Experienced Listeners

Bastiaan van der Weij

Experienced Listeners

Modeling the influence of long-term musical exposure on rhythm perception

ILLC Dissertation Series DS-2020-12

Institute for Logic, Language and Computation

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

Printed and bound by GVO printers & designers, Ede.

ISBN: 978-94-6332-664-3

Experienced Listeners

Modeling the influence of long-term musical exposure on rhythm perception

Academisch Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op donderdag 1 oktober 2020, te 13.00 uur

door

Bastiaan Jan van der Weij

geboren te Amsterdam

Promotiecommisie

Promotor:	Prof. dr. H.J. Honing	Universiteit van Amsterdam
Copromotor:	Dr. M.T.P. Pearce	Queen Mary University of London
Overige leden:	Dr. J.A. Burgoyne	Universiteit van Amsterdam
	Prof. dr. J.J.E. Kursell	Universiteit van Amsterdam
	Prof. dr. J.M. London	Carleton College
	Prof. dr. K. Sima'an	Universiteit van Amsterdam
	Dr. W.H. Zuidema	Universiteit van Amsterdam

Faculteit der Geesteswetenschappen

Contents

Acknowledgments

Just like the listeners described in this thesis, my actions cannot be seen separate from my environment. For having been able to produce this thesis, I would first and foremost like to thank Henkjan Honing, my main supervisor. He encouraged me to not only make things but also to investigate them, and he gave me enough freedom to develop my own ideas, while challenging me to turn them into questions. Without Henkjan, I would have stuck to the safety and comfort of writing computer programs, rather than daring to ask questions and attempting to answer them. I would also like to deeply thank Marcus Pearce, whose close involvement with the project began during a wonderful and inspiring visit to his lab in autumn 2015. I'm very happy to have found him willing to join the project as my co-supervisor and I'm thankful for his enthusiastic support, his mentorship, and the boost he gave to the project.

Thanks are due to Ashley Burgoyne, Julia Kursell, Justin London, Khalil Sima'an, and Jelle Zuidema for kindly agreeing to serve on my doctorate committee and for taking the time to read this thesis.

For inspiring me to pursue a PhD in music cognition, there are two people I want to thank in particular. Remko Scha, who I unfortunately can no longer thank in person, kindly supervised my bachelor's project and later mentored me while I was writing my research proposal. His fascinating ideas, which masqueraded as gentle musings on (artificial) art, computational linguistics, and artificial intelligence and his general attitude toward research were a great inspiration to me. For similar reasons, I'm thankful for the mentorship provided by Mark Steedman, who supervised my master's thesis. I fondly remember our regular discussions in his office in Edinburgh and have been inspired by the genuine excitement for research he showed. The willingness of Remko and Mark to take my (probably naive) ideas serious and the enjoyment I received out of doing these projects played a very large role in motivating me to pursue a PhD.

I also want to thank the members of the music cognition group in Amsterdam, Ashley, Berit, Carlos, Fleur, Joey, and Maki, for being great colleagues over the past years. Thank you for the fun group meetings, social outings, and interesting reading groups.

This thesis has benefitted from conversations and discussions with many people. In particular, I want to credit Peter Harrison for great conversations about research and life, Jelle Bruineberg for showing me glimpses of an exciting world of ideas that smoothly interpolate between philosophy of mind and physics, and the organizers of the Worlding the Brain conference for successfully building bridges between neuroscience and the humanities. The formal parts of this thesis have benefitted from the generous and patient help of Frederik Möllerström Lauridsen, Nadine Theiler, and Levin Hornischer. Thanks to Peter, Frederik, Levin, and Thomas Brochhagen for commenting on earlier versions of parts of this thesis. Thanks to Jan-Willem van der Weij for proofreading the Dutch translation of the thesis summary. No one but me is to blame for any remaining issues.

The typesetting of this thesis and the research in it has been performed almost entirely using free and open-source software. I want to thank and acknowledge those who have developed or contributed to such software for making their work available to be used, modified and improved for the benefit of anyone anywhere in the world.

A shout-out to three fantastic flatmates, Mrinalini, Frederik, and Levin, is in order. Mrinalini's taste for music and life provided much-needed relief from academic seriousness. The late-night conversations with Frederik about music, philosophy, law, and the joys of good pilsner really brightened up the last few years. In the category of companions while working from home during a lockdown, I could not have wished for a better one than Levin. Cheers to Sophie, Annika, and Maki for being great and patient climbing buddies to someone who often could not decide until the last moment whether the night should be spent working or climbing.

Life at the ILLC was made memorable by a bunch of excellent people I met there over the years. Thanks for microwave-side chats, the coffees, the brunches and lunches, the drinks, and the kindness. In particular, I want to mention Nadine, Arnold, Dieuwke, Thomas, Iris, Ronald, and Giovanni, who became dear friends. I would like to thank Carlos for being great company for many years in one of the ILLC's brutalist offices and for enduring my loud typing. For the family-like atmosphere they created, I want to thank the ILLC administrative staff.

For, among many other things, enduring my frequent states of sorrow over the past years, I would like to thank my friends. Thanks for the winter adventures in Scotland, concerts, cocktails, island parties, climbing trips, companionship, wine sampling, and rediscovery of Amsterdam. (And, of course, to more in the future!)

Most importantly, I want to thank my family. Their love, support, and trust in me during this project and before has been heartwarming. Dankjewel lieve Nannie, Jan-Willem en Lisa.

Amsterdam
August, 2020

Bastiaan van der Weij

List of publications

Chapter 2 of this thesis will be published as

van der Weij, B., Pearce, M. T., & Honing, H. (submitted). Computational modeling of rhythm perception and the role of enculturation. In *Oxford Handbook of Music and Corpus Studies*.

Author contributions: BvdW wrote the chapter and conceived of the idea for the chapter together with HH. HH and MTP provided feedback to the chapter and suggested revisions.

Chapter 6 of this thesis has previously been published as

van der Weij, B., Pearce, M. T., & Honing, H. (2017). A probabilistic model of meter perception: simulating enculturation. *Frontiers in Psychology*, 8 (824).

Author contributions: BvdW and MTP conceived of the model. BvdW, HH, and MTP designed the simulations. BvdW implemented the model, performed the simulations and analyses, and wrote the paper. MTP and HH provided assistance with the implementation of the model, provided feedback to the paper and suggested revisions.

Chapter 1

Introduction

The way that musical rhythm sounds to us—the result of our *rhythm perception*—is contingent on our history of experiences and interactions with music (London, 2012). In this thesis, I develop and apply computational modeling techniques in an attempt to better understand how these experiences and interactions contribute to rhythm perception.

It is not straightforward to define what exactly is being perceived when we hear the rhythm of music, nor is it immediately clear how our understanding of rhythm perception can benefit from developing computational models, or what such models should look like. Below, I will first give a brief characterization of what the term "rhythm perception" refers to and point out which of its aspects are relevant to this thesis. Then, to contextualize the approach taken in this thesis, I discuss the role that computational models of cognition might play in understanding perception and cognition and discuss different conceptions of what such models might look like.

Rhythm perception can be decomposed into several aspects that contribute to the experience of the listener. At a basic level, time intervals between events in a rhythm (such as the beginning of a note) are *categorized* into a small number of perceptual categories (Clarke, 1987; Desain & Honing, 2003; Jacoby & McDermott, 2017): the listener perceives for example that one note is half the length of the previous note, or that it is three times the length of the next note, even when actual ratios between the time intervals do not precisely reflect this.

Beat and *meter* are perceived regularities in a rhythm related to listeners' ability to coordinate their movements with rhythms (Repp, 2005; Repp & Su, 2013). The perceived regularity causes some events in a rhythm to become "marked for

consciousness" (Cooper & Meyer, 1960, p. 8) and tends to persist in the "mind and musculature of the listener" (Cooper & Meyer, 1960, p. 3), even in the absence of sounded notes that reinforce it. While beat (or sometimes *pulse*) perception refers to the perception of one level of regularity, meter perception refers to the perception of multiple levels of regularity, resulting in a recurring pattern of strong and weak beats.

Tempo is related to a sense of motion that is perceived in a rhythm, making it possible for a rhythm to sound as if it is speeding up or slowing down. *Timing* relates to small adjustments in the timing of notes that can be used expressively and makes it possible for the same rhythm to be played in different ways that listeners may perceive as, for example, "laid back", "rushed", or "stiff" (Iyer, 1998; Honing & Bouwer, 2019).

The above characterization of the main aspects of rhythm cognition is very condensed (for a more elaborate description, see Honing & Bouwer, 2019) but serves to provide a sense of the variety of perceived aspects of rhythms. This thesis is concerned only with a narrow sense of rhythm perception, namely the perception of meter in what might be called a categorized representation of time intervals in rhythms: symbolic representations corresponding to the different note values found in Western music notation. Often when we use the term "rhythm perception" in this thesis, we mean to refer to this narrow interpretation of the term.

That brings us to computational cognitive modeling, to which I will devote a slightly longer discussion. While generally considered a worthwhile endeavor, it is not easy to define precisely the roles that computational modeling can play in furthering understanding of cognition. This is partly because it is not entirely agreed upon what a model of cognition should look like.

A dominant tradition in cognitive science considers computer models to be theories of the "computations" performed by the mind in order to bring about the perceptions of the world of which we are aware. These computations can be described independently, without considering the particular way in which they are implemented, as formulated most famously by Marr (1982). Christopher Longuet-Higgins, who is credited for having first used the term "cognitive science" (Hünefeldt & Brunetti, 2004), describes an example of this perspective in a fictional conversation between a physicist and a biologist in which the biologist tells the physicist: "Ask yourself, what kind of thing do we really want to know about the brain? I suggest that what we would like is a detailed account, among other things, of the 'software' " (Longuet-Higgins, 1981, p. 12). This line of thinking licenses computer programs to serve as explanations of the mental operations that give rise to perception. A cognitive model, in this view, is a precise description of the information-processing steps involved in a certain task, mapping a set of inputs to a certain set of outputs. This reasoning has intimately connected the

field of cognitive science to early artificial intelligence.

A compelling motivation for using computational models to formulate theories of perception and cognition is that it "sets new standards of precision and detail in the formulation of models of cognitive processes" (Longuet-Higgins, 1973, 1987, p. 46). The very process of trying to formulate a precise and working solution to some problem often leads to a deeper understanding of the problem itself, as well as the proposed solutions. More recent echoes of this line of thinking can be found in characterizations of the role of computational models in cognitive science as "reverse engineering" the mind (Tenenbaum, Kemp, Griffiths, & Goodman, 2011, p. 1279).

Another motivation relates to a certain frustration with theories that, often based closely on empirical observations, could be said to *describe* perception but not to explain *how* perception actually works, or *why* it works in that way (Longuet-Higgins, 1981, 1973, 1987; Marr, 1982). How, for example, is a person with arms, hands, and a brain (among other body parts) to know where the coffee cup is on the table and how to grasp for it? To identify the physiological processes involved in this task, such mechanisms involved in the activation of muscles and increased blood flow to certain areas of the brain, does not by itself answer this question. Computer models, theories precise enough to be implemented as such, or mathematical models provide an attractive framework in which possible answers to how and why questions about cognition can be formulated. The neuroscientist Horace Barlow once made the following analogy, which elegantly illustrates the necessity of understanding a complex system at multiple levels: [1]

> A wing would be a most mystifying structure if one did not know that birds flew. One might observe that it could be extended a considerable distance, that it had a smooth covering of feathers with conspicuous markings, that it was operated by powerful muscles, and that strength and lightness were prominent features of its construction. These are important facts, but by themselves they do not tell us that birds fly. Yet without knowing this, and without understanding something of the principles of flight, a more detailed examination of the wing itself would probably be unrewarding. (Barlow, 1961, p. 217).

Returning to music perception, a perceptual "problem" in rhythm perception may be stated in computational terms as follows: how does a trained musician, tasked with transcribing a melody they are hearing for the first time, know where to place the bar lines and which time signature to place at the beginning of the staff (loosely after Longuet-Higgins & Lee, 1982)? While concepts like bar lines

[1] Marr (1982) later, and more famously, formulated a strikingly similar analogy: "trying to understand perception by studying only neurons is like trying to understand to understand bird flight by studying only feathers: it just cannot be done" (p. 27).

and time signatures are foreign to most listeners who have not received formal music education, listeners are usually at some level aware of the difference between a march and a waltz. Longuet-Higgins and Lee (1982) draw a comparison to language: just like being able to speak a language grammatically does not require the ability to describe its structure, so too does the perception of melodies not require the ability to read and write music. Concepts like bar lines and time and key signatures are thus proposed to be descriptions of the intuitive understanding that listeners have of music, similar to how a linguist might use a syntax tree to represent a person's intuitive grammatical understanding of a sentence. A theory of a listener's rhythmic understanding of a melody might therefore take the form of a computer program that is able to perform the transcription task of the trained musician described above. A significant part of the research in modeling rhythm perception can be seen as addressing variants of this problem (e.g., Longuet-Higgins and Steedman, 1971; Longuet-Higgins and Lee, 1982; Lerdahl and Jackendoff, 1983; Temperley and Sleator, 1999; Temperley, 2007; see also Temperley, 2013).

The elegant and clear formulations in which Longuet-Higgins and colleagues defined various computational problems involved in music perception have inspired a vast amount of research and led to many valuable insights that played a crucial role in bringing about the area of research that today is known as music cognition. A significant part of this research focused on a search for "the rules" of beat and meter perception (Povel & Essens, 1985; Lerdahl & Jackendoff, 1983; Steedman, 1977; Longuet-Higgins & Steedman, 1971; Longuet-Higgins & Lee, 1982, 1982; Lee, 1991; Desain & Honing, 1999). However, among the insights of this research is perhaps the realization that it is very hard to find a set of rules whose results agree precisely with the musical intuitions of trained musicians for any input. There always appear to be cases where the rules do not quite produce the desired result (see for example Lee, 1991).

The computational approach applied by these models represents a particular kind of view of cognition that is introspective and represents perception as *information processing*, involving computations whose goal it is to produce the right representations from sensory input. This way of thinking has spilled over from early artificial intelligence, in which symbol representation and heuristic search were dominant (Newell & Simon, 1976). Brooks (1991a) argued that this dominance is mainly a result of technological constraints of the time but has biased the field to consider thought and reason, and not perception and motor skills, as the aspects of intelligence most worthy of investigation. Philosophers have called the information processing view of perception and cognition "cognitivism" (see e.g., Anderson, 2003). They have argued that cognitivist approaches run the risk of losing sight of the role that both the environment and the body play in perception and cognition by focusing narrowly on cognitive mechanisms in the brain.

A very different view has been described by J. J. Gibson (1979), who argued

that perception involves active exploration of the environment by walking around, touching, hearing, and moving one's head and eyes. These actions cause sensory stimulation to change as a function of one's movements in a way that is informative about the environment, implying that perception and action are inextricably intertwined. Instead of focusing on how various cognitive mechanisms might infer information about the environment from sensory input, Gibson dedicates almost half of his book on vision to a description of the *environment* and how it is specified by stimulation of the sensory surfaces of animals like humans. Inspired in part by these views, some philosophers and cognitive scientists have stressed that cognition is *embodied* and that describing it in abstract terms that are independent from the body and the environment misses out on important parts of the picture (Chemero, 2009; Wilson & Golonka, 2013).

Music perception is often described in entirely abstract terms that sometimes appear to bear no connection to physical reality. What, for example, *is* meter, besides a regular pattern of strong and weak beats (Lerdahl & Jackendoff, 1983), or a grammar from which a rhythm is generated (as proposed by Longuet-Higgins & Lee, 1984). Iyer (1998) has elaborately described how aspects of rhythm perception may be grounded in physical embodied interaction with the environment. The observation that "musical motion is, first and foremost, audible human motion" (p. 25) by which Iyer describes a point made by Shove and Repp (1995), compactly summarizes that what we are aware of when we perceive music is likely to be much more than abstract structure.

Two characteristics of early cognitivist models of rhythm perception stand out in this light. First, they often focus on *theory*. For example, Longuet-Higgins and Lee's (1984) account of metrical interpretation is based on a computational formulation of the music-theoretic concept of syncopation. Lerdahl and Jackendoff (1983) propose a very detailed (but informal and not strictly computational) set of well-formedness rules that specify constraints on perceived structure in music, and a set of preference rules that describe which properties of music tend to give rise to which interpretations. Temperley and Sleator (1999) and Temperley (2007), in turn, propose computational models of music perception constrained strongly by the theoretical ideas of Lerdahl and Jackendoff (1983). Second, cognitivist models tend to pay little attention to the ways in which music perception might be shaped by training, practice, and experience in different musical environments. This is perhaps understandable, because slow and gradual shaping of perception through long-term exposure is not easily captured by symbolic rules.

The lack of consensus, mentioned earlier, about the "rules" for beat and meter perception may arguably be seen as symptomatic of the cognitivist approach. Van Gelder (1995) showed that the paradigm of symbolic computation is not the only way in which cognition can be described; for some phenomena, the language of dynamical systems theory seems much more appropriate. If beat and meter

perception is more appropriately described as a dynamical system, as some have argued (e.g. Eck, Gasser, & Port, 2000; Large & Kolen, 1994; McAuley, 1995), then it is not surprising that it is difficult to find a set of rules that adequately describe it. Furthermore, rhythm perception has been argued (Iyer, 1998; London, 2012; Clayton, 2000) and demonstrated (Stobart & Cross, 2000; Hannon & Trehub, 2005b; Soley & Hannon, 2010; Hannon, Soley, & Ullal, 2012; Jacoby & McDermott, 2017; Polak et al., 2018) to be shaped by practice, experience, and training that results from embedding in a musical environment (referred to as *enculturation* in this thesis). This suggests that even if rules can appropriately describe rhythm perception, there may not be a *single* set of rules that can fully account for it. [2]

There is a class of rhythm perception models that are formulated as dynamical systems. These models describe beat and meter perception as a form of coupled oscillation (Large & Kolen, 1994; McAuley, 1995), or as resonance in dynamical models of neural networks (Large, Herrera, & Velasco, 2015). Such approaches do not propose rules by which symbolic representations are inferred from input. Instead, they propose a mathematical description of a physical phenomenon as an explanation of beat and meter perception (Large, 2010b). However, while these models are to some extent capable of simulating effects that enculturation might have on rhythm perception (Large et al., 2015), the degree to which patterns and regularities in rhythms can influence the behavior of these models is at the moment limited.

A primary contribution of this thesis is a probabilistic generative model of rhythm perception that can learn from patterns and regularities in empirical samples of rhythms that represent musical environments. Although there are significant differences between probabilistic models and rule-based models, the work presented in this thesis is in many ways unashamedly cognitivist. It trades primarily in abstract representations of rhythm and meter and does not explicitly make contact with what is arguably one of the defining characteristics of rhythm: its ability to inspire and induce movement. However, there is one important sense in which it takes inspiration from the ecological and embodied views of cognition mentioned above: we argue that rhythm perception cannot be understood independently from the musical environment by which it has been shaped and view rhythm perception as a product of the patterns and regularities in a listener's musical environment.

Compared to earlier probabilistic approaches to rhythm perception (Temperley, 2007, 2010), the emphasis of this thesis lies less on music theory and more on what can be learned from patterns and regularities in the musical environment. We are therefore interested in the *diversity* of patterns and regularities in different

[2] It should be noted that the authors of the rule-based approaches stressed that their models are intended to reflect listeners familiar with Western classical music (e.g., Longuet-Higgins, 1979).

musical environments and adopt a cross-cultural approach. We model rhythm perception as a function of the musical environment by which it has been shaped and investigate how "enculturated" models perform on both culturally familiar rhythms and culturally unfamiliar rhythms. That is, instead of attempting to model *the listener*, we model *experienced listeners*.

This approach is motivated by a theory which suggests that *prediction-error minimization* is a more useful concept for describing perception and cognition than information processing. This theory is known as predictive processing (or predictive coding) and has been gaining popularity in recent years (Clark, 2013). It has roots in work in computational neuroscience (Rao & Ballard, 1999), the efficient coding hypothesis (Barlow, 1961; Simoncelli & Olshausen, 2001; Smith & Lewicki, 2006), and Bayesian theories of perception (Weiss, Simoncelli, & Adelson, 2002; Knill & Pouget, 2004). According to the predictive processing theory, perception is not just influenced by patterns and regularities in the environment but is fine-tuned by them in a principled way. Perception, it is proposed, relies on internal probabilistic generative models that describe, or learn to describe, how objects and events in the environment (including body movement) cause sensory stimulation. By minimizing prediction error between this model and sensory stimulation, the parameters of the generative model will come to reflect the causes of sensations. Prediction-error minimization is achieved by probabilistic (Bayesian) inference on the parameters of the generative model.

Predictive processing explains both perception and perceptual learning (E. J. Gibson, 1963) as prediction-error minimization and therefore naturally accommodates modeling the influence of the musical environment on rhythm perception of culturally embedded listeners. Although the work in this thesis does not engage with embodiment, Clark (2016) argues that the predictive processing theory leads to an embodied view of the mind. The primary model presented in this thesis may be regarded as a theory of the kind of generative model that listeners might, if the predictive processing theory is accurate, employ when they perceive rhythms.

1.1 Outline of this thesis

This thesis discusses topics that may be categorized as theoretical, methodological, and empirical. First, it presents a theory based on predictive processing, in the form of a probabilistic generative model, of how rhythm perception is influenced by the musical environment. Second, regarding methodology, it presents a framework in which a variety of music perception models can be represented as dynamic Bayesian networks: generative models that iteratively perform probabilistic inference on a sequence of observations. Finally, it describes model simulations in which different probabilistic generative models are tested on empirical samples of rhythms from

different cultures to investigate how statistical regularities in rhythms may influence the way enculturated listeners perceive meter.

The main contribution is a novel probabilistic model of meter perception. Rhythm and meter, according to this model, are related in a different way than other models of beat and meter perception propose (e.g., Povel & Essens, 1985; Large & Palmer, 2002; Large et al., 2015; Temperley, 2007). Most other models describe meter as a periodic pattern of event expectations, in which events at metrically strong positions establish or reinforce the meter and the absence of events at metrically strong positions serves as counterevidence to the meter. These models provide no means by which internalized rhythmic patterns can bias the metrical interpretation of a rhythm. The model presented in this thesis learns to infer meter based on empirical samples of rhythms. Specifically, it learns the sequential statistics of rhythmic patterns expressed in relation to the *metrical cycle* (London, 2012, p. 96).

The model, which is introduced in Chapter 6 (a more technical description is given in Chapter 5), can be viewed as a generator of *predictions* that evaluates a rhythm note by note. Before encountering a note, the model generates a probability distribution that assigns probabilities to different times at which that note may occur (these are the predictions). These predictions are based on a probability distribution over metrical interpretations that has been inferred from the previous notes encountered in the current rhythm.

Consider, for example, the rhythm shown in two different representations in Figure 1.1. The top half of the figure represents the rhythm as a set of evenly-spaced time-points, visualized by dots, some of which contain a sound (e.g., a drum stroke) indicated by vertical lines. The bottom half shows one possible transcription of the rhythm, in which it has been metrically interpreted in 2/4 time. If we would evaluate the model on this rhythm, it would process the rhythm sound by sound. Before observing each sound, it would generate a probability distribution over the possible times (with respect to the present) at which the sound could occur. After observing each sound, it would revise its probability distribution over possible metrical interpretations of the sounds based on where the sound actually occurred.

If the model, after evaluating the first five sounds, has inferred that the rhythm is likely to be in 2/4 time, as shown in the transcription of the rhythm in the bottom half of the figure, its predictions of where the sixth sound will occur are based on where onsets have tended to occur in previous times that the model has encountered the preceding pattern of five sounds in a 2/4 meter. In this way, the predictions of the model are based on its previous encounters with rhythms. If the timing at which the next sound (the sixth) occurs confirms the model's prediction, that confirmation reinforces the model's inferred 2/4 interpretation of the rhythm. If the prediction is not fulfilled, other interpretations of the rhythm might become

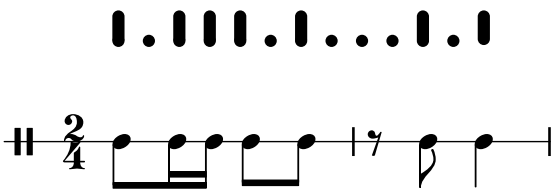


Figure 1.1: Two representations of a rhythmic cliché. The top part of the figure depicts the rhythm as dots that represent evenly-spaced time points. Moments at which a sound occurs are indicated by vertical lines. The bottom part shows the same rhythm, notated in Western music notation, including a time signature and bar lines that indicate its metrical interpretation.

more likely.

