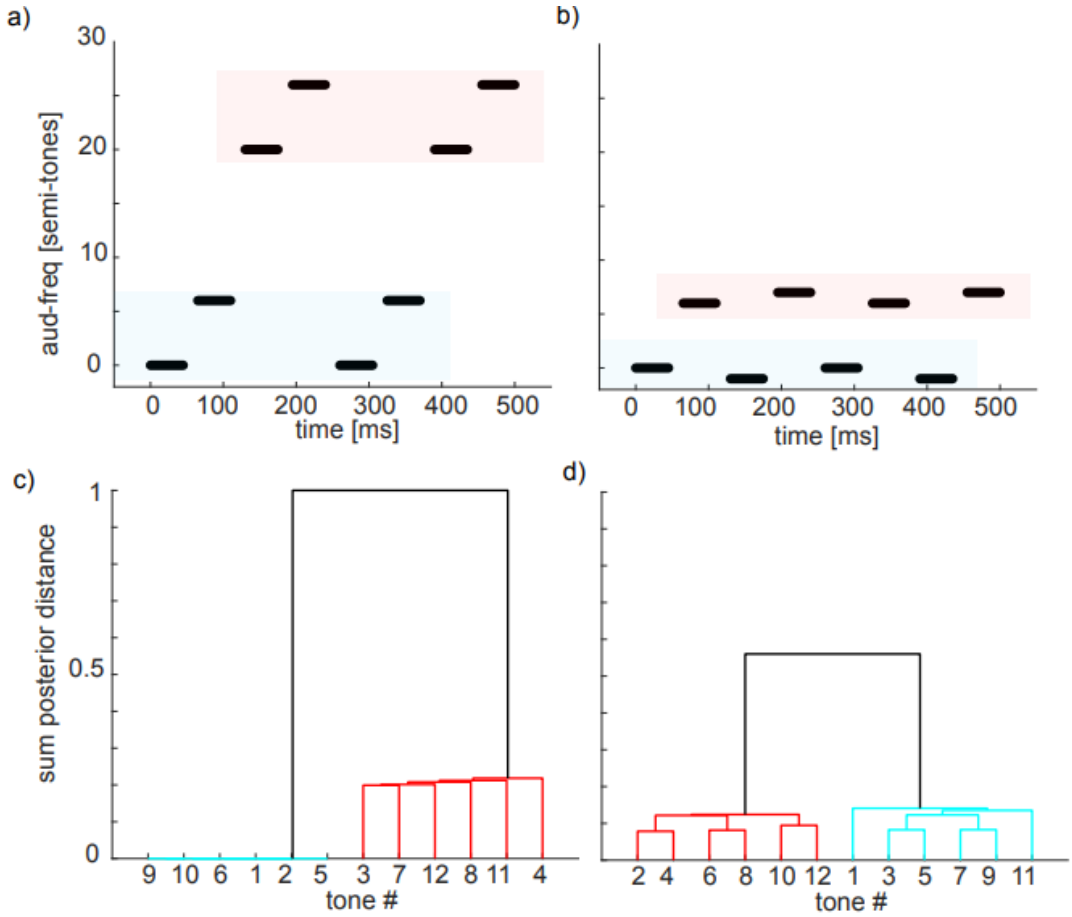


Figure 2.6: Context matters for the clustering of tones. Stimuli are shown at the top, at the bottom are dendrogram tree-plots based on the posterior distribution over clustering. A unique colour is assigned to clusters with more than 50 percent distance from other clusters. a) Two low tones , two high tones, leading to low tones segregated from high tones (posterior mode 1122) b) While the two low tones have been kept constant, the context of the two other tones now causes them to be clustered separately with the other tones (posterior mode 1212). Long sequence ISI 26.6ms, tone sequence repeated eight times. c-d) show the modelled sum posterior of the two sequences. Modelling parameters were the same as in Figure 2.3.

Modeling example - Crossing

As shown by Tougas and Bregman (1985) interleaving a decreasing and increasing series of tones gives the illusory percept of the two streams 'bouncing' i.e. the lower set of tones are clustered and segregated from the higher set of tones. Figure 2.7 recreates this experiment with 2 x 10 interleaved decreasing and increasing tones. The model recreates the perceptual phenomenon, with the lower frequency tones grouped together, separate from the higher frequency tones, thus implying a perceived 'bounce'.

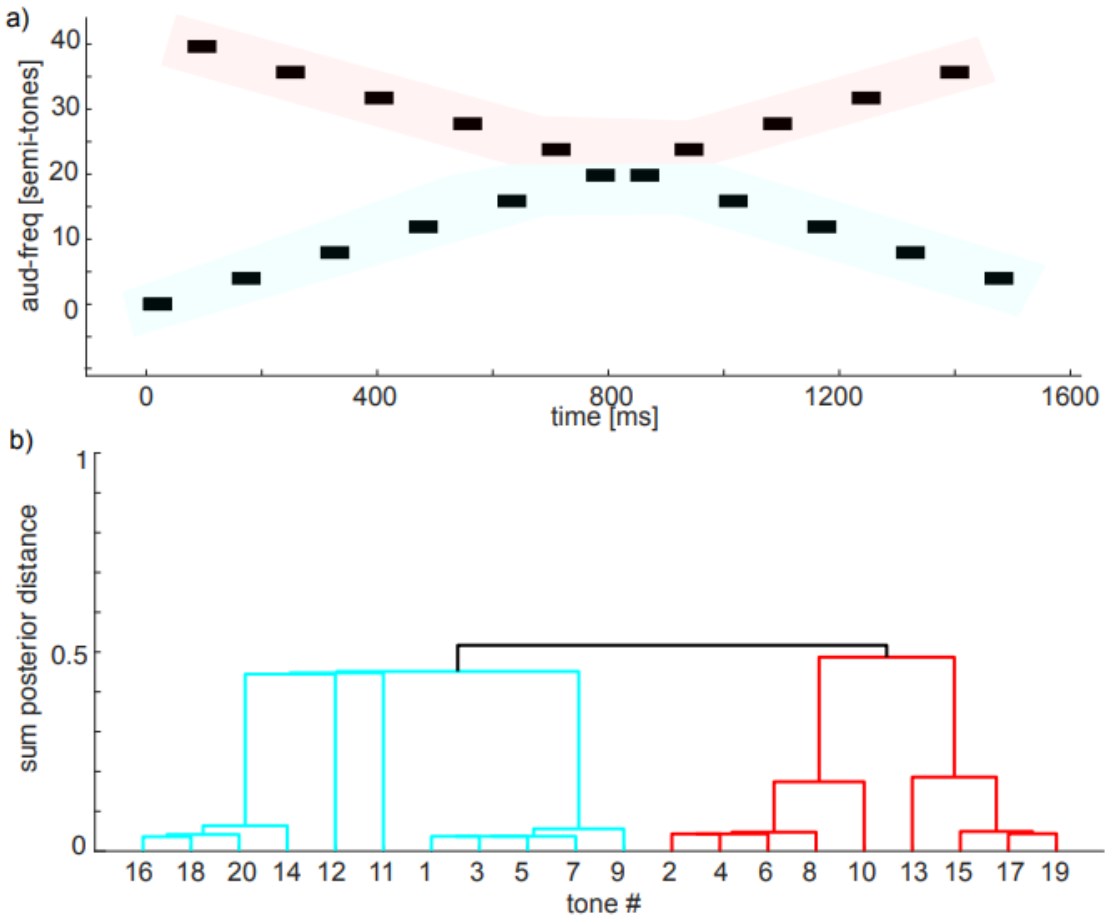


Figure 2.7: a) Interleaved increasing and decreasing series of tones, ISI 26.6ms. b) show the modelled sum posterior. Modelling parameters were the same as in Figure 2.3.

Overall, the model is able to recreate several phenomena in the experimental literature.

Experiment 1

To quantitatively compare the model to human performance we conducted a psychophysics experiment, in which fifteen participants with normal hearing listened to simple auditory sequences and performed a subjective judgement task (a variant of the galloping experiment by van Noorden, 1975). Given a series of Low-High-Low sequence of tones, subjects would respond whether they perceived one or two streams. Across trials the separation between low and high tones, and the inter-stimulus-interval, were varied (see Methods for details).

Model performance and comparison

As an example, response data from six subjects is shown in Figure 2.8. As expected, when the ISI was short, or when the difference in frequencies was large, subjects were more likely to report two streams than one. Figure 2.8 also shows the model fit for one participant.

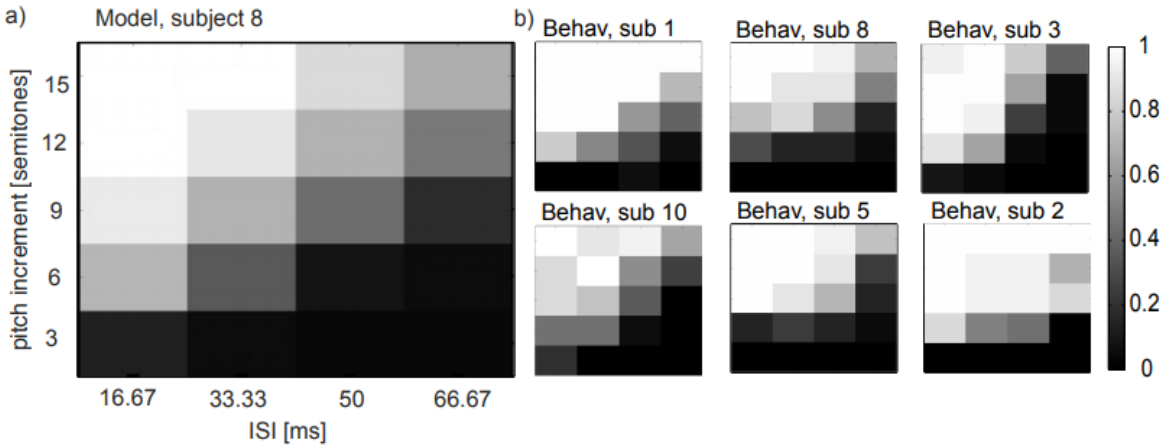


Figure 2.8: a) Model fit, based on fitted parameters from a typical subject, giving the fraction of trials in which the participant responded '2' for the number of streams perceived: 1 (perfectly white rectangle) means the number of streams perceived was reported to be '2' in all trials. Axes give the pitch difference for the middle tone and the inter stimulus interval (ISI): the time between the offset of one tone and the onset of the next. b) The behavioural results from 6 subjects.

While visually the model approximates the subject behaviour we used model-comparison to rule out other hypotheses.

The model with the non-parametric prior was compared against three alternatives that used different priors to constrain the number of possible streams:

A. When the stream combination comprised only one stream (repeated), the prior probability of the next stream being 1 or 2 was allocated according to the CRP, but if the combination already contained two streams, the prior probability of allocating stream 1 or 2 was simply the fraction of previous tones that were allocated to stream 1 or 2 respectively.

B. The prior probabilities of a new tone being allocated to stream 1 or stream 2 was given by $P1$, and $1 - P1$ respectively, where $P1$ is a fixed parameter.

C. The prior probabilities of a new tone being allocated to stream 1 or 2 were fixed at 0.5.

Because alternative model C has only one free parameter (all others have two), we use the Bayesian Information Criterion ($BIC = -2\log P(resp|tones) + k * \log(n)$, where k is the number of parameters and n is the number of data points fitted over) to compare model performance in Figure 2.9. The BIC is a measure of efficiency of a parameterized model in predicting the data. As adding more parameters to a model generally improves prediction whatever their effective usefulness and can lead to

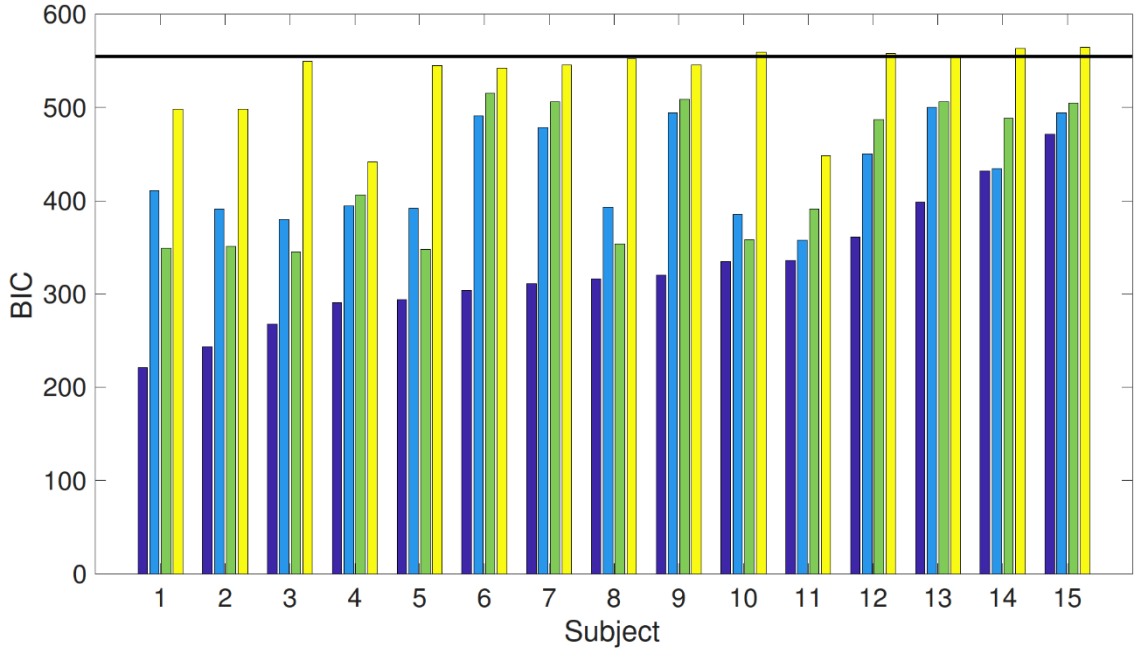


Figure 2.9: Model performance on experiment 1 in terms of Bayes Information Criterion (BIC) for each subject with the CRP model (dark blue), alternative A (light blue), alternative B (green), and alternative C (yellow). The black line indicates the performance of a purely random model that assigns 0.5 probability to either response for every condition. Subjects are ordered based on CRP model BIC values

overfitting, the BIC also introduces a penalty for each parameter added, to ensure that the gain in performance of the model is worth the increased complexity. A lower BIC is preferred.

The Bayesian model with the non-parametric process prior gives a better fit (smaller BIC) than all the alternatives considered. The mean $\pm SEM$ of the optimised parameters for the unconstrained model are $\alpha = 3.01 \pm 0.25$ (equivalently $P(11) = 0.273 \pm 0.030$) and $\sigma = 123.8 \pm 6.2$ [semitones/sec].

Experiment 2

While the model above theoretically allows an unlimited number of tones to be segregated into an unrestricted number of streams, Experiment 1 (presented above) only allows a repeated sequence of 3 tones to be separated into 1 or 2 streams. However, the model predicts that subjects should generally segregate based on frequency and temporal distances between tones with a possibly infinite number of streams. To test this further we performed a novel follow-up experiment in a broader auditory environment. The experimental setting was inspired by Barsz (1988) and was specifically designed to allow for a 3-streams situation to emerge, and to replace the explicit measure of stream segregation by an implicit one (see Methods for details).

Participants were asked to judge if two consecutive melodies comprised of 4 repeated tones were similar or different. In every condition, 1 tone was considered being in the low frequency range, 2 tones in the medium range, and 1 tone in the high range. According to previous experiments showing that order information is intact when tones are inside the same stream but lost when segregated into different streams (Barsz, 1988; Bregman & Campbell, 1971; Demany, 1982), if medium tones were to be part of a stream excluding low and high tones, participants should be unable to detect the difference between two sequences if only medium tones were inverted. Thus, subjects should only be

able to detect the inversion of the middle tones if the tones are placed in the same stream as the upper or lower stream (hence all tones clustered into 1 or 2 groups).

Conditions were created with varying discrepancies between frequency ranges (ISIs were kept constant for this experiment). See Figure 2.11 for a schematic representation of a typical stimulus with inversion.

As our model predicts that the probability of being assigned to different streams is dependent on the frequency difference, our first prediction is that participants should have a reasonable degree of performance in detecting a difference between two melodies when medium tones are inverted and there is only a small frequency difference between those medium tones and at least one of the high and low tones. This would reflect the fact that medium tones are in a stream also including other tones. Conversely, our other prediction is that as the frequency difference increases, participants should perform significantly worse at detecting the medium tones inversion. This would reflect the fact that medium tones are in a stream not including other tones.

Analysis preparation

Individual responses on perceived differences between sequences were transformed into D-prime scores to obtain a single measure of signal detection for each pair of frequency differences, while taking into account possible response biases. Two participants with a negative D-prime score on the easiest condition (*3-3*) were considered unable to perform the task correctly and were therefore excluded from further analysis, leaving a total of 24 participants. No data was missing in the dataset. A one-way repeated measures ANOVA was conducted on these D-prime scores, with FREQUENCY DIFFERENCES (*3-3* vs. *3-9* vs. *3-15* vs. *9-9* vs. *9-15* vs. *15-15*) as the only within-subject factor, along with seven subsequent paired-sample t-tests in line with our hypotheses. No correction for multiple comparisons has been applied.

Mauchly's test indicated that the assumption of sphericity had not been violated [$\chi^2(14)=0.355$, $p=.084$].

Data analysis

Although inferential statistical tests were conducted on D-prime scores only, Table 2.1 also includes summarizing statistics of proportions of "similar" responses in every experimental condition.

	3-3	3-9	3-15	9-9	9-15	15-15
Different stimuli presented	0.2708 ± 0.2089	0.2882 ± 0.2918	0.2396 ± 0.2581	0.4861 ± 0.2526	0.4965 ± 0.2839	0.5382 ± 0.3166
Similar stimuli presented	0.7986 ± 0.2304	0.7708 ± 0.2398	0.8333 ± 0.1966	0.7708 ± 0.2015	0.8542 ± 0.1985	0.8958 ± 0.1759
D-prime	1.4764 ± 0.7647	1.3706 ± 1.0161	1.7129 ± 0.8627	0.7483 ± 0.8433	0.9759 ± 0.7765	0.9604 ± 0.9401

Table 2.1: Mean proportions of "similar" responses and mean D-prime scores across all conditions. Reported errors are ± 1 standard deviation.

The one-way repeated measures ANOVA (FREQUENCY DIFFERENCES) showed that D-prime scores differed as a function of FREQUENCY DIFFERENCES [$F(5,115)=6.659$, $p<.001$, $\eta_p^2=0.225$].

Seven subsequent paired-samples t-tests were conducted to decompose the main effect of FREQUENCY DIFFERENCES over D-prime scores. Three one-tailed paired-samples t-tests revealed that D-prime scores were significantly higher in the *3-3* condition than in the *9-9* condition [$t(23)=3.003$, $p=.003$, $d_z=0.613$], in the *9-15* condition [$t(23)=2.556$, $p=.009$, $d_z=0.522$] and in the *15-15* condition [$t(23)=2.535$, $p=0.009$, $d_z=0.517$]. Two two-tailed paired-samples t-tests revealed no significant difference between the *3-3* and the *3-9* conditions [$t(23)=0.681$, $p=.503$, $d_z=0.139$] and between the *3-3* and the *3-15* conditions [$t(23)=-1.568$, $p=.131$, $d_z=-0.32$]. Another two two-tailed paired-samples t-tests revealed no significant difference between the *9-9* and the *9-15* conditions [$t(23)=-1.02$, $p=.318$,

d_z=-0.208] and between the *9-9* and the *15-15* conditions [t(23)=-0.949, p=.353, d_z=-0.194]. These results are summarized in Figure 2.10.

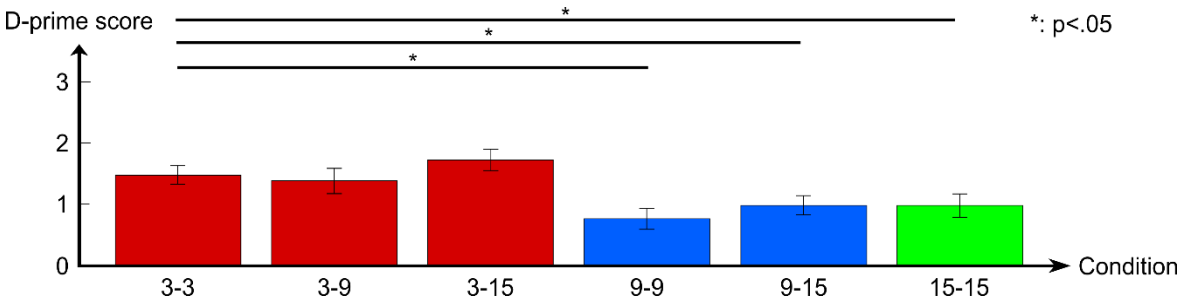


Figure 2.10: D-prime scores as a function of frequency difference. Red bars indicate conditions with a small minimum frequency difference, blue bars indicate conditions with an intermediate minimum frequency difference and green bars indicate conditions with a large minimum frequency difference. Error bars are ± 1 standard error

Discussion

We have presented a simple Bayesian perceptual model that is able to assign stimuli to an unrestricted number of sources, through clustering of stimuli. We have applied the model to the specific case of auditory stream segregation, an area where Gestalt psychology has long emphasized the need for grouping.

Utilizing a non-parametric Bayesian prior the model iteratively updates the posterior distribution over the assigned group of each stimuli and provides an excellent description of the perceptual interpretation of simple auditory sequences in human observers.