An interesting perceptual phenomenon occurs when a listener encounters the rest at the beginning of the second bar (corresponding to the second dot after the fifth sound in the top half of Figure 1.1). Even though there is no sound at this moment, the silence is unusually noticeable. This phenomenon has been called a *loud rest* (London, 1993) and is a direct result of a listener's metrical interpretation of a rhythm. Loud rests are commonly described as a strong violation of expectation: a silence where a sound was expected based on the perceived meter (e.g., Honing, 2009). Accordingly, most existing theories of beat and meter perception treat loud rests as counter evidence to the metrical interpretation that gave rise to them.

Perhaps depending on the reader's cultural background, they are likely to have heard the rhythm in Figure 1.1 more often than they have ever wished for. If the sample of rhythms from which the model presented in this thesis has learned is anything like this reader's experience, it will strongly predict the sixth sound to occur precisely where it does in this example, namely at the second eighth-note position of the second bar. That means that to this model, the loud rest that occurs after the fifth sound is strongly *expected*. In fact, the confirmed prediction of the timing of the sixth sound will *reinforce* the metrical interpretation under which the loud rest occurred (in a similar vein, Witek et al., 2020 recently suggested syncopations can reinforce a sense of beat). This conflicts with the predictions of most existing models of meter perception (Longuet-Higgins & Lee, 1984; Povel & Essens, 1985; Large & Palmer, 2002; Large et al., 2015) in which the absence of a sound on a metrically strong position, under some metrical interpretation, cannot reinforce that metrical interpretation (but cf. Desain & Honing, 1994).

That is not to say that the model assumes that listeners do not experience loud rests, but this experience is, in some sense, outside the scope of the model. It has been suggested that listeners maintain different kinds of expectations regarding the timing of sounds in musical rhythms (Teki, Grube, & Griffiths, 2012; Bouwer, Honing, & Slagter, 2020). Vuust, Gebauer, and Witek (2014) and Vuust and

Witek (2014) suggested that loud rests may simultaneously be a violation of expectations at one level and a confirmation of expectations at another level by appealing to different levels of the predictive processing hierarchy: loud rests may be violations of (more embodied) expectations at lower levels in the hierarchy, while they may simultaneously confirm (more abstract) expectations at higher levels. In the context of this view, the proposition explored by the current model is that the expectations at higher levels may influence the metrical interpretation that drives expectations at lower levels.

This thesis is also about computational modeling of music cognition (Pearce, 2005; Honing, 2006; Temperley, 2013) and a significant part of it (Chapters 3 and 4) is dedicated to developing a framework in which such models can be defined and demonstrating the use of this framework. The framework enables different music perception models to be represented in a unified way, namely as dynamic Bayesian networks with deterministic constraints. Computational cognitive models are sometimes presented in ways that are not completely formal and explicit, which harms the reproducibility of modeling research and makes it difficult to build and improve upon existing models. The framework may alleviate these problems, since it facilitates formal definitions of models that can straightforwardly be translated into executable implementations.

Dynamic Bayesian networks are probabilistic generative models that describe processes that develop dynamically over time (such as music). They can be used to model perception as a cyclical process in which observations occur in a sequence of temporally indivisible moments: In each moment, a model generates a prediction (a probability distribution over possible observations), after which an observation is made. Based on the observation, probabilistic inference is performed to update the parameters of the model, such that each observation influences subsequent predictions. This gives rise to a dynamic interaction between observations and expectations unfolding and evolving across time, making dynamic Bayesian network models well-suited for modeling music perception.

At each event in a sequence, both the uncertainty of a prediction and the discrepancy between the prediction and observation (the prediction error) can be quantified. Such quantifications of prediction and uncertainty have been recently been applied in music perception research using statistical models of melodies (e.g., Omigie, Pearce, Williamson, & Stewart, 2013; Egermann, Pearce, Wiggins, & McAdams, 2013; Hansen & Pearce, 2014). Models formulated as dynamic Bayesian networks are consistent with this approach and can be used to generate theoretical predictions for experiments investigating musical expectations and uncertainty (see also Pearce, 2018).

For any computational model, "the proof of the pudding is in the eating." Even apparently simple models may interact with empirical data in unexpected and

unanticipated ways (Longuet-Higgins, 1981, 1987; Desain & Honing, 1999). [3] This is especially true for models that learn from large amounts of empirical data. When different models and different empirical datasets are involved, these simulations can be set up like an experiment in which we investigate how characteristics of the empirical data and characteristics of a model contribute to the model's behavior.

In Chapter 7 we describe such experiments. Here, we consider the ability of a listener to perceive meter as a function of three variables: different ways in which the listener might be open to shaping by a musical environment (modeled by different probabilistic models), the musical environment itself (represented by empirical samples from music corpora), and the statistical properties of a current rhythm (drawn from one of the samples representing different musical environments). These three factors give rise to what we call *statistical affordances for meter*. In the experiments, we compare two models defined in Chapter 5 using three different empirical samples of rhythms: one sample of Turkish makam music and two samples of German and Dutch folksongs. One of these models is the primary model presented in this thesis and the other is based on traditional theories of meter perception (Temperley, 2007). As such we compare both the effect of generative models representing different theories of meter perception and the effect of different musical environments by which perception might be shaped.

We suggest such a comparison can be enlightening. For one, it may reveal patterns that remain stable between different musical environments as well as patterns that vary between different musical environments. This topic has received considerable attention in recent years (Savage, Brown, Sakai, & Currie, 2015; Jacoby et al., 2019; Mehr et al., 2019; Jacoby et al., 2020). The consideration of different models representing different theories of perception, however, illustrates that similarities and differences also depend on the model used to characterize them. This comparative approach is furthermore interesting, we suggest, because it shifts the focus from evaluating the performance of different computational models relative to what has been labeled as the "correct" interpretation of a rhythm (i.e., the "ground truth" [Gouyon and Dixon, 2005]) to comparing the *relative* performance of models on materials from different cultures.

Based on the results, we find evidence that there are statistical patterns in rhythms that listeners *could* use to infer meter from rhythms. These patterns are more nuanced than the statistical patterns to which a model derived from Temperley's (2007) model is sensitive. Furthermore, we find evidence that there are idiom-specific patterns and regularities that potentially provide only serve as cues for meter to enculturated listeners familiar with these idiom-specific patterns. These findings may serve as theoretical predictions that could be tested in cross-cultural

[3] Such surprises may themselves be enlightening as they may reveal overlooked consequences of theories and models of cognition that come to light only when they are implemented algorithmically and tested on empirical data.

experiments in the future.

1.2 Summary of chapters

Chapter 2 discusses ways in which rhythm and meter perception has been found to be influenced by prior experience, practice, and training. Against this backdrop, it reviews some of the different rhythm perception modeling approaches that have been pursued in the past decades. It shows that these approaches can be associated with different philosophies of perception and cognition: cognitivism, embodied cognition, and predictive processing. The aim of the chapter is to draw attention to the remarkable diversity displayed by modeling approaches proposed previously and to suggest that predictive processing is an attractive framework for modeling the influence of experience, practice and training.

Chapters 3, 4, and 5 are methodological. In Chapter 3, a framework for the specification of dynamic Bayesian networks with deterministic constraints is introduced. We discuss how deterministic constraints can be encoded by *congruency constraints*, functions that encode the assumption that some values of a random variable have zero probability of occurring. We discuss the consequences that these assumptions have for calculating marginal probabilities and performing inference in dynamic Bayesian networks. We then propose a model-definition framework in which the variables and congruency constraints that constitute a dynamic Bayesian network with deterministic constraints are expressed in a single table.

To illustrate how such tables can compactly define music perception models, Chapter 4 presents an adaptation of Temperley's (2007) rhythm perception model. Since this adaptation is a dynamic Bayesian network, it can be evaluated in temporally incremental way. This requires a few relatively minor technical and conceptual changes to the original model. The chapter draws attention to how the use of the model-definition framework results a detailed yet concise definition of the model.

Chapter 5 uses the model-definition framework to define two rhythm perception models: a simplification of Temperley's model, as described in Chapter 4, and an adaptation of the model introduced in Chapter 6. Common Lisp implementations of these models, based on an implementation of the framework of Chapter 3, can be found in Appendix A. We describe the notion of a rhythm space, which is a set of unique rhythms over which the two models define a complete probability distribution. We furthermore describe how empirical distributions of rhythms in this space can be used to estimate the parameters of the two models. In Chapter 7, we apply this process in a pair of experiments in which we compare the two models using different empirical samples of rhythm spaces.

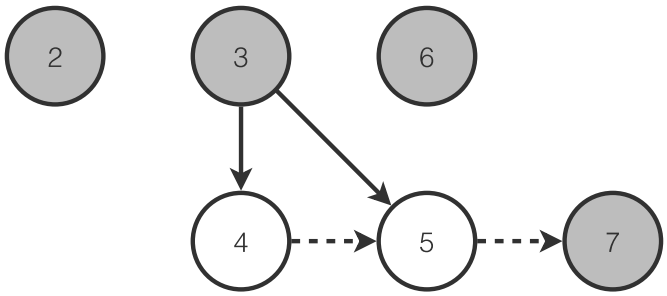


Figure 1.2: A graphical representation of how to read this thesis. Chapters corresponding to the numbers in the gray shaded nodes are self-contained and can be read independently. Before reading chapters corresponding to numbers in nodes that have incoming arrows, it is recommended to first read the chapters corresponding to numbers in nodes from which the arrows originate. If an incoming arrow has a dotted line, reading the chapter corresponding to the node from which it originates is not a prerequisite for understanding the chapter but will allow the chapter to be understood in more depth.

Chapter 6 introduces and motivates the primary probabilistic rhythm perception model proposed in this thesis. The model is intended to describe effects that long-term exposure to rhythms from a certain musical environment may have on rhythm perception. This chapter describes the model as an extension of the IDyOM modeling framework (Pearce, 2005), using the terminology of multiple viewpoint systems (Conklin & Witten, 1995). A technical description and self-contained description of this model is given in Chapter 5. Simulation results are reported which assess the model's ability to classify the meter of rhythms in an empirical corpus and its ability to predict the timing of onsets in rhythms, based on the preceding onsets in the rhythm. It furthermore presents exploratory results that identify rhythms for which inferential biases towards different meters can tip the model's metrical interpretation. The main difference between the presentation of the model in Chapter 5 and in this chapter is that Chapter 5 defines the model in its own right, rather than in terms of multiple viewpoint systems. Furthermore, courtesy of the congruency-constraint definitions, the definition in Chapter 5 is more concise and explicit and can be compared directly to the adaptation of Temperley's (2007) meter perception model defined in the same chapter.

Finally, Chapter 7 describes three experiments, two of which use the models described in Chapter 5. Here, we investigate the effects that (1) the statistical learning capabilities of a model, (2) the musical idiom from the model learns (its patterns and regularities) and (3) the musical idiom on which the model is evaluated have on the model's ability to infer meter and to predict the timing of onsets in rhythms. Each of these factors is varied independently. Datasets of Turkish makam music and German and Dutch folksongs represent the different musical idioms. The results suggest that there are statistical patterns in rhythms

that may contribute to meter perception and are more complex than the degree to which a rhythm aligns its onsets with metrically strong beats. Furthermore, we find evidence that some of these patterns are specific to Turkish makam music and others to Dutch and German folksongs but also that the differences between rhythms of these two musical idioms are limited.

Figure 1.2 shows a graph representing which chapters in this thesis rely on concepts developed in preceding chapters. Chapter corresponding to numbers shown in gray nodes, that is Chapters 2, 3, 6, and 7, are self-contained and can be read independently. Chapters 4 and 5 build on the concepts developed in Chapter 3. While Chapter 7 is self-contained, readers interested in the technical details of the models and methods may want to read Chapters 3 and 5 first. Readers not familiar with Temperley's (2007) meter perception model may furthermore benefit from reading 4 before reading Chapter 5.

Theoretical motivation for the work pursued in this thesis can be found most prominently in Chapter 2, and also in Chapters 6 and 7. The technical details of the modeling approach are discussed primarily in Chapters 3, 4, and 5. Empirical results based on model simulations performed with samples of rhythms can be found in Chapters 6 and 7.

Chapter 2

Computational modeling of rhythm perception and the role of enculturation

2.1 Introduction

In the music cognition literature, a conceptual distinction is often drawn between rhythm and meter. Rhythm refers to a temporal pattern of sounds, while meter refers to a subjective mental phenomenon (Honing & Bouwer, 2019). Listening to rhythms tends to induce a sense of pulsation in listeners (Povel & Essens, 1985). This pulsation, known as *beat* or *tactus* (Lerdahl & Jackendoff, 1983), provides a temporal reference with which movements can be coordinated (Repp, 2005; Repp & Su, 2013). We speak of *meter* when some beats appear as more accented than others and these accented beats recur more or less regularly (Cooper & Meyer, 1960). Pulse and meter form the basis of temporally coordinated musical activities such as clapping, dancing, singing or playing an instrument. While these characteristics of meter are generally regarded as uncontroversial among music cognition scholars, two aspects that elude consensus are the precise nature of the mental phenomenon known as meter, and the degree to which it is shaped by a listener's history of prior musical experiences and activities.

Regarding the nature of the mental phenomenon, a range of approaches have been proposed. Some of these highlight abstract hierarchical structures (Longuet-Higgins, 1978; Longuet-Higgins & Lee, 1984; Lerdahl & Jackendoff, 1983), others entrainment of attention (Jones & Boltz, 1989; Large & Jones, 1999), neural resonance (Large & Snyder, 2009), or embodied and ecological aspects of rhythm perception (Shove & Repp, 1995; Iyer, 1998; Clarke, 1987; Todd & Lee, 2015).

More recently, approaches based on predictive processing have been proposed (Vuust & Witek, 2014; Van der Weij, Pearce, & Honing, 2017).

Computational cognitive models of rhythm and meter perception (for brevity, we refer to such models collectively as rhythm perception models) are the focus of this chapter. By computational models, we mean models that are described—ideally in a formal language—with a level of precision that allows them to be implemented as a computer program, without the need to fill in many details (see also Temperley, 2013). Such models may be distinguished from verbal-conceptual models, which are expressed in prose or as conceptual diagrams, and may be consistent with multiple computational models.

We discuss rhythm perception models in the context of three broad theoretical perspectives, namely cognitivism (cf. Anderson, 2003), embodied cognition (Brooks, 1991b; Van Gelder, 1995; Anderson, 2003; Chemero, 2009), and predictive processing (Clark, 2013). Each of the above approaches can be associated with one of these perspectives. In turn, these perspectives can be associated with different computational modeling principles that underlie the models discussed in this chapter.

Briefly, cognitivism views cognition as being primarily involved in rule-governed information processing. This perspective is associated strongly with classical artificial intelligence approaches (e.g., see Newell & Simon, 1976). Among rhythm perception models, classic rule-based models (e.g., Longuet-Higgins & Steedman, 1971; Longuet-Higgins & Lee, 1982) and preference-rule models (Temperley & Sleator, 1999; Temperley, 2001) may be associated with this perspective (see Section 2.3). Embodied cognition may be characterized by a rejection of the idea that information processing and abstract representation provide the most appropriate explanation of many behaviors. It instead emphasizes the role of continuous dynamic interaction between brain, body and environment. Many characteristics of adaptive oscillator (e.g., McAuley, 1995; Large & Palmer, 2002), and neural resonance (e.g., Large et al., 2015) models harmonize well with this perspective (see Section 2.4). Finally, the term predictive processing (introduced by Clark, 2013) covers a class of theories that build on the Bayesian brain hypothesis (Knill & Pouget, 2004) and predictive coding (Rao & Ballard, 1999). These theories propose that perception and cognition can be understood as prediction-error minimization in a probabilistic generative model. Probabilistic generative approaches to rhythm perception (Temperley, 2007; Van der Weij et al., 2017) are consistent with this perspective (see Section 2.5).

The question of the degree to which rhythm perception of individual listeners is shaped by their history of prior experiences and activities is often considered in the context of cultural background. Cultural background is one predictor of stable tendencies in histories of prior musical experiences and activities of individual listeners. If these stable tendencies have the power to influence rhythm

perception, cultural background may predict certain individual characteristics of rhythm perception in listeners from different cultural backgrounds. We use the term *enculturation* to refer to the acquisition of implicit cultural knowledge by exposure to, and participation in, cultural activities. The predictive processing perspective most prominently draws attention to the role that enculturation might play in the shaping of perception and cognition. Although neither cognitivism, nor embodied cognition are explicitly incompatible with this role, predictive processing accounts for it normatively, namely as a consequence of a domain-independent prediction-error minimization mechanism.

While there is considerable evidence, some of which is discussed in Section 2.2 of this chapter, suggesting that enculturation shapes rhythm perception, enculturation plays little to no role in the majority of existing rhythm perception models. Many of these models have been inspired by Western music theory, and have been evaluated only on Western tonal music. A related lack of diversity can be identified in the materials and participants used in empirical and experimental music cognition research (Huron, 2008; Jacoby et al., 2020). This situation is problematic, because, assuming rhythm perception undergoes shaping by enculturation, it results in a biased understanding of rhythm perception.

Among models that do not account for effects of enculturation, some explicitly limit their scope to Western tonal music (e.g., Longuet-Higgins, 1979). Others aim to reflect universal constraints on perception and cognition (Povel & Essens, 1985; Large, 2010b). Parameters of such models are typically determined by musical intuition (e.g., Longuet-Higgins, 1976; Povel & Essens, 1985; Temperley & Sleator, 1999), or by optimal fit to experimental data (e.g., Shmulevich & Povel, 2000). On the other hand, there are models that aim to simulate the effects of prior *exposure* (one aspect of enculturation) to certain kinds of music on rhythm perception (Van der Weij et al., 2017; Tichko & Large, 2019). Parameters of these models are derived from empirical samples of rhythms that are intended to represent previous exposure to rhythms (see also Patel & Demorest, 2013; Pearce, 2018; Morrison, Demorest, & Pearce, 2019). Some of these models are probabilistic generative models, which are consistent with the mechanisms posited by the predictive processing perspective.

In summary, this chapter discusses computational cognitive models of rhythm perception and aligns them with three broad theoretical perspectives on cognition. It furthermore considers the role that enculturation may play in rhythm and meter perception, and the degree to which models take this role into account. The next section reviews research related to the role of enculturation in shaping rhythm perception. The remaining sections of this chapter are dedicated to each of the three broad theoretical perspectives on cognition: cognitivism, embodied cognition, and predictive processing. Each section opens with a brief discussion of the theoretical perspective, before turning to the rhythm perception models that

are consistent with it.

Finally, we note that evaluating the performance of the discussed models, and the connection between model predictions and empirical observations receive less attention in this chapter than the reader might have expected. This is partly because the emphasis lies on theoretical differences between cognitive models, and partly because extensive comparisons between computational models, especially on culturally diverse datasets are simply not available (but for an exception, see Desain & Honing, 1999).

2.2 A cross-cultural perspective on rhythm perception

Perhaps in part due to its high level of pervasiveness,[1] Western tonal music has, explicitly or implicitly, played a significant role in the formation of theories and models in music cognition (see also Jacoby et al., 2020). However, this musical tradition represents only a small slice of the variety in musical cultures that exists around the world (Trehub, Becker, & Morley, 2015; Savage et al., 2015; Mehr et al., 2019). If rhythm perception shaped by enculturation, it can be expected to vary for individual listeners depending on kind of music they are familiar with.

Below we discuss studies that suggest rhythm perception is shaped by enculturation. The discussion is considers three aspects of rhythm perception: the relation between metrical hierarchies and the likelihood of events, constraints for tactus beats to be isochronous, and the shape of perceptual categories for temporal intervals.

2.2.1 Metrical hierarchies and event likelihood

Arguably due to the familiarity of theorists with the Western musical idiom, metrical hierarchy plays a prominent role in theories of rhythm perception. The way such hierarchies are commonly conceptualized can be attributed to the influential work of Lerdahl and Jackendoff (1983), Longuet-Higgins (1978), Longuet-Higgins and Lee (1984). These authors describe meter as a hierarchy of *metrical levels*, popularly depicted as *metrical grids* by Lerdahl and Jackendoff (see Figure 2.1). Each metrical level consists of *beats* that are spaced evenly in time (isochronous). Beats are described as duration-less points in time, represented abstractly in the mind of the listener. The resulting representation imposes a pattern of alternating

[1] Huron (2008, p. 457) illustrates this point anecdotally, describing an episode in which he, joining an expedition of biologists, encountered subsistence hunters in the western Amazon that, thanks to their transistor radios, were familiar with Western popular music.

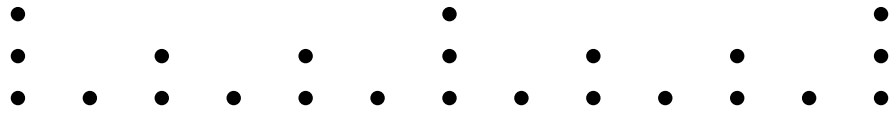


Figure 2.1: A metrical grid visualizing the putative hierarchical organization of two bars of a ternary time signature (such as 3/4 time). The hierarchy contains three metrical levels. The dots represent beats, the horizontal dimension represents time (which flows from left to right), and the vertical dimension represents metrical salience (towards the top of the figure is more salient). The top level usually indicates the bar-level periodicity. Note that between each pair of dots on a higher level, two or three dots occur at a lower level. For each top-level beat, there are three middle-level beats, and for each middle-level beat there are two beats on the lowest level, indicating ternary subdivision of the top level and binary subdivision of the middle level. Also note that beats at each level are equidistant in time.

strong and weak beats onto a perceived rhythm, where the metrical strength (or *accent*, or *salience*) of a beat is determined by the highest metrical level in which it occurs. Lerdahl and Jackendoff (1983) distinguish between *phenomenal accents*, which are due to way a piece of music is performed, and *metrical accents*, which are due to the metrical interpretation of the music by a listener, and occur on metrically strong beats. This distinction highlights the conceptual difference between the rhythm and its metrical interpretation by a listener.

It is commonly assumed in computational and verbal-conceptual theories of rhythm perception that the metrical strength of a beat represents the strength of prediction or expectation that an event will occur (Temperley, 2007; Large, 2008). The metrical phenomena of "loud rests" (London, 1993) and syncopation are commonly related to this assumption. A *loud rest* occurs when an event unexpectedly does not occur at a metrically salient beat. *Syncopation* occurs when a metrically salient beat passes silently or unaccented and is preceded by an onset or accent at a metrically weaker beat (Longuet-Higgins & Lee, 1984). These phenomena are generally described as deviations from the norm, or as violations of expectation (Fitch & Rosenfeld, 2007; Bouwer, Burgoyne, Odijk, Honing, & Grahn, 2018). Consequently, measures of syncopation are sometimes used to estimate the perceptual complexity of rhythms based on the idea that rhythms whose constituent events have unpredictable timing will be experienced as more complex (e.g., see Witek, Clarke, Wallentin, Kringelbach, & Vuust, 2014). However, recent studies have suggested that the presumed correlation between metrical strength and event expectation may not hold for all listeners.

Palmer and Krumhansl (1990) hypothesized that the "frequency with which

musical events in a piece occur in a given metrical context may provide important perceptual cues to meter". Using a set of Western classical music compositions by four different composers, Palmer and Krumhansl constructed *event-frequency distributions* based on the relative frequency of onsets at different positions in a bar. Such distributions were constructed separately for different meters and composers. In support of their hypothesis, Palmer and Krumhansl found that metrical salience more or less predicts the relative frequency of events, and that this effect is stable for different composers. Palmer and Krumhansl furthermore conducted a pair of behavioral experiments, the results of which indicated that expectations of listeners (especially if they are musicians) for notes to occur in different metrical contexts correlated with metrical salience of those contexts.

While Palmer and Krumhansl highlight the role of statistical regularities in music, they interpret this role in the context of multileveled representations of metrical hierarchies. They suggest that observed frequency distributions of musical events in different metrical contexts result from the presence of such metrical hierarchies in the minds of composers and listeners, rather from stylistic constraints in the music. However, Palmer and Krumhansl qualify this finding by noting that their observations are limited to Western classical music. Indeed, subsequent corpus studies applying the same methodology to rhythms from different musical idioms suggest a more prominent role for stylistic constraints in the shaping of frequency distributions of event timing.

Holzapfel (2015) analyzed event-frequency distributions derived from a corpus of Turkish makam music (Karaosmanoğlu, 2012). Turkish makam music is a style of both classical and folk music in which rhythmic organization is centered around the notion of an *usul* (Marcus, 2001). Usuls are rhythmic modes, characterized by a pattern of drum strokes. Holzapfel derived these distributions for Turkish makam music by collapsing over usul cycles, which are annotated in the corpus. The results show that compared to Western music, onsets in Turkish makam music were more spread out over different positions in the metrical cycle, and that usul patterns could be used to classify the usul underlying makam compositions.

London, Polak, and Jacoby (2017) examined a set of Malian djembe ensemble recordings using event-frequency distributions. This music does not make use of music notation, so London and colleagues relied on onset annotations of the recordings. After correcting for tempo changes, the observed onsets were collapsed over metrical cycles. Results show that the relative frequencies of events at different positions in the metrical cycle do not appear to be structured by metrical salience patterns, even though the the rhythms are metrically structured and the consistent timing subdivisons suggests the presence of a regular beat.

The above findings are consistent with the idea that the distribution of events over positions in the metrical cycle provides a cue for meter. However, metrical hierarchy, as predicted by theories based on Western music theory (Longuet-Higgins, 1978;

Longuet-Higgins & Lee, 1984; Lerdahl & Jackendoff, 1983), appears not to be the only predictor of those patterns. Based on their findings in Malian djembe ensemble recordings, London et al. (2017) claim that "the shared presumption that onset frequency is correlated with metrical accent holds only contingently, that is, for the corpora of Western classical and popular music that were used in these studies, and for which these models were developed" (p. 478). This also calls into question the view that syncopations necessarily reflect violations of expectation. Iyer (1998) anticipated this, suggesting that one "should not regard the global musical preponderance of 'syncopation' (off-beat accents) as a vast set of exceptions to the 'normal' accentual rules of meter, but rather as convincing counterexamples to such proposed accentual rules." (p. 44).

2.2.2 Meters with non-isochronous tactus beats

Theories metrical structure often include constraints for the beats at the tactus level to be evenly spaced (isochronous). Longuet-Higgins (1978), Longuet-Higgins and Lee (1984) described meter generatively as the recursive subdivision of intervals into two or three evenly spaced beats. Similarly, Lerdahl and Jackendoff (1983) suggested that beats in well-formed metrical hierarchies must be more or less evenly spaced. The authors cited here have indicated that their theories apply primarily to Western music, but their ideas have nevertheless shaped subsequent research, which does not always acknowledge this qualification.

Meters with uneven (non-isochronous) intervals between tactus beats, while relatively uncommon in Western classical music, are prevalent in many musical styles (e.g., see London, 1995; Polak et al., 2018). Cross-cultural studies have suggested that these meters are readily processed by listeners familiar with these structures. London (1995) calls such non-isochronous meters "complex", and argues that a non-isochronous tactus beat needs to be anchored in a faster, and isochronous, underlying pulse, such that tactus beats are measured by either two or three of these faster pulses. This suggestion has been challenged by observations that non-isochronous tactus beats do not always adhere to an underlying isochronous grid (Kvifte, 2007). Furthermore, it has been suggested that the purported complexity of non-isochronous meters is overridden by familiarity: adults and infants with exposure to non-isochronous meters are able to detect violations of the meter while adults with limited exposure to such meters can only do this for isochronous meters (Hannon & Trehub, 2005b; Soley & Hannon, 2010; Hannon et al., 2012). For such listeners, rhythms in non-isochronous meter are no more complex than those in an isochronous meter.

2.2.3 Perceptual categories for temporal duration ratios

It is thought that ratios between continuous time intervals in rhythms are perceived as a small number of discrete perceptual categories (Clarke, 1987; Desain & Honing, 2003). These categories appear to be centered around small-integer ratios (such as 1:2:1 and 1:2:3), but their size and shape varies, resulting in perceptual biases (Desain & Honing, 2003; Jacoby & McDermott, 2017). It has been hypothesized that small-integer ratios are a universal constraint on perceptual categories for duration intervals (Mehr et al., 2019). In support of this hypothesis, Mehr et al. found that simple-integer duration ratios are prevalent in a culturally diverse sample of music recordings. Furthermore, Savage et al. (2015) found evidence for the widespread occurrence for binary and ternary subdivision as well as isochronous beats to occur in music.

Jacoby and McDermott (2017) found evidence that biases in categorical perception of temporal intervals, that is, the size and shape of perceptual categories, may be attributable in part to enculturation. In a cross-cultural study involving adult members of a native Amazonian society (the Tsimané) and North American adults, they found that size and shape of perceptual categories for temporal intervals differed significantly between these two groups, but also that in both groups, perceptual categories appeared centered around small-integer ratios. Since the musical practices of the Tsimané and North American participants are apparently different enough to cause the observed differences in perceptual biases, this commonality is especially remarkable and consistent with the potential universality of small-integer ratio categories for ratios between temporal intervals in rhythms.

Polak, London, and Jacoby (2016), Polak et al. (2018), however, present work that challenges the hypothesized universality of (perceptual categories for) small-integer duration ratios in rhythms. Malian djembe ensemble performances commonly contain "swung" subdivisions—intervals subdivided into intervals related by a complex ratio. Polak and colleagues show that these non-isochronous subdivisions afford the production of precise and consistent timing patterns in an ensemble context. This seems to suggest that these Malian musicians are able to entrain to complex-ratio beat subdivisions. Polak et al. (2018) suggest that the production of such non-isochronous subdivisions may be supported by non-isochronous perceptual categories that depend on experience and training and, in a cross-cultural study, found evidence for the presence of a such a category expert musicians from Mali, but not in expert musicians from Germany or Bulgaria.

Musical features that are surprisingly prevalent in music from all over the world, known as statistical universals (Savage et al., 2015), are commonly interpreted as evidence for innate cognitive constraints. However, while the statistical universality of a musical feature may be a necessary condition for such constraints, it is not a sufficient condition. Individuals from different cultures share more than

genes: bodily constraints and stable properties of natural environments that are independent of geographic location may influence development in universal ways and could therefore also underlie universal tendencies. The link between innate constraints and statistical universals is perhaps further weakened by the dynamics of cultural transmission. Based on simulations with a Bayesian model of cultural transmission, Thompson, Kirby, and Smith (2016) argue that it is possible for strong universals to arise from weak and defeasible cognitive biases.

2.2.4 Reconciling enculturation with universal tendencies in rhythm perception

Some authors, such as Temperley (2000) and Agawu (1995), warn that there exists a tendency, primarily in the ethnomusicological literature, to overstate differences in rhythmic practices and rhythm perception between cultures. Others, such as Iyer (1998), Huron (2008), and, recently, Jacoby et al. (2020) lament the sparsity of cross-cultural work in music cognition and caution against interpreting the idiosyncrasies of a familiar musical (usually Western) culture as the norm, or of culture-specific perceptual constraints as universal.

A theoretical account of rhythm perception that can potentially reconcile these views has been proposed by London (2004). In accord with ideas of Jones and Boltz (1989) and Large and Jones (1999), London suggests that meter perception is a form of entrainment behavior, which serves to guide our attention over time in synchrony with musical rhythm. However, its openness to shaping—by experience, practice, music education and other forces of influence that an individual's embeddedness in a cultural environment entails—makes metrical entrainment a *skilled behavior*. Thus, meter perception is regarded more than a passive response to music, or a bottom-up analysis of sensations. However, London also argues that although the structure of metrical entrainment behavior is plastic, it simultaneously is constrained by a set of well-formedness conditions which he sets forth (echoing Lerdahl and Jackendoff's [1983] approach). London explicitly avoids proposing universal preference rules, since, he argues, the relation between rhythm and meter is malleable and ambiguous.

This theoretical account thus argues that certain aspects of rhythm perception can be shaped by enculturation while other aspects are less adaptable and can be captured by well-formedness constraints and occupies a middle-ground between work emphasizing cultural differences in rhythmic practices and work emphasizing universal constraints on perception. Specifically, constraints on metrical entrainment behavior are argued to arise universally, while an individual's capacity for metrical entrainment depends on their previous experiences and activities and is therefore subject to the influence of enculturation.

2.2.5 Enculturation and embodiment

A more radical reading of some of the above literature suggests that the enculturation of rhythm perception may involve information that cannot be gleaned from (symbolic or recorded) music corpora alone. London et al. (2017) suggest that

> [...] while the frequency of onset occurrence of events doubtless plays a role in our acquisition of rhythmic and metrical knowledge, those frequencies occur in holistic contexts that include timing, timbre and other auditory, visual, and sensorimotor channels of perception. Combinations of these cues forge associations between statistically common rhythms and their characteristic metrical orientations. (p. 479).

Some information about such holistic contexts may be encoded in music corpora, but these annotations provide no substitute for tightly coupled sensing and acting involved in participation in music-related cultural practices such as dancing, attending a concert, or singing in a group. These experiences involve coordinated movements and sensations in, most prominently, the auditory, visual and proprioceptive modalities.

An embodied view of rhythm perception acknowledges the role that these aspects might play in musical experiences. An action-oriented interpretation of predictive processing (Clark, 2013), however, in addition to emphasizing the role of embodiment in perception, suggests how embodied experience *shapes* perception (see also Clark, 2016). A theory of rhythm perception based on action-oriented predictive processing might therefore be consistent with the theoretical account described by London (2012), in that it may describe the role of practice and training in shaping rhythm perception.

The above considerations, if true, appear troublesome for computational models of rhythm perception that aim to simulate enculturation using samples from music corpora. Nevertheless, stable probabilistic properties of the music to which enculturated individuals are exposed may still play a significant role in shaping rhythm perception. Samples drawn from music corpora are likely to reflect these probabilistic properties. Models that use these samples to simulate the effects of enculturation on rhythm perception may therefore successfully capture some of these effects.

To conclude, the research reviewed above suggests that the experience of meter depends to a large extent on being situated in a cultural environment. If so, it seems that rhythm perception models that aim to be applicable to music from different cultures cannot be rely on a bottom-up analysis of a rhythm based on hypothetically universal mechanisms for rhythm perception. They must also account, in some way, for the effects of being intimately familiar with certain musical idioms. For computational modeling, empirical samples from music

corpora may go some way toward simulating the musical exposure of enculturated individuals. However, such corpora do not capture the holistic context in which exposure to music occurs, possibly leaving some aspects of enculturated rhythm perception unaddressed.

2.3 The cognitivist perspective

We now turn to rhythm perception models associated with the first of the three different broad theoretical perspectives discussed in this chapter, namely cognitivism. The cognitivist perspective is characterized by the view that cognition is most appropriately explained as pure *information processing*, involving rule-based computation performed on symbolic representations. These representations are derived from sensory input through bottom-up perceptual processes (e.g., Newell and Simon, 1976; Marr, 1982, cf. Anderson, 2003). This emphasis on information processing is typically reflected in terminology used to motivate and describe cognitivist models. Perceptual and cognitive phenomena are described as involving "problems" or "tasks" that cognition must "solve" or "decide". Modeling a cognitive process entails identifying the task it performs, identifying the appropriate representations of input and output, and finding an algorithm that generates the appropriate output given an input.

One motivation for the epistemological value of such models is that designing an algorithm to solve a specific cognitive task may provide insight into the cognitive process itself. Commonly, the process of designing such algorithms reveals unanticipated intricacies and complexities of the task itself that were overlooked by verbal-conceptual theories. The cognitivist approach was especially popular in the early days of cognitive science and its methodology was significantly influenced by contemporary developments in artificial intelligence. These influences have been noted by authors like Longuet-Higgins (1978), Newell and Simon (1976) and Bundy (1990).

2.3.1 Rule-based models of rhythm perception

The sections below describe a set of cognitivist rhythm perception models, proposed by Longuet-Higgins and colleagues, who pioneered computational modeling of music cognition in the 1970s and 1980s. These models propose specific mechanisms for various computational problems that are hypothesized to be involved in rhythm perception. Longuet-Higgins and colleagues point out at several occasions (Longuet-Higgins & Steedman, 1971; Longuet-Higgins, 1978, 1979) that their work aims to account for perception of Western tonal music by listeners familiar with such music. Therefore, the models described below are not intended as universal

accounts of rhythm perception, but rather as reflections of the perception of enculturated listeners. Clarke (1999) more extensively discusses these models, and many that followed in this early period of music-cognition modeling. Some of these models are still actively used in empirical studies (e.g., Fitch & Rosenfeld, 2007; Grahn & Brett, 2007; Song, Simpson, Harte, Pearce, & Sandler, 2013; Witek et al., 2014; Bouwer et al., 2018).

Longuet-Higgins and colleagues described a number of key issues that still inspire modelers of music perception to this day. The central issue is to understand listeners' ability to reconstruct the rhythmic and tonal relations, intended by the composer, between sounds from a performance of Western classical music. Western tonal music notation contains considerable information about the tonal and temporal relations in music. Trained musicians can reconstruct this information from "even a mediocre performance" (Longuet-Higgins & Steedman, 1971, p. 221). Therefore, it is argued, scores are likely to provide strong clues towards the kind of rhythmic and tonal relations that listeners infer from a performance. This ability is furthermore argued to be available to anyone "familiar with the composer's language" (Longuet-Higgins, 1978, p. 149). In the models described below, the inference of rhythmic relations (meter) and tonal relations (key) are treated independently. The discussion below considers only the parts of these models relevant to the interpretation of rhythm.

Longuet-Higgins and colleagues divided the central issue into a set of sub-problems, which were addressed individually by the computational models that we describe below. These models describe the inference problem from the perspective of the listener, who is processing a piece of music note by note. This listener must, from the first few notes, infer the phase and period of the beat. This problem is addressed by Longuet-Higgins and Lee (1982). Then, the established beat must be subdivided recursively until each note initiates a metrical unit at some level of beat subdivision. Sometimes, notes are played slightly earlier or later than expected. In these cases, the listener must figure out whether these deviations represent a change in tempo, a subdivision of the beat, or expressive timing of the performer. Such tracking and subdivision of a beat is addressed by Longuet-Higgins (1976). Finally, to find the correct time signature, beats must be grouped into higher-level metrical units such as bars. This problem is addressed by Longuet-Higgins and Steedman (1971). The sections below discuss these models in chronological order.

2.3.1.1 Grouping metrical units

Longuet-Higgins and Steedman (1971) propose an algorithm that addresses how, based on a pattern of note durations in a deadpan performance, a listener may

identify metrical units and group them into bars. [2] Motivated by "the progressive character of musical comprehension" (p. 223), Longuet-Higgins and Steedman propose a fundamental principle, which they call *rule of congruence*, by which the other rules of the model are motivated. The intuitive motivation for this principle is described eloquently by the observation that "music would be a dull affair if all notes had to be in the key and all accents on the beat, but it would be incomprehensible if the key and metre were called into question before they were established" (p. 224). The rule of congruence stresses the important role played by temporal order of musical events. This emphasis on the *temporal incrementality* of music listening sets this early approach apart from later approaches, which ignore the temporal order of events (e.g., Povel & Essens, 1985; Palmer & Krumhansl, 1990).

The rules of the model contain many subtleties, but can be summarized approximately as follows: The duration of the first or second note (whichever is shorter) establishes the smallest *metrical unit*. By the rule of congruence, an established metrical unit is never abandoned. The metrical hierarchy is progressively constructed from this smallest unit by means of grouping. Such grouping is prompted by one of three cues: (1) the occurrence of a long note beginning on an already established metrical unit, (2) a *dactyl*, a pattern consisting of two long followed by one short interval, or (3) a long note followed by a short note. If any of these is encountered, the current metrical unit is multiplied in length by two or three (depending on the length of the cue-pattern) to form a metrical unit at the next level of the hierarchy.

2.3.1.2 Beat tracking and subdivision

Longuet-Higgins (1976) proposes an algorithm addressing a different issue: how to track and subdivide a beat in a performance with a changing tempo? While Longuet-Higgins and Steedman's (1971) model and its extension (Steedman, 1977) assume deadpan performances, listeners are able to follow along with a beat despite tempo changes and expressive timing. Another way of stating the problem that this work aims to address is as follows: when an onset occurs close to where a beat is expected, how does a listener decide whether the onset marks a subdivision of the beat, a change of tempo, or an expressive deviation?

Longuet-Higgins proposes the following procedure: [3] Assuming a given beat

[2] A deadpan (or mechanical) performance is one that exactly reproduces the note duration ratios dictated by a notated musical score. For musicians performing music from a score, the goal is rarely to produce such (mechanical-sounding) performances. Instead, expressive tempo changes and deviations deadpan timing are the norm (Clarke, 1989; Repp, 1995).

[3] This procedure is used in a computer program that can transcribe a melody, played on an organ console connected to a high-speed paper tape punch (a MIDI keyboard would nowadays

interval, it can be determined where, assuming deadpan timing, the next beat is expected. Based on this, the amount of time by which the next note deviates from this expected beat can be determined. A temporal window around the expected next beat location is created by a parameter called *tolerance*. If the next note falls within the tolerance window, the next beat interval is increased or decreased by half the amount of deviation of the note from deadpan timing. If the next note instead occurs before the tolerance window, the beat interval is subdivided by two or three. The upshot is that, once processing is complete, each note occurs at the beginning of a metrical unit.

This mechanism for adapting beat duration based on deviation from deadpan timing bears some resemblance to beat perception models based on coupled oscillation (Large & Kolen, 1994) that are discussed in Section 2.4. The resemblance is notable because coupled oscillation models align with a theoretical perspective that is rather different from cognitivism.

2.3.1.3 Beat finding

Longuet-Higgins and Lee (1982) address another puzzle: assuming that a beat can be tracked and subdivided once established, how is the tactus beat established to begin with? How do we know whether the rhythm begins with an anacrusis? If it does, where does the first beat occur? How do we know the interval between the first and the second beat? Borrowing terminology used by Desain and Honing (1999), Longuet-Higgins and Lee's model maintains a *current beat hypothesis*. This hypothesis is specified by two variables representing virtual points in time: a 'first beat', t_1, and a 'second beat', t_2. These variables are initialized by the onset times of the first and second note. Subsequent notes revise and update the current beat hypothesis by subjecting it to two types of transformations: lengthening (stretching) it or shifting its position. These transformations are triggered by a set of rules that, in the interest of brevity, we will not attempt to summarize here. The algorithm aims to output values of t_1 and t_2 that encode the position and duration of the first tactus beat interval in a performance, thereby providing an answer to the questions at the beginning of this paragraph.

Desain and Honing (1999) note that this model and some of its successors were evaluated only qualitatively on small toy domains. Furthermore, it is difficult to derive general properties of their functioning since the rules in these models interact

suffice), into musical notation. The program has to be supplied an initial beat interval, similar to a drummer's count off before a performance. It then tries to track and, if necessary, subdivide this beat throughout a performance. No detailed description of the program is provided, but its source code (written in the POP2 programming language), was made available. The program, as well as a translation into the LISP programming language, can be found in Desain and Honing's (1992) book, *Music, Mind and Machine*.

with input in complex and unpredictable ways. Desain and Honing propose a unified framework in which the models can be expressed, and systematically analyze the behavior of the models using an empirical dataset and Monte-Carlo samples from the models' *input space*: the set of all possible input rhythms of up to thirty-five grid points. Desain and Honing's results show that these simple models perform surprisingly well. Their work represents one of the few existing systematic comparisons between computational models on the same dataset.

2.3.2 Optimization and preference-rule models

We describe optimization models here as cognitivist models, but differ significantly from the rule-based models described above. Instead of what are sometimes called "hard and fast" rules, optimization models employ soft constraints that can be satisfied to various degrees. In some ways, which we return to below, these models are similar to probabilistic generative models (described in Section 2.5). Unlike these models, however, optimization models provide no theoretically motivated interpretation of the metric that is optimized.

The well-known clock model of Povel and Essens (1985) is an optimization model that also contains some rule-based aspects. This model operates by generating a combinatorically exhaustive set of "clocks", defined by a unit (period) and location (the phase of the first event in the rhythm relative to the clock's period), calculating a score for each clock given a rhythm. Input rhythms are first preprocessed to mark events that, according to a set of rules, are predicted to be perceived as accented. The score that is calculated for each clock is based on how well the clock's ticks align with events marked as accented. The model selects the clock that optimizes this score, and its corresponding score is used to predict the degree to which the rhythm induces the clock.

Temperley and Sleator (1999) introduce another optimization approach, which they call a preference-rule model. Preference rule models are intended to be a computational implementation of the system of preference rules proposed by Lerdahl and Jackendoff (1983). Lerdahl and Jackendoff's ideas were influential, but lacking in formal rigor (see Hansen, 2010, for an extensive discussion), and preference-rule models are an attempt to address this. Temperley and Sleator propose independent models for meter and harmony. As before, our discussion considers only the meter model.

Preference rule models operate by generating an exhaustive set of analyses of a rhythm, specified by a set of well-formedness constraints. Each of these analyses receives a score based on a set of preference rules. Given an analysis and a piece of music, a preference rule yields a score representing the degree to which the analysis is preferred for the piece of music. The total score of the analysis is calculated

as a weighted linear combination of the scores of the individual preference rules. The analysis with the highest total score is the analysis that the model predicts to be correct.

For their meter model, Temperley and Sleator formalize three preference rules: the regularity rule, which prefers analyses in which beats are equally spaced, the event rule, which prefers analyses in which beats are aligned with events, and the length rule, which prefers analyses that align strong beats with the onsets of longer durations. Well-formed metrical hierarchies are constrained to contain exactly five metrical levels. Generalizing from Lerdahl and Jackendoff (1983), who based their theory primarily on music as notated in scores, Temperley and Sleator allow beats to be irregularly spaced. Analyses in which beats are regularly spaced are nevertheless preferred by the regularity rule.

The regularity rule is the only preference rule that depends only on the analysis, and not on its relation to a piece of music. This type of rule resembles the concept of a *prior probability* in probabilistic generative models. This probability represents the *a priori* probability of an analysis, which is independent of the rhythm that is analyzed. Preference-rule as well as generative approaches entail a trade-off between the *a priori* preferability of an analysis and its congruence with a piece of music: the more unlikely an analysis is *a priori*, the more strongly it needs to be supported by the piece of music.

Preference rule models have some advantages compared to rule-based models. Because preference rules represent soft constraints, they naturally allow for a certain degree of deviation from the norm: decreased congruence in one aspect (e.g., the degree to which beats are spaced evenly), may be compensated for by increased congruence in another (e.g., alignment of strong beats with notes). Furthermore, some rule-based approaches have been criticized for being opaque: it is difficult to describe regularities in their behavior based on the formulation of their rules, because the rules interact in complex ways (Desain & Honing, 1999). Preference-rule models, by contrast, have the benefit of being easy to interpret, because preference rules represent aspects of the relation between music and interpretation that are to be preferred.

A limitation of the optimization model of Povel and Essens (1985), but not necessarily of preference rule models (see Temperley, 2001, pp. 205–376), is that it does not consider the dynamic interplay between the unfolding music and the listener's perception and expectations. It is commonly emphasized (Longuet-Higgins & Steedman, 1971; Longuet-Higgins, 1978; Lee, 1991; Large & Kolen, 1994) that this interplay should be central to any account of rhythm perception.

2.4 The embodied perspective

The term embodied cognition carries a variety of connotations (Wilson & Golonka, 2013). Here we interpret it as emphasizing continuous dynamic interaction between brain, body, and environment, from which various behavioral and cognitive phenomena are *emergent* (Brooks, 1991b; Van Gelder, 1995; Chemero, 2009). This poses a contrast with the emphasis that cognitivist approaches place on strict information processing, which downplays the role of an agent's physical interaction with its environment. The emphasis on dynamic interaction is reflected in the type of models typically associated with this perspective, namely dynamical systems models (Chemero, 2009).

There is a class of cognitive models of rhythm perception that proposes that pulse and meter perception is based on *coupled oscillation* (Large & Kolen, 1994; McAuley, 1995). These models posit that constraints on meter, and how it is induced, emerge jointly from the dynamics of coupled oscillation. While cognitivist approaches describe meter perception as a cognitive mechanism that infers an abstract representation (meter), often in a bottom-up fashion, from perceptual input (rhythm), coupled oscillation models pose no sharp distinction between representations and a cognitive mechanisms that infer representations from sensory input.