With just two parameters, the model gives a good account of the basic characteristics of auditory stream segregation – the variation in the probability of perceiving a single sound source as a function of the repetition rate and pitch difference of the sounds. The basic model (with a softmax decision function) gave a better fit to the data than alternative models that were constrained to interpret the sounds as being produced from just one or two streams. Qualitative predictions from the model were also in accordance with results from a novel experiment with larger number of tones (Exp. 2).

Importantly the model goes beyond giving just the number of sources, but says which sounds are produced by each source. While the combinatorial space of the posterior distribution in experiment 1 was collapsed to give a marginal distribution in a continuous 1-d response space (leading to an estimate of response probability), the maximum a posterior (MAP) for all participants was always located at either 111-111... or 121-121..., depending on the stimulus condition (Figure 2.4). This is reassuring as it is consistent with the anecdotal evidence that participants always perceive either a galloping rhythm (streams 111-111...) or a high-pitch and a low pitch stream (121-121...), i.e. the percept is always at the MAP. Indeed, the percept cannot in general be at the mean because the space of possible percepts is discrete: there is no percept in between 111 and 121.

One consequence of the inference model that is not addressed by mechanistic models of stream segregation is that when a percept changes from say 111-111 to 121-121, the source allocation of previous sounds is changed. Ironically, this 'non-causal' effect is essentially a feature of causal inference – when an observer decides that the percept has changed to 121-121, this is based on previous evidence, and yet at the time that the previous tones were heard, they were all associated with one source. A similar effect is commonly encountered when mis-interpreted speech (perhaps mis-heard due to background noise) suddenly makes sense when an essential word is heard – the previous words are reinterpreted, similar to the letters in predictive text message systems.

The framework of the model is very general, and allows for the incorporation of other factors into the likelihood to describe other aspects of auditory stream segregation. Adding terms in the likelihood function may be able to explain other effects seen in the literature, such as segregation based on bandwidth (Cusack & Roberts, 2000), or build-up and resetting of segregation (Roberts et al., 2008). Furthermore, in the current study we assume that there is no ambiguity in the percept of the pure tones, the uncertainty arises from lack of knowledge about the underlying generative structure of the data. In a realistic

situation perceptual ambiguity would have to be taken into account using an approach such as suggested by Turner and Sahani (2011). Nevertheless, we should emphasize that even though we are dealing with a Markov property (each tone within a stream only depends on the previous tone within the stream), the mixture of streams makes the problem very different from work on e.g. Hidden Markov Models (or even Infinite Hidden Markov Models) for which the goal would be to infer underlying states despite perceptual ambiguity. Note also that while there are algorithms developed to separate audio signals (e.g. Roweis, 2001), these are not meant to mimic human perception, although a future comparison would certainly be very interesting.

In the current implementation we had to make numerical approximations in order to handle the complexity of the model. As an alternative to calculating our results analytically we could use Monte Carlo techniques (e.g. Markov Chain Monte Carlo sampling, i.e. MCMC, a different type of approximation), which have become a standard tool for solving complex statistical models. While not presented here, a MCMC version of the model has also been implemented with similar results.

The proposed model of auditory stream segregation is a specific instantiation of an iterative probabilistic approach towards inference of perceptual information. A major issue for this approach is the problem of dealing with multiple sources, as represented by the work done on causal inference (Körding et al., 2007; Shams & Beierholm, 2010). Until now models of causal inference have been unable to handle more than two sources, due to the escalating number of parameters needed for parametric priors. The use of a non-parametric prior allows a complex of many stimuli to be interpreted without running into this problem, potentially allowing for an arbitrary number of causes in the world. This approach is very general - it can be applied to any set of discrete sequential cues involving multiple sources - and it gives a simple, principled way to incorporate natural signal constraints into the generative model.

However, our generalizability argument is still mainly theoretical, as for now we have only applied the model to one variable of interest in the second experiment. Despite frequency certainly being the most studied dimension in the Auditory Scene Analysis literature, it is still necessary to investigate predictions made about other perceptual cues. This would also allow us to increase the ecological validity of the model by considering multi-cue situations and see how their interactions modulate the way we cluster auditory information both in terms of the model and from a behavioural point of view. Follow-up experiments of this type are clearly called for as the paradigm used in our second experiment can easily incorporate a wide range of variables.

The auditory streaming model also suffers from a lack of direct consideration to attentional mechanisms. As it is, it assumes that our perceptual system automatically groups or segregates cues into a potentially infinite number of streams based on stimuli characteristics, and that attention only has the role of selecting one or several clusters of interest in a given situation, or is being driven by the posterior probability of the different clusters. In either case, attentional processes would be entirely independent of the mechanisms simulated by the model. Nevertheless, it has already been proposed that attention could interact with the cluster formation process itself and therefore have an effect on their overall configuration (Sussman, 2017). If that is the case, the model's responses could significantly differ from behavioural results in situations where top-down attention is being manipulated.

In conclusion, we have shown that auditory scene perception of streams of single frequency tones can be explained by a simple Bayesian model utilizing a non-parametric prior. This highlights the importance of clustering in auditory perception, although the approach is applicable to any combination of stimuli and perceptual cues.

Together with advances in visual perception (Froyen et al., 2015b), this hints at clustering being a general property of perception.

Method

Model response

To determine the response of the model to a tone sequence, the posterior for each possible sequence, $S_{1:n}$, is calculated tone-by-tone until all 30 tones (maximum) have been presented. To relate the final posterior over sequences to response r_k of subject k, ('1 or 2 streams') $P_{model}(r_k|tones)$, we assume that subjects maximise the expected utility:

$$\underset{r_k}{\text{argmax}} < U > = \underset{r_k}{\text{argmax}} \sum_{i,j} U(r_k, S_i, S_j) P(S_i = S_j | f, t)$$

where the utility of a response given two tones being in same stream is counted as 1 if they are in the same stream, and zero otherwise. Note that the absolute values of S_i do not matter, just whether they are in same stream or not.

$$U(r_k, S_i, S_j) = 1 \text{ if } (r_k = 1, S_i = S_j) \text{ or } (r_k = 0, S_i \neq S_j)$$

$$U(r_k, S_i, S_j) = 0 \text{ if } (r_k = 0, S_i = S_j) \text{ or } (r_k = 1, S_i \neq S_j)$$

The best response is then to choose 1 (single stream) if

$$\sum_{i,j} P(S_i = S_j | f, t) > \sum_{i,j} \left(1 - P(S_i = S_j | f, t)\right)$$

If the observer believes all tones came from the same stream they should choose $r_k = 1$, if they are convinced half the tones are from one stream, half from another they should choose $r_k = 2$.

We assume soft-max, a variant of probability matching similar to Luce's law (1959) to explain variability in data and allow us to fit our models:

$$P(r_k|f,t) = \frac{\left(\exp\left(\beta * \sum P\left(S_i = S_j|f,t\right)\right)\right)}{\exp\left(\beta * \sum P\left(S_i = S_j|f,t\right)\right) + \exp\left(\beta * \sum\left(1 - P\left(S_i = S_j|f,t\right)\right)\right)}$$

The parameters par of the model (as well as for the alternative models) were optimised using the BADS toolbox (as a more robust alternative to MATLAB's fminsearch routine, Acerbi & Ma, 2017), to maximise the log-likelihood of the data, $\log\left(P_{model}(r_k|tones, par)\right)$ independently for each subject k. During each iteration of the search, a sequence of 30 tones was presented to the model for each condition, and the probability of response '1' was calculated per condition.

Model posterior approximation

Using the iterative scheme above we can calculate analytically the possible combinations of tones, but as the tone sequence progresses the number of possible source combinations - and hence the size of the posterior distribution - increases exponentially. To prevent combinatorial explosion two methods were used to generate an approximation of the full posterior distribution. The first limits the number of tones that are retained when using the previous posterior as the next prior, i.e. the algorithm only retains e.g. the last 30 tones and their potential allocations to sources.

Limiting the number of tones eases the computational load and can also be seen as a crude model of a limited memory capacity. Although the iteratively constructed prior retains some stream information of all previous tones, when a very short memory is used this may not be sufficient to generate stable stream allocation as the CRP prior probabilities fluctuate greatly when the number of previous tones is small. Furthermore, if the structure of the sequence is an important cue for streaming, a larger memory may be necessary to determine regularities in the sequence.

Even when the memory is limited to (e.g.) the previous six tones, allocating a stream to the seventh tone requires a posterior distribution taking 858 values, most of

which must necessarily have very small probabilities. A second method to limit the size of the posterior is simply to select only the most probable stream combinations by imposing a probability threshold, hence we only propagated stream combinations with $P(S_{1:n}|x_{1:n}) > 0.001$. Together these approximation methods allow a reasonable memory length of 30 tones (to avoid instability), while avoiding combinatorial explosion.

Experimental setup

Two experiments were conducted to test different aspects of the model. Experiment 1 was a replica of the 'galloping' stimuli experiment (van Noorden, 1975) performed to test the quantitative performance of the model, while the second experiment was a novel task designed to test one of the qualitative predictions based on the experimenters' knowledge of the model.

Subjects for both experiments were under-graduate students and received course credits for their participation, except for one of the authors who participated in Experiment 1. Each subject was fully briefed, provided informed consent and was given brief training on the task they performed. Experiments were completed by different subjects. No personal data was kept, ensuring participants' anonymity. The two experiments were approved by University of Birmingham and Durham University's Ethics Committees respectively.

Stimuli were dynamically programmed using Matlab on a PC desktop computer. Both experiments used the Psychtoolbox extension to ensure timings were accurate (Kleiner et al., 2007). Stimuli were played through Sennheiser 280 headphones at a comfortable supra-threshold level plugged into an external sound card (Behringer UCA20). No stringent calibration was made, although experimenters did check beforehand that sounds of different frequencies seemed of similar perceived sound level. The experiment was carried out in a special sound-attenuated room.

Experiment 1

This experiment replicated the study from van Noorden (1975), using fifteen participants. Figure 2.4 shows a schematic of the stimuli used – each sequence comprised 30 tones in repeated LHL- triplets, where the dash represents a silent gap. Each tone was 50 ms in duration, including 10 ms raised cosine onset and offset ramps. A 4x5 factorial design was used: the pitch of the high tones taking values of 3, 6, 9, 12 and 15 semitones above the low tone, which had a fixed frequency of 1000 Hz, and the offset to onset interval taking values 17, 33, 50 and 67 ms. The duration of the silent gap was equal to the tone duration plus the offset-onset interval. Conditions were ordered randomly – each condition was tested 20 times over 5 runs, each run lasting approximately 7 minutes.

At the end of the sequence participants pressed a key to report whether the percept at the end of the sequence was most like a single stream (a galloping rhythm) or two separate streams of notes.

Experiment 2

Participants

Twenty-six participants were enrolled for this study. All participants were Durham University students from undergraduate to postgraduate levels. Participants were asked if they had any known hearing impairment and were only allowed to participate if they reported a normal hearing. No personal data was kept, ensuring participants' anonymity.

Material and stimuli

Each testing trial consisted of 2 sequences of 4 pure tones in repeated Low-Medium-High-Medium (L-M1-H-M2 or L-M2-H-M1) quadruplets. The first sequence was always repeated 22 times for a total of 88 tones presented. The second one was always repeated a total of 11 times for a total of 44 tones presented. Tones between sequences within a same trial were always the same and only their order of presentation could differ.

Each tone was 100ms in duration, including 10ms raised cosine onset and offset ramps. The offset to onset interval between tones inside a sequence was 16.67ms. Each sequence also had general 500ms long raised cosine onset and offset ramps. The offset to onset interval between sequences inside a trial was 2s. The lowest tone had a fixed frequency of 440Hz across trials. The highest possible tone had a frequency of 2960Hz. The lowest frequency was specifically chosen to correspond to a common tone, and to control for differences in perceived loudness as much as possible across the range of played frequencies. Indeed, the 440-2960Hz range presents a low variability in equal-loudness (International Organization for Standardization [ISO], 2003). The frequency of M1 was calculated in semitone increases from the lowest tone, according to experimental conditions. M2 was always 3 semitones higher than M1. As was the case for the difference between L and M1, the frequency of H was calculated in semitone increases from M2, in relation to experimental conditions (see Figure 2.11 for a representation of a typical trial).

Training trials were similarly designed and consisted in 2 sequences of 3 pure tones in repeated Low-Medium-Medium (L-M1-M2 or L-M2-M1) triplets.

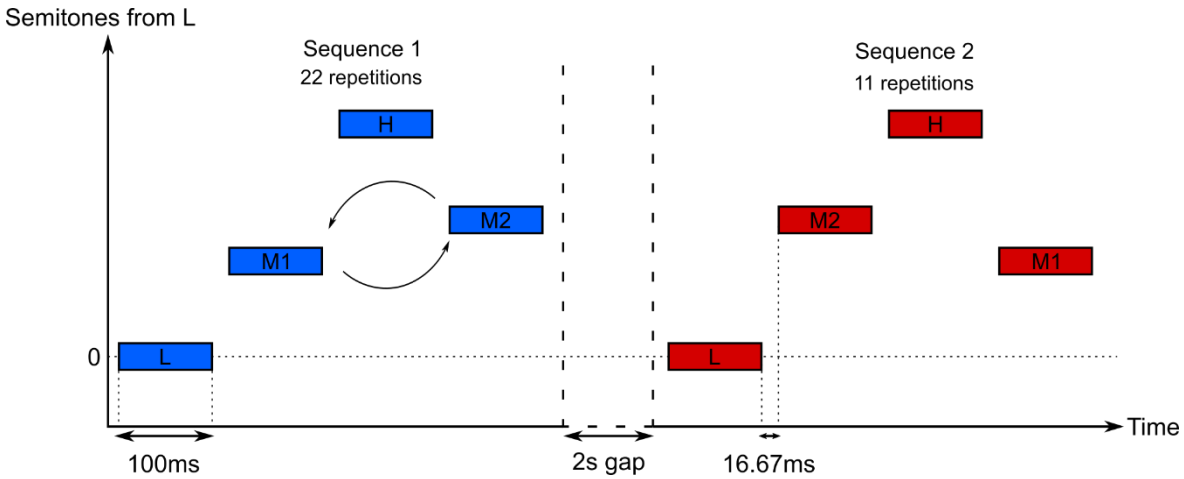


Figure 2.11: Visual representation of a trial with inversion in a 9-9 frequency difference condition.

Design

A 2x6 within-subjects factorial design was used. The first independent variable was the medium tones inversion, which could be either present or absent between the 2

sequences of a trial. This resulted respectively in trials objectively comprised of a pair of different sequences, and trials objectively comprised of a pair of twin sequences. The second independent variable was pairs of frequency differences between L and M1, and between M2 and H, counted in semitones. Possible values were 3-3, 3-9, 3-15, 9-9, 9-15 and 15-15. Conditions were presented randomly, and the order with which baseline and inverted tones sequences were presented was counterbalanced. Each identical pair of sequences was presented 6 times, while each different pair of sequences was presented 12 times for a total of 108 trials. Training trials consisted of 3 identical pairs and 3 different pairs repeated twice, for a total of 12 training trials. These were used to make participants familiar with the procedure, responses and stimuli through simplified trials.

The first dependent variable was the perceived difference between sequences (different vs. similar). The second dependent variable was the level of confidence in this judgement (on a scale from 1 to 4, 4 being “very confident”).

We did not analyse the confidence judgments here.

Procedure

Participants were greeted in a small sound-attenuated room and were asked to sit at a desk, approximately 60cm from a computer screen. They were handed an information and a privacy notice sheet, stating the general aims of the experiment, their rights as a subject and how their data would be handled. After reading and asking any question they may have to the experimenter, they were asked to sign a consent form. They were then asked to put the headphones on once they confirmed they understood the task’s instructions.

Participants were asked to listen to pairs of melodies presented sequentially, and judge after each pair if melodies were similar or different by pressing the right key on a keypad (“1” for different, “2” for similar). They were also asked to rate their confidence

about the judgement they just made, by pressing a key from 1 to 4 on the same keypad. Participants were also warned that tones between sequences had the same frequency, and that they should focus on the order of tones within the melody. Although it was clearly stated that both melodies consisted of the same tones and that they should focus on the order only, the word “similar” was used instead of “same” to ensure subjects still did not respond “different” in case they had a subjective sensation that tones differed in anything other than the order. Instructions were first given orally, then repeated on the screen at the beginning of the experiment. On each trial, a white dot was displayed in the middle of the screen for a brief period to signify a new trial is about to start. A grey dot was then displayed in place of the white one along with the instruction “listen” while melodies were being played. A black dot, along with a reminder of response keys, then replaced them as soon as melodies were finished, meaning they could enter their responses. Participants had no time limit to respond, as the next trial would only start after they did. Once the experiment was over, the experimenter gave oral feedback explaining the aims, design and experimental background of the study. The whole experiment lasted about 55 minutes.

Interchapter

This second chapter gave a mathematical description of our non-parametric Bayesian model and checked its usefulness in the context of ASA. Concretely, the final output from our model is a list of stream combinations, each tone being assigned to a single stream represented by a number, with a probability assigned to each of them, representing their respective credibility (see Table 2.2 for an example).

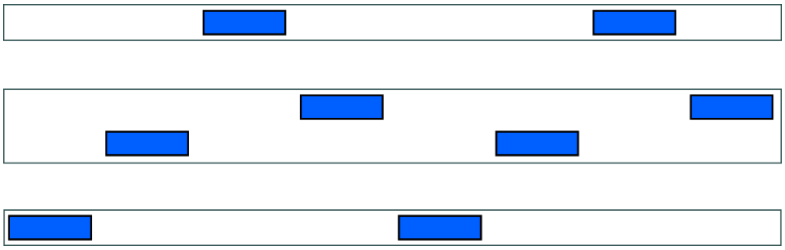
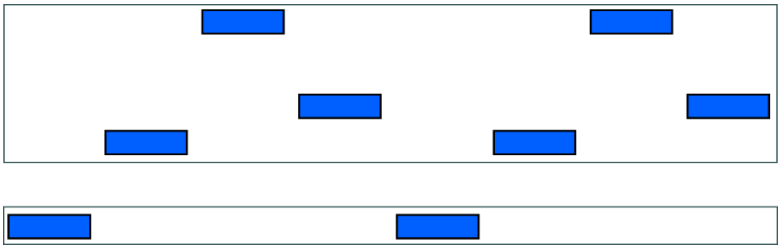
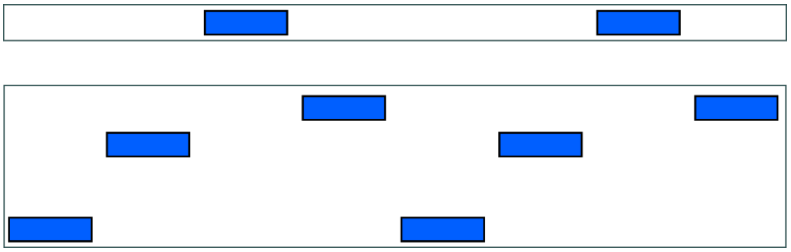
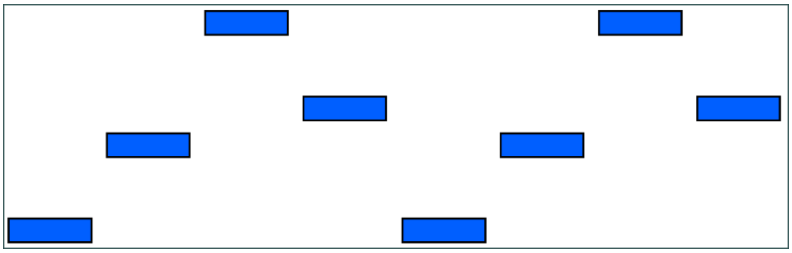
Stream combination	Probability	Graphical representation
1 2 3 2 1 2 3 2	0.5	
1 1 1 1 1 1 1 1	0.2	
1 1 2 1 1 1 2 1	0.2	
1 2 2 2 1 2 2 2	0.1	

Table 2.2: examples of an output from the model applied to an 8-tones melody, along with a graphical representation of the stream combinations

For now, our model has only been used to predict stream combinations based on the frequency difference over time ratio, but its predictions are theoretically similar on other dimensions based on the same rationale (e.g. two tones rapidly played from distant locations will more likely be assigned to different streams than from the same one), and effects are expected to be additive (tones are even more likely to be assigned to different streams as they cumulate fast differences on several dimensions).

If the brain really does this kind of causal inference, the question of how to actually interpret this in terms of phenomenology and behaviour is raised. One possibility is that the perceptual system generates all these combinations, assigns a credibility to each one of them, takes the most likely one, and discards all the others. In the example from Table 2.2, the listener would therefore have the sensation to be confronted with 3 different melodies being played simultaneously, and would then have to select which one he attends to. This means that if the two middle tones were to be inverted in a second presentation, it would be difficult for a listener to tell the difference as he cannot focus his attention on a melody containing tones relevant to the task. Conversely, should the most likely stream combination be the one where all tones are clustered together in a single melody, this inversion becomes easier to detect.

This seems like a reasonably simple solution in terms of cognitive load, and it is in line with both our results and the Gestalt literature in ASA as the process is fully pre-attentional and can deal with an unlimited number of streams.

The next chapter will expand on the behavioural insights given by this model and existing ASA literature by examining other tone characteristics thought to influence the grouping combination, in a somewhat more ecological situation where several of them can interact. Specifically, two variables will be under scrutiny: spatial location and timbre of tones.