Although coupled oscillation models do not necessarily emphasize a role for the body and environment in rhythm perception, they are compatible with two central tenets of embodied cognition: a rejection or downplay of the importance of cognitive representations (Anderson, 2003; Wilson & Golonka, 2013) and a rejection of the idea that cognition is most appropriately described in terms of computation and symbol manipulation (see Van Gelder, 1995). This sentiment is reflected strongly in the fragment below, which appears in an introduction to neural resonance models (a type of coupled oscillation models) of music perception (Large, 2010b).

> The brain does not "solve" problems of missing fundamentals, it does not "compute" keys of melodic sequences, and it does not "infer" meters of rhythmic input. Rather, it *resonates* to music. (p. 201, italics occur in original).

Coupled oscillation models can account for a remarkable number of phenomena in rhythm perception, without resorting to domain-specific constraints. [4] Among

[4] Coupled oscillation models do arguably incorporate some domain-specific constraints: Adaptive and neural oscillators need to be tuned to frequencies that are relevant to musical rhythms. Period coupling and temporal receptive fields in adaptive oscillator models are explicitly introduced to account for music perception and do not occur in physical coupled oscillation systems such as clocks suspended from the same beam or metronomes on the same moving platform.

these phenomena are aspects that are argued to pose challenges for other approaches, namely tracking a beat in rhythms with tempo changes (Large & Jones, 1999) or expressive timing (Large & Palmer, 2002), and entraining to syncopated rhythms in which the pulse frequency is absent from the Fourier spectrum of the rhythm (Velasco & Large, 2011). Coupled oscillation models have therefore emerged as popular models of rhythm perception. Below, two types of coupled oscillation models are discussed: *adaptive oscillator models* and *neural resonance models*.

2.4.1 Adaptive oscillator models

McAuley (1994) proposes the term *adaptive oscillator* for a class of oscillators that adapt their period in response to external rhythms. McAuley (1994, 1995) and Large and Kolen (1994) independently (McAuley, 1995, p. 67) proposed oscillators of this type as models for rhythm perception. McAuley (1995) describes the theoretical status of these models as somewhere in between a "single-neuron model and that of a psychological theory". Large and Kolen (1994) describe their model as representing "a single abstract processing unit, amenable to connectionist implementation". Thus, both proposals describe these models as abstract, rather than mechanistic, accounts of rhythm perception (in contrast to neural resonance models).

Different aspects of the behavior of coupled oscillators may be connected to different aspects of rhythm perception. Large and Kolen (1994) describe an oscillator that synchronizes with a periodic component of a rhythmic pattern as "embodying the notion of musical pulse, or beat". Similarly, McAuley (1995, p. 12) writes that oscillators model "global dynamics of perceptual mechanisms involved in the processing of rhythmic patterns". Metrical hierarchy is proposed to emerge from two or more internal oscillators entraining to each other, as well as to an external rhythm (Large & Kolen, 1994; McAuley, 1995). McAuley (1995) also raises dimensionality reduction, or efficient memory encoding, as a motivation for the approach: an oscillator may be seen as an efficient memory representation of where the pulse is. Finally, it is often stated that the oscillators encode a *prediction* or *expectation* of when events are expected.

An oscillator produces periodic behavior that is described by two state variables, period, p, and phase, ϕ. Period represents the amount of time required for an oscillator to complete its cycle. Phase represents the relative position of the oscillator within its cycle and evolves from zero to one, at which point it wraps back to zero. An oscillator is sometimes said to "fire" when its phase reaches zero.

The paragraphs below describe a general mathematical framework for adaptive oscillator models. This framework is limited to describing how one oscillator is

influenced by another. Extensions to two endogenous oscillators that entrain to different periodicities in a rhythm are described by Large and Jones (1999) and Large and Palmer (2002). The oscillator that is being influenced is the *endogenous* oscillator: a source of endogenous oscillations. The other is an *external* "oscillator", which, in adaptive oscillator models, is not really an oscillator, but an external rhythm. Causation is unidirectional: the external oscillator (the rhythm) exerts influence on the endogenous oscillator, but the endogenous oscillator has no influence on the behavior of the external oscillator. This influence is called *coupling* and causes the phase and period of endogenous oscillator to be perturbed by activity of the external oscillator.

Adaptive oscillator models can be evaluated in a sequence of discrete time steps. Because the phase and period are only perturbed by the firing of the external oscillator (or the presence of an onset, since the external oscillator is a rhythm), the dynamic behavior of the system can be described entirely by considering only the *relative phase* of the two oscillators at moments when the external oscillator fires. The relative phase is the (circular) difference between the endogenous and external oscillator's phase. The relative phase at the $(n+1)$th firing of the external oscillator, given the relative phase at the nth firing and the periods of the endogenous and external oscillators, is described by

$$\phi_{n+1} = \left(\phi_n + \frac{q}{p}\right) \mod 1, \tag{2.1}$$

where q represents the period of the external oscillator and p represents the period of the endogenous oscillator This equation is known as a *circle map*. In the above form, it describes the relative phase of two uncoupled oscillators (e.g., two metronomes ticking away independently at their own tempos).

Coupling is introduced by allowing the external oscillator to perturb the phase of the endogenous oscillator. Such *phase coupling* is incorporated by adding a coupling term to the circle map.

$$\phi_{n+1} = \left(\phi_n + \frac{q}{p} + \eta_\phi F_\phi(\phi_n)\right) \mod 1. \tag{2.2}$$

Here, η_ϕ is a parameter that controls the coupling strength. The term $F_\phi(\phi_n)$ is the coupling function, which, given the relative phase, calculates the amount by which the phase is perturbed.

To model the relative phase of an endogenous oscillator and a rhythm, the "period", q of the external oscillator is replaced by the nth inter-onset interval, i_n, in a rhythm. The equation below illustrates this.

$$\phi_{n+1} = \left(\phi_n + \frac{i_n}{p} + \eta_\phi F_\phi(\phi_n) \right) \mod 1. \tag{2.3}$$

If t_m is the mth onset in a rhythm, then the nth inter-onset interval is $i_n = t_{n+1} - t_n$.

Musical rhythms tend to fluctuate in tempo in a way that listeners can track (Repp, 2005). To account for this, adaptive oscillator models implement *period coupling*. The ability of oscillators to adapt their period to a rhythm motivates McAuley (1994) to call these oscillators *adaptive*.

The function below calculates a new period at each onset t_n.

$$p_{n+1} = p_n + p_n \eta_p F_p(\phi_n). \tag{2.4}$$

The function is parameterized by a coupling-strength parameter, η_p, and a period coupling function, $F_p(\phi_n)$, which calculates the change in period given a relative phase ϕ_n.

Given suitable coupling functions, the endogenous oscillator will *entrain* (be driven to fire in synchrony) to an approximately periodic rhythm. The degree to which this happens depends on how close the period p of the oscillator is to a (sub)harmonic of the period of the rhythm, q.

The dynamic behavior of the system can be visualized by *regime diagrams*. Such illustrations (e.g., see Large & Kolen, 1994) visualize the time it takes for an oscillator settle into a mode-locked state as a function of the coupling strength and the ratio between the endogenous oscillator's period and a driving pulse. [5] Regime diagrams reveal regions centered around p/q values, where p and qs are small integers, in which stable phase-locked (entrained) states emerge readily. These *entrainment regions* are wider around points where the ratio between p and q can be expressed by small integers (e.g., 1:1, 1:2, 2:3), and increase in width as coupling strength increases. Entrainment regions describe the constraints on pulse and meter perception predicted by adaptive oscillator models.

The oscillator described so far is easily disturbed by rhythms that contain onsets far from where the oscillator "expects" the onset. To allow an oscillator to entrain to a single periodic component in a rhythm that contains more onsets apart from periodic ones, Large and Kolen (1994) propose period and phase coupling functions for which the strength of their effect depends on how close the onset occurs to where the endogenous oscillator predicts it to occur. The further an onset deviates from the prediction, the smaller the influence it exerts on the endogenous oscillator's phase and period. This endows the oscillator with what

[5] Mode-locking is a generalization of phase-locking that describes states in which one oscillator aligns its phase with another oscillator exactly every n cycles (where n is an integer).

Large and Kolen call a *temporal receptive field*. The width and sharpness of the temporal receptive field are parameterized. In Large and Kolen's model, these parameters remain fixed throughout a simulation, but Large and Palmer (2002) propose a model in which the temporal receptive field sharpens as onsets occur closer to where they are predicted.

In Large and Kolen's model, another pair of parameters specifies a lower and upper bound on the oscillator's period. The oscillator's initial period, called its resting period, lies halfway between the lower and upper bound. When no onsets are encountered within its temporal receptive field, the oscillator maintains its current period. Large and Kolen associate this behavior with the tendency of a pulse percept to be sustained in the absence of rhythmic events (Cooper & Meyer, 1960).

A set of adaptive oscillator models proposed by McAuley (McAuley, 1993, 1994, 1995), are similar to Large and Kolen's adaptive oscillator. Instead of gradual phase adaptation, McAuley's models reset their phase based on the relative phase. Furthermore, these oscillators have a resting period to which they gradually return in absence of inputs.

Both Large and Kolen (1994) and McAuley (1995) have associated their models with dynamic attending theory (Jones & Boltz, 1989). Large and Jones (1999) present an adaptive oscillator model where self-sustained oscillations take on the role of *attending rhythms* (Jones & Boltz, 1989). An attending rhythm consists of periodic pulses of attention, defined as periods of sharp perceptual acuity during which an event is anticipated. In Large and Jones' model, an attentional pulse is implemented by a bell-shaped probability density function centered around phase zero. The width of this distribution is governed by a concentration parameter, reflecting attentional focus. As synchronization (due to entrainment implemented by sinusoidal phase and period coupling) increases, the pulse of attention sharpens, focusing attention in time.

The adaptive oscillator of Large and Jones is further developed by Large and Palmer (2002). In this extension, phase and period coupling depend on the strength of attention (as indexed by the attentional pulse), creating a temporal receptive field. However, since the width of the attentional pulse depends on the degree of synchrony between the oscillator and the rhythm, this temporal receptive field narrows as synchrony increases. Increased synchrony thus leads the oscillator to become less sensitive to events deviating far from where the beat is expected and more sensitive to events close to where the beat is expected.

Large and Jones report simulation results of a model in which two adaptive oscillators, both driven by a rhythm, are bidirectionally coupled to each other. Inter-oscillator coupling is defined such that the oscillators are driven towards either a 2:1 or 3:1 period ratio, to ensure metrical entrainment between them. Simulations

carried out by Large and Palmer (2002) show that such inter-oscillator coupling can improve entrainment stability in tracking expressive piano performances, and that the models can be used to detect phrase-boundaries based on phrase-final lengthening, and also to detect melody notes in chords which are accentuated by being timed slightly early (melody leads). Furthermore, Large and Palmer show that in performances performed with strong rubato, detection of melody leads can in certain situations improved beat tracking performance of the adaptive oscillator model.

Large and Palmer (2002) also describe some of their adaptive oscillator model's limitations. First, the model is not suitable for finding the initial beat and has to be provided with this information. Furthermore, for inter-oscillator coupling, functions that actively drive the oscillators to the desired metrical ratios are required, introducing an asymmetry between rhythm-to-oscillator and oscillator-to-oscillator coupling.

2.4.2 Neural resonance models

Interactions between excitatory and inhibitory populations of neurons can give rise to neural oscillations (Large et al., 2015). It has been hypothesized that pulse and meter perception are emergent phenomena of such oscillations (Large, 2008; Large & Snyder, 2009). The neural resonance theory of rhythm and meter perception proposes an explanation based on a mathematical description of a biological mechanism, rather than an abstract model like an adaptive oscillator.

At a high level of mathematical abstraction, the dynamics of neural oscillations may be described by a *canonical model* (Large, 2008), which describes dynamical behavior that is shared by a large class of more detailed models. *Neural resonance models* of pulse and meter perception are based on a canonical model of neural oscillation.

These models share some characteristics with adaptive oscillator models. Both approaches propose explanations of pulse and meter perception based on coupled oscillation. The phase dynamics (but not amplitude dynamics) of neural resonance models can also be described by a circle map (Large, 2008). Unlike adaptive oscillator models, however, neural resonance models posit a specific neural mechanism from which oscillations arise. Furthermore, neural oscillators do not exhibit period coupling: they oscillate near their natural frequency, which does not adapt to tempo fluctuations. Instead neural resonance models posit *gradient-frequency networks* of neural oscillators, in which oscillators with natural frequencies close to (harmonics of) periodicities in the rhythm resonate to the rhythm (Velasco & Large, 2011; Large et al., 2015).

Neural resonance models exhibit three behaviors associated with different aspects

of pulse and meter perception (Large, 2010b). *Spontaneous oscillation* relates to the perception of pulse, and, in particular, the tendency for pulse to persist in the absence of external events. *Entrainment* of neural oscillations to external rhythms reflects the perception of a periodic pulse in rhythms that are not strictly periodic. *Higher-order resonance*—the capacity of neural oscillation to resonate at harmonics or sub-harmonics of periodicities in a rhythm—is posited to account for meter induction and for the perception of pulse at frequencies that are absent from the Fourier spectrum of a rhythm (Velasco & Large, 2011).

2.4.3 Coupled oscillation models and enculturation

Large (2010a) describes how neural resonance models can be extended with plastic inter-oscillator connections that adapt their strength based on the principles of Hebbian learning. Large et al. (2015) note that this makes it possible to simulate the effect of enculturation on neural resonance models. Recently, Tichko and Large (2019) proposed a simulation of effects of exposure to music with non-isochronous meters in infants observed by Hannon and Trehub (2005a, 2005b), using a gradient-frequency network of neural oscillators inter-connected by plastic connections. This model resembles to a model proposed by Large et al. (2015), which consists of two networks connected to each other. One represents a sensory network that receives input from a rhythm, the other represents a motor network that is connected via bidirectional coupling to the sensory network. Citing findings of limited development of movement-to-rhythm synchronization in infants, Tichko and Large only use a sensory network.

To represent exposure to non-isochronous meters in Balkan music and isochronous meters in Western tonal music, Tichko and Large expose two instantiations of their network to a different rhythm. One network is exposed to a 4/4 rhythm, intended to represent exposure to Western tonal music, the other to a 7/8 rhythm, intended to represent exposure Balkan music in a non-isochronous meter. Both networks, and a third network without prior exposure, are exposed to six rhythms: the two training rhythms, and two modified versions of each training rhythm: one that preserves the meter and one that violates it. Analyzing the response of the networks to the different rhythms, Tichko and Large show an effect of training that resembles the results obtained by Hannon and Trehub (2005a, 2005b). However, because these results are based on simulations involving a single training and a single test rhythm per condition, it remains unclear how robust they are to variation in the specific rhythms used for training and testing. Furthermore, the results appear to be influenced by the frequencies of periodicities in the training rhythm. It could be that , which may have influenced the results more than the type of meter (isochronous or non-isochronous) used.

Large (2010b) suggests that innate constraints on music perception may emerge

from the intrinsic dynamics of the brain. An important question for these models is therefore whether the dynamic behavior of neural resonance models sufficiently explains empirically observed variance in metrical entrainment behavior. Both adaptive oscillators and neural resonance models predict that perceived pulses are constrained to be isochronous. Coupled oscillation models can entrain to periodic components related by simple integer ratios, which may be polyrhythmic (such as 3:2 or 4:3), but they cannot entrain to a non-isochronous beat. For example, while the networks described by Tichko and Large (2019) resonate to rhythms in (isochronous) 4/4 and (non-isochronous) 7/8 meters, they do not entrain to the non-isochronous tactus level of the 7/8 meter. It could be that, as Large (2008, p. 221) suggests, non-isochronous meters are "compelling specifically because they thwart an intrinsic expectation of periodicity". However, as discussed in Section 2.2, there is evidence suggesting that non-isochronous meters are readily processed given familiarity with music in which they are prevalent (Hannon & Trehub, 2005b; Soley & Hannon, 2010; Hannon et al., 2012). There also is evidence that, given the right kind of training, metrical entrainment to non-isochronous beat-subdivisions is possible (Polak et al., 2016; Polak et al., 2018). Whether rhythm perception shows more flexibility than the constraints of coupled oscillation models allow remains an active topic of discussion.

2.5 The predictive processing perspective

The predictive processing perspective (described elaborately by Clark, 2013) builds on a class of theories that notably include predictive coding (Rao & Ballard, 1999) and the Bayesian brain hypothesis (Knill & Pouget, 2004). It has recently come to be associated with a number of different theoretical perspectives that range from cognitivist to radically embodied (for discussions, see Allen & Friston, 2018; Wiese & Metzinger, 2017). Nevertheless, these perspectives share a commitment to the idea that perception is based on minimizing prediction error, which in turn is based on Bayesian inference. Predictive processing theories propose a domain-general mechanism that underlies both perception and perceptual learning.

More specifically, predictive processing posits that perception and cognition involve an internal, multilayered generative model of sensations. This model can be represented as a Bayesian network: a directed acyclic graph that may be interpreted to reflect causal dependencies between random variables (Pearl, 2000). Sensations are considered to be the result of a stochastic generative process (the environment), that is predicted by the outcomes generated by leaf nodes of an (internal) generative model. The better the generative model resembles the generative process underlying sensations, which involves the underlying environmental causes of sensations, the more accurately sensations can be predicted. Prediction error, which is continuously generated by the discrepancy between observed and

predicted outcomes, revises the generative model to better predict future sensations. These changes, that are driven by prediction error, are hypothesized to underlie both perception and perceptual learning.

Note, however, that this process, and its implied outcome (convergence to a perfect, barring physiological limitations, generative model of the environment), is argued to be significantly altered when the role of *action*—the ability of an organism to influence the flow of sensory stimulation and to shape its environment—is considered (see Clark, 2016). This role can be integrated into predictive processing to create what Clark (2013) calls *action-oriented* predictive processing. In any case, sensitivity to the statistical structure of the environment plays a significant role in all predictive processing accounts (action-oriented or not). The probabilistic generative models discussed in this chapter are passive models that do not account for effects that action may have on their input. Such effects remain an important topic for future research, together with the question of how they interact with effects of passive exposure simulated by probabilistic generative models.

Prediction-error minimization in predictive processing is equivalent to probabilistic inference in a probabilistic generative model. In such a model, random variables upon which outcomes are conditioned are called *latent* variables. Latent variables cannot directly be observed, but their probability distribution may be inferred through probabilistic inference. The marginal distribution of observed variables corresponds to the generative model's predictions of stochastic outcomes. This distribution assigns a probability to every possible outcome of the generative process. The probability of a given observation is known as the *model evidence* for that observation.

Prediction error is operationalized by the negative logarithm of the model evidence, such that minimizing prediction error corresponds to maximizing model evidence. This quantity corresponds to a measure of *information* defined in information theory (Shannon, 1948). A system that minimizes prediction error thus also minimizes information transmission. The intuition behind this is that only parts the sensory signal that have not already been predicted by the generative model need to be considered.

The approaches discussed below estimate their parameters directly from samples of empirical data, annotated with underlying structure (meter), using a maximum likelihood approach. These samples are called the *training data* of the model. The maximum likelihood approach ensures that the estimated parameters cause the model to assign the maximum possible probability to the training data. This training process has been used to simulate the effects prior exposure to music on rhythm perception. Section 2.5.6 describes in more detail how probabilistic generative models can be used to simulate enculturation.

It is worth noting that the maximum likelihood approach is different from the

so-called "fully Bayesian" approach, in which a model describes its own parameters as random variables. This allows the models to infer their own parameters from data using probabilistic inference, and eliminates the need for annotated (also known as labeled) training data, allowing the model to "bootstrap" itself off mere observations. The distinction between a training phase in which the model is parameterized and a testing phase in which the model is evaluated thereby also disappears. Furthermore, the fully Bayesian approach accounts in a principled way for uncertainty that the model has about its own parameters. Such uncertainty plays an important role in predictive processing (see Clark, 2016), but is beyond the scope of this chapter.

2.5.1 Probabilistic generative modeling of rhythm perception

Temperley (2007) makes a strong case for probabilistic approaches to music perception based on the hypothesis that knowledge of musical style is probabilistic in nature and inferred by listeners from regularities in the music they have been exposed to. The basic framework outlined by Temperley applies to all models described in this chapter, and also corresponds to the predictive processing framework described above. In this framework, observed variables represent the *musical surface* and latent variables represent its *underlying structure*. For rhythm models, the musical surface corresponds to a pattern of event times, and its structure corresponds to some conceptualization of meter. In predictive processing terms, the musical surface is the outcome of a generative process involving latent variables that represent perceptual concepts like meter.

A generative model of rhythm perception may represent rhythms and meter by multiple random variables, but to obtain a compact representation, these variables can be merged into two variables, namely R (for rhythm) and M (for meter). According to the product rule of probability, the joint distribution of these variables, $p(R, M)$, can be written in one of the following ways:

$$p(R, M) = p(M \mid R)p(R) = p(R \mid M)p(M). \tag{2.5}$$

Since the goal is to describe the *generative* process underlying rhythms, generative models of rhythm perception aim to estimate the factors $p(M)$—the *a priori* probability of a meter—and $p(R \mid M)$—the probability of a rhythm given a meter. When these factors appear in Bayes' theorem, as shown below, they are known as the *prior distribution* and the *likelihood function*. It follows from Equation 2.5 that the probability of a meter given a rhythm, $p(M|R)$, can be expressed in terms of the generative model as follows:

$$\overbrace{p(M|R)}^{\text{posterior}} = \frac{\overbrace{p(R|M)}^{\text{likelihood}}\overbrace{p(M)}^{\text{prior}}}{\underbrace{p(R)}_{\text{model evidence}}}. \tag{2.6}$$

This equation is known as Bayes' theorem, and forms the basis of *probabilistic inference* in generative models. Since it enables inferring the distribution of latent variables given an observed outcome, inference is sometimes called the *inversion* of a generative model (MacKay, 2003).

Equation 2.6 reveals some similarities between probabilistic generative models and preference-rule models (described in Section 2.3.2). Since model evidence is independent of latent variables, the posterior probability of a meter is influenced only by the two factors in the numerator of the fraction on the right-hand side of the equation. Meters that are probable *a posteriori* strike a balance between *a priori* probability and the probability of the rhythm given the meter. Meters that are *a priori* improbable require strong bottom-up evidence to be *a posteriori* probable, compared to meters that are *a priori* probable. A similar dynamic interaction occurs in preference-rule models, which postulate rules that apply only to a given metrical analysis (e.g., the regularity rule), comparable to a prior distribution, and rules that measure the fit between an analysis and a rhythm (e.g., the event and length rules), comparable to a likelihood function.

Model evidence, recall, is the probability that a generative model assigns to an outcome. In (2.6), it is given by the denominator of the fraction, which may also be written as

$$p(R) = \sum^{M} p(R|M)p(M). \tag{2.7}$$

Differences between generative rhythm perception models, which are all compatible with this general framework, reside in how the prior distribution and likelihood function are implemented. The sections below describe different possibilities that have been explored in the literature. Were applicable, we describe how these possibilities are applied in different generative models of rhythm proposed by Temperley (2007, 2009) and Van der Weij et al. (2017).

2.5.2 Rhythmic outcomes: grids, intervals, and phases

How a generative model represents a rhythm corresponds to how the stochastic outcomes of the model should be interpreted. All of the models we discuss below represent rhythms as temporally ordered sequences of outcomes. These sequences

depend only on *note-onset times*: the times at which note events begin (i.e., when they are played, struck, plucked, or sung). Temporal intervals are always represented as integer multiples of some atomic temporal unit, which may either be an absolute duration (e.g., 50 milliseconds), or a symbolic score-duration (e.g., a sixteenth note). However, the models differ in whether they represent rhythms by grids of temporal bins, sequences of temporal intervals, or more abstractly in terms of the metrical functions of notes. Below, we introduce a distinction between four types of models: grid, interval, phase and metrical salience models.

In *grid* models, stochastic outcomes represent temporally adjacent grid cells, each of which represents an atomic temporal interval. Outcomes in such models are binary variables representing *whether* an onset occurs within (or at) the current grid cell, or whether it remains silent. Grid models, in other words, predict *what* happens at the current moment.

Interval models, by contrast, predict *when* an onset occurs relative to the last onset. In interval models, outcomes represent the temporal interval between two note-onsets: the inter-onset interval. When a model is temporally discrete, this intervals is often one out of prespecified set of possibilities that may occur with non-zero probability. This set of possibilities is sometimes called an alphabet (Conklin & Witten, 1995).

Phase and *metrical salience* models predict a more abstract property of the next event, namely its metrical function. By the *phase* of an onset, we mean its position in a metrical cycle denoted by bars notated in a score. By *metrical salience*, we mean the highest metrical level in which a beat associated with the current onset occurs. Phase and metrical salience representations are *variant* with respect to meter: how a given note-onset event is represented, depends its metrical interpretation. Given a meter and the position of bar lines, predictions of metrical salience or phase do not predict a unique point in time, but constrain the possible points in time at which an event may occur in order to be in agreement with the prediction. Like interval models, phase and metrical salience models predict *when* onset occurs, but do so more abstractly.

A final important aspect of representation, which is of relevance generative rhythm models, is that the *granularity* of a representation affects expected prediction error. Predictions that have a low temporal granularity are more likely to be correct, since they are consistent with a large number of events. When the granularity of a representation depends on the value of a latent variable, as is the case for phase and salience models, this introduces a (possibly unintended) bias into the model. Since in phase and salience models, the temporal granularity of a prediction depends on the period of the metrical cycle, such models are susceptible to biases favoring meters with shorter metrical cycles.

2.5.3 The prior probability of meters

The prior distribution of meters, $p(M)$, describes the probability of meters independently, that is, without considering observations (which represent bottom-up sensory input). Van der Weij et al. (2017) employ a categorical distribution that reflects the relative frequency of meters derived from notated time signatures in empirical training data. This approach makes no assumptions about the internal structure of meter, assuming that this structure may be culture-specific. On the other hand, it has no way of estimating the probability of meters that do not occur in its training data.

Models proposed by Temperley (2007, 2009) employ prior distributions based on a hierarchical view of meter consistent with ideas of Lerdahl and Jackendoff (1983). In these prior distributions, a meter is generated by a set of stochastic outcomes, represented by different random variables, such as the duration of a tactus interval, whether tactus beats are grouped by two or three, and whether tactus beats are subdivided into two or three sub-tactus beats. Compared to the approach of Van der Weij et al., this prior requires fewer parameters, and can, due to its compositional nature, estimate the probability of meters not occurring in training data. On the other hand, it makes assumptions about the structure of meter that may be specific to the Western musical idiom.

Priors may also be based on abstract theoretical measures. Studying the production and categorical perception of interval ratios, Sadakata, Desain, and Honing (2006) assign prior probabilities to interval ratios that are proportional to a theoretical quantification of the ratio complexity. Such priors are consistent with the hypothesis that due to cognitive constraints, some meters may be generated more readily than others.

2.5.4 Likelihood functions: generating rhythms from meters

To illustrate how the design of the likelihood function affects which cues for meter a generative model is sensitive to, we discuss six models described by Temperley (2010) in a model comparison study investigating the probabilistic principles underlying what the study calls "common practice rhythm". Unlike the multilayered generative models of Temperley (2007, 2009) and Van der Weij et al. (2017), these models assume that the meter is known and fixed. [6] The six models can be distinguished by two aspects of their design: the representation of rhythms and metrical structure, and the probabilistic independence assumptions

[6] By a multilayered generative model we mean a generative model that conditions observations on underlying latent variables.

Table 2.1: An overview of six likelihood functions discussed by Temperley (2010). The middle column indicates whether each model uses a grid, interval, or phase representation. The right-most column indicates which representation of metrical context observations are conditioned on. In the right-most column, the empty set symbol $\emptyset$ indicates that outcomes are modeled independently.

Model	Representation	Metrical context
Uniform Position Model	Grid	N/A
Zeroth-Order Duration Model	Interval	N/A
Metrical Position Model	Grid	Salience
Fine-grained Position Model	Grid	Phase
Hierarchical Position Model	Grid	Salience, Metrical anchoring
First-Order Metrical Duration Model	Phase	Previous phase

they make. These aspects are summarized in Table 2.1 for the six models that Temperley (2010) presents, and the paragraphs below discuss them in more detail.

Regarding representation, Temperley distinguishes between "position models" and "duration models". As Table 2.1 shows, four different position models, and two duration models are discussed. The four position models correspond to what we call grid models. Grid cells, in this case, correspond to eighth-notes. Of the two duration models discussed, one corresponds to what we call an interval model, while the other is a phase model. [7]

The number of probabilistic independence assumptions made by a model must fall somewhere in between two extremes. At one extreme, each random variable depends on all other random variables, which corresponds to a fully connected Bayesian network. At another extreme, the outcome of each variable is assumed to be independent of all other outcomes, which corresponds to an unconnected Bayesian network. The Uniform Position Model, which models the independent probability of an onset at a grid cell, and the Zeroth-Order Duration Model, which models the independent probability of inter-onset intervals, posit only a single variable at each time step. In the search for a model that balances prediction performance with complexity, these models may be seen as baselines against which the effect of progressively removing independence assumptions from the models may be compared.

Some generative models can be evaluated incrementally over a sequence of time steps. This is possible only when variables are conditioned on no other variables other than those occurring in the same or the preceding time steps. Models in which variables in each time step are conditioned on variables in the n immediately

[7] We use the term *phase* for what Temperley calls *metrical position*.

preceding time steps are called nth-order *Markov models*. For example, in zeroth-order Markov models, outcomes are independent of preceding outcomes, while in first-order Markov models, outcomes depend on variables in the preceding time step. Except for the First-Order Metrical Duration Model and the Hierarchical Position Model, all models compared by Temperley are zeroth-order. The Hierarchical Position Model is not a Markov model: it conditions outcomes at a given metrical level on outcomes at higher metrical levels. Rhythms are generated hierarchically, rather than in temporal order, by this model.

Within a time step, the presence or absence of independence assumptions may incorporate sensitivity to metrical structure into a model. The Uniform Position model and Zeroth-Order Duration Model are not sensitive to metrical structure, that is, the probability of their outcomes is independent of meter. The Metrical Position Model, Fine-Grained Position Model, and Hierarchical Position Model, however, condition outcomes on the metrical status of a grid cell. Of these, the Fine-Grained Position Model differs from the other two in the representation of metrical status: outcomes are conditioned on the phase of a grid cell, while in other two models they depend on the metrical salience of a grid cell. In the Hierarchical Position Model outcomes depend on the metrical salience of the current grid cell and on whether the surrounding metrically stronger beats contain onsets. In Table 2.1, this situation is referred to as *metrical anchoring* (Temperley, 2009).

Another means of introducing sensitivity to meter into a model is by choosing a representation of outcomes that is itself sensitive to meter. This strategy is employed in the First-Order Duration Model, which is a phase model: it predicts the *phase* of an outcome. Since this is a first-order Markov model, the probability of a phase is additionally conditioned on the previous outcome.

The Metrical Position Model is a grid model which most faithfully embodies the theory that the frequent occurrence of onsets on metrically strong beats is a strong cue for meter (Palmer & Krumhansl, 1990). The model conditions the probability of an onset at a grid cell on the metrical salience of that grid cell. The metrical salience representation has a lower temporal granularity than the phase representation: for example, the second and fourth beat of a 4/4 bar have different phases, but are indistinguishable by metrical salience. If the assumption that metrical salience, rather than phase, most strongly predicts onset likelihood is true, then models based on metrical salience would more compactly capture statistical patterns in common-practice rhythms than models based on phase, and the Metrical Position Model should perform as well as the Fine-Grained Position model, despite having fewer parameters.

A phase representation, on the other hand, assumes periodicity of meter, but otherwise makes few theoretical commitments its organization. For example, the phase representation of a rhythm does not depend on whether the underlying meter is 3/4 or 6/8. A phase model may be able to distinguish between these meters, but

differences between them must be encoded in a probability distribution of phases that is conditioned on meter. These difference may be learned during model training, where the parameters of the model are estimated from empirical training data. In any case, this aspect is irrelevant in Temperley's model comparison study where all considered rhythms have a 4/4 meter.

Temperley evaluates the performance of these six models in terms of the per-rhythm cross-entropy (the negative logarithm of model evidence). [8] The definition cross-entropy is identical to that of prediction error. Results can therefore be interpreted as representing how well the models predict rhythms in the style of the chosen samples. The objective is to investigate which general principles underlie the composition of what Temperley calls "common-practice rhythm". Accordingly, the models are trained and evaluated on empirical samples of European folksongs and first-violin parts of string quartets by Mozart and Haydn.

The results show that in general, the four models sensitive to metrical structure achieve better prediction performance than those not sensitive to such structure. Overall, the First-Order Metrical Duration Model achieves the best performance, and the Fine-Grained Position Model outperforms the Metrical Position model. Both of these models are based on a phase representation, suggesting that, even in common-practice rhythm, phase may provide greater predictive power for the timing of notes in the empirical samples of Western music than metrical salience. However, the model comparison does not include a first-order salience model, which would allow for a more elaborate comparison of phase and salience representations.

While the best performance is achieved by the First-Order Metrical Duration Model, the Hierarchical Position Model achieves comparable performance using significantly fewer parameters. Taking this into account, Temperley concludes that the Hierarchical Position Model most accurately captures statistical properties of common-practice rhythms. Findings of Holzapfel (2015), and London et al. (2017), however, suggest that the relatively strong performance of this model might not generalize well to non-Western musical idioms, in which metrical salience sometimes is less predictive of onset probability.

Some of the likelihood functions described above are used in multilayered generative

[8] It should be noted that there are subtle issues, not mentioned by Temperley, involved in comparing these results between different (grid, interval, or phase) representations. For example, a grid representation of a rhythm is (1 0 1 0 1), which is a sequence of five binary outcomes. This rhythm is one of $2^5 = 32$ possible rhythms. In an interval representation each outcome is one of x possible intervals. For a model that considers $x = 8$ intervals per outcome with non-zero probability, the same rhythm, represented as (2 2), is one of $8^2 = 64$ possible outcomes. In a phase representation, assuming atomic temporal units of quarter notes and a meter with a period of four quarter notes, the same rhythm is represented as (0 2 0), which is one of $4^3 = 64$ possibilities. Predicting one out of sixty-four possibilities is more difficult than predicting one out of thirty-two possibilities. Grid models are thus likely to have an advantage over interval and phase models.

rhythm models. In particular, Temperley (2007, 2009) proposes grid models, in which the grid cells (or *pips*) represent not symbolic score-units but real absolute durations. The model described by Temperley (2007) uses the Metrical Position Model as its likelihood function. The Hierarchical Position Model is used as the likelihood function in the model described by Temperley (2009). However, this model is not a Markov model and violates temporal incrementality. Accordingly, the model is presented primarily as a music analysis model, rather than a music perception model. Both models contain a number of variables that accommodate a certain degree of freedom in tempo and timing, but these aspects are beyond the scope of this chapter. Another multilayered generative model of rhythms described recently by Van der Weij et al. (2017) uses a different representation and a different likelihood function. The next section describes this model in more detail.

2.5.5 Modeling sequential structure in rhythms

The models described so far are based on zeroth- or first-order Markov models, or on hierarchical models (Temperley, 2009, 2007, 2010). Van der Weij et al. (2017) instead propose a probabilistic generative model using a *variable-order* Markov model. In this model, events (outcomes) are conditioned on all preceding events in a sequence that represents a rhythm. This is achieved using a modeling technique called prediction by partial match (PPM), proposed originally as a data compression method (Cleary & Witten, 1984).

Instead of a grid or phase representation, Van der Weij et al. (2017) propose a representation of outcomes that is sensitive to meter, but maps one-to-one onto inter-onset intervals. The representation combines the phase of an onset with the number of metrical cycles elapsed since the last onset: it encodes the temporal interval between the current event and the bar-level downbeat preceding the last event. The representation is referred to here as the *downbeat distance*. This representation ensures that the temporal granularity of predictions is independent of meter. It therefore avoids biases that depend on the period of the metrical cycle into the model (see Section 2.5.2).

Compared to zeroth-order models, a variable-order Markov model of events represented by downbeat distances widens the range of cues for meter that Van der Weij et al.'s model is sensitive to. The probability of an event given a meter depends not only its metrical context, but also on the downbeat distances of previous events. This changes to role of meter from a periodic template of onset probabilities (Palmer & Krumhansl, 1990; Temperley, 2007) into a periodic temporal reference with respect to which patterns of events are interpreted and remembered. It allows a model to learn rhythmic patterns that occur predictably in particular metrical contexts. For example, it may be the case that in a hypothetical

musical sample, syncopations occur predictably in certain contexts, even though in the same style, notes generally begin on metrically strong beats. Such predictable deviations from the norm would be undetectable in event-frequency distributions (Palmer & Krumhansl, 1990; Holzapfel, 2015; London et al., 2017), which are sensitive to only to zeroth-order statistical properties of rhythms.

Simulations performed by Van der Weij et al. suggest that variable-order Markov modeling improves prediction performance of rhythms derived from German folksongs. Applying different variants of their model, in which the maximum order of the variable-order Markov model (the order bound) varies between zero and four, they find that the prediction of rhythms derived from German folksongs improves when the order bound is increased. The performance gain is most pronounced between zeroth-order and first-order modeling, but small improvements occur beyond first-order models.

The increased complexity of the relation between rhythm and meter supported by Van der Weij et al.'s model may improve the model's applicability to music from different cultures. In music from, for example, regions in western Africa (Locke, 1982) and the African diaspora (Iyer, 1998), it is common for onsets to occur consistently on beats that, according to a Western theoretical understanding of meter (Longuet-Higgins, 1978; Lerdahl & Jackendoff, 1983), are metrically weak instead of on metrically strong beats. Findings presented by London et al. (2017) illustrate this quantitatively for Malian djembe music, and Holzapfel (2015) shows that rhythms in Turkish makam music also deviate from norms based on patterns of metrical salience. It nevertheless remains an open question whether these observations warrant the level of flexibility in the relation between rhythm and meter afforded by Van der Weij et al.'s model. Comparing the performance of different probabilistic generative rhythm models on culturally diverse samples of rhythms may provide more insight into this matter.

The applicability of Van der Weij et al.'s model to rhythms from diverse musical cultures is somewhat hampered by its reliance on Western music notation. Music notation plays little or no role in many musical traditions around the world, and transcribing music from these traditions in Western music notation may not be appropriate. For example, Western music notation's emphasis on temporal intervals related by small-integer ratios cannot naturally express so called "swung" beat subdivisions, as found, for example, in jazz music (Honing & De Haas, 2008) and Malian djembe music (Polak et al., 2016; Polak et al., 2018).

It appears, in any case, that the model partially fulfills a set of requirements that Iyer (1998) proposes for rhythm perception models, namely that

> [...] any model of rhythm perception and cognition must include stages at which incoming rhythms are compared to known rhythms, matched against known meters, and situated among broader expectations about

> musical events. It also must involve some degree of what may be called active perception, by which is meant the assessment of various alternative readings of the musical signal, and the switching among them, all carried out *in time* and continually revised and updated. (p. 55, italics occur in original).

2.5.6 Simulating enculturation with probabilistic generative models

Probabilistic generative models offer a principled way of simulating the effects of prior exposure on perception. To understand this, it is helpful to consider the exhaustive set of possible outcomes of a generative rhythm perception model, namely the entire set of event-timing patterns that the model can generate. For each item in this set, there is an unknown probability of encountering it as a musical rhythm. For some items, this probability is low, because they are unlikely rhythms, for others, it is high, for example because they correspond to stereotypical rhythms. There is, in other words, an unknown probability distribution of musical rhythms. To minimize prediction error, a generative rhythm perception model aims to approximate this distribution as closely as possible.

The approximation is performed by inferring the model's parameters model from a (relatively) small sample drawn from the target distribution. How well the model has approximated the unknown target distribution is usually evaluated by testing the model on another small sample from this distribution. How well the model generalizes based on the small sample depends on the way that the designers of the model have constrained it. Evaluating which constraints improve the model's approximation of the target distribution may provide insights into the probabilistic constraints of rhythms.

However, the target distribution of relevance to *culturally situated* individuals depends on their cultural environment. A generative model aiming to simulate the perception of such individuals should derive its parameters from a sample that represents the music they are likely to encounter. Music corpora, such as the Essen folksong collection (Schaffrath & Huron, 1995), may be used for this purpose. Parameterizations that result from training a generative model on such a sample can be seen as a simulation of an enculturated listener (Van der Weij et al., 2017). The success of the enculturated model in predicting perceptual idiosyncrasies resulting from such biased sampling, may provide evidence as to whether learning mechanisms of listeners resemble those posited by predictive processing. This approach is entirely compatible with the cultural distance hypothesis of Demorest and Morrison (2016), Morrison et al. (2019), according to which the degree of overlap in statistical structure between the music of two cultures predicts the

ability of listeners from those cultures to process music from the other culture.

2.6 Summary

A great variety of rhythm perception models exists in the music cognition literature. Some of these models propose incremental changes to other models, but others propose radically different principles. This chapter reviews a selection of previously proposed rhythm perception models and associates them with three broad perspectives—cognitivism, embodied cognition, and predictive processing—that each entail a different view on the nature of perception and cognition.

The cognitivist perspective describes cognition as information processing involving the rule-based manipulation of symbolic representations. Rule-based models of rhythm perception, such as those proposed by Longuet-Higgins and Steedman (1971), Longuet-Higgins (1976), Longuet-Higgins and Lee (1982), can be associated with this perspective. Embodied cognition instead emphasizes the role of continuous dynamic interaction between brain, body, and environment. Coupled oscillation models (McAuley, 1994; Large & Kolen, 1994; Large & Snyder, 2009), although they do not always emphasize an explicit role of embodiment, are consistent with this view. Finally, predictive processing views perception and perceptual learning as the result of a single underlying mechanism, namely prediction error minimization based on Bayesian inference. Probabilistic generative models, such as those proposed by Temperley (2007, 2009), are consistent with this perspective.

Additionally, this chapter reviewed literature that studies the role of enculturation in shaping rhythm perception. Although the extent to which rhythm perception is constrained by enculturation and by universal principles remains a topic of debate, it seems uncontroversial that experience, training, and practice play a role. Despite this consensus, few models of rhythm perception account for the possible effects of enculturation. Instead, some models aim to represent universal aspects of perception (such as Povel & Essens, 1985; Large, 2010b), while others aim to model the perceptual processes of listeners enculturated in a musical idiom (such as Longuet-Higgins, 1979).

Probabilistic generative models, which are consistent with the principles of predictive processing, can simulate the effect of previous exposure to rhythms by deriving their parameters from empirical samples of music. Neural resonance models have recently been extended to use plastic connections and Hebbian learning enabling them to adapt based on previous exposure to rhythms (Large, 2010a; Large et al., 2015). In general, rhythm perception models have primarily been evaluated on datasets of Western tonal music. It would therefore be a fruitful avenue for future research to evaluate and compare these models on culturally diverse samples of rhythms.

Chapter 3

Definition of dynamic Bayesian networks with deterministic constraints

3.1 Introduction

Probabilistic generative models have been an established modeling tool in the cognitive sciences for some time (Chater, Tenenbaum, & Yuille, 2006). Two significant advantages of these models are their ability to deal with uncertainty and their ability to learn from data. Formalisms like Bayesian networks (see for example Pearl, 2000) have made it easier to define probabilistic generative models involving high-level abstract concepts in terms of which theories in cognitive science are often stated (Chater et al., 2006).

In the cognitive science of music, well-known examples of theories expressed in terms of such abstract concepts are Longuet-Higgins and Lee's (1984) theory of meter perception and syncopation and Lerdahl and Jackendoff's (1983) generative theory of tonal music. Instead of engaging with sound signals, these theories operate on symbolic representations related closely to concepts in Western music notation such as notes, rests, bars, and time signatures. Probabilistic generative models of music perception, such as those described by Pearce (2005), Temperley (2007), and Van der Weij et al. (2017 [Chapter 6]), also make use of such abstract concepts. These models commonly involve a mixture of probabilistic and deterministic constraints.

Consider, for example, Temperley's (2007) probabilistic model of rhythm and meter perception. This model represents meter as abstract hierarchical structure based on cognitive theories of meter (Lerdahl & Jackendoff, 1983). These structures are

generated by a set of probabilistic decisions regarding the hierarchical structure of meter. For example, the number of tactus beats in a bar, how the tactus level is subdivided, and so on. The possible outcomes of some of these decisions are constrained by the outcome of other decisions. For example, which tactus beat aligns with the first note is constrained by the number of tactus beats in a bar.

While the constraints described above are straightforward, other constraints in the model are significantly more subtle and interact in complex ways (see Chapter 4, where we describe these constraints). In publications where cognitive models are proposed, these more detailed constraints are sometimes considered *implementation details*: details that are explicit only in algorithmic implementations of the model. There often is, in other words, a gap between the description of a model in a publication and its algorithmic implementation. This means that researchers wishing to significantly extend or change a model often need to consult these algorithmic implementations in order to fully understand the model. This tends to be impractical and cumbersome because implementations are often complex and tailored specifically to a specific model. Furthermore, understanding these implementations requires familiarity with the programming language in which they are stated.

It is in general useful and desirable to have an algorithmic implementation of a model available, as it enables other researchers to reproduce simulation results, experiment with different parameter settings, and improve or build upon the model. These advantages are tempered, however, by the problems described above. The framework developed in this chapter is intended to improve this situation. It affords formal and concise definitions of discrete probabilistic models that map closely to algorithmic implementations. These definitions are based on *congruency constraints*: functions associated with a random variable that encode deterministic constraints on the values that it may assume with non-zero probability as a function of the values of other random variables.

Music and its perception evolve dynamically over time. The relevance of these across-time aspects of music to its perception has long been noted (Meyer, 1957; Longuet-Higgins & Steedman, 1971) and has been of significant interest to modelers (e.g., Longuet-Higgins & Steedman, 1971; Pearce & Wiggins, 2012). *Dynamic Bayesian networks* (see for example Murphy, 2002) are a broad class of probabilistic models suitable for modeling phenomena that evolve over time. The framework described in this chapter is tailored to describing discrete dynamic Bayesian network with deterministic constraints. It can be used to express a variety of music perception models. These model definitions, in a sense, disentangle the deterministic aspects of a model from its probabilistic aspects.

The central concept in the framework is a *model-definition table*, which enumerates random variables and their congruency constraints. These tables formally define the deterministic constraints at play in a probabilistic model. Combined with a

definition of the conditional probability distributions, they provide a complete specification of a probabilistic model. These specifications can straightforwardly be translated into algorithmic implementations, as illustrated in Appendix A for two models defined in Chapter 5. The framework aims to facilitate the implementation, exploration, and comparison of probabilistic generative models of music perception. It has been implemented as a Common Lisp package, which the author aims to make available as free and open-source software as soon as possible at https://osf.io/z4389/.

We demonstrate the use of the framework in Chapters 4 and 5. Chapter 4 defines an adaptation of Temperley's (2007) rhythm model and Chapter 5 defines adaptations of two rhythm perception models that we evaluate in Chapter 7. These chapters illustrate how distilling the deterministic constraints of probabilistic models clearly reveals conceptual differences and commonalities between different models.

Following a brief discussion of related work below, Section 3.2 enumerates relevant concepts in probability and Bayesian network theory. Section 3.3 introduces congruency constraints and shows how model evidence is calculated and inference is performed in Bayesian networks with congruency constraints. Section 3.4 describes dynamic Bayesian networks and the consequences that congruency constraints have for calculating model evidence and performing inference in these models. Section 3.5 describes a framework for defining the congruency constraints of a dynamic Bayesian network. In this section, model-definition tables are introduced. Finally, Section 3.6 wraps up by summarizing the introduced concepts.

3.1.1 Related work

The mixing of deterministic constraints with probabilistic inference in Bayesian networks has been well studied. Dechter and Larkin (2001) and Mateescu and Dechter (2008) describe a framework for performing inference on *mixed networks*: Bayesian networks in which deterministic constraints may be represented as zero-probability values in conditional probability distributions or as formulas in first-order logic. The congruency constraints described in this chapter are of the former type: they imply that certain values in conditional probability distributions have zero probability. As such, they do not impose a separate set of deterministic constraints onto a Bayesian network but only describe constraints encoded in its probability distributions.

In some ways, the merits of the framework we describe here are similar to those of IDyOM (Pearce, 2005), a modeling framework for sequence prediction models based on the framework of multiple viewpoint systems (Conklin & Witten, 1995). This framework allows a set of representations of sequences to be defined, different subsets of which can be used in parallel for a sequence-prediction task. The

framework described here is similar to multiple viewpoint systems and IDyOM in the sense that it supports the definition of sequence prediction models but different in the sense that it supports the definition of a more general class of dynamic Bayesian network models. In fact, the present framework emerged out of an attempt to generalize multiple viewpoint systems to include latent variables.