Chapter 3

–

Auditory stream segregation in a complex and controlled environment: how do space and timbre interact with frequency differences?

Abstract

Auditory stream segregation has been studied extensively throughout the years, in either strictly controlled environments or more ecological settings such as in the cocktail party problem. Computational modelling has the potential to give more stringent explanations of these Gestalt-type experiments. However, controlled studies observing the interaction between several perceptual cues are rather scarce yet needed for a fine validation of such models. This study used a newly developed paradigm to allow for an easy study of an auditory stream segregation situation with several perceptual cues and with more than the usual dichotomy one/two streams. A set of low, medium, and high tones were manipulated by changing the frequency distance between these bands, along with either a spatial distance or a difference in timbre, to try to create a sensation of up to three simultaneous auditory streams in a 2AFC task. Participants were expected to have better performance telling two sequences apart if only one stream is heard than three, and it was expected that cues interact together in an additive way. Results show that performance does indeed degrade when more streams are being heard, but the interaction between cues did not follow predictions.

Introduction

Studying perceptual grouping mechanisms is of importance in order to understand how humans can have access to a coherent and well-categorized perception of the world surrounding them. Many early psychology studies, particularly the Gestalt school of thought, made these mechanisms their main preoccupation (Wertheimer, 1923). Although the law of Prägnanz and its underlying set of gestalt laws have been very successful at qualitatively describing how humans naturally perceive objects as organized patterns in several sensory modalities (Jäkel et al., 2016), stringent definitions allowing for accurate predictions are still lacking.

Recent endeavours to surpass these limitations have given birth to promising computational models (see for instance Froyen et al., 2015; Shams & Beierholm, 2010). One such model, developed recently and applied to the auditory modality, strives to explain general perceptual grouping mechanisms via a very small set of mathematically well-defined yet meaningful assumptions (Larigaldie, Yates & Beierholm, reviewing in progress). The model is essentially a custom clustering algorithm, sequentially grouping together similar stimuli while segregating dissimilar ones. It considers the grouping process to be pre-attentive, capable of producing an unlimited number of clusters, but trying to keep things as simple as possible by not unnecessarily multiplying the number of clusters. These assumptions are commonly accepted in the traditional Auditory Scene Analysis (ASA) literature (Bregman, 1994), even if there is still debate as to the possible role of attention on the grouping process (Kaya & Elhilali, 2017). Using those, this model successfully replicated several classical phenomena in the ASA literature and allowed for several qualitative predictions in the context of multi-cues unisensory perception.

Many studies describe how changes in different perceptual cues affect the way auditory streams are being created. The frequency was among the first and the most studied perceptual cue to show a consistent segregation effect of sounds (see for instance van Noorden, 1977). Indeed, these studies show that rapid sequences of alternating low- and high-pitched sounds can be heard as a unique melody when the frequency difference between them is small. However, the bigger this frequency difference, the more likely participants reported hearing two different melodies being played simultaneously instead of a single one. On top of this, whenever participants had this 2 streams sensation, they also lost the order information between the tones, effectively rendering them unable to report how tones were alternating. The reciprocity of this effect is since then often used as an implicit measure of stream segregation: if the order information between two tones is not accessible, then they were perceived as belonging to different streams (see Barsz, 1988,

for an example). Several other perceptual cues are known to influence this auditory stream segregation process, such as volume (van Noorden, 1977), intervals & tone durations (Bregman et al., 2000; Bregman & Dannenbring, 1973; Dannenbring, 1976; Dannenbring & Bregman, 1976) or spatial location of sounds (Eramudugolla et al., 2008). A more ecological version of this process is often called “The cocktail party effect”, in which one can rather easily segregate all the relevant sound information coming from a single individual from the other sounds in a noisy room in order to have a discussion (Bronkhorst, 2000). However, studies trying to take a middle ground and create a setting where several variables can interact in a tightly controlled environment are still scarce.

Working on the auditory modality gathers several advantages over the more studied visual one (Jäkel et al., 2016). Notably, the sequential nature of audition allows for stimuli less polluted by undesired variables (Bregman, 1994). Grouping mechanisms being also often considered in this context to be pre-attentive (Bregman & Rudnicky, 1975), the more sequential nature of this modality should make attention easier to control for than in other modalities. For instance, groups of dots presented visually may allow participants to purposely scan and focus on some areas, modifying the way groups were originally formed by the perceptual system. Furthermore, it has already been suggested that grouping mechanisms follow the same or very similar laws (Bregman & Achim, 1973; Warren & Gregory, 1958). Last but not least, the new paradigm introduced by Larigaldie, Yates & Beierholm (reviewing in progress) provides a convenient experimental setting to study perceptual grouping with several variables and their interactions. Interestingly, it creates a favourable setting to study clustering to up to 3 streams instead of the usual 2 streams most auditory experiments are limited to.

The model from the aforementioned article allowed for several qualitative predictions using this paradigm and different perceptual cues. Those predictions were made without actual simulation, based on the experimenters’ knowledge of the model. The

first one, in line with the ASA literature, was that as the difference in any perceptual cue becomes larger between two tones, the likelihood of them being segregated into different streams should also become larger. The second one was that larger likelihoods of tones being segregated should stack if several perceptual cues were to present a difference – for instance, two sounds with very different frequencies and played from very different spatial locations should more likely belong to different streams than two sounds with the same frequency difference being played from the same spatial location. Another prediction is that, given the right circumstances, it should be possible to observe segregation into at least 3 different streams, which the paradigm previously mentioned should allow.

Two experiments were conducted for this study, both using this paradigm and based on the qualitative predictions from the model. In both experiments, participants were presented with 2 melodies comprised of 4 repeated tones whose characteristics were manipulated, in a 2AFC task. They had to judge whether the melodies were the same or different. In half of the trials, two tones were inverted between the first and the second presentation, effectively creating different melodies. Using past observations stating that order information is lost between perceptual groups but kept within (Bregman & Campbell, 1971), it was assumed that the objective difference between melodies would be much easier to detect when each melody is perceived in a single perceptual group, and conversely.

The first experiment was set to study how the spatial location of sounds, their frequencies, and the interaction of those cues, influence the way our perceptual system groups or segregates sounds in a melody. The second experiment follows the same logic, using timbre (in this context, taking the form of layers of harmonics) and frequency as sensory cues being manipulated.

Predictions from the model, and consequently for this study, are that bigger discrepancies in any sensory cue should lead to a stronger tendency to segregate sounds

into different streams and that effects from different cues should be additive, resulting in stronger segregation effects if several cues present discrepancies.

Hypotheses for the first experiment are that accuracy in detection of a difference should be higher for small frequency differences than for large frequency differences. Similarly, accuracy should be higher when all sounds are played from the same spatial location, and lower as the spatial location increases. Cumulating large frequency differences and large spatial differences should result in a lower accuracy than any one of these differences alone.

Hypotheses for the second experiment are the same, the only difference being that timbre difference is a dichotomic variable (with/without): the accuracy in detection of a difference should also be higher for small frequency differences than for large frequency differences. Accuracy should also be higher when all tones are played without any timbre added than when only some of them do. Cumulating large frequency differences and the presence of timbre on some tones should result in a lower accuracy than any one of these differences alone.

Method

Experiment 1

Participants

Participants in this study were 29 Durham University volunteer students in Psychology, ranging from Undergraduate to Postgraduate level recruited through convenience sampling. Undergraduates were recruited through a compulsory program requiring them to participate in studies in exchange for course credits. No personal data was kept, ensuring participants' anonymity. The experiment is in accordance with GDPR and was approved by Durham University's Ethics Committee. Participants were asked if

they had any known hearing impairment and were only allowed to participate if they reported a normal hearing.

Material and stimuli

Each testing trial consisted of 2 sequences of 4 pure tones in repeated Low-Medium-High-Medium (L-M1-H-M2 or L-M2-H-M1) quadruplets (see Figure 3.1 for a representation of a typical trial). The first sequence was always repeated 22 times for a total of 88 tones presented. The second one was always repeated a total of 11 times for a total of 44 tones presented. Tones between sequences within the same trial were always the same and only their order of presentation could differ: either they were perfectly identical under one condition, or M1 and M2 were swapped in the second sequence under another condition. Each tone was 100ms in duration, including 10ms raised cosine onset and offset ramps. The offset to onset interval between tones inside a sequence was 16.67ms. Each sequence also had general 500ms long raised cosine onset and offset ramps. The offset to onset interval between sequences inside a trial was 2s. The lowest tone had a fixed frequency of 440Hz across trials. The highest possible tone had a frequency of 1480Hz. The lowest frequency was specifically chosen to correspond to a common tone and to control for differences in perceived loudness as much as possible across the range of played frequencies. Indeed, the 440-1480Hz range presents a low variability in equal-loudness (International Organization for Standardization [ISO], 2003). The frequency of M1 was calculated in semitone increases from this one, according to experimental conditions. M2 was always 3 semitones higher than M1. As was the case for the difference between L and M1, the frequency of H was calculated in semitone increases from M2, in relation to experimental conditions.

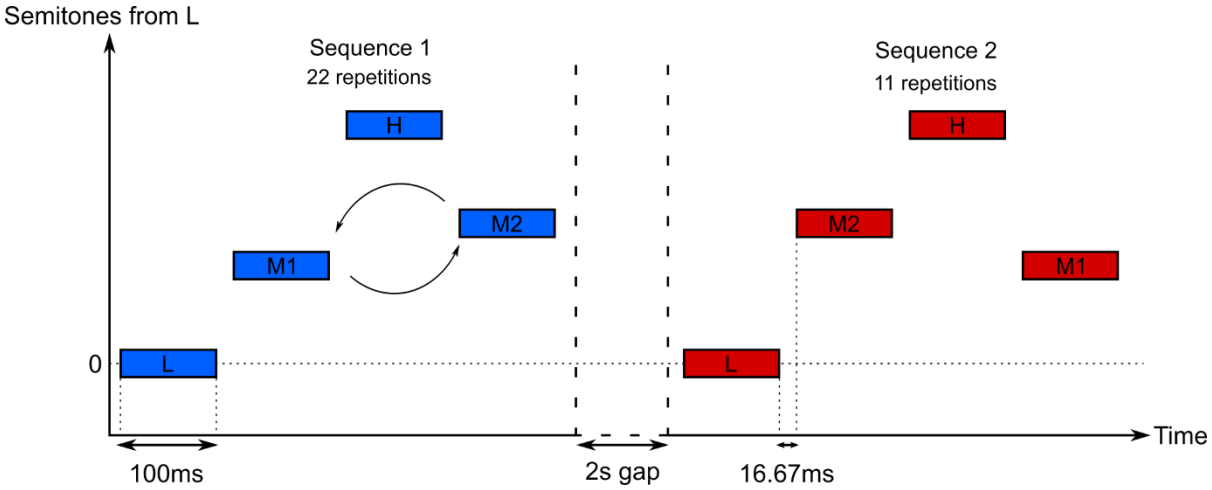


Figure 3.1: Visual representation of a trial with inversion in a 9-9 frequency difference condition

Tones were played through an array of speakers at a comfortable supra-threshold level through an external sound card (Focusrite Scarlet 20i18). In one condition, all tones were played from the central speaker, directly ahead of the participant. In another condition, L was played on a speaker 10 degrees of visual angle to the left of the central one, M1 and M2 were played on the central speaker, and H was played on a speaker 10 degrees of visual angle to the right of the central one. In the last condition, L was played on a speaker 30.5 degrees of visual angle to the left of the central one, M1 and M2 were played on the central speaker, and H was played on a speaker 30.5 degrees of visual angle to the right of the central one. Speakers were tested and their individual volumes increased or decreased by software to ensure that sounds arriving at the participants' location displayed the same physical sound level. No other calibration was made although experimenters did check beforehand that sounds from different frequencies seemed of a similar perceived sound level.

Training trials were similarly designed and consisted of 2 sequences of 3 pure tones in repeated Low-Medium-Medium (L-M1-M2 or L-M2-M1) triplets. However, tones in these trials were all played by the same central speaker.

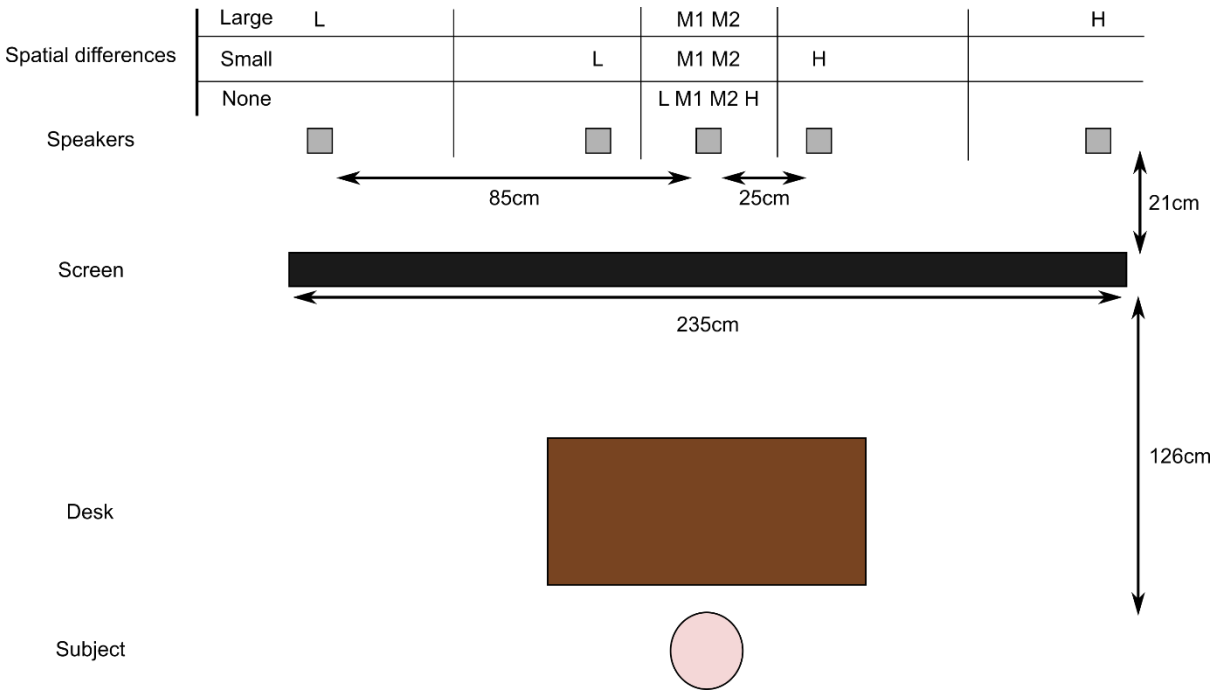


Figure 3.2: visual top-down representation of the spatial organization in experiment 1

Stimuli were dynamically programmed in Matlab on a PC desktop computer. The whole experiment used the Psychtoolbox extension to ensure timings were accurate (Kleiner et al., 2007). The experiment was carried out in an echo-attenuated room. Echo attenuation was achieved by all walls having been treated with 10-inch acoustic foam. Visual information was presented on a 235x131cm acoustically transparent projection screen. The speakers were hidden 21cm behind the screen. A chinrest was used to stabilize the participants' head position with regard to the central speaker. See Figure 3.2 for a visual representation of the overall setup in this experiment.

Information sheets, privacy notice sheets, and consent forms in accordance with GDPR were prepared in advance and given to participants.

Design

A 2x3x3 within-subjects factorial design was used. The first independent variable was the inversion of the medium tones, which could be either present or absent between the 2 sequences of a trial. This resulted respectively in trials objectively comprised of a pair of different sequences, and trials objectively comprised of a pair of twin sequences. The second independent variable was pairs of frequency differences between L and M1,

and between M2 and H, counted in semitones. Possible values were 3-3, 6-6, and 9-9. The last independent variable was the spatial difference between extreme frequencies and medium ones, which could either be absent, small, or large (respectively ±0, ±10, and ±30.5 degrees of visual angle). Conditions were presented randomly, and each identical and different pairs of sequences were presented 6 times per possible spatial location, for a total of 108 trials. Training trials consisted of 3 identical pairs and 3 different pairs repeated twice, for a total of 12 training trials.

The dependent variable was the perceived difference between sequences (different vs. similar).

Procedure

Participants were greeted in a small soundproof room and were asked to sit at a desk, approximately 126cm away from a projector screen. They were handed an information and a privacy notice sheet, stating the general aims of the experiment, their rights as a subject, and how their data would be handled. After reading and asking any question they may have to the experimenter, they were asked to sign a consent form.

Participants were then asked to listen to pairs of melodies presented sequentially, and judge after each pair if melodies were similar or different by pressing the right key on a keypad ("1" for different, "2" for similar). Participants were also warned that tones between sequences had the same frequency and that they should focus on the order of tones within the melody. They were also asked to rate their confidence about the judgement they just made, by pressing a key from 1 to 4 on the same keypad (data not analysed in this paper). Although it was clearly stated that both melodies consisted of the same tones and that they should focus on the order only, the word "similar" was used instead of "same" to ensure subjects still did not respond "different" in case they had a subjective sensation that tones differed in anything other than the order. Instructions were

first given orally, then repeated on the screen at the beginning of the experiment. On each trial, a white dot was displayed in the middle of the screen for a brief period to signify a new trial is about to start. A grey dot was then displayed in place of the white one along with the instruction “listen” while melodies were being played. A black dot, along with a reminder of response keys, then replaced them as soon as melodies were finished, meaning they could enter their responses. Participants had no time limit to respond, as the next trial would only start after a response.

Participants did training trials for 5-10 minutes and were systematically asked if they understood what was being asked and if they noticed a difference in at least a few trials. When this was not the case, they were proposed to do another set of the same training trials before doing the main task. These training trials were used to make participants familiar with the procedure and responses with simplified stimuli. If the task was done correctly the first time, participants only did one set of training trials before doing the main task, which lasted 40-45 minutes. Since the experiment was long and required a lot of attention, they had 2 opportunities for pauses during the main task.

Once the experiment was over, the experimenter gave oral feedback explaining the aims, design, and experimental background of the study. The whole experiment lasted about 55 minutes.

Experiment 2

Participants

Participants in this study were 31 volunteers, recruited through voluntary sampling on online discussion forums. No personal data was kept, ensuring participants’ anonymity. The experiment was in accordance with GDPR and was approved by Durham University’s Ethics Committee. Participants were asked to only participate if they had no known hearing impairment.

Material and stimuli

Stimuli in this experiment were designed in a similar way to the first one. However, they were now played through participants' headphones, at the experimenter's request. Before training trials started, a looped melody was played indefinitely. Participants were asked to adjust the volume to a comfortable level. No further calibration was attempted.

Fundamental frequencies were the same as in the first experiment. However, medium tones in half of the stimuli were no longer pure tones, as they also included second, third, and fourth harmonics.

Training trials were designed the same way as they were in the last experiment, but medium tones included second, third, and fourth harmonics in half of these trials.

Stimuli were this time generated beforehand using Matlab, but the experiment was programmed using PsychoPy (Peirce, 2007), then hosted online and ran through Pavlovia.org (Peirce et al., 2019).

Information sheets, privacy notice sheets, and consent forms in accordance with GDPR were prepared in advance and put in an online form.

Design

A 2x3x2 within-subjects factorial design was used. The first two independent variables were the same as in the former experiment: the first independent variable was the inversion of the medium tones, which could be either present or absent between the 2 sequences of a trial. This resulted respectively in trials objectively comprised of a pair of different sequences, and trials objectively comprised of a pair of twin sequences. The second independent variable was pairs of frequency differences between L and M1, and between M2 and H, counted in semitones. Possible values were 3-3, 6-6, and 9-9. The last independent variable was the presence of a timbre addition on middle tones, which could be either absent or present (in the form of second, third, and fourth harmonics added to the

middle pure tones). Conditions were presented randomly, and the order with which baseline and inverted tones sequences were counterbalanced. Each identical and different pairs of sequences were presented 10 times per timbre condition, for a total of 120 trials. Training trials consisted of 4 identical pairs and 4 different pairs repeated, for a total of 8 training trials. Half of the training trials included a timbre difference for its middle tones.

The dependent variable was the perceived difference in tones order reported by participants between sequences (different vs. similar).

Procedure

Participants could do the experiment from their personal computer. They were invited to click on a link that redirected them to an online form, containing an information and a privacy notice sheet, stating the general aims of the experiment, their rights as a subject, and how their data would be handled. After reading, they were asked to electronically sign the consent form. Once the form was submitted, the link to the actual study was displayed.

Participants were then asked to listen to pairs of melodies presented sequentially, and judge after each pair if melodies were similar or different by pressing the right key on a keypad (“1” for different, “9” for similar). Participants were also warned that tones between sequences had the same frequency and that they should focus on the order of tones within the melody. They were also asked to rate their confidence about the judgement they just made, by pressing a key from 1 to 4 on the same keypad (data not analysed in this paper). Although it was clearly stated that both melodies consisted of the same tones and that they should focus on the order only, the word “similar” was used instead of “same” to ensure subjects still did not respond “different” in case they had a subjective sensation that tones differed in anything other than the order. Instructions were written on the screen at the beginning of the experiment. On each trial, a white square was displayed in the middle of the screen for a brief period to signify a new trial is about to

start. A grey square was then displayed in place of the white one along with the instruction "listen" while melodies were being played. A black square, along with a reminder of response keys, then replaced them as soon as melodies were finished, meaning they could enter their responses. Participants had no time limit to respond, as the next trial would only start after they did.

Participants did training trials for 5-10 minutes. These were used to make participants familiar with the procedure and responses with simplified stimuli. If their accuracy in these training trials was lower than 50%, the program warned the subject that they would be presented with these trials a second time before doing the experiment. The majority of participants only did one set of training trials before doing the main task, which lasted 40-45 minutes. Since the experiment was long and required a lot of attention, they had 2 pauses during the main task.

Once the experiment was over, the experimenter gave written feedback to interested participants explaining the aims, design, and experimental background of the study. The whole experiment lasted about 55 minutes.

Results

Experiment 1

Analysis preparation

Individual responses on perceived differences between sequences were transformed into D-prime scores to obtain a single measure of signal detection for each crossed condition of frequency and spatial differences while taking into account possible response biases. One participant with a negative D-prime score on the easiest FREQUENCY DIFFERENCES condition (*3-3*) was considered unable to perform the task correctly and was therefore excluded from further analysis, leaving a total of 28 participants. No data were missing from the dataset. A two-way repeated-measures ANOVA was conducted on these

D-prime scores, with FREQUENCY DIFFERENCES (*3-3* vs. *6-6* vs. *9-9*) and SPATIAL DIFFERENCES (*none* vs. *small* vs. *large*) as within-subject factors, along with three paired-samples t-tests in line with our hypotheses. No correction for multiple comparisons was applied.

Shapiro-Wilk normality tests revealed that D-prime scores from the *6-6* condition with a *large* spatial difference, and in the *9-9* condition with a *large* spatial difference were not following a normal distribution [respectively $W(28)=0.915$, $p=.025$; $W(28)=0.910$, $p=.02$]. Even though removing outliers from these conditions restored normality, the impact on effect sizes and significances was unnoticeable. Participants were therefore kept in the analysis.

Mauchly's test indicated that the assumption of sphericity had not been violated for the FREQUENCY DIFFERENCES factor [$\chi^2(2)=0.960$, $p=.586$], the SPATIAL DIFFERENCES factor [$\chi^2(2)=0.977$, $p=.734$], nor the interaction [$\chi^2(9)=0.628$, $p=.224$].

Data analysis

Although inferential statistical tests were conducted on D-prime scores only, Table 3.1 also includes summarizing statistics of proportions of "similar" responses in every experimental condition.

The two-way repeated-measures ANOVA (FREQUENCY DIFFERENCES*SPATIAL DIFFERENCES) revealed that D-prime scores differed as a function of the factor FREQUENCY DIFFERENCES [$F(2,54)=12.63$, $p<.001$, $\eta_p^2=0.319$]. However, no statistically significant difference as a function of the SPATIAL DIFFERENCES factor was observed [$F(2,54)=2.579$, $p=.085$, $\eta_p^2=0.087$]. Similarly, no statistically significant interaction of the FREQUENCY DIFFERENCES and SPATIAL DIFFERENCES factors was observed [$F(4,108)=1.993$, $p=.101$, $\eta_p^2=0.069$].

Three paired-samples t-tests were conducted to decompose the main effect of FREQUENCY DIFFERENCES over D-prime scores. Two one-tailed paired-samples t-tests revealed that D-prime scores were significantly higher for the *3-3* condition than for the *6-6* condition [$t(27)=4.188$, $p<.001$, $d_z=0.791$] and the *9-9* condition [$t(27)=4.729$, $p<.001$, $d_z=0.894$]. However, there was no statistically significant difference in D-prime scores between the *6-6* and the *9-9* conditions [$t(27)=0.883$, $p=.385$, $d_z=0.167$].

These results are summarized in Figure 3.3.

	3-3			6-6			9-9		
	None	Small	Large	None	Small	Large	None	Small	Large
Different	0.230 ± 0.269	0.448 ± 0.316	0.391 ± 0.325	0.466 ± 0.293	0.570 ± 0.225	0.523 ± 0.330	0.552 ± 0.302	0.563 ± 0.276	0.546 ± 0.285
Similar	0.805 ± 0.189	0.828 ± 0.211	0.753 ± 0.192	0.707 ± 0.243	0.868 ± 0.129	0.741 ± 0.207	0.782 ± 0.179	0.816 ± 0.150	0.730 ± 0.211
D-prime	1.585 ± 0.860	1.065 ± 0.895	0.959 ± 0.817	0.656 ± 0.896	0.810 ± 0.662	0.553 ± 0.883	0.585 ± 0.857	0.631 ± 0.775	0.522 ± 0.811

Table 3.1: Mean proportions of "similar" responses and mean d-prime scores across all conditions in experiment 1. Reported errors are ± 1 standard deviation

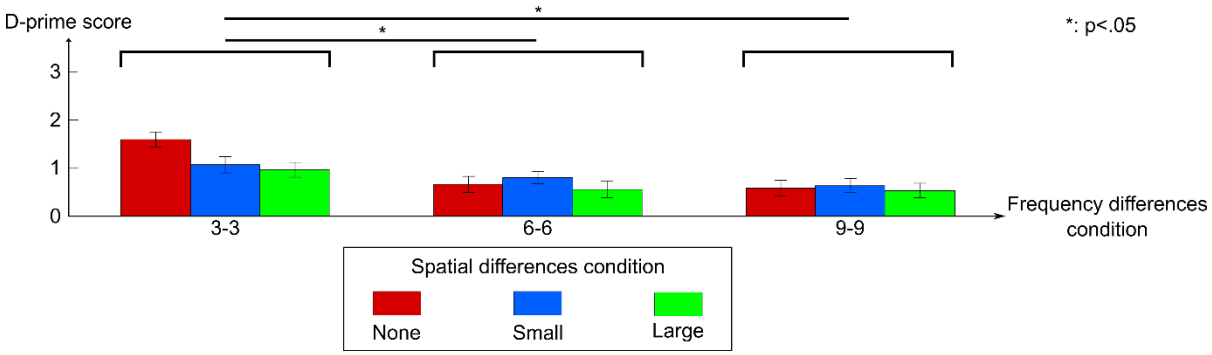


Figure 3.3: D-prime scores as a function of frequency difference and spatial difference in experiment 1. Error bars are ± 1 standard error

Experiment 2

Analysis preparation

Individual responses on perceived differences between sequences were transformed into D-prime scores to obtain a single measure of signal detection for each crossed condition of frequency differences and timbre presence while taking into account possible response biases. No data were missing from the dataset. A two-way repeated-measures ANOVA was conducted on these D-prime scores, with FREQUENCY DIFFERENCES (*3-3* vs. *6-6* vs. *9-9*) and TIMBRE (*absent* vs. *present*) as within-subject factors, along with three paired-samples t-tests in line with our hypotheses. Six post-hoc paired-samples t-tests were also conducted to decompose the interaction. No correction for multiple comparisons was applied.

Mauchly's test indicated that the assumption of sphericity had not been violated for the FREQUENCY DIFFERENCES factor [$\chi^2(2)=0.813$, $p=.051$] or the interaction [$\chi^2(2)=0.464$, $p=.464$]. Although the test approached significance for the FREQUENCY DIFFERENCES factor, applying the Greenhouse-Geisser correction had no noticeable effect on the overall results.

Data analysis

Although inferential statistical tests were conducted on D-prime scores only, Table 3.2 also includes summarizing statistics of proportions of "similar" responses in every experimental condition.

	3-3		6-6		9-9	
	Absent	Present	Absent	Present	Absent	Present
Different	0.123 ± 0.167	0.665 ± 0.309	0.361 ± 0.270	0.613 ± 0.332	0.439 ± 0.297	0.519 ± 0.353
Similar	0.813 ± 0.173	0.861 ± 0.138	0.761 ± 0.235	0.855 ± 0.177	0.797 ± 0.194	0.865 ± 0.154
D-prime	2.106 ± 0.915	0.592 ± 0.897	1.198 ± 0.791	0.726 ± 1.029	1.026 ± 0.763	1.037 ± 0.989

Table 3.2: Mean proportions of "similar" responses and mean d-prime scores across all conditions in experiment 2. Reported errors are ± 1 standard deviation

The two-way repeated-measures ANOVA (FREQUENCY DIFFERENCES*TIMBRE) revealed that D-prime scores differed as a function of both the FREQUENCY DIFFERENCES factor [$F(2,60)=9.352$, $p<.001$, $\eta_p^2=0.238$] and the TIMBRE factor [$F(1,30)=24.165$, $p<.001$, $\eta_p^2=0.446$]. Similarly, the interaction of the FREQUENCY DIFFERENCES and TIMBRE factors was statistically significant [$F(2,60)=24.724$, $p<.001$, $\eta_p^2=0.452$].

Three paired-samples t-tests were conducted to decompose the main effect of FREQUENCY DIFFERENCES over D-prime scores. Two one-tailed paired-samples t-tests revealed that D-prime scores were significantly higher for the *3-3* condition than for the *6-6* condition [$t(30)=4.947$, $p<.001$, $d_z=0.888$] and the *9-9* condition [$t(30)=3.471$, $p<.001$, $d_z=0.623$]. However, there was no statistically significant difference in D-prime scores between the *6-6* and the *9-9* conditions [$t(30)=0.613$, $p=.545$, $d_z=0.11$].

Six paired-samples t-tests were conducted to decompose the interaction of FREQUENCY DIFFERENCES and TIMBRE factors over D-prime scores.

The first three compared TIMBRE conditions within each FREQUENCY DIFFERENCES condition using two-tailed tests. Within the *3-3* condition, D-prime scores were significantly lower when timbre was *present* [$t(30)=7.392$, $p<.001$, $d_z=1.327$]. Similarly, within the *6-6* condition, D-prime scores were significantly lower when timbre was *present* [$t(30)=2.644$, $p=0.014$, $d_z=0.475$]. However, within the *9-9* condition, D-prime scores in the *absent* condition was not significantly higher than in the *present* condition [$t(30)=-0.064$, $p=0.95$, $d_z=0.011$].

The following three compared all pairs of FREQUENCY DIFFERENCES conditions when TIMBRE was *present* using two-tailed tests. Analysis showed the *3-3* condition was not significantly different to the *6-6* one [$t(30)$=-0.993, p=0.328, d_z=0.178]. The *3-3* condition was significantly lower than the *9-9* one [$t(30)$=-2.583, p=0.014, d_z=0.464]. The *6-6* condition was not significantly different to the *9-9* one [$t(30)$=-1.8428, p=0.076, d_z=0.331]. These results are summarized in Figure 3.4.

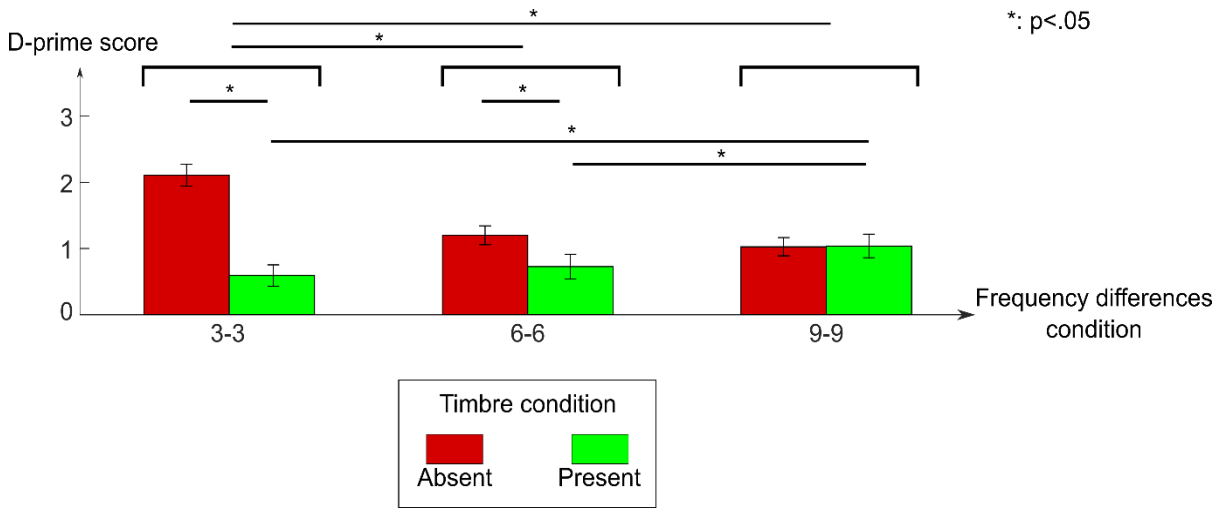


Figure 3.4: D-prime scores as a function of frequency difference and spatial difference in experiment 2. Error bars are ± 1 standard error

Discussion

Hypotheses for the first experiment were that an increasing difference in spatial locations between tones should increase tone segregation into 3 different streams and in return reduce participants' accuracy. The same was hypothesised for frequency differences. An additive effect of both variables was expected, with no visible interaction.

The hypotheses were partially validated. Indeed, the principal effect of frequency differences from classic ASA literature was replicated (Bregman, 1994). Medium and large frequency differences were both making it harder for participants to tell when sequences were different than in the low frequency difference, effectively meaning that subjects were more likely to hear several melodies rather than one. This further validates the

experimental paradigm from Larigaldie, Yates & Beierholm (reviewing in progress). However, results were less clear regarding the other hypotheses.

No main effect of spatial difference was observed and no interaction between space and frequency was observed. However, a clear trend can be seen in the lower frequency difference condition, in which increasing the spatial difference seems to decrease performance. Nevertheless, there did not seem to be any observable additive trend of both variables in the other conditions. Instead, there seemed to be a floor effect reached quite easily and suddenly, as if once a certain threshold in cue differences was passed, whatever the cue considered, performance would stabilize at low accuracy. In other words, the task can either be easy with cues allowing for an easy grouping of the percepts or suddenly hard when any of the cues make participants perceive several streams, with little to no middle-ground. On an individual level, this sudden switch from the perception of one stream to several streams has already been described (Bregman et al., 2000). This floor effect could explain why the principal effect of spatial location cannot be observed in this first experiment. Other possibilities could be that the gaps between the conditions chosen were too big to observe any interaction effect at some critical level, or that individual differences in the breaking point for each variable could mask the effect. Finally, another possibility could be that the effect of spatial discrepancies is pretty low in the auditory stream segregation process. Spatial localization is much noisier than visual localization, and it seems plausible that the perceptual system tends to rely less on spatial cues when it comes to auditory situations. In fact, even if such spatial effects have already been observed (McAnally & Martin, 2007), there have been conflicting results and it is known at best to have a modest effect compared to frequency cues (Eramudugolla et al., 2008). Finally, the spatial manipulation may have not been as powerful as could have been, as previous observations reported that it is generally difficult for human listeners to localize pure tones

(Blauert, 1997). However, insight from the second experiment opens to other alternative hypotheses.

The hypotheses for the second experiment were that the presence of a difference in timbre between tones should increase tone segregation into different streams and in return reduce participants' accuracy. The same was hypothesised for frequency differences. An additive effect of both variables was expected, with no visible interaction.

Hypotheses were once again partially validated. First, the principal effect of frequency differences was once again replicated. Second, timbre did have a strong effect on tone segregation into different streams. Both results are in line with classic ASA literature (Bregman, 1994). Unexpectedly and contrary to experimental hypotheses, an interaction between frequency difference and the presence of a timbre difference was present, and there was no cumulative effect observed throughout the conditions. Instead, when middle tones had timbre added to them, participants performed better for larger differences in frequency. When frequency differences were high, there was no longer a difference in performance with or without timbre.

These results are hard to reconcile with the classic ASA view of a grouping based solely on the perceptual cues and an infinite number of possible streams. In this situation, increasing the frequency distance between low and medium tones, and between medium and high tones, on top of creating a different timbre for these medium tones should only make those even less likely to be clustered with either low or high tones. Instead, it would seem that increasing this distance somehow makes them more likely to be clustered with either the low or high tones, as this is the only way to complete the task correctly.

There is however an alternate explanation coming from the attentional literature. Some authors (Mack et al., 1992) have already proposed that attention plays a crucial role in the stream formation process. Their study suggests that whenever subjects are busy with

a task, even obvious occurrences of Gestalt laws in their visual field do not lead to segregation into different groups. This suggests that grouping may usually not happen at all under a condition of inattention. In another set of studies, Carlyon et al. (2001) observed that focusing on tones played in one ear tended to decrease or even cancel auditory streaming in the contralateral ear. Perhaps even more interestingly, patients with unilateral neglect also showed this pattern of decreased/absent auditory streaming on their impaired side only. A possible explanation is therefore that at any given time, the maximum number of streams is in fact limited to a maximum of 2, as was already suggested (Bregman & Rudnicky, 1975; Mack & Rock, 1998): the one stream the attention is focused on and whose construction is strongly influenced by both top-down attentional processes and the percept's characteristics, and another one where all the remaining percepts are grouped, in which only a basic bottom-up attentional monitoring takes place. In fact, this very possibility has already been proposed in the past (Brochard et al., 1999) in an experiment where all unattended auditory percepts also seemed fused whenever they were not attended to. On top of this, the ability to at least subconsciously recognize a change in a melody that is not being attended to has already been observed in EEG studies (Sussman, 2017; Thomassen & Bendixen, 2018). Our own experiment would then suggest that this information is also somewhat consciously accessible. According to this idea, most of the stream segregation and grouping process would no longer be pre-attentional. It could in fact be mainly attentional: the possibility or not to create an attended stream comprising of certain stimuli would still depend on their cues and would still be following the Gestalt laws, but all the other percepts would simply be grouped together in an unattended stream.

It is possible our results from the second experiment could be explained by this phenomenon. In the second experiment, in the low frequency difference condition and without added timbre on medium tones, everything could easily fall into the single attended stream, making the task quite easy. By adding a timbre difference on the middle

tones, the very low performance reveals that participants most probably have a hard time keeping them in this single, coherent stream. Instead, middle tones are now too different from low or high tones to form a stream still containing the relevant order information. However, it is at this point impossible to know what exact streams participants are forming and focusing on. There are 4 main streams participants could focus their attention on while having such a poor performance: middle tones, low tones alone, high tones alone, or low & high tones (see Figure 3.5 a)). However, as we increase the frequency difference, results show that performance seems to rise until it reaches a similar performance to a high frequency difference without a timbre difference. This indicates that the stream formation process is impacted by this frequency difference increase. From the 4 possibilities aforementioned, the only one most likely impacted according to Gestalt laws is the one with a stream containing low and high tones, as these are now more dissimilar. This stream must therefore be harder to form and/or focus attention on. As a result, participants will more likely focus on low tones alone, medium tones, or high tones alone (see Figure 3.5 b)). As focusing on medium tones has no reason to increase performance, it follows that focusing on low tones alone or high tones alone somehow does. If the leftover tones are now fused together in the unattended stream as the attentional literature suggests, it is coherent that increasing the frequency differences makes the task difficult but possible again as long as they did not focus on the middle tones. This pattern of increased performance in the presence of large discrepancies in both perceptual cues would not be visible in our first experiment simply because there is no situation in which low and high tones are pushed to be fused together by having only the middle tones stand out: contrary to the second experiment, in all conditions, a difference is made between all frequency ranges at once.

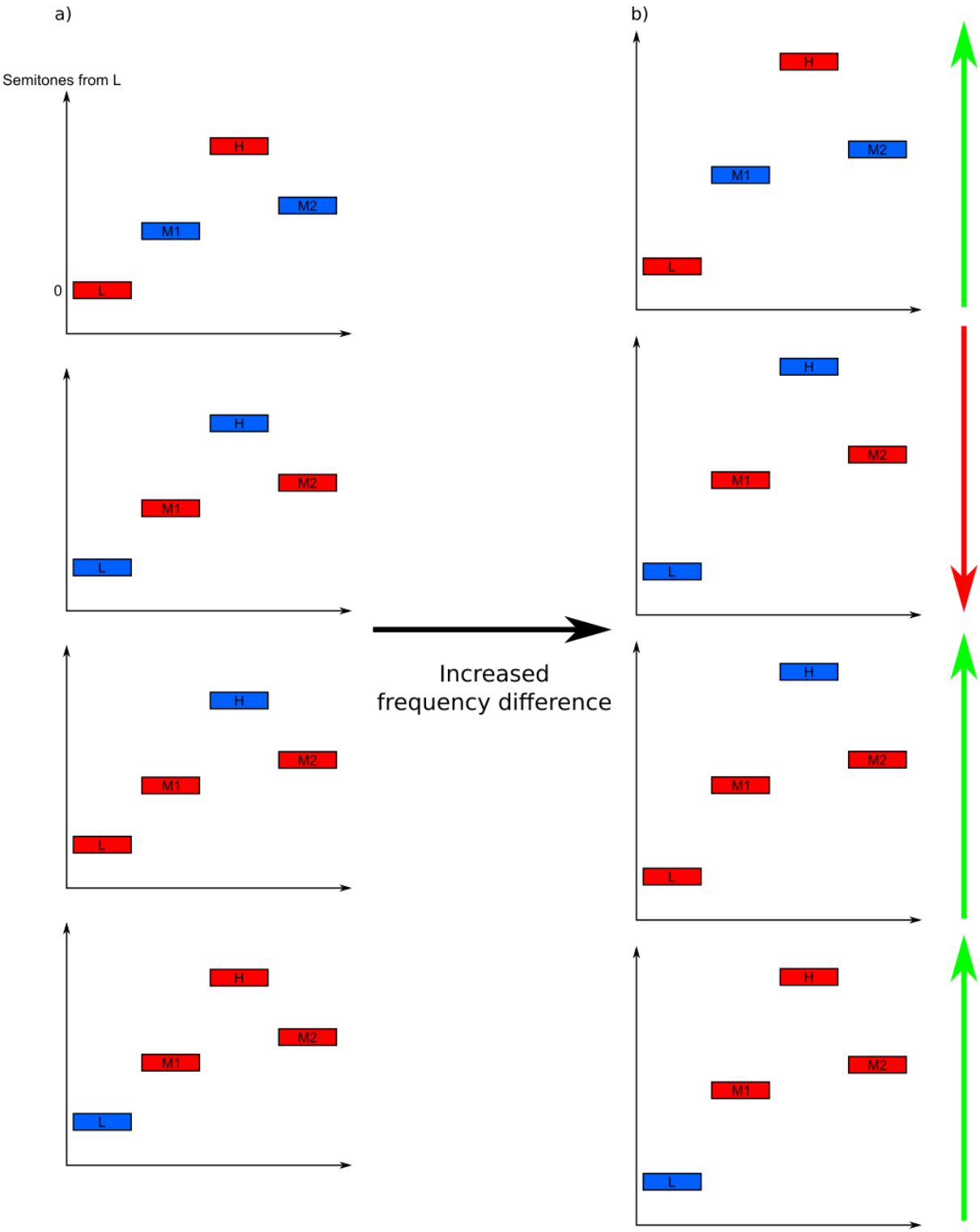


Figure 3.5: Main possible streams on which participants could focus their attention when timbre was added to the middle tones. Blue tones are tones attended to; red tones are background tones. Increasing the frequency difference decreases the likelihood of participants attending to a stream formed of low and high tones together. a) streams with a small frequency difference b) streams with a large frequency difference

This hypothesis is also in line with the fact that throughout all our experiments following this paradigm, we found that our subjects are still somewhat able to perform above chance even in the most extreme conditions, while other authors have usually found

that this is not the case when working on 2 streams (Barsz, 1988). In our paradigm, focusing on only some tones (namely, the low or the high ones) often ensures that the temporal information is in fact still available in the unattended stream, making the task hard but possible.