With regard to providing a flexible representation language for probabilistic models that facilitates the exploration of different probabilistic models, the merits provided by the framework are similar to those provided by first-order probabilistic languages (e.g., BLOG, Milch, 2006) and probabilistic programming languages (see e.g., Gordon, Henzinger, Nori, & Rajamani, 2014). First-order probabilistic languages are more expressive than Bayesian networks and support types of probabilistic reasoning that Bayesian networks do not (see Russell, 2015, for a recent overview of such approaches).

Congruency constraints are more limited in scope than both mixed networks (Mateescu & Dechter, 2008) and first-order probabilistic languages. They represent a subset of the types of deterministic constraints described by Mateescu and Dechter (2008). For the cognitive models of music perception of relevance to this thesis, however, they are sufficiently expressive and provide a convenient means for expressing their deterministic constraints.

3.2 Preliminaries

Below we provide an overview of those aspects of probability and Bayesian network theory relevant to the development of the current framework. The primary goal of this section is to introduce the notation used in this chapter, and not to rigorously discuss these topics. Elaborate treatments of Bayesian networks and probability theory can be found elsewhere (regarding Bayesian networks, see for example Pearl, 2000, or Bishop, 2006 and regarding probability theory, see for example Hoel, Port, and Stone, 1971).

3.2.1 Probabilities and random variables

We closely follow the notation conventions used by Pearl (2000) for probabilities and random variables. A random variable represents or measures an aspect of a momentary state of affairs. Following Pearl, we call the set of possible values that a random variable may assume its *domain* and allow this domain to be any set of symbols. Additionally, we use the word *state* to refer to a particular value of a random variable. The symbol D_X denotes the domain from which the values of a random variable X are drawn. The values that a random variable may assume are

mutually exclusive and *exhaustive*. By exhaustive, we mean that for any possible state of affairs, there is a corresponding value that a random variable assumes. By mutually exclusive, we mean that each possible state of affairs uniquely determines the value of a random variable. In this chapter, we are concerned only with discrete random variables that can assume a finite number of values.

We consistently use capital letters to refer to random variables and lower case letters to refer to their states. For example, if X, Y, and Z are random variables the symbols x, y, and z are used to refer to their respective values.

The notation $\Pr(X = x)$ denotes the total probability of all states of affairs in which X assumes the value x. Probabilities are always larger than or equal to zero and smaller than or equal to one. $\Pr(X = x) = 1$ indicates that X equals x with absolute certainty, while $Pr(X = x) = 0$ indicates that X does not equal x with absolute certainty. Both of these cases represent *deterministic constraints*. Probabilities in between zero and one describe intermediate degrees of certainty about the value of X. We usually abbreviate $\Pr(X = x)$ to $\Pr(x)$, except in some cases where this could cause confusion.

If X and Y are both random variables, then the values (x, y) that may be assumed by X and Y simultaneously for a given state of affairs are described by a *compound random variable* with the domain $D_X \times D_Y$. We denote a compound random variable as a set of variables $\{X_0, X_1, \cdots\}$. A compound random variable is itself a random variable whose states are denoted by tuples $(x_0, x_1, \cdots)$.

A *probability distribution* is the set of probabilities corresponding to the possible values of a random variable. A probability distribution of a random variable X is denoted as $\Pr(x)$. A *conditional distribution* describes the probabilities that a random variable assumes values in its domain given the values of other random variables. The conditional probability that $X = x$ given that $Y = y$ is denoted by $\Pr(X = x \mid Y = y)$, which we usually abbreviate to $\Pr(x \mid y)$.

3.2.2 Directed acyclic graphs

Bayesian networks are defined as *directed acyclic graphs*. A directed acyclic graph is defined by a pair $G = (V, E)$, where V is a finite set of *vertices* and $E \in V \times V$ is a set of *directed edges*. An edge $e \in E$ is a pair (X, X') that represents a directed edge from X to X'. There is said to be a *path* from $X \in V$ to $X' \in V$ if X' can be reached by starting at X and following directed edges in E to X'. The *parents* of a vertex X are the set of vertices from which edges run to X_i, given by $\{X' \in V \mid (X', X) \in E\}$. A vertex that has no parents is called a *root* of G.

In a directed acyclic graph, paths do not form cycles. That is, it is impossible to follow directed edges from one vertex and arrive at the same one. If a graph is

a directed acyclic graph, then it is possible to sort their vertices in a way that whenever there is an edge from $X \in V$ to $X' \in V$ and $X \neq X'$, X must occur before X' in the ordering. Vertices arranged in this way are said to be *sorted topologically*.

3.2.3 Bayesian networks

We describe a discrete Bayesian network as a set of n random variables $V = \{X_i\}_{i=0}^{n-1}$ that are simultaneously the vertices of a directed acyclic graph $G = (V, E)$. If V is a Bayesian network with respect to G, then the graph encodes for each variable, $X_i \in V$, the minimal set of variables that have to be observed in order to render the probability distribution of X_i independent of all its ancestors in G. Variables in this set are also known as the *Markovian parents* of X_i (Pearl, 2000). PA_i denotes the compound random variable consisting of the Markovian parents of X_i. The symbol pa_i denotes a state of this random variable. The defining property of a Bayesian network is that its joint distribution, the distribution of V, can be written as the following product of conditional distributions of $X_i \in V$ given pa_i:

$$\Pr(v) = \prod_{i=0}^{n} \Pr(x_i \mid pa_i).$$

In this chapter, we commonly refer to the Markovian parents, PA_i, as the *dependencies* of X_i.

Since G is a directed acyclic graph, its vertices can be sorted topologically. In this chapter, whenever we refer to the variables of a Bayesian network, V, we assume that every $X_i \in V$ has been assigned a unique index $i \in \mathbb{N}$ of consecutive natural numbers beginning at 0 such that sorting X_i by their indices in ascending order ensures that the variables are sorted topologically with respect to G.

The structure of the network graph can be exploited to more efficiently compute marginal probabilities, which are required to perform exact inference. The graph can furthermore be used to more efficiently specify the parameters of V, since only parameters of the conditional distributions require specification in order to define the joint distribution.

3.3 Congruency constraints for Bayesian networks

Deterministic constraints in a Bayesian network may be encoded as zero-probability values of its conditional probability distributions. Congruency constraints describe these constraints by defining the set of states of a random variable that are *not* a priori known to have zero probability as a function of its dependencies in the network graph. These states are called *congruent states* and the states which are a priori known to have zero-probability states are called *incongruent states*.

When a Bayesian network has deterministic constraints, we can deterministically infer from observing the state of one variable that some states of the joint distribution that were possible before the observation are no longer possible after the observation. This gives rise to a distinction between *a priori congruent states* and *a posteriori congruent states*: the states congruent respectively before and after an observation. The a posteriori congruent states represent a deterministic inference that can be made from observations. When computing the marginal probability of an observation (its *model evidence*), only a posteriori congruent states need to be considered.

Below, we define congruency constraints and the a priori and a posteriori congruent states of a Bayesian network. We then show how these deterministic constraints can be exploited when calculating model evidence and performing inference. First, however, we will clarify the concept of congruency constraints with a concrete example.

3.3.1 Example

Consider a Bayesian network that describes the probability that a light bulb is on depending on whether that light bulb is broken and whether there is a power outage. These states of affairs are described by three random variables: L, B, and O. Each variable has a Boolean-valued domain: $\{t, f\}$, for true and false. The outcome $L = t$ should be interpreted to mean that the bulb is emitting light, $B = t$ that the bulb is broken, and $O = t$ that there is a power outage. While the probability that the light bulb is on depends both on whether it is broken and whether there is a power outage, the latter two events are independent. These causal dependency relations can be described intuitively in a Bayesian network graph, as shown in Figure 3.1.

The Bayesian network guarantees that joint distribution $\Pr(l, b, o)$ can be written as $\Pr(b) \Pr(o) \Pr(l \mid b, o)$. Six parameters are required to define this joint distribution: the probability that a light is broken, that there is a power outage, and four

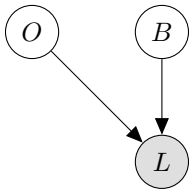


Figure 3.1: A Bayesian network model of a light bulb represented as a Bayesian network.

conditional probabilities that a light is on given whether it is broken and whether there is a power outage. We will assume a priori probabilities that there is an outage and that the light is broken to be respectively 0.001 and 0.05. The probability that the light is on depends on whether there is an outage and whether the light is broken. We assume that the probability that the light is on is 1 if it is not broken and there is no power outage. Otherwise, the probability that the light is on is 0. (For simplicity, we assumed that there is no light switch.)

If the states of both O and B are known (or if one of them is known to be true), there is no uncertainty about L. These deterministic constraints can be described by a congruency constraint. Deterministic constraints commonly represent simplifications of a model that constrain its dimensionality: in the real world, there are many situations not included in our model that can cause a light to be off.

The states of the joint distribution are enumerated fully in Table 3.1. Notice that, due to the deterministic constraint, some states have zero probability. We will call such states *incongruent* and the remaining states *congruent*. Table 3.1 shows that there are four incongruent states, corresponding to situations where the light is on while there is a power outage or while the light is broken. Rows corresponding to congruent states are highlighted in gray.

Let us assume that we can observe whether the light is on, but we cannot observe whether it is broken or whether there is a power outage. That is, L is an *observed variable* (indicated, as is customary, in Figure 3.1 by a node with a shaded background) and B and O are *latent variables*. *Before* observing whether the light is on, there are four congruent states. These are a priori congruent states. *After* observing whether the light is on, the number of states that remain possible reduces to one if the light is on or three if the light is off. These states are a posteriori congruent states. In Table 3.1, a posteriori congruent states corresponding to the observation that the light is on are highlighted in dark gray, while those corresponding to the observation that the light is off are highlighted in a lighter shade of gray.

The model evidence is the marginal probability that a probabilistic model assigns to an observed value of one of its variables. The model evidence of an observed

Table 3.1: The joint probability distribution of three random variables representing whether a light is on, whether the light is broken, and whether there is a power outage. A priori congruent states are highlighted in different shades of gray. The state shaded dark gray is a posteriori congruent given the observation that the light is on, while light gray states are a posteriori congruent given the observation that the light is off.

Variable			Probability
L	B	O	
t	t	t	$0 \times 0.05 \times 0.001 = 0$
t	t	f	$0 \times 0.05 \times 0.999 = 0$
t	f	t	$0 \times 0.95 \times 0.001 = 0$
t	f	f	$1 \times 0.95 \times 0.999 \approx 0.949$
f	t	t	$1 \times 0.05 \times 0.001 \approx 0.000$
f	t	f	$1 \times 0.05 \times 0.999 \approx 0.050$
f	f	t	$1 \times 0.95 \times 0.001 \approx 0.001$
f	f	f	$0 \times 0.95 \times 0.999 = 0$

value of L is the total probability of all states of the joint distribution in which L has this value. When a model has congruency constraints, the model evidence corresponds to the sum of the probabilities of a posteriori congruent states. For example, if we observe the light to be on, the model evidence corresponds to the probability of the only congruent state in which the light is on.

While we cannot observe whether the light is broken, or whether there is a power outage, observing the state of the light tells us something about the state of these variables. The *posterior distribution* of the latent variables is given by:

$$\Pr(b, o \mid l) = \frac{\Pr(l \mid b, o) \Pr(b, o)}{\Pr(l)}$$

Note that the denominator, $\Pr(l)$, is the model evidence (the sum of probabilities of a posteriori congruent states). The numerator, $\Pr(l \mid b, o) \Pr(b, o)$ may also be written as $\Pr(l \mid b, o) \Pr(b) \Pr(o)$ since the Bayesian network defines L and O to be independent. The numerator can now be recognized as the joint distribution. Therefore the posterior probability of each state (b, o), given an observed value l, is given by the joint probability, $\Pr(l, b, o)$ divided by the model evidence of the observed state of l. Obtaining the posterior distribution by conditioning the model on observation is known as *probabilistic inference*.

3.3.2 Congruency constraints

We will now formally define congruency constraints. Let X_i be a random variable in a Bayesian network and let PA_i be the set of Markovian parents of X_i. A congruency constraint of X_i is a function $\kappa_{X_i}\colon D_{PA_i} \to \mathcal{P}(D_{X_i})$, where $\mathcal{P}(D_{X_i})$ denotes *powerset* of D_{X_i} (the set of all possible subsets of D_{X_i}). The function κ_{X_i} generates a set of states that are congruent given a state, pa_i, of another (possibly compound) random variable. The congruent states of a variable X_i are in general a subset of its domain, D_{X_i}. A probability distribution of X_i, $\Pr(x \mid pa_i)$, is said to *conform to* congruency constraint κ_{X_i} if it is zero for all incongruent states of X_i. That is, if $x \notin \kappa_{X_i}(pa_i)$ then $\Pr(x \mid pa_i) = 0$ for any $x \in D_{X_i}$ and $pa_i \in D_{PA_i}$.

Congruency constraints can restrict uncertainty in a probabilistic model to a variable degree: At one extreme, the congruent states of X_i may be a singleton given pa_i, that is, $\kappa_{X_i}(pa_i) = \{x\}$. In that case, X_i is deterministically known given pa_i. At the other extreme, the congruent states of X_i given pa_i may be equal to its domain, that is, $\kappa_{X_i}(pa_i) = D_{X_i}$, which is identical to the situation without congruency constraints.

Returning to the light bulb example, we may specify a single congruency constraint for the conditional distribution of L as a function of the states of O and B:

$$\kappa_L((o,b)) = \begin{cases} \{f\} & \text{if } t \in \{o,b\}, \\ \{t\} & \text{otherwise.} \end{cases}$$

3.3.3 A priori congruent states

The states of a variable that are congruent prior to an observation are called a priori congruent states. Below, we show how a priori congruent states of a Bayesian network are generated using the congruency constraints of its constituent variables. This process resembles ancestral sampling.

We again use V to denote a set of n random variables, $\{X_i\}_{i=0}^{n-1}$, that constitute a Bayesian network. That is, each variable, X_i, is also a vertex in a directed acyclic graph $G = (V, E)$ that encodes the conditional independence relations among variables in V. Furthermore, there is a congruency constraint, κ_{X_i}, for each variable $X_i \in V$ except if X_i is a root of G, since the congruent states of variables corresponding to root vertices are identical to their domains.

The symbol V^i denotes the compound variable constituted by $\{X_j\}_{j=0}^{i-1} \subseteq V$. That is, $V^n = V$ and V^i, for $0 \leq i < n$ consists of all variables X_j in V for which $0 \leq j < i$. The a priori congruent states of V^i, for $1 \leq i \leq n$, are a set that is defined recursively using the given congruency constraints as follows:

$$K_{V^i} = \begin{cases} D_{V^1} & \text{if } i = 1, \\ \{(x_0, \cdots, x_{i-1}, x_i) \mid (x_0, \cdots, x_{i-1}) \in K_{V^{i-1}},\ x_i \in D_{X_i}\} & \text{if } X_i \text{ is a root of } G, \\ \{(x_0, \cdots, x_{i-1}, x_i) \mid (x_0, \cdots, x_{i-1}) \in K_{V^{i-1}},\ x_i \in \kappa_{X_i}(pa_i)\} & \text{otherwise.} \end{cases}$$

The congruent states of a Bayesian network with n variables $V = \{X_i\}_{i=0}^{n-1}$, are generated by evaluating K_{V^n} and expanding the recursion to lower i until $i = 1$. Recall that indices i have been assigned such that they define a topological ordering of the variables X_i. Therefore, X_0 must be a root of G and the congruent states of $V_1 = \{X_0\}$ correspond to the domain of $\{X_0\}$. The congruent states of V^i for $1 < i < n$ are constructed by using each congruent state of V^{i-1} to generate the congruent states of X_i using $K_{X_i}(pa_i)$ and combining each of these with the congruent state of V^{i-1} from which they were generated. [1] Note that the state pa_i to which the congruency constraint of X_i is applied is always contained in a congruent state of V^{i-1}, since X_i are sorted topologically.

For any state $v \in K_V$, $(x_0, \cdots, x_i, \cdots, x_{n-1})$, and any given value x_i' of a variable $X_i \in V$, it is the case that if $x_i' \notin \kappa_{X_i}(pa_i)$, then $\Pr((x_0, \cdots, x_i, \cdots, x_{n-1})) = 0$. That is, if the value of X_i, x_i', in a state $v \in K_V$ does not occur in the set of states that are congruent given the values of its Markovian parents, pa_i, as encoded in a given state v, then the joint probability of that state must be zero if the probability distribution of V conforms to the congruency constraints. We will say that a state, x_i' of a variable $X_i \in V$ is *congruent with* a state, s, of $V \setminus X_i$ if s and x_i' occur together in an a priori congruent state of V.

3.3.4 A posteriori congruent states

When the value of one of the variables in V is observed, a subset of a priori congruent states remains possible. That is, if we observe that the value of a variable $X_i \in V$ is x_i', we can deterministically infer that only the states of other variables $X_j \in V$ that are congruent with x_i' remain possible. We call these states the a posteriori congruent states given that $X_i = x_i'$. This set is given by a subset of the congruent states of V, K_V, in which the value of X_i is equal to x_i'.

The function $\hat{\kappa}_V^{X_i} \colon D_{X_i} \to \mathcal{P}(D_V)$ generates a posteriori congruent states of a Bayesian network, V, given an observation x_i' of $X_i \in V$. This function is defined as

$$\hat{\kappa}_V^{X_i}(x_i') = \{(x_0, \cdots, x_i, \cdots, x_{n-1}) \in K_V \mid x_i' = x_i\}.$$

[1] This procedure can be made more efficient by evaluating the congruent states of X_i only for the unique states of pa_i contained in the congruent states of V^{i-1}.

Note that if $x'_i \notin K_V$, then $\hat{\kappa}_V^{X_i}(x'_i) = \emptyset$.

3.3.5 Model evidence

The model evidence, or marginal probability, of a state x'_i of a variable $X_i \in V$ is the marginal probability that $\Pr(X_i = x'_i)$. By the product rule of probability, $\Pr(v)$ may be written as $\Pr(x_i \mid V \setminus X_i = s) \Pr(V \setminus X_i = s)$ where s is a state of of $V \setminus X_i$ (the compound variable consisting of all variables in V except X_i). The marginal probability $\Pr(X_i = x'_i)$ is given by

$$\Pr(X_i = x'_i) = \sum_{s \in D_{V \setminus X_i}} \Pr(x'_i \mid V \setminus X_i = s) \Pr(V \setminus X_i = s).$$

That is, the model evidence is the sum of probabilities of all states of the joint distribution in which the value of X_i corresponds to its observed value x'_i. This involves a summation over the states of $V \setminus X_i$, which correspond to the Cartesian product $\prod_{X_j \in V \setminus X_i} D_{X_j} = D_{V \setminus X_i}$.

Infamously, this summation makes model evidence computationally expensive to compute, since the number of such states is exponential in the number of variables in V. Since V is a Bayesian network, there exist exact inference algorithms that exploit the network graph to optimize this calculation (see e.g., Bishop, 2006). However, if the conditional probability distributions of the Bayesian network also conform to a set of congruency constraints $\{K_{X_i}\}$, where X_i are non-root vertices of G, the incongruent states of V are guaranteed to have zero probability and do not contribute to model evidence. Therefore, when congruency constraints are given, model evidence is given by the total probability of a posteriori congruent states:

$$\Pr(X_i = x'_i) = \sum_{v \in \hat{\kappa}_V^{X_i}(x'_i)} \Pr(V = v)$$

3.3.6 Inference

Given an observation of the state x'_i of a variable $X_i \in V$, the posterior distribution over the states of the remaining variables in V is given by

$$\Pr(V \setminus X_i = s \mid X_i = x'_i) = \frac{\Pr(V \setminus X_i = s, X_i = x'_i)}{\Pr(X_i = x'_i)}.$$

While the above equation uses the explicit notation $\Pr(V \setminus X_i = s \mid X_i = x_i')$ to indicate the posterior distribution, in the remainder of the chapter, we simply refer to the posterior distribution of V conditioned on x_i': $\Pr(v \mid x_i')$.

The denominator in the equation above corresponds to the model evidence, defined previously as the sum of the probabilities of a posteriori congruent states. As such the posterior distribution may be written as

$$\Pr(v \mid x_i') = \frac{\Pr(v)}{\sum_{v \in \hat{\kappa}_V^{X_i}(x_i')} \Pr(v)}.$$

3.4 Congruency constraints for dynamic Bayesian networks

In this section, we extend the definition of the congruent states of Bayesian networks developed above to dynamic Bayesian networks. Dynamic Bayesian networks encompass a broad class of probabilistic models that include, for example, hidden Markov models (see e.g. Russel & Norvig, 2003). A dynamic Bayesian network evolves over a series of time steps. We refer to these as *moments*. Each moment corresponds to an observation of the value of the observed variables of the dynamic Bayesian network. Given a sequence of observations, we show how the model evidence of each observation and the posterior distribution given each observation is obtained by considering only moments relative to the present, namely the *current moment*, the *previous moment*, and the *first moment*.

Congruency constraints reflect facts that a model assumes to deterministically hold about possible states of affairs. Dynamic Bayesian networks model a state of affairs that undergoes change from moment to moment. Therefore, congruency constraints of a Bayesian network can reflect deterministic constraints on the ways in which states of affairs evolve over time.

Since observations constrain the congruent states of a Bayesian network that has deterministic constraints, the congruent states of a dynamic Bayesian network evolve over moments (time steps) as a function of the observations made in each moment. This evolution can be described by a finite-state automaton that processes observations of variables of the Bayesian network as input. We call the set of possible sequences of observations accepted by this finite-state automaton the *congruent input sequences* of a model.

Below, we first define dynamic Bayesian networks. We then show how the evolution of a dynamic Bayesian network from moment to moment can be described as a cyclical process. First, we sketch an *absolute-time* version of this process in which

we use absolute indices, t, to refer to different moments. Next, we show how the explanation can be expressed in terms of moments relative to the present: t and $t-1$ and introduce notation to refer to these moments. This *present-relative* formulation allows all absolute indices, except for zero, to be dropped from the notation. Finally, after describing the present-relative formulation of the cyclical process, we show how congruent input sequences are described by a finite-state automaton.

3.4.1 Dynamic Bayesian networks

A dynamic Bayesian network is defined by two Bayesian networks: an *initialization model* and a *transition model*. The initialization model describes the distribution of a set of random variables V in the first moment. The transition model describes the probability distribution of V conditioned on the previous moment (which may be the first or any subsequent moment). In order to distinguish between V in different moments, we temporarily give V an absolute-time index, $t \in \mathbb{N}$, indicating the number moments elapsed since the first moment, such that V_0 represents the first moment. The transition model is a conditional distribution $\Pr(V_t \mid V_{t-1})$ for $t > 0$. Note that while there is an arbitrary number of moments, there is only one transition model.

The initialization model is a regular Bayesian network and the dependency relations between its variables are described by an acyclic directed graph $G_0 = (V_0, E_0)$. The transition model is also a Bayesian network, involving the variables $V_{t-1} \cup V_t$. Its dependency relations are described by $G_t = (V_{t-1} \cup V_t, E_t)$. The variables V_t represent the *current moment*, while the variables in V_{t-1} represent the same variables in the *previous moment*. The edges of G_t, E_t can be partitioned into a set E^H and a set E^V, indicating respectively the *horizontal dependencies* and the *vertical dependencies* of the transition model, such that $G_t = (V_t \cup V_{t-1}, E^H \cup E^V)$. Horizontal dependencies represent temporal (across-moment) dependencies, which may run only from V_{t-1} to V_t. Vertical dependencies represent instantaneous (within-moment) dependencies, which may run only between variables in V_t.

The following example illustrates horizontal and vertical dependencies: Imagine that we wish to infer the location of a moving vessel from a sequence of imprecise GPS measurements. We can consider each measurement to be an observed variable in a generative transition model in which the vessel's actual location is a hidden (latent) variable. This latent variable has a horizontal dependency on the variable representing vessel's actual location in the previous moment. The observed variable, representing a measurement, has a vertical dependency on the latent variable representing the actual location of the vessel in the current moment (this model is an example of what is known as a hidden Markov model). The probability distribution over the vessel's current location is constrained, via a

horizontal dependency, by its previous location, and the measurements in the current moment are constrained via a vertical dependency by the vessel's current location.

3.4.2 Absolute-time moment transitions

From here onward, we denote the state of a compound variable, $X \cup Y$, consisting of two disjoint sets of variables X and Y as xy. This represents the concatenation of the two tuples: if $x = (x_0, x_1, \cdots)$ and $y = (y_0, y_1, \cdots)$, then $xy = (x_0, x_1, \cdots, y_0, y_1, \cdots)$.