Future research should explore in more detail the role attention plays in this particular paradigm, for instance by explicitly orienting which tones participants should try to focus on, and see how performance is impacted. Should our proposal be correct, suggesting to focus only on high tones instead of high and low tones in a high frequency condition should increase the general performance, as the temporal information becomes available in the unattended stream.

Another follow-up study could include individual calibrations to points of subjective equality using a staircase procedure for quantitative variables (e.g. frequency and spatial location). This could allow observing potential additive effects across variables through finer conditions.

Overall, this study has shown a clear replication of auditory streaming segregation effects resulting from frequency and timbre differences between tones, and to a lesser extent, a visible trend from spatial localization differences. Additive effects between those variables were expected but not observed, and the interaction between timbre and frequency even showed an inverse pattern. All in all, a strong limitation in the maximum number of streams formed by the perceptual system can potentially account for the unexpected observations, especially regarding the interaction between frequency and timbre in the second experiment. Results from the first experiment could also suggest the existence of a floor effect, preventing the observation of an additive effect between frequency and space on stream segregation.

Interchapter

This third chapter explored the effects of spatial location and timbre of tones and their interactions with frequency on the participants' formation of stream combinations. It concluded on observable effects of these cues on the way they combine tones together, but unexpectedly, without any additive effect with the frequency effect observed in the second chapter. As mentioned in the discussion, the stream formation process in this dissertation has been mainly considered as a pre-attentive process for now. However, results suggested that contrary to traditional Auditory Scene Analysis suggestions, the auditory stream segregation process may in fact be strongly dependent on attention and that forming more than two streams may not be possible.

During the course of our studies, several participants and experimenters also reported being able to switch between different stable stream combinations in some conditions (and so, hear either one or two melodies for instance), even if it required some conscious effort to do so. This qualitative observation is in direct contradiction with the idea that our perceptual system only keeps track of the most likely stream combination, and that the only role of top-down attention is to select the stream that the listener wants to attend to. At the very least, less likely combinations may be kept in memory (and therefore an attentional focus could allow for a specific selection of different combinations), or attention could be part of the stream formation process itself.

If attention indeed plays an active part in the stream formation, it could have important repercussions in the way we interpret both our results and the output from our model. This could mean that several stream combinations could potentially be available to listeners at any time, should they focus their attention correctly. Furthermore, our new paradigm is a good opportunity to explore this matter, which is still debated today. This is why the next chapter will investigate the possible role of attention on the stream formation process.

Lab vs. Online experiments comparison

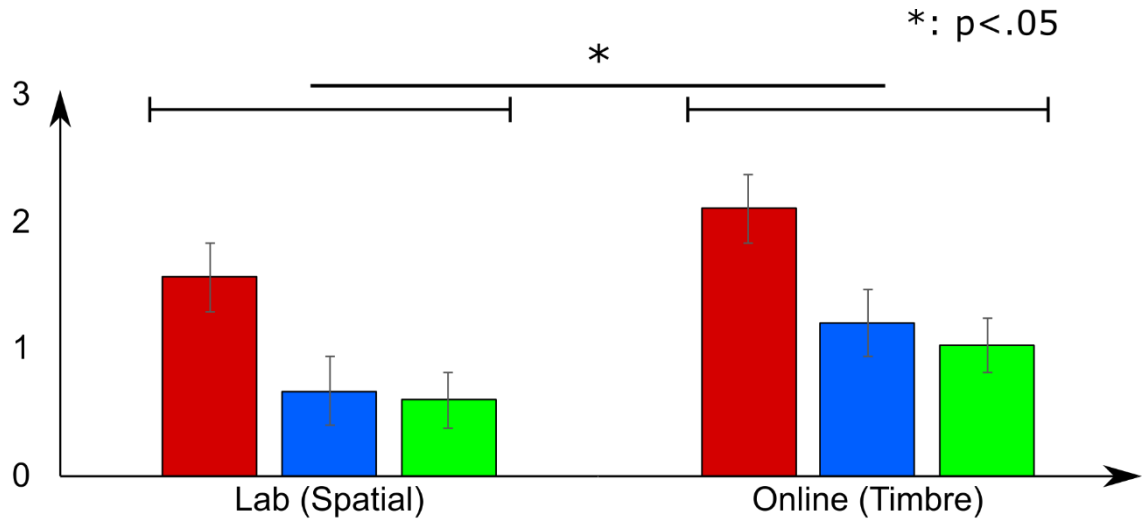
The two experiments from this chapter were quite similar in their design, but were conducted in very different environments: the first one was made in a controlled lab environment, while the second one was done online by participants from home, with their own equipment. Fortunately, the *3-3*, *6-6* and the *9-9* FREQUENCY DIFFERENCES conditions across the two experiments were directly comparable as stimuli were exactly the same respectively in the *none* SPATIAL DIFFERENCES condition from the first experiment, and the *absent* TIMBRE condition from the second one. This allowed us to perform an analysis to see if the online setting produced comparable results, and therefore did not introduce too many confounding variables.

A two-way mixed-model ANOVA was conducted on these D-prime scores, with FREQUENCY DIFFERENCES (*3-3* vs. *6-6* vs. *9-9*) as the within-subject factor, and SETTING (*lab* vs. *online*). Normality and sphericity tests were conducted, and made no differences in the results.

A well-reproduced experiment should not display any interaction between those factors (as the influence of the frequency on D-prime scores should not depend on the experimental setting), and ideally not display a principal effect of SETTING (as general performance should not be better/worse between settings).

The ANOVA showed that D-prime scores differed as a function of FREQUENCY DIFFERENCES [$F(2,114)=39.88$, $p<.001$, $\eta_p^2=0.412$] and SETTING [$F(1,57)=198.635$, $p<.001$, $\eta_p^2=0.777$]. However, no significant interaction was shown [$F(2,114)=0.119$ $p=.888$, $\eta_p^2=0.002$]. A graphical representation of these results, along with D-prime scores, can be found in Figure 3.6.

The absence of interaction between the two factors along with its negligeable effect size make for a strong argument in favour of a successful port of the experiment in an



3-3	1.562	2.105
6-6	0.668	1.198
9-9	0.593	1.026

Figure 3.6: D-prime scores as a function of frequency differences and experimental setting (experiments 1 & 2). Error bars are ± 1 standard error

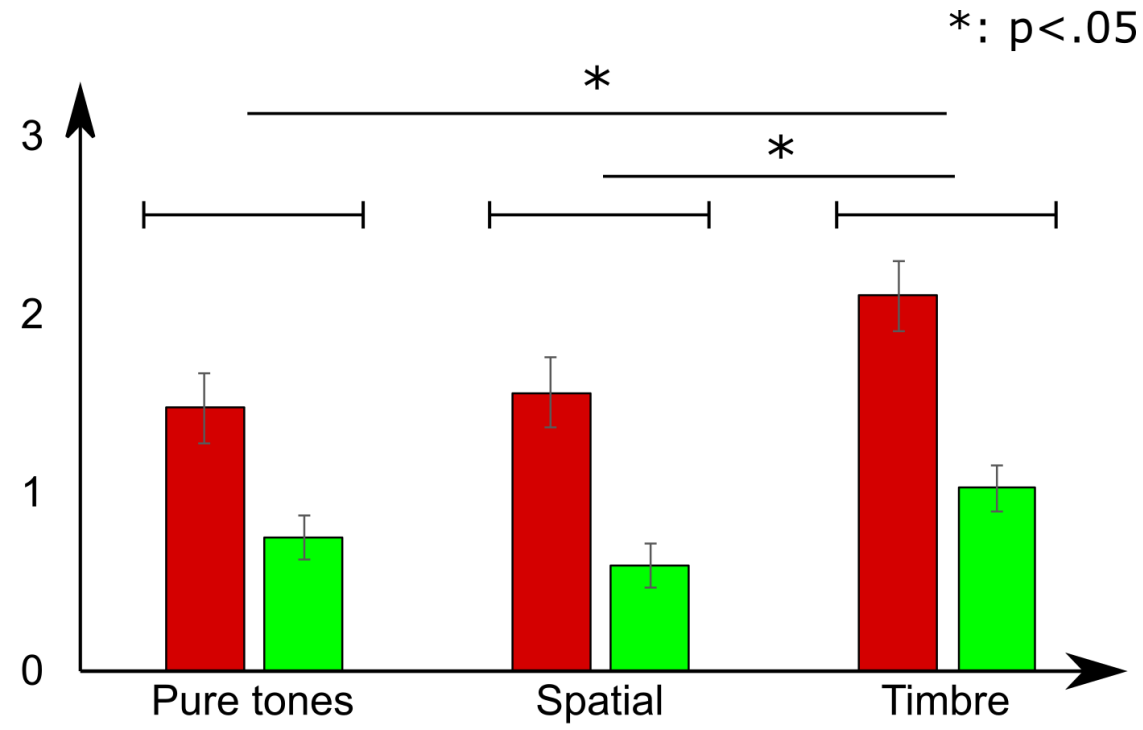
online setting, as the effect of frequency differences on the D-prime scores was the same across settings. However, it seems that the general performance of participants was better in all conditions in the online experiment than in the lab one. This could be explained by worse-quality equipment used by subjects (e.g., different output volumes for different frequency bands), or varying individual settings, making the task easier to perform. However, it could also be explained by a generally higher motivation to do the task, as subjects were recruited voluntarily in science-enthusiastic forums, as opposed to a compulsory university program. Overall, the absence of interaction is the most important information, and a general increase in performance in these conditions is most certainly beneficial and could even be viewed as a strong argument in favour of the online setting.

Cross-experiments comparisons

All experiments in this thesis have also been designed so that we could replicate some conditions and compare them to some extent. The objective was not only to potentially strengthen our argument regarding the observed experimental effects, but also to check if the change of settings or material used had no deleterious impact on the experiments. As of the end of this chapter, two conditions have remained constant across three experiments using our novel paradigm: *3-3* and *9-9* FREQUENCY DIFFERENCES conditions, in the absence of additional interacting variables.

A two-way mixed-model ANOVA was conducted on these D-prime scores, with FREQUENCY DIFFERENCES (*3-3* vs. *9-9*) as the within-subject factor, and EXPERIMENT (*pure tones* vs. *spatial* vs. *timbre*). *Pure tones* stands for the second experiment of Chapter 2, *spatial* and *timbre* stand for the first and second experiment from chapter 3, respectively. Three post-hoc independent-samples t-tests were also conducted. No correction for multiple comparisons was applied. Normality and sphericity tests were conducted, and made no differences in the results.

A well-reproduced experiment should not display any interaction between those factors (as the influence of the frequency on D-prime scores should not depend on the experiment), and ideally not display a principal effect of EXPERIMENT (as general performance should not be better/worse between experiments). This analysis can be viewed as a generalization of the one made just previously. However, since the objective and the conditions changed substantially, this called for a different analysis.



3-3	1.477	1.562	2.105
9-9	0.748	0.593	1.026

Figure 3.7: D-prime scores as a function of frequency differences and experiment. Error bars are ± 1 standard error

The ANOVA showed that D-prime scores differed as a function of FREQUENCY DIFFERENCES [$F(1,80)=72.287$, $p<.001$, $\eta_p^2=0.475$] and EXPERIMENT [$F(2,80)=277.027$, $p<.001$, $\eta_p^2=0.776$]. However, no significant interaction was shown [$F(2,80)=0.879$ $p=.419$, $\eta_p^2=0.021$]. A graphical representation of these results, along with D-prime scores, can be found in Figure 3.7.

Three independent-samples t-tests were conducted to decompose the main effect of EXPERIMENT over D-prime scores. Two of them revealed that D-prime scores were significantly higher for the *Timbre* condition than for the *Pure tones* condition [$t(53)=2.466$, $p=.017$, $d_z=0.67$] and the *Spatial* condition [$t(57)=2.573$, $p=.013$, $d_z=0.671$]. However, there was no statistically significant difference in D-prime scores between the *Pure tones* and the *Spatial* conditions [$t(50)=0.206$, $p=.838$, $d_z=0.057$].

The absence of interaction between the two factors suggests the effect of frequency over D-prime scores did not depend on the experiment, therefore strengthening further its generality. The Timbre/online experiment also showed the general increase of performance compared to the *pure tones* experiment, as was already the case compared to the *spatial* one in the previous analysis, while those last two seemed to display similar participants performance. As argued before, this could be due to a difference in equipment or a difference in motivation. All in all, as of this chapter, our new paradigm used across three experiments seems to reliably capture the effect of frequency on the general performance, and therefore the way participants cluster tones together.

Chapter 4

–

Testing the role of attention in Auditory Scene Analysis

Abstract

The traditional Gestalt Psychology literature considers perceptual grouping mechanisms to be completely preattentive. Within this assumption, the perceptual system is supposed to generate a theoretically infinite number of possible groups, after which the role of attention is simply to select the cluster relevant to the task at hand. However, recent studies and models have challenged this view and now usually suggest that attention plays an important role in the process and that the number of streams may potentially be far more limited. This study is designed to explore the role of attention and how perceptual groups are segregated in an auditory modality setting. Melodies comprised of low, medium, and high tones were manipulated to create a 1 vs. 3 streams situation according to the traditional Gestalt model of Auditory Scene Analysis. Low vs. High-frequency differences between low & medium, and medium & high tones were being used to manipulate the possible number of streams perceived by subjects in a 2AFC task that should be easier in the 1 stream than in the 3 streams situation. In parallel, participants had to focus their attention on either all, high only, or low and high tones. It was hypothesized that the different attentional conditions would modify the way streams were created and that a maximum of 2 streams only could be reached, making the task easier than what would be expected if 3 streams were reached. Results suggest that attention indeed plays a significant role in the stream formation process and that subjects never reached more than 2 streams. We propose that the perceptual system may use Gestalt principles to create a list of possible single streams from unattended elements that the attentional system could then pick. However, this list of coexisting mental representations from unattended elements also automatically includes an item within which all unattended elements are fused.

Introduction

At any given time, humans are presented with an astonishing amount of perceptual information and need to quickly analyse and react to the content of their perception.

Among numerous mechanisms to make sense of the environment, the perceptual grouping of these elements into meaningful clusters across different sensory modalities has been extensively studied by Gestalt psychologists (Jäkel et al., 2016).

Auditory Scene Analysis (ASA) has been proposed by Bregman (1994) as a model explaining grouping and segregation of sounds into streams for the auditory modality. This model proposes that, given a mixture of sounds recorded from a complex environment, our perceptual system is able to extract and use sensory cues in order to create sensible streams of sounds, specifically not to mix up information coming from different sources. In order to do that, laws of Gestalt Psychology in the visual modality such as proximity or continuity (Wagemans et al., 2012) are usually considered to have their auditory counterpart (for examples concerning these two laws, see Bregman & Campbell, 1971; Bregman & Dannenbring 1973). A large body of literature on the matter has shown the ASA model not only is a good fit to existing observations but has also been able to produce useful predictions for decades.

However, parts of this model are still theoretical and under debate. One such aspect is the role that attention plays in the stream formation process. Traditionally, ASA literature considers this process to be preattentive: the perceptual system would first group or segregate sounds into a theoretically infinite number of streams based on Gestalt laws to infer which sounds came from the same source, and the listener would then just have to use their attention to select the stream they want to attend to. In fact, the preattentive nature of the phenomenon is often tightly linked to the idea of an infinite number of streams: if top-down attention's role is to pick which stream is to be selected, then all possible grouping combinations have to already be worked out before. Conversely, should the number of streams be limited to only a few, then the stream formation process would probably require some form of attention, as it cannot know in advance which stream it should focus on constructing, and attentional processes would not be able to pick unconstructed streams

without requiring first a stream formation from the perceptual system, making it de facto a postattentive process. Despite early recognition of the possibility that the number of streams being held at once may not be more than one (Bregman & Rudnicky, 1975), this hypothetical vision of a preattentive process has remained the dominant one in the ASA literature throughout the years. As most studies in this domain during the first decades have been limited to the expected segregation between two streams, whether or not these assumptions were true was of little consequence to their results.

Nevertheless, several studies have challenged this idea by either emphasizing the importance of top-down and bottom-up attention on the stream formation process itself or by challenging the maximum number of streams being constructed by the perceptual system. Importantly, one such study has suggested that even within the visual modality, a maximum of two perceptual clusters could be held at once: one cluster on which the attention is focused on, and another one within which no segregation happens and all percepts are fused (Mack et al., 1992). In each of their experiments, participants had to perform a demanding task regarding a cross in the middle of a screen, that was surrounded by patterns of obviously ungrouped elements that were not attended to by subjects. When later asked about those elements, they were unable to report the existence of different clusters in the background, suggesting either that they were fused in a unique unattended cluster, or that the information was not encoded in memory because of the lack of attention on this specific task. In the auditory modality, Sussman (2017) proposed several EEG experiments not only suggesting that both bottom-up and top-down attention played crucial roles in the stream formation process, but also that non-attended elements could under certain circumstances elicit a response indicating that participants could detect melody changes even without focusing their attention on it. Their findings were that an EEG response to a change in melody was detectable when passively listening to tones while doing another task, but not when the change was in a background melody while actively

listening to other sounds. However, previous results (Larigaldie & Beierholm, writing in progress) suggest that background melody changes could influence participants' responses and that it could also happen when actively listening to another melody. Even if Sussman's experiments tried to demonstrate that unattended elements were being segregated into different streams, results were also compatible with a unique unattended stream. Taken together, these studies could suggest that the perceptual and the attentional systems work hand in hand to create one stream that is being attended to, and another one where all the other elements are being fused.

This is also in line with a set of experiments (Carlyon et al., 2001) where inattention in healthy subjects prevents stream segregation from a well-known ASA experiment (van Noorden, 1975). Perhaps more interestingly, experimenters also observed that patients with unilateral left neglect displayed the stream segregation pattern when listening to sounds in their right ear, but less so in their left ear. All in all, several authors already proposed models either stating that attention is necessary to group elements in the foreground, that the elements in the background are being fused together, or both, whether it is in the auditory modality or in general (see for instance Mack & Rock, 1998; Shamma et al., 2011; Shamma et al., 2013; and Treisman, 1998). Of course, even if all these results suggest that attention modulates the way percepts are being grouped, this does not mean that Gestalt laws no longer play an active part in the process. On the contrary, they would rather set boundaries within which groups can or cannot be formed, based on the percepts' features, as predicted by these guiding principles.

However, it is important to note that other experiments have repeatedly shown some form of low-level perceptual grouping processes in situations of inattention, whether it was in the visual (Montoro et al., 2014) or the auditory (Winkler et al., 2005) modality. Overall, the maximum number of streams held by our perceptual system and the destiny of unattended perceptual features still seem to be matters of debate. But one thing however

remains certain: there is indeed an influence of attention on the stream formation process, and in the more general perceptual grouping mechanisms, that can no longer be overlooked when considering more and more ecological experimental settings.

In a previous set of experiments, Larigaldie, Yates & Beierholm (under review) devised a novel ASA paradigm specifically designed to reach a theoretical maximum of 3 streams, with a task capable of implicitly measuring the number of streams participants were constructing. However, unexpected results (Larigaldie & Beierholm, writing in progress) seemed to point at a possible attentional effect, with one interpretation being that at any point participants only had access to two streams, including a single unattended one where all percepts were fused together.

The present experiment was designed to further explore the role of attention in the auditory stream formation process, and whether unattended elements are being fused into a unique background stream, or segregated into several streams. It consisted of pairs of melodies of 4 tones being repeated in a 2AFC task, where participants had to judge the similarity of both melodies. Tones could be either close or far away from each other in terms of frequency. In half of the trials, the two melodies were indeed different, as the order of the tones was changed. In previous studies (Larigaldie & Beierholm, writing in progress), this paradigm has shown that it was far easier for participants to detect an objective difference between melodies when the tonal range was small than when it was large. Bregman & Campbell (1971) have already shown that order information is lost between perceptual groups but kept within, and these results therefore suggest that small frequency differences allow participants to keep all tones in a single stream. It was hypothesized that attention would strongly influence this process and that no more than two streams would be able to coexist. Therefore, subjects were also asked to focus their attention on specific subparts of the melodies in small and large frequency range conditions. In order to make sure subjects did indeed direct their attention correctly, they

were also asked to press a key every time they heard when a tone in the frequency range they were asked to focus on was being played louder. In a setting similar to that of Sussman (2017), the experiment was designed so that the main task asked of participants – judging the similarity of the melodies – becomes harder if unattended percepts are being segregated into different streams, and conversely, easier if unattended percepts are all fused into a single stream.

Operational hypotheses were that when frequency differences are small, it should be easy for participants to cluster every tone together. They should therefore be able to perform the task correctly when asked to focus their attention on the entire melodies. However, it is expected that asking participants to focus on subparts of the melody should slightly decrease the performance compared to making no such request, as the task becomes harder to complete. Indeed, requesting to specifically focus on low and high tones should force subjects to create a cluster comprised only of these tones, and have another cluster comprised of the middle tones only. This configuration does not allow the task to be completed successfully, and it is therefore expected that this condition should be the worst of all. Finally, requesting to focus on high tones only should force subjects to create a cluster comprised only of these tones and another one with low and middle tones. In this configuration, as the order information is still available in the unattended cluster,

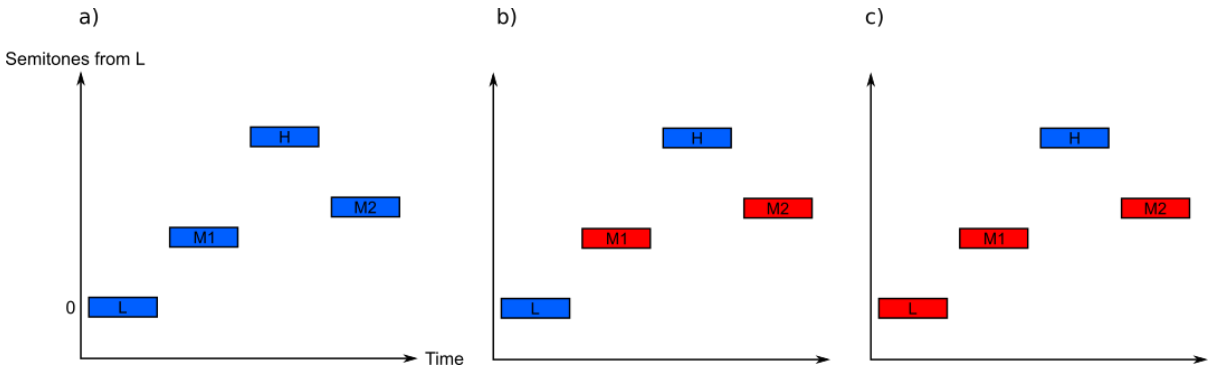


Figure 4.1: Main expected attended/unattended tones in the small frequency difference condition. Blue tones are being attended to, red tones are not. a) subjects are asked to focus on the whole melody b) subjects are asked to focus on low and high tones c) subjects are asked to focus on high tones

performance should be somewhere in between the other two situations.

When frequency differences are large, it should be generally harder for participants to cluster tones together. In contrast with the case of small frequency differences, it is not expected that asking participants to focus on low and high tones should decrease the performance. Indeed, participants should either be unable to perform this correctly because of the large frequency difference, making them only able to focus on low or high tones in both situations (see Figure 4.2 a) and b) top), or be able to create this cluster, in which case they would in both situations (see Figure 4.2 a) and b) bottom). However, whenever they are required to focus solely on high tones, they should form a cluster comprised only of these tones, and (if unable to form more than two clusters) have another cluster comprised of the low and medium tones. In this situation, the "unattended" cluster contains the order information, and the performance should increase and be above both the condition where the attentional instruction is to focus on all tones and the one where participants are asked to focus on low and high tones.

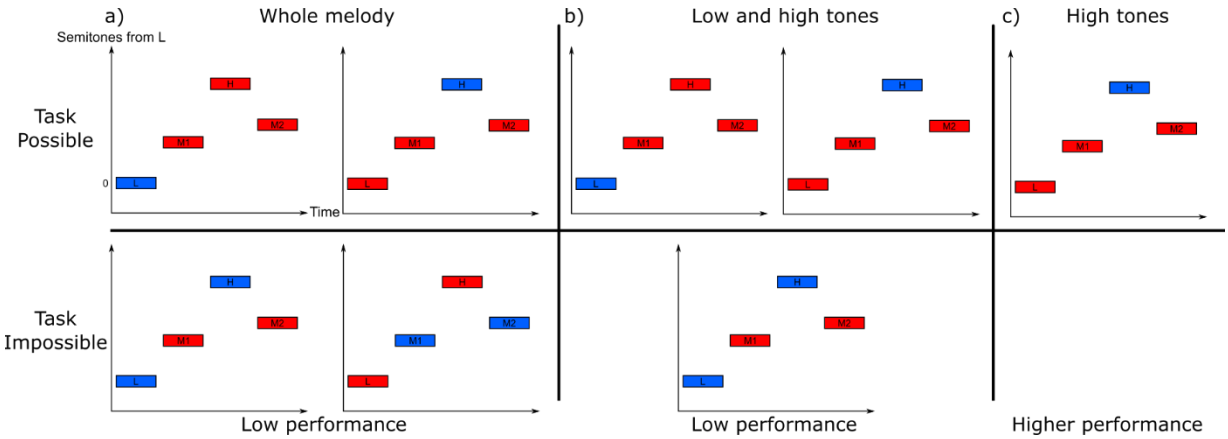


Figure 4.2: Main expected attended/unattended tones in the large frequency difference condition. Blue tones are being attended to, red tones are not. Due to the large difference in frequencies it is not possible to integrate the whole melody. Top configurations should keep the task possible to a certain extent, while bottom one makes the task impossible a) subjects are asked to focus on the whole melody b) subjects are asked to focus on low and high tones c) subjects are asked to focus on high tones

Method

Participants

Participants in this study were 45 volunteers (20 in a small frequency differences condition, 25 in a large one), recruited through voluntary sampling on online discussion forums. No personal data was kept, ensuring participants' anonymity. The experiment is in accordance with GDPR regulations and was approved by Durham University's Ethics Committee. Participants were asked to only participate if they had no known hearing impairment.

Material and stimuli

Each testing trial consisted of 2 sequences of 4 pure tones in repeated Low-Medium-High-Medium (L-M1-H-M2 or L-M2-H-M1) quadruplets. The first sequence was always repeated 22 times for a total of 88 tones presented. The second one was always repeated a total of 11 times for a total of 44 tones presented. Tone frequencies between sequences within the same trial were always the same although their order of presentation could differ. Each tone was 100ms in duration, including 10ms raised cosine onset and offset ramps. The offset to onset interval between tones inside a sequence was 16.67ms. Each sequence also had general 500ms long raised cosine onset and offset ramps. The offset to onset interval between sequences inside a trial was 2s. The lowest tone had a fixed frequency of 440Hz across trials and experiments. The highest possible tone had a frequency of 740Hz within the condition with small frequency differences, and 1480Hz within the condition with large ones. The lowest frequency was specifically chosen to correspond to a common tone and to control for differences in perceived loudness as much as possible across the range of played frequencies. Indeed, the 440-1480Hz range presents a low variability in equal-loudness (International Organization for Standardization [ISO], 2003). The frequency of M1 was calculated in semitone increases from this lowest tone: 3 in the first condition, 9 in the second. M2 was 3 semitones higher than M1 in both. As was

the case for the difference between L and M1, the frequency of H was calculated in semitone increases from M2, and was also 3 in the first experiment and 9 in the second (see Figure 4.3 for a representation of a typical trial).

For one-third of the first sequences, 3, 4, or 5 of all tones were selected pseudo-randomly to be 3.5 times louder. Two louder tones could never be one right after another. For another third, these tones were selected only out of the low and high tones. For the last third, they were selected only out of the high tones. The same logic was followed for the second sequences, but with only 1, 2, or 3 tones being louder. These louder tones were designed to be targets for the 3 possible attentional conditions.

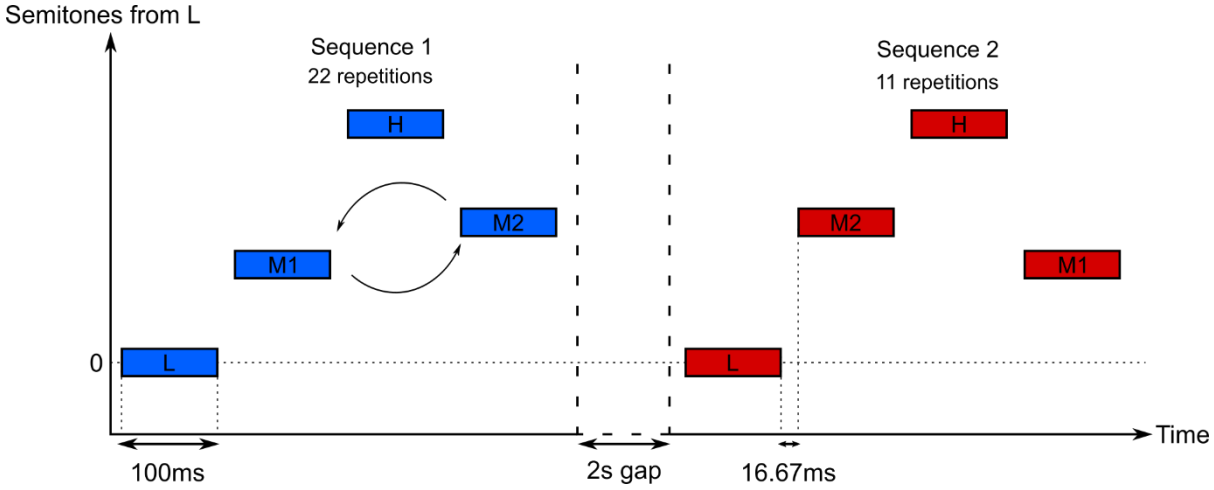


Figure 4.3: Visual representation of a trial with inversion used within both experiments

The two sets of training trials were similarly designed and consisted of 2 sequences of 3 pure tones in repeated Low-Medium-Medium (L-M1-M2 or L-M2-M1) triplets. One set had no louder tones. In the second set first sequences always had 4 louder tones and the second sequences had either 0, 1, or 2.

Tones were played through participants' headphones, at the experimenter's request. Before training trials started, a looped melody was played indefinitely. Participants were asked to adjust the volume to a comfortable level. No further calibration was attempted.

Stimuli were generated beforehand using Matlab, but the experiment was programmed using PsychoPy, then hosted online and ran through Pavlovia.org (Peirce et al., 2019).

Information sheets, privacy notice sheets, and consent forms in accordance with GDPR regulations were prepared in advance and put in an online form.

Design

A 2x2x3 mixed design was used. The first independent variable was the pairs of frequency differences between L and M1, and between M2 and H, counted in semitones. Possible values were 3-3 and 9-9. This variable was between subjects. The second independent variable, within-subjects, was the inversion of the medium tones, which could be either present or absent between the 2 sequences of a trial. This resulted in sets of trials objectively comprised of a pair of different sequences, and trials objectively comprised of a pair of twin sequences. The third independent variable was the attentional focus, which could be either on all tones (L, M1, M2, and H), low and high tones (L and H), or high tones (H). This last variable was also within subjects. Conditions were presented randomly and the order with which baseline and inverted tones sequences were presented was counterbalanced. Each identical and different pairs of sequences were presented 12 times (6 before counterbalancing) per possible attentional focus, for a total of 72 trials per pair of frequency differences. Training trials without louder tones consisted of 2 identical pairs and 2 different pairs repeated three times, for a total of 12. Training trials with louder tones consisted of the same trials, with the pseudo-random addition of louder tones.

The dependent variable was the perceived difference in tones order reported by participants between sequences (different vs. similar).

Procedure

Participants could do the experiment from their personal computer. They were invited to click on a link that automatically redirected them with an equal chance to one of two online forms, corresponding to each experiment (with frequency differences of either 3 or 9 semitones). Each online form contained an information and a privacy notice sheet, stated the general aims of the experiment, their rights as a subject, and how their data would be handled. After reading, they were asked to electronically sign the consent form. Once the form was submitted, the link to the experiment was displayed. Subjects only did one condition of frequency differences and had no knowledge of the other one.

Participants were then asked to listen to pairs of melodies presented sequentially and told they would have for each pair to either listen to all the tones normally or focus on specific tones. They were warned that they would have to perform two tasks simultaneously: press space every time one of the tones they were asked to focus on was louder, and judge after each pair if melodies were similar or different by pressing the right key on a keypad (“1” for different, “9” for similar). Participants were also warned that tones between sequences had the same frequency and that only the order of tones within the melody would matter. Although it was clearly stated that both melodies consisted of the same tones and that they should focus on the order only, the word “similar” was used instead of “same” to ensure subjects still did not respond “different” in case they had a subjective sensation that tones differed in anything other than the order. General instructions were written on the screen at the beginning of the experiment. Before each trial, a small text would instruct if subjects had to focus on the whole melody, low and high tones, or high tones only. After that, a white square was displayed in the middle of the screen for a brief period to signify a new trial is about to start. A grey square was then displayed in place of the white one along with the attentional instruction while melodies were being played. A black square, along with a reminder of response keys, then replaced

them as soon as melodies were finished, signalling that they could enter their responses. Participants had no time limit to respond, as the next trial would only start after they responded.

Participants did two sets of training trials for a total of 5-10 minutes. In the first set, training trials were a simplified version of the task, without any mention of attentional focus or instruction to press space when tones were louder. Once this set was finished, they had to do the second one with attentional focus and louder tones. If their accuracy in these sets of training trials was lower than 50%, the program warned the subject that they would be presented with these trials a second time before doing the experiment. The rest of the participants only did these training trials once before doing the main task, which lasted 20-25 minutes. These training trials were used to make participants familiar with the procedure and responses with simplified stimuli. Since the experiment required a lot of attention, they had 2 pauses during the main task.

Once the experiment was over, the experimenter gave written feedback to interested participants explaining the aims, design, and experimental background of the study. The experiment lasted about 35 minutes.

Results

Analysis preparation

Individual responses on perceived differences between sequences were transformed into D-prime scores to obtain a single measure of signal detection for each attention condition while taking into account possible response biases. No data was missing in the dataset. A two-way mixed model ANOVA was conducted on these D-prime scores, with ATTENTIONAL FOCUS (*all tones* vs. *low and high tones* vs. *high tones*) as the within-subject factor and FREQUENCY (*3-3* vs. *9-9*) as the between-subject factor, along with six paired-

samples t-tests (three for each frequency condition) in line with our hypotheses. No correction for multiple comparisons was applied.

Mauchly's test indicated that the assumption of sphericity had not been violated [$\chi^2(2)=0.992$, $p=.837$].

Data analysis

Although inferential statistical tests were conducted on D-prime scores only, Table 4.1 also includes summarizing statistics of proportions of "similar" responses in every experimental condition.

	3-3			9-9		
	All tones	High & low tones	High tones	All tones	High & low tones	High tones
Different	0.688 ± 0.266	0.604 ± 0.229	0.604 ± 0.251	0.697 ± 0.178	0.687 ± 0.214	0.75 ± 0.197
Similar	0.258 ± 0.199	0.275 ± 0.191	0.329 ± 0.278	0.45 ± 0.214	0.503 ± 0.213	0.433 ± 0.248
D-prime	1.214 ± 0.194	0.909 ± 0.149	0.83 ± 0.219	0.681 ± 0.116	0.527 ± 0.09	0.898 ± 0.127

Table 4.1: Mean proportions of "similar" responses and mean d-prime scores across all conditions in both frequency conditions. Reported errors are ± 1 standard deviation

The two-way repeated measures ANOVA (ATTENTIONAL FOCUS*FREQUENCY) for the small frequency differences condition revealed that D-prime scores did not differ as a function of ATTENTIONAL FOCUS [$F(2,86)=2.904$, $p=0.06$, $\eta_p^2=0.063$] nor FREQUENCY [$F(1,43)=2.485$, $p=0.122$, $\eta_p^2=0.055$]. However, the interaction between the two factors proved significant [$F(2,86)=5.291$, $p=0.007$, $\eta_p^2=0.11$].

Six paired-samples t-tests were conducted to decompose the interaction over D-prime scores, three for each FREQUENCY conditions.

In the *3-3* condition, two one-tailed paired-samples t-tests revealed that D-prime scores were significantly higher for the *all tones* condition than for the *low and high tones* condition [$t(19)=2.222$, $p=0.019$, $d_z=0.497$] and the *high tones* condition [$t(19)=2.257$, $p=0.018$, $d_z=0.505$]. However, there was no statistically significant difference in D-prime

scores between the *low and high tones* and the *high tones* conditions [$t(19)=0.482$, $p=0.636$, $d_z=0.108$].

In the *9-9* condition, two one-tailed paired-samples t-tests revealed that D-prime scores were significantly higher for the *high tones* condition than for the *all tones* condition [$t(24)=2.03$, $p=0.027$, $d_z=0.406$] and the *high and low tones* condition [$t(24)=3.085$, $p=0.003$, $d_z=0.617$]. However, there was no statistically significant difference in D-prime scores between the *all tones* and the *high and low tones* conditions [$t(24)=1.249$, $p=0.224$, $d_z=0.25$].

These results are summarized in Figure 4.4.

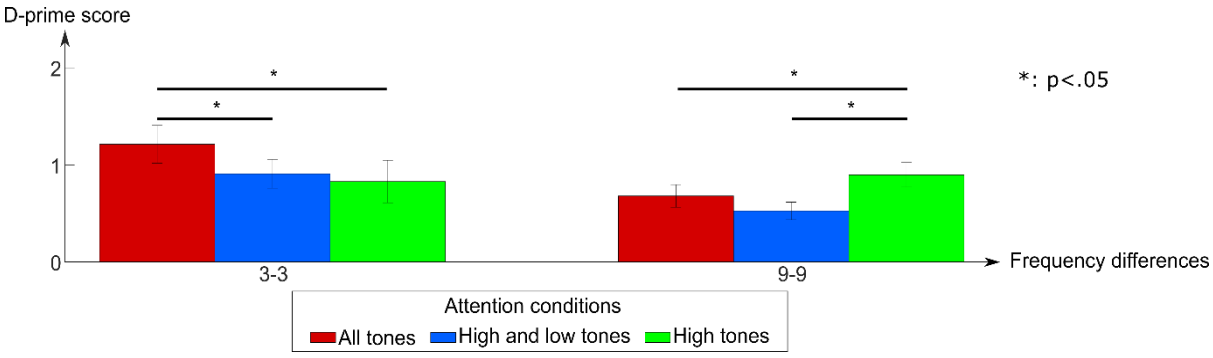


Figure 4.4: D-prime scores as a function of attentional focus and frequency difference. Error bars are ± 1 standard error

Discussion

Hypotheses for the small frequency difference were that directing attention on anything but the entire melody should make subjects have two streams instead of one, decreasing general performance. A stronger impact for the high and low tones attentional focus was expected.

These hypotheses were partially validated. Focusing on whole melodies did indeed display a higher level of performance than the other two attentional conditions. Since in all three conditions, subjects are asked to perform simultaneously the similarity judgement

task and the reaction to loudness task, it seems reasonable to infer that this decrease in performance is due to a difference in the way subjects cluster tones together. However, the attentional focus on low and high tones did not seem to decrease the performance further compared to focusing on the high tones only. The rationale behind this expectation was that if subjects did create a stream comprised only of low and high tones by specifically focusing their attention on those only, then the other stream would only contain middle tones. As tones order information is lost between streams (Bregman & Campbell, 1971), such a stream combination would prevent the similarity judgement from being possible. On the other hand, by focusing on high tones only and therefore creating a stream comprised only of those, the unattended stream would include both low and middle tones, and therefore the order information necessary to complete the task. Although the information is certainly made harder to use by being in an unattended stream, as the decrease in performance compared to the whole melody attention condition seems to show, previous research has shown that order changes can still be detected in such a situation (Thomassen & Bendixen, 2018). The apparent similarity between the high and low attention and the high attention conditions could be explained by the simplicity of the task in this small frequency difference situation. Indeed, despite being explicitly requested to direct their attention on some tones, the task requiring their attention could still be completed while focusing on all tones in every condition. Perhaps the decrease in performance from these two conditions compared to the one asking to listen to whole melodies only shows a marginal tendency to form 2 clusters instead of 1, but that the easiness to complete both tasks by not following the attention instruction flattened the difference. Another possibility is that the stream segregation process is indeed at least partly pre-attentional. Even if subjects did indeed focus their attention on a stream combination where low and high tones are clustered together, the unattended stream combination in which all tones are clustered together may have been built and considered by the perceptual system. Therefore, the information is still harder to access than in the

situation where the focus is on the melody as a whole, but no more than in the situation where a cluster of high tones only is formed.