A moment represents an atomic (indivisible) unit of time corresponding to the observation of a state $x'_{i,t}$ of one of the variables in $X_{i,t} \in V_t$. Evaluating a model on a sequence of observations—that is, obtaining model evidence and posterior distributions—on a sequence of observations, $(x'_{i,0}, x'_{i,1}, \cdots)$, of arbitrary length can be described as a cyclical process in which for each observation, congruent states are generated, model evidence is calculated and a posterior distribution is derived. Finally, the posterior distribution is marginalized to serve as the prior distribution in the next moment.

The cyclical process is initialized in the first moment by observing the state $x'_{i,0}$ of a variable X_i in the initialization model V_0. The posterior distribution, $\Pr(v_0 \mid X_{i,0} = x'_{i,0})$, is a probability distribution over a posteriori congruent states of V_0, which are of the form $\{v_0, v'_0, v''_0, \cdots\}$. In the second moment, the transition model defines the a priori congruent states of $\Pr(v_1 \mid v_0, x'_{i,0})$. Observing $x'_{i,1}$ yields the posterior distribution $\Pr(v_0 \mid X_{i,0} = x'_{i,0})$, defined for a posteriori congruent states of $\Pr(v_0 \mid x'_{i,0}) \Pr(v_1, \mid v_0, x'_{i,0}, x'_{i,1})$. These are of the form $\{v_0 v_1, v_0 v'_1, v_0 v''_1, \cdots\}$. Before transitioning to the third moment, we can calculate the marginal posterior distribution over states in V_1 as follows:

$$\Pr(V_1 \mid x_{i,0}, x_{i,1}) = \sum_{v_0 \in \hat{\kappa}_{V_0}^{X_{i,0}}} (x'_{i,0}) \Pr(v_0 \mid x'_{i,0}) \Pr(v_1, \mid v_0, x'_{i,0}, x'_{i,1}).$$

The congruent states of the marginal distribution are of the form $\{v_1, v'_1, v''_1, \cdots\}$. This marginal posterior distribution serves as the prior distribution in the next step.

To summarize: in the second moment, at which $t = 1$, we used the posterior distribution of $V_{t-1} = V_0$ as a prior distribution for the transition model. Posterior to the second observation, we obtain the posterior distribution over the states of $V_{t-1} \cup V_t$ given the first and second observations. This posterior distribution is then marginalized to a marginal posterior distribution of V_t by summing over a posteriori (with respect to the first observation) congruent states of V_{t-1}. This

last step completes the cycle, because in the third moment, at which $t = 3$, we can once again use the marginal posterior distribution of V_1 as a prior distribution for the transition model and repeat the process of observation and marginalization.

3.4.3 Present-relative formulation

The process described above can be described by referring only to the first ($t = 0$), previous ($t-1$) and current moment (t). If we define the moment indicated by t as *the present*, we can describe a *present-relative* formulation of the cycle described above. In order to refer to a variable in the previous moment, we decorate it with a hat, such that $\hat{x}_i$ and $\hat{X}_i$ refer respectively to $x_{i,t-1}$ and $X_{i,t-1}$, while plain variable symbols x and X refer to x_t and X_t. The only other temporal distinction that is relevant is whether the present is the first moment since the initialization model applies at that moment. Therefore, we use X_0 and x_0 to refer to variables and states of variables in the first moment.

We assume that whenever we reference V, we do so from the perspective of a specific moment and that $\hat{V}$ refers to the marginal posterior distribution of V in the previous moment given all observations preceding the current moment. To minimize clutter, we omit the dependency of the posterior distribution on all these preceding observations. In this way, in each moment, we perform inference on a generative model $\Pr(\hat{v}) \Pr(v \mid \hat{v})$ by observing the state x'_i of a variable X_i in V.

The emphasis on describing the model from the perspective of the current moment has two motivations: First, we stress the continuity of V, which describes a system (e.g., a model of a perceiver) the state of which is updated by observations that follow each other in temporal order. Second, in the cognitive models that we will define later, we are primarily interested in the posterior distribution of latent variables in a given moment and the model evidence as a measure of "expectedness" or surprisal of observations.

3.4.4 Present-relative moment transitions

A present-relative formulation of the cyclical process described in Section 3.4.2 is summarized in Table 3.2 as a four-step process, divided into two a priori (before observation) and two a posteriori (after observation) steps. The probability distributions that describe the state of the system are shown in each step. The congruent states of these distributions (the states that are not a priori known to have zero probability) are shown next to each distribution. Below, we describe each step in detail.

Step one is not so much a step as it is a state of affairs: we have a prior

Table 3.2: The cycle in which the congruent states of a dynamic Bayesian network are updated by an observation and marginalized in order to transition to the next moment. The dashed line indicates the moment at which an observation occurs.

Step	Distribution	Congruent states	Orientation
1 (prior)	$\Pr(\hat{v})$	$K_{\hat{V}} \subseteq D_V$	A priori
2 (prediction)	$\Pr(v \mid \hat{v}) \Pr(\hat{v})$	$K_{\hat{V} \cup V} \subseteq \{\hat{v}v \mid \hat{v} \in K_{\hat{V}}, v \in D_V\}$	
3 (inference)	$\Pr(v \mid \hat{v}, x'_i) \Pr(\hat{v})$	$\hat{\kappa}^{X_i}_{\hat{V} \cup V}(x'_i) \subseteq K_{\hat{V} \cup V}$	A posteriori
4 (marginalization)	$\Pr(v \mid x'_i) \rightarrow \Pr(\hat{v})$	$\hat{\kappa}^{X_i}_{V}(x'_i) \subseteq D_V$	

distribution, indicated by $\Pr(\hat{v})$, which is a probability distribution over a set of a priori congruent states $K_{\hat{V}}$. The state of affairs in step one can be seen as either the result of the fourth step of the four-step process in the previous time step or as the result of observing one of the variables of the initialization model. In the latter case, $\Pr(\hat{v})$ is equated to the posterior distribution of the initialization model: a probability distribution over a posteriori congruent states of V_0 given an observation x'_i, given by

$$\Pr(V_0 \mid x'_i) = \frac{\Pr(V_0)}{\sum_{v_0 \in K_{V_0}} \Pr(V_0)}.$$

Since V_0 is a normal Bayesian network, the procedure described in Section 3.3.6 can be used to obtain the above posterior distribution.

In **step two**, a priori congruent states of the current moment are generated using the transition model. The result is a probability distribution over a priori congruent states of the transition model, $K_{\hat{V} \cup V}$. In order to define this set, we introduce some additional notation to indicate horizontal and vertical dependencies of variables in the transition model.

Horizontal dependencies are denoted by $PA^H_i \in \hat{V}$ and defined as $PA^H_i = \{X \in \hat{V} \mid (X, X_i) \in E^H\}$. Vertical dependencies are denoted by $PA^V_i \in V$ and defined as $PA^V_i = \{X \in V \mid (X, X_i) \in E^V\}$. Similarly, pa^H_i and pa^V_i denote states of PA^H_i and PA^V_i respectively. Again, $V^i \subseteq V$ is a compound variable constituted by $\{X_i\}^{i-1}_{i=0} \subseteq V$.

The a priori congruent states of the transition model depend on a set of congruent states, $K_{\hat{V}}$. These can be either the a posteriori congruent states of the initialization model, or the congruent states of a marginal posterior distribution of the previous

time step, as described in step four. Given this set, a priori congruent states of the transition model are given by

$$K_{\hat{V}\cup V^i} = \begin{cases} K_{\hat{V}} & \text{if } i = 0, \\ \{\hat{v}(x_0, \cdots, x_{i-1}, x_i) \mid \hat{v}(x_0, \cdots, x_{i-1}) \in K_{\hat{V}\cup V^{i-1}}, x_i \in D_{X_i}\} & \text{if } X_i \text{ is a root of } G, \\ \{\hat{v}(x_0, \cdots, x_{i-1}, x_i) \mid \hat{v}(x_0, \cdots, x_{i-1}) \in K_{\hat{V}\cup V^{i-1}}, x_i \in \kappa_{X_i}(pa_i^H pa_i^V)\} & \text{otherwise.} \end{cases}$$

This definition is similar to that of a priori congruent states of a Bayesian network, except that there now is a set of congruent states from which the congruent states of the transition model branch. The most important point here is that the congruency constraint of each variable in the transition model has both horizontal and vertical dependencies (pa_i^H and pa_i^V), and may thus restrict the possible states of a variable based on states of variables in the previous moment. The horizontal dependencies are contained in the states $\hat{v} \in K_{\hat{V}}$.

Step three is the first step after observing x'_i. Here, inference is performed to obtain the posterior distribution of V given x'_i and $\hat{v}$. This distribution is given by

$$\Pr(v \mid \hat{v}, x'_i) \Pr(\hat{v}) = \frac{\Pr(v \mid \hat{v}) \Pr(\hat{v})}{\sum_{\hat{v} \in \hat{\kappa}^{X_i}_{\hat{V}\cup V}(x'_i)} \Pr(v \mid \hat{v}) \Pr(\hat{v})}$$

Note that the model evidence (in the denominator) corresponds to the sum of the probabilities of a posteriori congruent states of $\Pr(v \mid \hat{v}) \Pr(\hat{v})$ given x'_i. These are given by the states in $K_{\hat{V}\cup V}$ that are congruent with the observed value x_i of $X_i \in V$. The definition of a posteriori congruent states is given in Section 3.3.4.

In **step four**, we obtain the marginal posterior distribution. This distribution is obtained by summing across a priori congruent states of the prior distribution, $K_{\hat{V}}$ (see step one):

$$\Pr(v \mid x'_i) = \sum_{\hat{v} \in K_{\hat{V}}} \Pr(v \mid x_i, \hat{v}) \Pr(\hat{v}).$$

The cycle is completed by equating the resulting marginal posterior distribution to $\Pr(\hat{v})$ *in the next moment* and returning to step one.

Note that an upper-bound on the computational cost of inference in step three is linearly proportional to the maximum number of congruent states of $V \cup \hat{V}$.

3.4.5 Congruent input sequences

The congruent states of the model in each moment constrain the possible values that can be observed. How the set of congruent states changes as a function of observations may be described by a transition function, $\delta\colon \mathcal{P}(D_V) \times D_{X_i} \to \mathcal{P}(D_V)$, which takes as its input the set of congruent states of the prior, $\hat{V}$ (the congruent states in step one in Table 3.2), and an observation x_i' of one of the variables in V. As its output, the transition function produces the set of a posteriori congruent states of the marginal posterior distribution (the congruent states in step four of Table 3.2).

Given this function, we can define a *finite-state automaton* (FSA) (see e.g. Carroll & Long, 1989) that describes how the congruent states of a dynamic Bayesian network evolve by "processing" observations x_i' of a variable $X_i \in V$ as "input". A FSA is defined by a 5-tuple $(\Sigma, s_0, S, \delta, F)$, where Σ is the *input alphabet*, a finite set of input symbols, S is a finite and non-empty set of *states*, $s_0 \in S$ is a *start state*, $\delta\colon S \times \Sigma \to S$ is a *transition function*, which produces a new state given a state and an input, and $F \subseteq S$ is a possibly empty set of *final states*. A FSA that describes how the congruent states of a dynamic Bayesian network with congruency constraints evolve across moments has the transition function δ described above, its states correspond to $S = \mathcal{P}(D_V)$, its input alphabet is the domain of an observed variable $\Sigma = D_{X_i}$, its initial state, s_0, is the set of a posteriori congruent states of V_0 given an initial observation x_i^0: $s_0 = \hat{\kappa}_{V_0}(x_i^0)$, and its set of final states is empty.

Note that if, at any state s, an input x_i' that is not a priori congruent is provided to the transition function, then, by the definition of a posteriori congruent states, $\delta(s, x_i') = \emptyset$. Once this state is reached, the FSA cannot escape it. We call the set of sequences of observations of length n that the FSA can process without reaching this state, the *congruent input sequences*.

The set of congruent input sequences of a dynamic Bayesian network model of a given length n correspond to the set of possible sequences of observations over which the model defines a complete probability distribution.

3.5 Model-definition framework

In this section, we describe a framework that is used in Chapters 4 and 5 to define a variety of rhythm perception models. Section 3.4 showed that a dynamic Bayesian network is described by an initialization model, a transition model, and a set of congruency constraints. The framework described below provides a format in which the congruency constraints of the transition model can be

defined. The need to define an initialization model is eliminated by assuming a fixed initialization model in which each variable has one congruent state, namely $*$. That is, $\Pr(X_{i,0} = *) = 1$ for any $X_{i,0} \in V_0$. This allows us to focus on the transition model and avoid the need to specify custom initialization models for each model.

Strictly speaking, the fixed initialization model causes $*$ to occur in the domain of every variable in models defined this way. However, we consider this information redundant and will not explicitly include the value $*$ when defining variable domains. That is, when we define the domain of a variable as $D_X = \{a, b, c\}$, it should be understood that the actual domain of X is $\{a, b, c, *\}$.

A *model-definition table* defines the variables of a dynamic Bayesian network, their domains, their horizontal (across-moment) and vertical (instantaneous within-moment) dependencies, and their congruency constraints. This table fully defines the deterministic and representational aspects of a model and completing the specification of a model requires only the additional definition of the conditional probability distributions of each variable. In model-definition tables, we follow the notational convention used throughout this chapter that capital letters denote random variables and non-capitalized letters denote states of random variables indicated by their capitalized counterpart. Each row in the table defines one variable, and different columns in each row define the horizontal and vertical dependencies, the domain, and the congruency constraints of that variable. Below, we introduce model-definition tables by means of an example.

3.5.1 Example

Consider a hypothetical keyboard instrument that provides rather limited musical possibilities: it has only three keys and allows only one them to be played at the same time. Melodies that can be played on this keyboard are monophonic and consist of different sequences of the three notes. We will describe a simple probabilistic model of such melodies. We represent the notes of the three keys by the natural numbers 0, 1, and 2. In this melody model, it does not matter which notes are encoded by these numbers. Figure 3.2 shows the dependency graph of this model. We will explain it bit by bit in the following paragraphs.

In the melody model, a moment corresponds to a note being played. This note is described by a random variable N that can assume the values $\{0, 1, 2\}$. Any pair of notes played subsequently creates a pitch interval, which is described by a variable I. We assume that the probability of a melody depends on the probability of the pitch intervals that it contains. We furthermore assume that we can observe the value of N directly.

The domain of I—the set of possible pitch intervals—is $\{-2, -1, 0, 1, 2\}$. However,

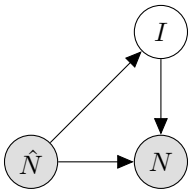


Figure 3.2: A network graph representing the transition model of a dynamic Bayesian network representing a model of notes played on a keyboard conditioned on the interval between notes. The variable $\hat{N}$ represents a note played in the previous moment, the variable I represents a pitch interval, and the variable N represents a note played in the current moment.

the set of pitch intervals that can occur in a given moment depends on which note was played previously. The conditional distribution of I therefore has a horizontal dependency: $\Pr(i \mid \hat{n})$ and a deterministic constraint stating that pitch intervals that cannot be created in a moment given the previous note must have a probability of zero. For example, if the previous note was 1, the set of pitch intervals that can be created in the current moment is $\{-1, 0, 1\}$.

We assume that given a pitch interval is deterministically known given two consecutive notes. However, we encode this constraint generatively: the next note is known deterministically given the previous note and the pitch interval. Therefore, the conditional distribution of N depends horizontally on the previous note and vertically on the current pitch interval: $\Pr(n \mid \hat{n}, i)$. The joint distribution of the transition model is given by $\Pr(n, i) = \Pr(n \mid i, \hat{n}) \Pr(i \mid \hat{n})$, in agreement with the graph in Figure 3.2.

One more issue requires addressing: in the first moment, there is no previous note, so the pitch interval is undefined. We will assume that, in the first moment, N has a uniform distribution and that I deterministically generates the symbol $*$. These constraints can be defined fully in the transition model. In the first moment, $\hat{n}$ and $\hat{i}$ refer to the values of N and I in the initialization model. Since we have assumed that, in the initialization model, these values deterministically are $*$, the congruency constraints can check whether the current moment is the first moment by checking whether $\hat{n} = *$ (I also generates $*$ in the first moment, so checking that $\hat{i} = *$ would cause I to forever generate $*$).

The deterministic constraints on the conditional distributions of I and N described above are implemented by congruency constraints. The congruency constraint of I is:

$$\kappa_I((\hat{n})) = \begin{cases} \{*\} & \text{if } \hat{n} = *, \\ \{i \in D_I \mid \hat{n} + i \in D_N\} & \text{otherwise.} \end{cases}$$

Table 3.3: The congruency constraints of a dynamic Bayesian network describing melodies that can be played on a small keyboard allowing for uncertainty about the pitch intervals between subsequent notes.

X_i	PA_i	D_{X_i}	$\kappa_{X_i}(pa_{X_i})$
I	$\{\hat{N}\}$	$\{-2,-1,0,1,2\}$	$\begin{cases}\{*\} & \text{if } \hat{n}=*,\\ \{i \in D_I \mid \hat{n}+i \in D_N\} & \text{otherwise.}\end{cases}$
N	$\{\hat{N}, I\}$	$\{0,1,2\}$	$\begin{cases}D_N & \text{if } \hat{n}=*,\\ \{\hat{n}+i\} & \text{otherwise.}\end{cases}$

The congruency constraint of N is:

$$\kappa_N((\hat{n}, i)) = \begin{cases} D_N & \text{if } \hat{n} = *, \\ \{\hat{n} + i\} & \text{otherwise.} \end{cases}$$

The constraints indeed ensure that I is deterministic in the first moment by generating the singleton $\{*\}$ and that N has multiple a priori congruent states, namely $\{0, 1, 2\}$. However, after observing, N, the state of the model is fully known. After observing that N is, for example, 2, the joint distribution of $\{I, N\}$ has only one a posteriori congruent state, namely $(*, 2)$. In the next moment, I is no longer deterministic and generates the following a priori congruent states: $\{1, 0, -1\}$. However, since N is deterministic given I, the value of both variables are deterministically known a posteriori. For example, if we would observe that N is 0, the only a posteriori congruent state of $\{I, N\}$ would be $(-1, 0)$.

A model-definition table concisely enumerates these constraints and simultaneously defines the dependency graph of a model. Table 3.3 shows the model-definition table of the example model described above. Horizontal and vertical dependencies are given in the column PA_i. They are specified as a single set in which horizontal and vertical dependencies can be distinguished by the present-relative notation conventions. The congruency constraints are listed as function definitions in the column $\kappa_{X_i}(pa_i)$. These definitions make use of the states of variables in pa_i, which are the arguments to the congruency constraint. The dependency graph of the model $G = (V, E)$, where $V = \{N, I, \hat{N}, \hat{I}\}$ and $E = \{(\hat{N}, I), (\hat{N}, N), (N, I)\}$ (shown in Figure 3.2), is specified completely by the model-definition table. Combined with a definition of conditional probability distributions (not given for this example) of each variable, the model-definition table completely specifies a dynamic Bayesian network. Note finally that we have highlighted the variable that can be observed, N, by giving its row a shaded background.

The behavior of the model described in this example is similar to the behavior

of multiple viewpoint systems (Conklin, 1990; Conklin & Witten, 1995). Such models predict sequences of symbols by constructing *derived representations* that are constrained by the symbol sequence from which they derive. The mechanism that constrains the possible continuations of a derived representation in a multiple viewpoint systems is similar to the mechanism described in this example.

In fact, the framework described here could be used to define a dynamic-Bayesian-network formulation of multiple viewpoint systems. While multiple viewpoint systems are fully observed models—that is, after an observation, there is no uncertainty about its state—a dynamic Bayesian network represents derived representations as latent variables in the manner illustrated in the above example (which is also fully observed). This would enable treating multiple viewpoint systems as a general-purpose Bayesian network, rather than a unique representational formalism. We leave the precise definition of multiple viewpoint systems as dynamic Bayesian networks with deterministic constraints as an opportunity for future work to address.

3.5.2 Terminology

Some general patterns of behavior arise in the definition of congruency constraints for music-cognition models in the next chapters. Below, we enumerate some of them and introduce a vocabulary for referring to them.

In the example described in the previous section, the variable N had a horizontal dependency on itself in the previous moment ($\hat{n}$). We will call variables with such dependencies *recursive variables*.

Furthermore, I deterministically generates the symbol $*$ in the first moment. We will say that variables whose only a priori congruent state in a given moment is $*$ are *inactive* at that moment. All variables are inactive in the initialization model. We will furthermore say that a variable *activates* the first time that it generates symbols other than $*$.

The fixed initialization model in a sense defers the definition of the initialization model to the congruency constraints of the transition model. This theoretically enables the initialization model to be spread out across multiple moments: Notice that I deterministically generates $*$ in the first moment and activates in the second moment. The variable encodes the first-order pitch derivative, which is defined only from the second moment onward. Another variable could be introduced that checks whether $\hat{I} = *$ and deterministically generates $*$ otherwise, resulting in a variable that activates in the third moment. This mechanism can be used to define variables that encode higher-order representations (such as a second-, or higher-order pitch derivatives) of sequences that activate once enough observations have occurred for the higher-order information to become available.

Another behavior that commonly occurs is when a latent variable generates a set of states once and retains its state in all subsequent moments. We call such variables *persistent variables*. An example of a congruency constraint of such a variable is

$$\kappa_X((\hat{x})) = \begin{cases} D_X & \text{if } \hat{x} = *, \\ \{\hat{x}\} & \text{otherwise.} \end{cases}$$

The condition $\hat{x} = *$ is deterministic, since it only holds in the first moment. A persistent variable generates multiple a priori congruent states in when it activates and is deterministic given its previous state in subsequent moments. Persistent variables need not be recursive: they may also rely on other deterministic conditions, such as whether the state of another variable equals $*$ to activate.

Persistent variables are useful for modeling latent variables that are assumed to apply to an entire sequence. The models presented in Chapters 4 and 5 use persistent variables to describe the meter of a rhythm.

3.6 Summary

We have shown how congruency constraints can describe deterministic constraints. When applied to conditional probability distributions of Bayesian networks, these constraints can be exploited to calculate marginals, which are required for exact inference, more efficiently. The deterministic constraints reduce uncertainty in probabilistic models and replace it by deterministic restrictions. These restrictions may be motivated, for instance, by a modeler's expert knowledge of a certain domain. Like Bayesian networks, congruency constraints provide a means of controlling the dimensionality of a probabilistic model. Congruency constraints are an example of how deterministic constraints can be mixed into Bayesian networks (Mateescu & Dechter, 2008).

We described the consequences that congruency constraints have for a class of models known as dynamic Bayesian networks. Here, we emphasized how such models evolve from moment to moment by updating their congruent states, performing inference, and marginalizing the congruent states. We showed that deterministic constraints cause the set of congruent states of a dynamic Bayesian network to change as a function of the sequence of observations. This process can be described by a finite-state automaton whose state corresponds to possible sets of congruent states.

Finally, we introduced a framework in which a dynamic Bayesian network and its deterministic constraints can be fully specified in a model-definition table.

Chapters 4 and 5 use this framework to present compact definitions of different rhythm perception models. Chapter 4 shows that the compact definitions can reveal complex and interacting deterministic mechanisms in such models, and the framework described here allows them to be disentangled from these models, revealing the underlying dynamic Bayesian network model.

Chapter 4

Deterministic constraints of a probabilistic rhythm perception model

4.1 Introduction

The previous chapter developed a framework in which discrete dynamic Bayesian networks with deterministic constraints can be defined compactly. This chapter aims to demonstrate both the expressive capacities of the framework and the advantages it brings to the definition of cognitive models. We do so by using the framework to formulate an adaptation of a probabilistic generative model of meter perception proposed by Temperley (2007).

Temperley's model is intended to simulate how listeners infer a rich and complex hierarchical structure, namely meter, from a melody. Temperley describes this model in a top-down fashion: first, the model generates a metrical grid that represents a hierarchical metrical structure, then, on top of this grid, it generates a rhythm represented by a sequence of "pips"—consecutive time points at which onsets may occur, separated by a fixed and small temporal duration. Here, we present a dynamic Bayesian network adaptation of the model. Since dynamic Bayesian networks are by definition temporally incremental, this adaptation can process a rhythm in a temporally incremental, "left-to-right" fashion. This requires some of the probabilistic decisions that the original model makes globally, while generating a metrical grid and before generating a rhythm, to be localized to a specific moment in time. A few minor changes to the original model are proposed to achieve this.