Hypotheses for the large frequency difference were that subjects would naturally not be easily able to cluster middle tones with either low or high tones, therefore focusing on those only should not decrease the performance. On the contrary, it was expected that focusing on high tones only would increase performance. These hypotheses were validated. Focusing on high tones only yielded better performance than both focusing on the whole melody, or on both the low and high tones. This result can be easily interpreted if there is indeed a hard limitation to a maximum of 2 streams being held simultaneously, and that the stream segregation process relies at least partly on attention. By focusing on high tones, subjects could create a stream comprised of only those, and therefore have another unattended stream within which all tones are fused. Since this stream now contains both low and middle tones, the order information is present within it and the tonal inversion can now be detected. Conversely, whenever the attention is focused on high and low tones, then the order information is present in none of the two streams, making the task harder to complete. When subjects are asked to focus on the whole melody, the large frequency difference makes subjects unable to cluster middle tones with either low or high tones in the attended stream, which is in fact an elaborate reproduction of the usual Bregman & Campbell (1971) study.

This is an important result that makes a convincing argument in favour of a limitation on the number of streams being held together, and a strong influence of attention on the stream formation process. Indeed, this increase in performance cannot be satisfyingly explained by the usual assumption in the ASA literature that the number of streams is unlimited and that their formation process is strictly pre-attentive: The classical ASA model would have predicted that low, middle, and high tones are all in their own stream because of their very different characteristics, and that participants would just have

to pick whichever of these streams they want to focus on. Doing so would then just either not change anything, or elicit a re-analysis of the background tones, which would lead to a segregation into two background streams (one for the low and one for the middle tones). Selecting the high tones should therefore have no impact on performance. Instead, it would seem that focusing on high tones somehow "fuses" low and middle tones together, making the task easier. This directly relates to the experiment from Mack et al. (1992), where unattended visual percepts do not seem to elicit any segregation process. Similarly, Carlyon et al. (2001) have shown in left hemineglect patients that listening to galloping sounds from a classical ASA experiment (L. van Noorden, 1975) with their contralateral ear induces a stream segregation process which is less present in their ipsilateral ear.

There are two main possible consequences for these results. The first one would be that the whole concept of "streams" without attention is flawed: without attention, all elements are simply fused together in a single group. Once a subject tries to combine several elements together, whether voluntarily through top-down attention, or because something surprising or threatening happened in this single group, the perceptual system would use Gestalt principles to allow or prevent the fusion of some elements together, with the idea that these principles are designed to make sure the limited attentional and cognitive resources are being directed to a single, coherent and meaningful event (generated by one source; e.g. focusing on a guitar). Everything absent from this cluster, created hand in hand by both attention and the perceptual system, stays fused in an unattended cluster. Within this hypothesis, the stream formation process necessitates attention and simply does not happen at all without it.

A second possibility is that, even if overall studies agree with the existence of a fusion phenomenon of unattended elements and this experiment shows a clear effect of attention on the way streams are clustered, parts of the stream segregation process could still happen preattentively. Some authors (Thomassen & Bendixen, 2018) have observed

from an EEG study in a similar setting as the present one, that EEG patterns are consistent with *both* stream segregation and stream integration from background elements. Focusing on a stream would therefore indeed create a background stream within which elements are fused. But at the same time, the perceptual system would analyse this background stream for future possible attentional shifts. That being said, our results seem to lean towards the fact that the perceptual system may only focus on individual events, instead of performing complex streaming combinations. If we consider the stream formation process as a source inference problem, the difference between considering an unlimited amount of streams and only two (the attended and the unattended one) would be the same as asking either "what is the most likely combination of sources that produced this complex scene?" or "what elements in this complex scene could likely be produced by a unique source?". In our experiment, this would mean that, for instance, without any focus on the melodies, the perceptual system would independently ask questions such as "could low tones be produced by a single source?" ; "could middle tones be produced by a single source?" ; "could high tones be produced by a single source?" ; "could all tones be produced by a single source?". In the small frequency difference condition, the answer to all these questions is "according to Gestalt laws, yes", which means a listener would be able to focus his attention on this particular stream if he wants to. In the large frequency difference condition, the answer to the last one being "no", attention cannot be focused on a stream containing all tones. However, at no point would the perceptual system try to answer the question "how likely is it that low, middle, and high tones are created by 3 different sources?", probably because the attentional system is unable to work on more than 2 things at the same time anyway (Strobach et al., 2018), so only individual events are of interest. Selecting a stream would then make all other elements irrelevant, and therefore fused; but from this fused stream, the perceptual system would then create a mental representation of all possible streams on which the attention could be shifted to.

Whichever of these hypotheses is true, this experiment also showed that it is possible to behaviourally observe signal change detections from the unattended background. Relying on EEG studies may therefore not always be necessary for such analysis, opening up to new possibilities of experimentation using our paradigm.

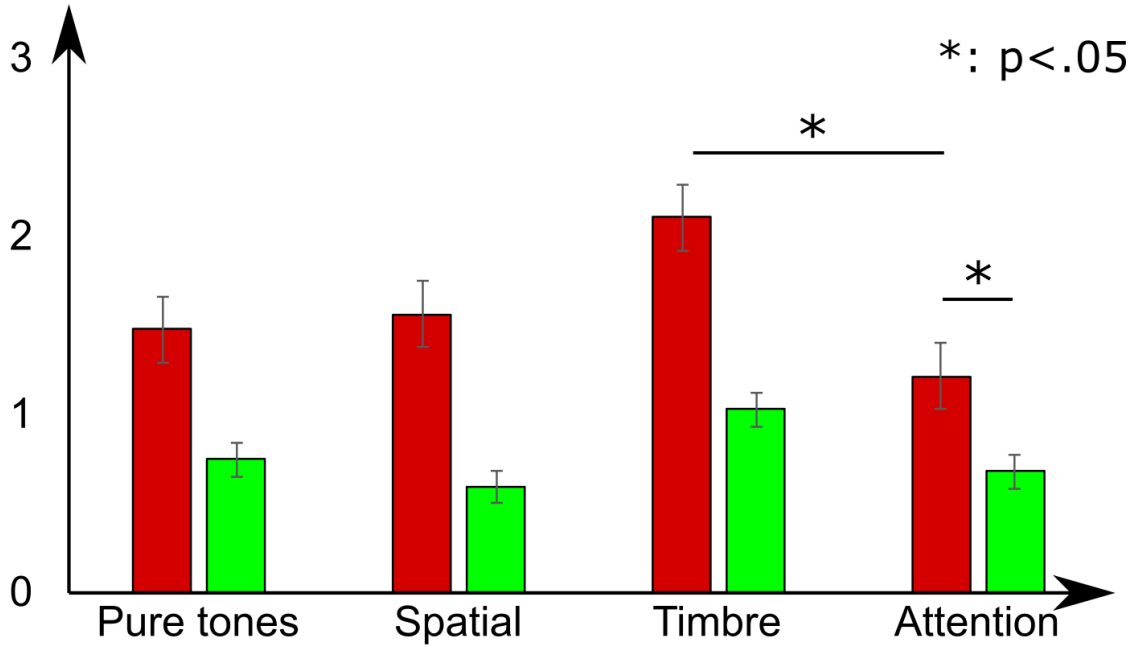
Overall, this study showed the influence of attentional focus on the general stream formation process, as it seems to constrain the perceptual system to work in the background on whatever elements are unattended for, fused in their own stream. Further research should try to investigate whether the perceptual system only works towards mental representations of credible single streams from the unattended cluster, or if it is already making complex analyses of the auditory scenes including several streams held at once.

Interchapter

Cross-experiments comparisons

As was already done between chapters 3 and 4 (see Figure 3.7), comparisons were conducted between similar *3-3* and *9-9* FREQUENCY DIFFERENCES conditions across all experiments. Since during this last experiment those conditions were performed by different subjects, it was not possible to run an ANOVA across all experiments and look for an interaction. However, the usual trend observed in the past three experiments can clearly be seen graphically (see Figure 4.5).

On top of this descriptive analysis, seven independent t-tests were conducted to check for general differences in performance between across experiments, as has been previously observed for the second experiment from Chapter 3 (*Timbre*). The D-prime scores of the *3-3* FREQUENCY DIFFERENCES condition of the *Attention* experiment differed significantly from scores of the *3-3* condition of the *Timbre* experiment [$t(49)=3.468$,



3-3	1.477	1.562	2.105	1.214
9-9	0.748	0.593	1.026	0.680

Figure 4.5: D-prime scores as a function of frequency differences and experiment. Error bars are ± 1 standard error

p=.001, d_z=0.995]. However, they did not differ significantly from the *3-3* condition coming from respectively the *Pure tones* experiment [t(42)=1.067, p=.381, d_z=0.112], nor the *Spatial* experiment [t(46)=1.366, p=.178, d_z=0.4]. Similarly, no significant difference has been observed in the *9-9* FREQUENCY DIFFERENCES conditions across the *Pure tones* experiment [t(47)=0.328, p=.744, d_z=0.096], the *Spatial* experiment [t(51)=0.428, p=.671, d_z=0.118], nor the *Timbre* experiment [t(54)=1.866, p=.068, d_z=0.502].

On top of those first six independent t-tests, another one was conducted to check if the trend observed between the *3-3* and the *9-9* FREQUENCY DIFFERENCES condition for the *Attention* experiment was indeed present, and as was allowed by the observed interaction between the two factors of the experiment (FREQUENCY DIFFERENCES*ATTENTIONAL FOCUS) reported in the previous results section. The t-test showed a significant difference in D-prime scores between those two conditions, as was observed in the three previous experiments [t(43)=2.462, p=.018, d_z=0.739].

The performance increase observed in the *Timbre* experiment does not seem present in this new one. It seems instead that it is once again comparable to what was observed in former studies. Overall, it seems that this second online experiment once again allowed for the replication of the frequency effect on the subjects' performance, which was therefore observed consistently on more than 100 participants in various experimental settings, further validating simultaneously the reliability of our novel paradigm, of the strong effect of frequency differences over how participants clustered tones together, and the online setting (as opposed to the lab setting).

Chapter 5

–

General discussion, conclusions and future directions

The field of Gestalt Psychology has declined in popularity at the end of the last century but is now starting to regain a lot of attention with promising prospects of mathematically stringent models allowing quantitative predictions, as opposed to previous verbal principles that are less accurate. A growing body of literature is trying to investigate the grouping principles mainly in the visual field, but auditory and multisensory grouping have always been slightly less investigated. This thesis presented three studies that investigated the Auditory Scene Analysis (ASA) from a computational point of view to bring a contribution to this timely endeavour.

The first study mainly focused on the design, implementation and application of a custom Bayesian clustering algorithm to classic ASA paradigms. The starting point was to select a few general assumptions concerning the way our perceptual system clusters elements together that were meaningful and compatible with the pre-existing literature. These assumptions had to be implemented in a stringent and coherent mathematical way in order to obtain a fully quantitative model able to capture known behavioural phenomena and to output useful predictions to orient towards new experimental directions.

This endeavour was mainly successful, as the model was a good fit for well-known ASA phenomena, and allowed for fruitful experimental studies based on its predictions, although they were only qualitatively formulated. These were indeed made without running a simulation, but only on the experimenters' knowledge of the model. However, it gives credit to the fact that the assumptions used as a base for the model may indeed be used by the brain.

The first and perhaps the most important of those assumptions was that the perceptual system tries to group elements using a generalized proximity principle. This idea comes from the observation that most Gestalt principles of grouping, although verbally defined as different, can be mathematically derived in the same way whatever the modality they are in. For instance, the well-known "common fate" principle can be

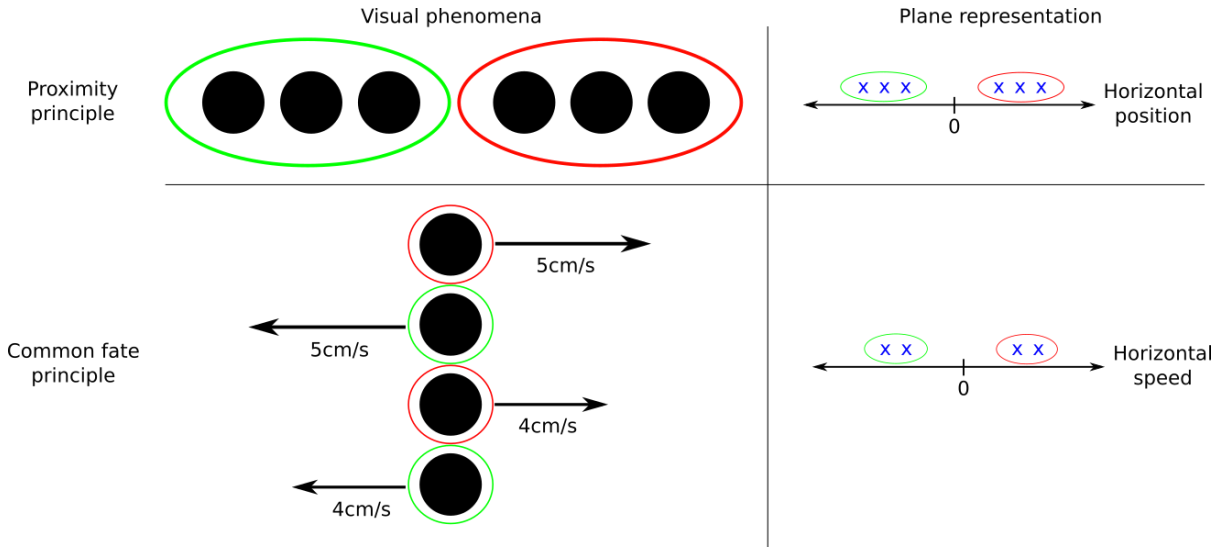


Figure 5.1: The "common fate" principle seen as a generalized proximity principle, in simplified 1-dimensional planes

regarded as a proximity principle applied in an oriented speed plane (see Figure 5.1). The principle of similarity can be modelled using either discrete/qualitative or continuous dimensions depending on the situation. The good continuation principle can be captured by consecutive angles, as was already successfully implemented in another model applied to the visual modality (Froyen et al., 2015b), etc… The model developed in this piece of work was designed to be general enough to be used in different sensory modalities, and ultimately in multisensory settings. For now, it has only been used in the auditory modality, which implies that it regularly used frequency as a grouping variable. The sequential nature of this modality was taken into account in the model by using the proximity of elements as a function of time (using the ratio of the difference in the variable over the difference in time). The likelihood of elements being grouped together followed a univariate Normal distribution. Theoretically, it should be possible to model most, if not all, Gestalt grouping principles in a multisensory setting simply by adding the relevant dimensions and the right parameters (such as the right standard deviations) for the clustering algorithm, effectively transforming the likelihood to a multivariate Normal distribution. This mathematical definition and application of this generalized proximity principle to Gestalt perceptual grouping modelling seems like a very promising direction considering the encouraging results found in our first study.

The other modelling assumptions were that the grouping process was pre-attentive, capable of producing an unlimited number of clusters and that the perceptual system generally tries to keep things as simple as possible by tending to keep the total number of clusters as low as possible (which was a mathematical way of implementing Ockham's razor).

The predictions from this model pushed towards the development of an experimental paradigm designed to go further than the usual 2 streams configurations used in the ASA literature. However, despite being in line with classical ASA assumptions hypotheses, the application of this new paradigm allowing for the observation of 3 simultaneous auditory streams contradicted the model's predictions. Indeed, cumulated differences over several dimensions (i.e. perceptual cues) were originally expected to increase the stream segregation process into a clearly visible 3 streams situation in our second study.

By looking for additive effects of frequency and spatial distances in one experiment, and frequency and timbre distances in another, the second study was unable to observe these effects. Instead, the pattern of results seemed to challenge the assumption of an unlimited number of simultaneous streams held by the perceptual system, along with the purely pre-attentional nature of the grouping process. Instead, it seemed as though attention was possibly playing an active role in the stream formation process, and/or that the total number of streams could be limited to 2: the foreground cluster (the one attended to) and the background cluster, within which no further clustering seemed to happen.

The role of attention in the stream formation process was therefore investigated in the final study. Results from this study seemed to confirm that those specific assumptions in the model may have been superfluous: controlling the participants' attention had a visible effect on the way they formed auditory streams, and all the patterns of results could be interpreted in foreground/background streams situations. However, this interpretation is

not incompatible with the existence of coarse clustering processes made by the perceptual system on the unattended stream. Several mental representations of possible streams could exist at once, including but not limited to the single fused background stream that can influence behaviour, as was suggested in recent literature on ASA (Thomassen & Bendixen, 2018).

It could be argued that these results are specific to the paradigm used, and that it might be possible that more streams are perceived in a more naturalistic setting. Indeed, a listener would be able to simultaneously detect a change in the melody a bird sing, the rhythm of a dog barking, and a tone of a person's voice. While it is a distinct possibility, it is still compatible with a dual-streams situation. As was seen in the literature and in the presented experiments, having several streams does allow to actively focus and work on subparts of the environment, but at the cost of pieces of information. Indeed, the order information is available within streams, but lost between streams. Even if the singing, the barking and the voice were all clustered in a same stream, a salient change in their characteristics could as well be detected as if they were in different stream – just as the order change was detected in an unattended stream during the experiment in Chapter 4. The change in rhythm of the dog's barking could therefore be detected not only relatively to itself, but also to each note produced by the bird. Conversely, if we imagine the worst-case scenario where absolutely every sound is being held in its own unique stream, then we would lose all ability to access the order of those tones, essentially losing the ability to detect the change in rhythm of the dog's barking. In short, having only two streams does not mean that there is no passive monitoring of the background cluster. Quite the contrary: this monitoring process (which would be related to bottom-up attention) could in fact have access to more information than if more clusters were present.

This has several implications on the way this new model can be used and interpreted. First of all, and as can be observed by the results from our first study, this

should have no negative repercussion in its applications in settings where a maximum of only 2 streams can be expected. As was briefly mentioned in chapter 2, the output from the model is a list of credible stream assignments for all elements, along with a credibility estimation for each combination. With the assumption that the process was purely pre-attentional and able to deal with an unlimited number of streams, it was proposed that the perceptual system may work out this full list of credible combinations and only keep the most credible one from which the listener can just select the stream they want to attend to. Knowing that attending to a specific subset of elements changes the way streams are formed, this explanation now seems insufficient.

The classical experiment from Bregman & Cambpell (1971) can be used to illustrate an asymmetry of the phenomenon captured by our model. When sounds are being played slowly, it predicts that all are coming from the same source with high credibility, and conversely, that their coming from two different sources has low credibility. Participants can indeed, in this situation, focus on the whole melody effortlessly. As sounds are being played faster and faster, the credibility of having two different sources gets higher and higher as the credibility of having only one source gets equally lower and lower and participants need more and more effort to keep a focus on all tones at once until it finally becomes impossible. In this configuration, probabilities output by the model can be regarded as a difficulty to maintain the focus on a particular stream combination. The problem is that across the whole experiment, it is always rather easy for any participant to only focus on a subset of the tones (for instance, only on high tones), and therefore create a two streams situation. This means that while a very low probability of one stream reveals that it is impossible to form such a stream, a very low probability of two streams is not particularly meaningful. Furthermore, results from other studies (Carlyon et al., 2001; Sussman, 2017; Thomassen & Bendixen, 2018) and Chapter 4 show that focusing on

something else than those tones could make even very fast galloping tones somehow fused again as they are not attended to.

Our model is not designed to explicitly take attention into account, especially when it comes to background fusion. However, it should be able to predict when subjects will start being unable to cluster elements together in the foreground. Thanks to its application of Ockham's razor, the model is biased towards the fusion of every new element to an already existing cluster. This means that, in order to infer a higher number of clusters than was already inferred, the likeliness of belonging to one of those pre-existing clusters has to be very low. While this has to be verified experimentally, perhaps the complexity of the most likely stream combination output by the model can reveal an impossibility to bring elements together in the foreground. In other words, if the most likely output from the model is a single cluster, then it means that subjects can choose to bring all tones in the foreground at once, or only some of them. But if the most likely output contains three clusters, they can only bring forward one of those or some of their subparts. However, once the subject manages to bring some tones to the foreground, all the others are automatically put in the background stream, where the clustering process happens again.

A general model that stays consistent with all results from this piece of work certainly has to dissociate the streams formed by attention and the mental representations that the perceptual system creates; but also how those two interact. We propose that at any point, two streams are available: the foreground stream (on which attention is actively maintained), and the background stream. Every element in the background stream is fused, but the perceptual system performs rudimentary clustering using Gestalt principles on background elements to determine on which elements the attention could potentially be focused on. All these mental representations coexist and can potentially influence behaviour.

Across all our studies and as was regularly observed (Bregman, 1994), participants reported that they sometimes could increase their attentional effort to maintain bigger streams as the difference between tones increased. Interestingly, this could mean that top-down attention allows us to go, to a certain extent, against what our perceptual system normally allows, by "stretching" its inferential boundaries. Each element being assigned probabilistically to each cluster, lower probabilities to belong to a particular cluster would mean more attentional resources to include the element inside the cluster. All in all, maybe the attended cluster is simply not just selected, but actively constructed. Elements that are closely related (via a generalized proximity law) could just be extremely easy to cluster, as would a very high likelihood of being created by the same source reflect. And as elements are less and less likely to be generated by the same source, clustering them together would require more and more conscious effort, up until a point of impossibility. Conversely, a likelihood of two elements being clustered together that is too high could lead to an

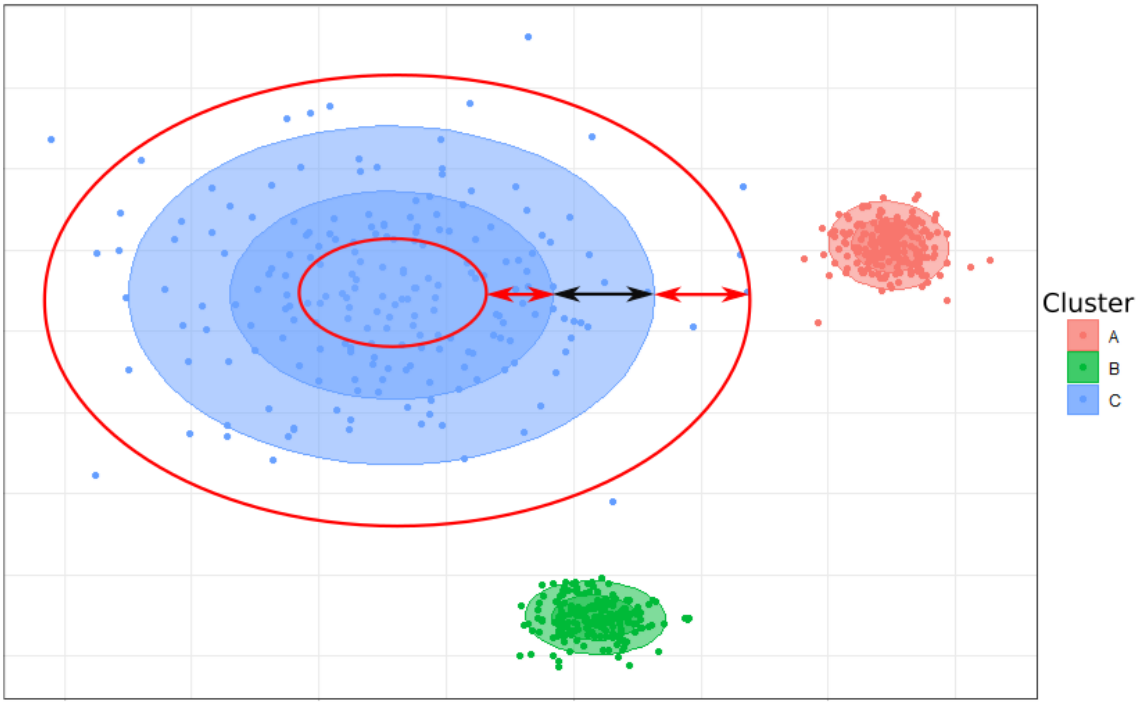


Figure 5.2: visual representation of the foreground stream formation process. While unattended, the perceptual system guessed that the perceptual elements may have been produced by 3 possible sources. When the subject tries to focus his attention on elements from Cluster C, he can do so with easiness in the area represented by the black arrow. However, trying to over-segregate (inner red arrow) or over-fuse (outer red arrow) elements require more and more attentional effort, up until it becomes impossible (red circles)

impossibility to put one in the foreground cluster and the other in the background cluster. This impossibility to separate several elements was already observed in the literature, although it has not been as studied as its counterpart (Bregman, 1994). See Figure 5.2 for a representation of how attention could stretch or reduce the clustering area on a plane similar to the one presented in Chapter 1.

While most of these ideas are just conjectures at this point, they seem to create a coherent model that is consistent both with the data gathered in our studies and the latest discoveries in the ASA and auditory attention literature.

This research project had other beneficial outcomes. Notably, the results from our studies show that the background stream fusion can directly influence performance in an implicit behavioural task, which has never been shown to our knowledge. This could lead to new experiments without needing EEG settings. The last two experiments were conducted fully online using a repository for online experiments (Peirce et al., 2019). This could be considered a limitation because of the weaker control on the experimental setting: it is virtually impossible to know how serious subjects were when doing the experiment, if the experiment ran smoothly for each one of them, if their computer hardware was adequate, etc… However, both of those experiments contained partial replications of former results that were successful, which leads to believe those fears are unfounded. Moreover, this may be one of the best candidates to overcome the usual WEIRD (Western, Educated, Industrialized, Rich and Democratic) sampling bias seen in the majority of behavioural experiments (Henrich et al., 2010), where most participants come from a very similar subpopulation while the intention is often to draw conclusions about the whole human population. This could also allow to increase the number of participants and engage more people in scientific research. Overall, the positive aspects may vastly surpass the negative ones.

The biggest limitation of this piece of work is that although our proposed model is fully quantitative, as of now all of its predictions on experimental settings where more than 2 streams were expected were only qualitative. Further development of the model would require to systematically link its outputs to a behavioural response from our experimental settings. Furthermore, while the model would theoretically be rather easy to derive on more than one dimension using a multivariate Normal distribution as a likelihood function, this remains to be done and experimentally validated.

Overall, this thesis aimed to use computational modelling techniques to make significant advancements and hypotheses in the field of auditory perception and work toward a unified theory and understanding of multisensory perception. The proposed model was successfully able to reproduce key observations in the field and to produce new predictions that led to more behavioural investigation. Even though the predictions were not verified and the model in its present state has shown its limitations, the overall research project led to interesting insights into the Auditory Scene Analysis.

Implementing fully quantitative models is still rather new and difficult in the field of psychology. However, the most powerful and convincing scientific models in the history of science have always been predictive models, as can still be seen in physics where experimental verifications sometimes only happen decades after mathematically derived model predictions. Pursuing this direction can only strengthen the field and our comprehension of the brain.

References

Acerbi, L., & Ma, W. J. (2017). Practical Bayesian Optimization for Model Fitting with Bayesian Adaptive Direct Search. *Advances in Neural Information Processing Systems*, *30*, 1836–1846.

Aldous, D. J. (1985). Exchangeability and related topics. *Lecture Notes in Mathematics*, *1117*, 1–198.

Anstis, S. M., & Saida, S. (1985). Adaptation to auditory streaming of frequency-modulated tones. *Journal of Experimental Psychology: Human Perception and Performance*, *11*(3), 257–271.

Barlow, H. (1961). Possible principles underlying the transformation of sensory messages. *Sensory Communication*, 217–234.

Barsz, K. (1988). Auditory pattern perception: The effect of tonal frequency range on the perception of temporal order. *Perception & Psychophysics*, *43*(3), 293–303. https://doi.org/10.3758/BF03207873

Beauvois, M. W., & Meddis, R. (1997). Time decay of auditory stream biasing. *Perception & Psychophysics*, *59*(1), 81–86. https://doi.org/10.3758/BF03206850

Beierholm, U. R. (2013). Bayesian Models of Perception. In D. Jaeger & R. Jung (Eds.), *Encyclopedia of Computational Neuroscience* (pp. 1–5). Springer. https://doi.org/10.1007/978-1-4614-7320-6_451-2

Blauert, J. (1997). *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press.

Bregman, A. S. (1978). Auditory streaming is cumulative. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(3), 380–387. https://doi.org/10.1037/0096-1523.4.3.380

Bregman, A. S. (1994). *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press.

Bregman, A. S., & Achim, A. (1973). Visual stream segregation. *Perception & Psychophysics, 13*(3), 451–454. https://doi.org/10.3758/BF03205801

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

Bregman, A. S., & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology*, *89*(2), 244–249.

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

Bregman, A. S., & Dannenbring, G. L. (1973). The effect of continuity on auditory stream segregation. *Perception & Psychophysics*, *13*(2), 308–312. https://doi.org/10.3758/BF03214144

Bregman, A. S., & Rudnicky, A. I. (1975). Auditory segregation: Stream or streams? *Journal of Experimental Psychology: Human Perception and Performance*, *1*(3), 263–267. https://doi.org/10.1037/0096-1523.1.3.263

Brochard, R., Drake, C., Botte, M.-C., & McAdams, S. (1999). Perceptual organization of complex auditory sequences: Effect of number of simultaneous subsequences and frequency separation. *Journal of Experimental Psychology. Human Perception and Performance*, *25*(6), 1742–1759. https://doi.org/10.1037//0096-1523.25.6.1742

Bronkhorst, A. W. (2000). The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multiple-Talker Conditions. *Acta Acustica United with Acustica*, *86*(1), 117–128.

Carlyon, R. P., Cusack, R., Foxton, J. M., & Robertson, I. H. (2001). Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(1), 115–127. https://doi.org/10.1037/0096-1523.27.1.115

Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979. https://doi.org/10.1121/1.1907229

Cusack, R., & Roberts, B. (2000). Effects of differences in timbre on sequential grouping. *Perception & Psychophysics*, *62*, 1112–1120.

Dannenbring, G. L. (1976). Perceived auditory continuity with alternately rising and falling frequency transitions. *Canadian Journal of Psychology*, *30*(2), 99–114.

Dannenbring, G. L., & Bregman, A. S. (1976). Effect of silence between tones on auditory stream segregation. *The Journal of the Acoustical Society of America*, *59*(4), 987–989.

Demany, L. (1982). Auditory stream segregation in infancy. *Infant Behavior and Development*, *5*(2), 261–276. https://doi.org/10.1016/S0163-6383(82)80036-2

Eramudugolla, R., McAnally, K. I., Martin, R. L., Irvine, D. R. F., & Mattingley, J. B. (2008). The role of spatial location in auditory search. *Hearing Research*, *238*(1), 139–146. https://doi.org/10.1016/j.heares.2007.10.004

Fetsch, C. R., DeAngelis, G. C., & Angelaki, D. E. (2013). Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons. *Nature Reviews. Neuroscience*, *14*(6), 429–442. https://doi.org/10.1038/nrn3503

Fetsch, C. R., Pouget, A., DeAngelis, G. C., & Angelaki, D. E. (2011). Neural correlates of reliability-based cue weighting during multisensory integration. *Nature Neuroscience*, *15*(1), 146–154. https://doi.org/10.1038/nn.2983

Froyen, V., Feldman, J., & Singh, M. (2015a). Bayesian hierarchical grouping: Perceptual grouping as mixture estimation. *Psychological Review*, *122*(4), 575–597. https://doi.org/10.1037/a0039540

Froyen, V., Feldman, J., & Singh, M. (2015b). Bayesian Hierarchical Grouping: Perceptual grouping as mixture estimation. *Psychological Review*, *122*(4), 575–597. https://doi.org/10.1037/a0039540

Gallace, A., & Spence, C. (2011). To what extent do Gestalt grouping principles influence tactile perception? *Psychological Bulletin*, *137*(4), 538–561. https://doi.org/10.1037/a0022335

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition* (3 edition). Chapman and Hall/CRC.

Gershman, Samuel J., & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*(1), 1–12. https://doi.org/10.1016/j.jmp.2011.08.004

Gershman, Samuel Joseph, & Niv, Y. (2013). Perceptual estimation obeys Occam's razor. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00623

Ghahramani, Z. (2013). Bayesian non-parametrics and the probabilistic approach to modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *371*(1984), 20110553. https://doi.org/10.1098/rsta.2011.0553

Hájek, A. (2012). Interpretations of Probability. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2012). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2012/entries/probability-interpret/

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29–29. https://doi.org/10.1038/466029a

Hoffman, D. D., Singh, M., & Prakash, C. (2015). The Interface Theory of Perception. *Psychonomic Bulletin & Review*, *22*(6), 1480–1506. https://doi.org/10.3758/s13423-015-0890-8

Hollensteiner, K. J., Pieper, F., Engler, G., König, P., & Engel, A. K. (2015). Crossmodal integration improves sensory detection thresholds in the ferret. *PloS One*, *10*(5), e0124952. https://doi.org/10.1371/journal.pone.0124952

International Organization for Standardization. (2016). Occupational health and safety management systems—Requirements with guidance for use (ISO/DIS Standard No. 226:2003). Retrieved from https://www.iso.org/standard/34222.html

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Jäkel, F., Singh, M., Wichmann, F. A., & Herzog, M. H. (2016a). An overview of quantitative approaches in Gestalt perception. *Vision Research*, *126*, 3–8. https://doi.org/10.1016/j.visres.2016.06.004

Jonas, E., & Körding, K. P. (2017). Could a Neuroscientist Understand a Microprocessor? *PLOS Computational Biology*, *13*(1), e1005268. https://doi.org/10.1371/journal.pcbi.1005268

Kaya, E. M., & Elhilali, M. (2017). Modelling auditory attention. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1714). https://doi.org/10.1098/rstb.2016.0101

Klapp, S. T., & Jagacinski, R. J. (2011). Gestalt principles in the control of motor action. *Psychological Bulletin*, *137*(3), 443–462. https://doi.org/10.1037/a0022361

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in psychtoolbox-3. *Perception*, *36*(14), 1–16.

Knill, D. C., & Richards, W. (1996). *Perception as Bayesian Inference*. Cambridge University Press.

Knill, D. C., & Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Research*, *43*(24), 2539–2558.

Koffka, K. (1935). *Principles of Gestalt psychology* (p. 720). Harcourt, Brace.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal Inference in Multisensory Perception. *PLOS ONE*, *2*(9), e943. https://doi.org/10.1371/journal.pone.0000943

Kruschke, J. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan* (2 edition). Academic Press.

Lee, M. D., & Habibi, A. (2009). A Cyclic Sequential Sampling Model of Bistable Auditory Perception. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Shonmaker (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2669–2674).

Lee, S.-H., & Blake, R. (1999). Visual Form Created Solely from Temporal Structure. *Science*, *284*(5417), 1165–1168. https://doi.org/10.1126/science.284.5417.1165

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review*, *66*(2), 81–95. https://doi.org/10.1037/h0043178

Mack, A., & Rock, I. (1998). *Inattentional Blindness*. https://doi.org/10.7551/mitpress/3707.001.0001

Mack, A., Tang, B., Tuma, R., Kahn, S., & Rock, I. (1992). Perceptual organization and attention. *Cognitive Psychology*, *24*(4), 475–501. https://doi.org/10.1016/0010-0285(92)90016-U

Marr, D., & Poggio, T. (1976). *From Understanding Computation to Understanding Neural Circuitry*. https://dspace.mit.edu/handle/1721.1/5782

McAnally, K. I., & Martin, R. L. (2007). Spatial Audio Displays Improve the Detection of Target Messages in a Continuous Monitoring Task. *Human Factors*, *49*(4), 688–695. https://doi.org/10.1518/001872007X215764

McCabe, S. L., & Denham, M. J. (1997). A model of auditory streaming. *The Journal of the Acoustical Society of America*, *101*(3), 1611. https://doi.org/10.1121/1.418176

McDermott, J. H., & Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron*, *71*(5), 926–940. https://doi.org/10.1016/j.neuron.2011.06.032

McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite Mixture Models. *Annual Review of Statistics and Its Application*, *6*(1), 355–378. https://doi.org/10.1146/annurev-statistics-031017-100325

Montoro, P. R., Luna, D., & Ortells, J. J. (2014). Subliminal Gestalt grouping: Evidence of perceptual grouping by proximity and similarity in absence of conscious perception. *Consciousness and Cognition*, *25*, 1–8. https://doi.org/10.1016/j.concog.2014.01.004

Moore, B. C. J., & Gockel, H. E. (2012). Properties of auditory stream formation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1591), 919–931. https://doi.org/10.1098/rstb.2011.0355

Orbanz, P., & Teh, Y. W. (2010). *Bayesian Nonparametric Models*. Springer.

Palmer, S. (1992). Common region: A new principle of perceptual grouping. *Cognitive Psychology*, *24*(3), 436–447.

Palmer, S., & Rock, I. (1994). Rethinking perceptual organization: The role of uniform connectedness. *Psychonomic Bulletin & Review*, *1*(1), 29–55. https://doi.org/10.3758/BF03200760

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made

easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Qian, Y., Weng, C., Chang, X., Wang, S., & Yu, D. (2018). Past review, current progress, and challenges ahead on the cocktail party problem. *Frontiers of Information Technology & Electronic Engineering*, *19*(1), 40–63. https://doi.org/10.1631/FITEE.1700814

Roberts, B., Glasberg, B. R., & Moore, B. C. J. (2008). Effects of the build-up and resetting of auditory stream segregation on temporal discrimination. *Journal of Experimental Psychology. Human Perception and Performance*, *34*(4), 992–1006. https://doi.org/10.1037/0096-1523.34.4.992

Rohe, T., & Noppeney, U. (2015). Cortical Hierarchies Perform Bayesian Causal Inference in Multisensory Perception. *PLoS Biology*, *13*(2). https://doi.org/10.1371/journal.pbio.1002073

Roweis, S. T. (2001). One Microphone Source Separation. *Advances in Neural Information Processing Systems*, *13*, 793—– 799.

Russell, S., & Norvig, P. (2016). *Artificial Intelligence: A Modern Approach* (3 edition). Pearson.

Sekuler, A. B., & Bennett, P. J. (2001). Generalized common fate: Grouping by common luminance changes. *Psychological Science*, *12*(6), 437–444. https://doi.org/10.1111/1467-9280.00382

Shamma, S., Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., Pressnitzer, D., Yin, P., & Xu, Y. (2013). Temporal Coherence and the Streaming of Complex Sounds. *Advances in Experimental Medicine and Biology*, *787*, 535–543. https://doi.org/10.1007/978-1-4614-1590-9_59

Shamma, S., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences*, *34*(3), 114–123. https://doi.org/10.1016/j.tins.2010.11.002

Shams, L., & Beierholm, U. R. (2010). Causal inference in perception. *Trends in Cognitive Sciences*, *14*(9), 425–432. https://doi.org/10.1016/j.tics.2010.07.001

Shams, L., Ma, W. J. W. J., & Beierholm, U. (2005). Sound-induced flash illusion as an optimal percept. *Neuroreport*, *16*(17), 1923–1927.

Strobach, T., Wendt, M., & Janczyk, M. (2018). Editorial: Multitasking: Executive Functioning in Dual-Task and Task Switching Situations. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00108

Summerfield, Q. (1992). Lipreading and Audio-Visual Speech Perception. *Philosophical Transactions: Biological Sciences*, *335*(1273), 71–78. JSTOR.

Sussman, E. S. (2017). Auditory Scene Analysis: An Attention Perspective. *Journal of Speech, Language, and Hearing Research: JSLHR*, *60*(10), 2989–3000. https://doi.org/10.1044/2017_JSLHR-H-17-0041

Thomassen, S., & Bendixen, A. (2018). Assessing the background decomposition of a complex auditory scene with event-related brain potentials. *Hearing Research*, *370*, 120–129. https://doi.org/10.1016/j.heares.2018.09.008

Tiippana, K. (2014). What is the McGurk effect? *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00725

Tougas, Y., & Bregman, A. S. (1985). Crossing of auditory streams. *Journal of Experimental Psychology: Human Perception and Performance*, *11*(6), 788–798. https://doi.org/10.1037/0096-1523.11.6.788

Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *353*(1373), 1295–1306.

Turner, B. M., & Sederberg, P. B. (2014). A Generalized, Likelihood-Free Method for Posterior Estimation. *Psychonomic Bulletin & Review*, *21*(2), 227–250. https://doi.org/10.3758/s13423-013-0530-0

Turner, R. E. (2010). *Statistical models for natural sounds* [PhD Thesis]. University College London.

Turner, R. E., & Sahani, M. (2011). Probabilistic amplitude and frequency demodulation. In J. Shawe-Taylor, R. S. Zemel, P. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing* (pp. 981–989). Red Hook.

van Noorden, L. (1975). *Temporal coherence in the perception of tone sequences* [PhD Thesis]. University of Technology, Eindhoven, The Netherlands.

van Noorden, L. P. A. S. (1977). Minimum differences of level and frequency for perceptual fission of tone sequences ABAB. *The Journal of the Acoustical Society of America*, *61*(4), 1041–1045. https://doi.org/10.1121/1.381388

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A Century of Gestalt Psychology in Visual Perception I. Perceptual Grouping and Figure-Ground Organization. *Psychological Bulletin*, *138*(6), 1172–1217. https://doi.org/10.1037/a0029333

Wang, D. (1996). Primitive auditory segregation based on oscillatory correlation. *Cognitive Science*, *20*(3), 409–456. https://doi.org/10.1016/S0364-0213(99)80011-1

Warren, R. M., & Gregory, R. L. (1958). An auditory analogue of the visual reversible figure. *The American Journal of Psychology*, *71*(3), 612–613.

Warren, R. M., Obusek, C. J., Farmer, R. M., & Warren, R. P. (1969). Auditory sequence: Confusion of patterns other than speech or music. *Science (New York, N.Y.)*, *164*(3879), 586–587.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*(6), 598–604. https://doi.org/10.1038/nn858

Weisswange, T. H., Rothkopf, C. A., Rodemann, T., & Triesch, J. (2011). Bayesian Cue Integration as a Developmental Outcome of Reward Mediated Learning. *PLOS ONE*, *6*(7), e21575. https://doi.org/10.1371/journal.pone.0021575

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt. II. *Psychologische Forschung*, *4*(1), 301–350. https://doi.org/10.1007/BF00410640. Translation published in Ellis, W. (1938). *A source book of Gestalt psychology* (pp. 71-88). London: Routledge & Kegan Paul.

Winkler, I., Czigler, I., Sussman, E., Horváth, J., & Balázs, L. (2005). Preattentive binding of auditory and visual stimulus features. *Journal of Cognitive Neuroscience*, *17*(2), 320–339. https://doi.org/10.1162/0898929053124866

Wood, F., Goldwater, S., & Black, M. J. (2006). A non-parametric Bayesian approach to spike sorting. *Conference Proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, *1*, 1165–1168. https://doi.org/10.1109/IEMBS.2006.260700