One advantage of the model-definition framework that we hope to demonstrate

is that the formal definitions that it produces are detailed yet relatively concise. In order to arrive at the definition of Temperley's model below, we sometimes needed to consult the algorithmic implementation provided by Temperley. [1] This brought to light a few constraints that are not explicit in the model's original description, illustrating that informal, or partially formal, definitions of theories and computational models, which are prevalent in the literature (e.g., Temperley, 2007; Van der Weij et al., 2017 [Chapter 6]), may leave a, sometimes surprising, number of details and subtle interactions unarticulated. [2] The model-definition framework introduced in Chapter 3, by contrast, requires a model's constraints to be specified formally and makes these constraints explicit.

Furthermore, compact but precise definitions of computational models facilitate their implementation and comparison. This is demonstrated in Chapter 5, where we present two rhythm models that are compared in Chapter 7. The differences between models are reduced to the differences between their model-definition tables and conditional distribution definitions, instead of being hidden in textual model descriptions. In Chapter 5, we define a constrained version of Temperley's model.

In Section 4.2 we give a high-level overview of Temperley's model. The dynamic Bayesian network formulation is presented in Section 4.3. Finally, in Section 4.4 we summarize the definition of the model and reiterate some of the observations we made along the way.

4.2 Model overview

Temperley's original model and the version described here are based on a discretized representation of continuous time. This representation describes rhythms as a sequence of consecutive time points, separated by a small and constant unit of duration. These time points are called *pips* and the duration separating them is defined to be fifty milliseconds by Temperley (2007, p. 31). Each pip represents a binary random event, namely whether or not an onset occurs at its time point. The random variable N describes this outcome. It can assume two values: t or f (for *true* if a note occurs and *false* if it does not). In Chapter 2, we called models employing this type of rhythm representation *grid models*, because they represent rhythms as a discrete grid of time points.

[1] The implementation that we consulted can, at the time of writing, be found in a file called `meter16.c` downloadable from http://davidtemperley.com/music-and-probability/.

[2] Desain and Honing (1999) made a similar observation about rule-based models of beat induction. In their effort to implement these models, they found that the rules of these models, for some of which only verbal descriptions were available, interacted in ways not made explicit in the descriptions, and sometimes unforeseen by their authors.

Meter is conceptualized as an abstract multileveled hierarchical structure, prescribed in notated music by a time signature. The model always generates three levels of metrical hierarchy: an *upper level*, a *tactus level*, and a *lower level*. Each level represents a more or less regular stream of perceived pulses or beats. The tactus level represents a pulse at a moderate rate to which the listener would, for example, tap their foot (Lerdahl & Jackendoff, 1983). The upper level corresponds to a measure—that is, a grouping of tactus beats by two or three. The lower level corresponds to faster pulses obtained by subdividing the tactus beats by two or three. A hierarchical relation between levels is established by requiring that points in time corresponding to a beat at one level must also be a beat at all levels below: for example, a beat at the upper level (indicating the beginning of a bar), must also be a beat at the tactus level and at the lower level. This is equivalent to metrical well-formedness rule (MWFR) 2 described by Lerdahl and Jackendoff (1983).

The model describes metrical structure by two random variables: U (UT in Temperley's description) describes the number of tactus beats per upper-level beat and can assume the values two or three. L (LT in Temperley's description) describes the number of lower-level beats per tactus beats and can also be two or three. This structure is determined globally and does not change throughout a rhythm.

The variable *Uph* describes how the metrical grid aligns with a rhythm. More precisely, it describes the phase of the first tactus beat with respect to the upper level. In other words, it encodes whether the first tactus beat is the first, second, or third beat in the bar (the latter is only possible if the upper level is triple). The dynamic Bayesian network version of the model that we describe below introduces an additional variable, *Tph*, that describes the offset of an onset from the last tactus beat.

Two aspects of the model engage with tempo and expressive timing aspects of rhythms: the duration of the tactus interval can fluctuate throughout a rhythm and lower-level beats may be unevenly spaced between tactus beats. The duration of the nth tactus interval is described by the variable T_n and the positions of lower-level beats are described, for the nth tactus interval, by three random variables: Db_n (for duple beat) describes the location of a lower-level beat when the lower level is duple ($L = 2$), $Tb1_n$ and $Tb2_n$ (for triple beat one and two) describe the locations of the first and second lower-level beat when the lower level is triple ($L = 3$). However, both irregular spacing of lower-level beats and changes to the tactus interval decrease the a priori probability of the metrical structure. That is, the model considers these irregularities as deviations from an abstract ideal posited by the model: a metrical analysis in which the tactus interval does not fluctuate and in which lower-level beats are spaced equally in between tactus beats.

The model outlined above is subject to deterministic constraints. Some of these are evident in the model's description. For example, if the number of tactus beats per bar is two, there are only two possible upper-level phases, namely zero and one, while if there are three tactus beats per bar, three upper-level phases are possible. Others are more subtle. For example, the positions of lower-level beats fall within a set of possible positions determined by the tactus interval. If there are two lower-level beats, the second one must occur after the first one, and the first one must leave room for the second one to occur. Additionally, lower-level beats must not deviate by more than three pips from their most probable location. The congruency constraints described in Section 4.3.1 formally defined these constraints.

4.3 A dynamic Bayesian network formulation

Below, we define Temperley's original model (*the original model*) as a dynamic Bayesian network with deterministic constraints (*the DBN model*). This model is based on an interpretation of the generative model that conceptually is different from the one Temperley presents: [3] The original model and its algorithmic implementation emphasize a generative view: a metrical grid, which must always begin and end with a tactus beat, is generated first, on top of which a rhythm is generated, with the restriction that the first onset must occur before the second tactus beat. The DBN model may be said to emphasize the perspective of a listener who listens to a rhythm that is revealed incrementally in time. The listening begins when the first onset of a rhythm occurs. Therefore, in the DBN adaptation, the first moment corresponds to the first onset of a rhythm, which may align with any position in a tactus interval. Practically, this necessitates the addition of a *tactus phase* variable, *Tph*, which encodes the offset in pips of the current pip from the last tactus beat.

A variable in the original model that is not included in the DBN model is the "another beat" variable, A, which decides whether another tactus beat is generated at the end of each tactus interval. The variable A makes the number of moments that occur a random outcome that determines the duration of a grid that is generated. In the DBN model, moments correspond to subsequent atomic temporal units and this progression of time is not part of the generative model.

The DBN model generates a rhythm moment by moment. A *moment* corresponds to a pip in Temperley's model. A Dynamic Bayesian network is a first-order Markov model. This means that in each moment, a probability distributions describing a state of affairs must be generated from information that is locally available. These restrictions force us pick precise moments in time at which

[3] We thank David Temperley for clarifying this to us in private communication.

certain probabilistic decisions are made. This decisions propose an answer to questions like: when are the positions of lower-level beats are decided and when is the duration of the next tactus interval decided? In Temperley's top-down formulation, these decisions are made when the metrical grid is generated, which happens before the rhythm is generated. In the DBN formulation, such decisions are made at specific moments in time (namely the first moment, and whenever a tactus interval ends, as explained below).

The deterministic constraints of the model are defined in Section 4.3.1, and discussed in Sections 4.3.1.1, 4.3.1.3, and 4.3.1.4. In Section 4.3.2, we discuss issues regarding the interpretation of the first and the final moment. Section 4.3.3 describes some possibly unintended inferential biases that are the result of deterministic constraints on conditional probability distributions. Finally, the model's parameters, and probability distributions of its variables are described in Section 4.3.4.

4.3.1 Deterministic constraints

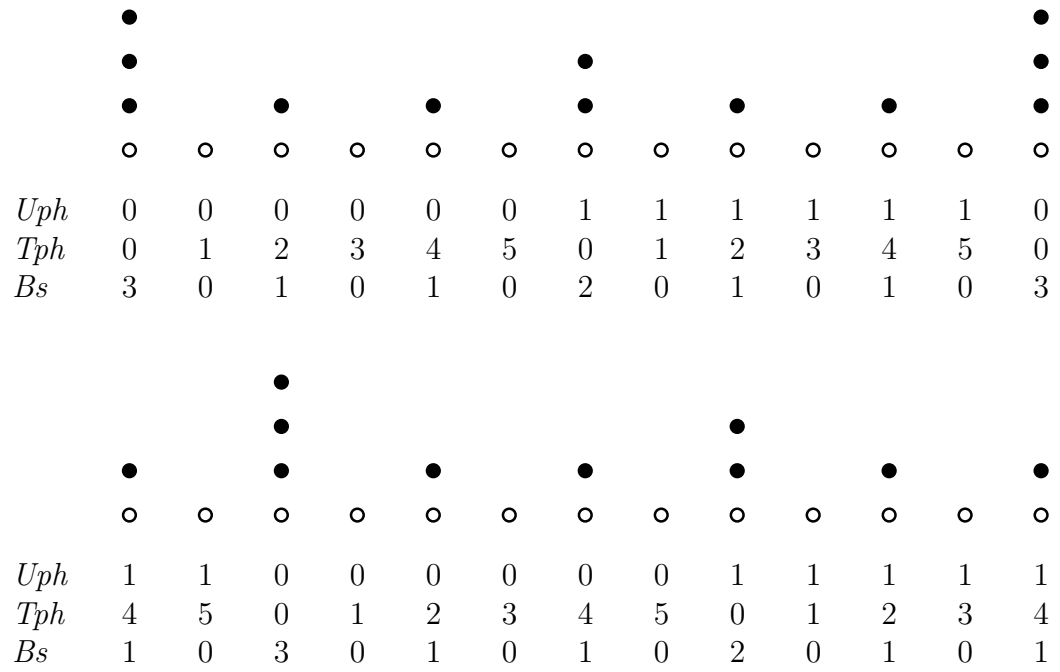
Table 4.1 defines the deterministic constraints of the DBN model. Two basic observations can be made upfront. First, the variables U and L are persistent (see Chapter 3), reflecting the fact that they are generated once and do not change their value throughout a rhythm. Second, the piece-wise function definitions in Table 4.1 are used to differentiate between two special states of the joint distribution: The first moment is identified by checking that the previous value of some variable is $*$. Every variable, except *Bs*, uses this pattern to define custom congruency constraints for the initial moment. The conditions $tph = 0$, and $\hat{tph} + 1 = \hat{t}$ both check for the situation that a tactus interval has ended. This situation is used by *Uph* to check whether the upper-level phase needs to be increased, by *Db*, *Tb1*, and *Tb2* to check whether new lower-level beat positions need to be generated, and by *Bs* to check whether the metrical salience of the current moment is higher than one. Below, we describe the congruency constraints related to phase (Section 4.3.1.1), tactus interval (Section 4.3.1.2), positioning of lower-level beats (Section 4.3.1.3), and determination of metrical salience (Section 4.3.1.4) in more detail.

4.3.1.1 Phase

Uph represents the offset of the current moment from the last upper-level beat, measured in tactus beats. *Tph* represents the offset of the current moment from the last tactus beat, measured in pips (which correspond to moments). We hope to be forgiven for the slight inconsistency in nomenclature that upper-level phase

Table 4.1: A dynamic Bayesian network version of the rhythm model described by Temperley (2007), with congruency constraints describing its deterministic constraints.

X_i	PA_i	D_{X_i}	$\kappa_{X_i}(pa_{X_i})$
U	$\{\hat{U}\}$	$\{2,3\}$	$\begin{cases} D_U & \text{if } \hat{u} = *, \\ \{\hat{u}\} & \text{otherwise.} \end{cases}$
L	$\{\hat{L}\}$	$\{2,3\}$	$\begin{cases} D_L & \text{if } \hat{l} = *, \\ \{\hat{l}\} & \text{otherwise.} \end{cases}$
T	$\{\hat{T}, \hat{Tph}\}$	$\{n \in \mathbb{N} \mid 9 \leq n \leq 22\}$	$\begin{cases} D_T & \text{if } \hat{tph} = * \text{ or } \hat{tph} + 1 = \hat{t} \\ \{\hat{t}\} & \text{otherwise.} \end{cases}$
Uph	$\{\hat{Uph}, Tph, U\}$	$\{0,1,2\}$	$\begin{cases} \{n \in \mathbb{N} \mid 0 \leq n < u\} & \text{if } \hat{uph} = *, \\ \{(\hat{uph} + 1) \mod u\} & \text{if } tph = 0, \\ \{\hat{uph}\} & \text{otherwise.} \end{cases}$
Tph	$\{\hat{Tph}, \hat{T}, T\}$	$\{n \in \mathbb{N} \mid 0 \leq n \leq 22\}$	$\begin{cases} \{n \in \mathbb{N} \mid 0 \leq n < t\} & \text{if } \hat{tph} = * \\ \{(\hat{tph} + 1) \mod \hat{t}\} & \text{otherwise} \end{cases}$
Db	$\{\hat{Db}, Tph, T, L\}$	$\{n \in \mathbb{N} \mid 1 \leq n < 22\}$	$\begin{cases} \{-1\} & \text{if } l = 3 \\ \text{LB}(0,1,t,l) & \text{if } tph = 0 \text{ or } \hat{db} = * \\ \{\hat{db}\} & \text{otherwise.} \end{cases}$
$Tb1$	$\{\hat{Tb1}, Tph, T, L\}$	$\{n \in \mathbb{N} \mid 1 \leq n < 21\}$	$\begin{cases} \{-1\} & \text{if } l = 2 \\ \text{LB}(0,1,t,l) & \text{if } tph = 0 \text{ or } \hat{tb1} = * \\ \{\hat{tb1}\} & \text{otherwise.} \end{cases}$
$Tb2$	$\{\hat{Tb2}, Tb1, Tph, T, L\}$	$\{n \in \mathbb{N} \mid 2 \leq n < 22\}$	$\begin{cases} \{-1\} & \text{if } l = 2 \\ \text{LB}(tb1, 2, t, l) & \text{if } tph = 0 \text{ or } \hat{tb1} = * \\ \{\hat{tb2}\} & \text{otherwise.} \end{cases}$
Bs	$\{Db, Tb1, Tb2, Tph, Uph\}$	$\{0,1,2,3\}$	$\begin{cases} \{3\} & \text{if } tph = 0 \text{ and } uph = 0, \\ \{2\} & \text{if } tph = 0 \text{ and } uph \neq 0, \\ \{1\} & \text{if } tph \in \{db, tb1, tb2\}, \\ \{0\} & \text{otherwise.} \end{cases}$
N	$\{\hat{N}, Bs\}$	$\{t,f\}$	$\begin{cases} \{t\} & \text{if } \hat{n} = *, \\ \{t,f\} & \text{otherwise.} \end{cases}$



Uph	0	0	0	0	0	0	1	1	1	1	1	1	0
Tph	0	1	2	3	4	5	0	1	2	3	4	5	0
Bs	3	0	1	0	1	0	2	0	1	0	1	0	3

Uph	1	1	0	0	0	0	0	0	1	1	1	1	1
Tph	4	5	0	1	2	3	4	5	0	1	2	3	4
Bs	1	0	3	0	1	0	1	0	2	0	1	0	1

Figure 4.1: Two metrical interpretations of a sequence of thirteen moments (pips). In both interpretations, the meter is given by $U = 3, L = 2, T = 6$. The moments themselves are indicated by unfilled circles below the metrical grid (a rhythm is not shown). The values of the upper-level phase, *Uph*, the tactus phase, *Tph*, and the beat salience, *Bs*, corresponding to each moment are shown below the grid. The way in which the grid aligns with the meter is specified by the first (left-most) values of *Uph* and *Tph* in the sequence of moments.

is measured in tactus beats (the level immediately below the upper level), while the tactus phase is measured in pips (rather than lower-level beats).

These variables serve a dual purpose: in the first moment, they generate a set of alternative hypotheses about the tactus and upper-level phase of the first moment, while in subsequent moments they deterministically track the metrical status of the current moment. In the first moment, the values of *Uph* are constrained by the value of U (the number of tactus beats per bar). *Tph* may assume values constrained by the value of T, the number of pips in the current tactus interval.

In non-initial moments, the state of *Uph* is deterministic and is incremented by one, modulo U, whenever a tactus interval ends (when $tph = 0$). Similarly, the state of *Tph* is deterministic given $\hat{tph}$ and $\hat{t}$ (the tactus phase and tactus interval in the previous moment) and is incremented by one, modulo $\hat{t}$ (the tactus interval prevailing in the previous moment). The reason that *Tph* uses the tactus interval in the previous moment is that the moment that a tactus interval ends (when $tph + 1 = t$) a new tactus interval is generated by T.

The interpretation of *Uph* and *Tph* is illustrated in Figure 4.1, which shows two metrical interpretations, represented by a *metrical grid* (Lerdahl & Jackendoff, 1983; Temperley, 2007), of a sequence of pips (represented by unfilled circles). The meter of both interpretations is the same, but the initial values of *Uph* and *Tph* are different. That is, the Figure illustrates how *Uph* and *Tph* encode the way in which a meter aligns with a grid of pips.

As mentioned before, the original model does not have a *Tph* variable. In the original model, the first pip represents the first tactus beat, which may occur at different moments with respect to the first onset. As such, the time point represented by a pip is interpreted relative to the first tactus beat and it is not possible to link a pip to an absolute moment in time. In the DBN model, the addition of the *Tph* enables a pip to be interpreted as an absolute moment in time, during which the model may have an exhaustive set of alternative hypotheses of the metrical status of that moment.

4.3.1.2 Tactus interval

T is deterministic and generates its previous value, $\hat{t}$, in in any moments except the first and whenever the previous tactus interval ends, as indicated by the previous tactus phase plus one being equal to the tactus interval: $\hat{tph} + 1 = \hat{t}$. In other words, a new tactus interval is generated in the first moment and whenever a tactus interval ends. Following Temperley's definition, the tactus interval is bound by a minimum and maximum interval, set to 9 and 22 pips (moments), corresponding to between 450 and 1100 milliseconds.

4.3.1.3 Lower-level beats

Like the tactus interval, the locations of lower-level beats, *Db*, *Tb1*, and *Tb2*, are generated whenever the tactus interval changes (when $tph = 0$) and in the first moment. These positions are restricted by several interacting constraints that can be isolated into a single function, LB. The need for three different variables described by similar constraints and probability distributions reflects a limitation of the expressive power of Bayesian networks. While the position of a lower-level beat could be described by one conditional probability distribution, the number of lower-level beats that occur is contingent on the outcome of another random variable, namely L. Since such contingencies cannot be expressed by a Bayesian network (see Russell, 2015), separate variables for each lower-level beat are required. Since whether *Db* or *Tb1* and *Tb2* apply to a metrical analysis depends on the value of L, these variables can "disable" themselves by deterministically generating an arbitrary symbol that does not occur in D_{Tph} (we chose -1, as can be seen in Table 4.1).

The definition of LB, which describes the constraints on the positions of lower-level beats, is somewhat detailed because lower-level beats are subject to multiple constraints: Lower-level beats must occur before the end of tactus interval, they must not deviate from their most probable location by more than three pips, and, if $L = 3$, they must leave room for the next lower-level beat to occur.

The most probable locations of lower-level beats are the points that divide the tactus interval into two or three equal intervals, rounded down to the nearest pip. The function DEV: $D_{Tph} \times \{1, 2\} \times D_T \times D_L \to \{n \in \mathbb{N} \mid 0 \leq n < 21\}$ describes how far a given lower-level beat deviates from this point. Given the location $b \in D_{Tph}$ and phase, $bph \in \{1, 2\}$ of a lower-level beat, a tactus interval, t, and the number of lower-level beats per tactus interval, l, DEV is defined as

$$\text{DEV}(b, bph, t, l) = \text{ABS}\left(b - \text{FLOOR}\left(bph \cdot t/l\right)\right).$$

where the function ABS(r) returns the absolute value of r, and the function FLOOR(r) returns the highest integer below r. Note that *bph*, unlike a value of *Tph*, encodes the phase of a lower-level beat with respect to the tactus interval *measured in lower-level beats*, rather than pips.

The function LB: $D_{Tph} \times \{1, 2\} \times D_T \times D_L \to D_{Tph}$ generates a set of possible locations of a lower-level beat. Given the location of the previous (lower-, tactus-, or upper-level) beat, $\hat{b} \in D_{Tph}$, its phase, $bph \in \{1, 2\}$, a tactus interval, t, and the number of lower-level beats per tactus interval, l, LB is defined as

$$\text{LB}(\hat{b}, bph, t, l) = \{b \in D_{Tph} \mid \hat{b} + 1 \leq b \leq t - (l - bph), \text{DEV}(b, bph, t, l) < 4\}.$$

For *Db* and *Tb1*, the position of the previous beat is zero, while for *Tb2* it is *tb1* (see Table 4.1). Each beat must occur at least one pip after the previous beat (i.e., it must be greater or equal to $\hat{b} + 1$), it must leave room for any subsequent beats (i.e., it must be smaller than or equal to $t - (l - bph)$), and it must not deviate by more than three pips from its most probable location.

The description of the original model does not make explicit that the distribution of *Tb2* is conditioned on *Tb1* since *Tb2* must occur after *Tb1*. The constraints described above do occur in the algorithmic implementation provided by Temperley. As such, they are an example of how implementation details are made explicit by congruency constraints.

4.3.1.4 Metrical salience

Finally, a deterministic variable *Bs* determines the metrical salience of a moment given the tactus and upper-level phase and the locations of the lower-level beats. The possible levels of metrical salience are $\{0, 1, 2, 3\}$. Moments that do not align with any beat have a metrical salience of zero, and moments that align with a lower-level, tactus-level, or upper-level beat have a metrical salience of respectively one, two, or three. The variable N, which is conditioned on *Bs*, describes whether or not a note onset occurs within the current moment. An example of how a metrical interpretation specifies the metrical salience of a sequence of moments is provided by Figure 4.1, where the beat salience of each moment given two different interpretations of a sequence of moments is shown.

4.3.2 Interpretation of the first and the final moment

An issue of some delicacy is how to interpret the meaning of "the first moment" and "the final moment". It seems reasonable to let the first moment correspond to the first onset in a rhythm. This leads to a deterministic constraint that the first pip always contains an onset, as reflected by the congruency constraint of N in Table 4.1. [4]

One possibility for the interpretation of the final pip is that it corresponds to the final onset in a rhythm. This represents a deterministic constraint: the final pip always contains an onset. However, this constraint cannot be captured by congruency constraints of a dynamic Bayesian network since it would violate temporal order: it is not possible to know a priori when the last onset occurs. Alternatively, one could consider only rhythms represented on a grid of a predetermined fixed duration. In this case, the final onset could occur anywhere within this sequence. Since the best solution seems to be application-dependent, we will leave the issue of how to interpret the final pip open.

4.3.3 Period-dependent biases

Both the original model and the DBN model exhibit biases that depend on the duration of tactus intervals. First, analyses positing long tactus intervals are favored a priori. This can be seen by considering two alternative analyses of a rhythm represented by n pips. An analysis that posits short tactus intervals will contain more moments in which a tactus interval ends ($\hat{tph} + 1 = \hat{t}$), and a new tactus interval has to be selected from the congruent states of T (see the

[4] Interestingly, dropping this deterministic constraint leads to the the "rhythm" of John Cage's $4'33''$ to be included in the set of possible rhythms that can be generated by the model.

congruency constraint of T in Table 4.1), compared to an analysis positing long tactus intervals. Therefore, the short-tactus-intervals analysis is a priori less likely than the long-tactus-intervals analysis. This bias could be addressed in future work by allowing a prevailing tactus interval to shrink or expand within each moment, rather than only when a tactus interval ends.

Second, a bias in the reverse direction is present for the first tactus interval. If the first tactus interval is long, the probability of any given value of *Tph* (except *Tph* = 0), due to the definition of its probability distribution (see Section 4.3.4) is less likely than if the tactus interval is short. In general, models with deterministic constraints seem especially susceptible to such biases: when the number of congruent states of a latent variable depends on the outcome of another variable, biases are likely.

4.3.4 Model parameters

Parameters of probability distributions of the congruent states of each variable are given by maximum likelihood estimates derived from an empirical sample of rhythms. Temperley (2007) uses a subset of the Essen folksong collection (Schaffrath & Huron, 1995) for this purpose. Since this collection does not represent tempo and timing aspects of musical performances, a few distributions are based on musical intuition or prior research.

The distribution of the tactus interval, T, given the previous tactus interval, is given by

$$\Pr(t \mid \hat{tph}, \hat{t}) = \begin{cases} \theta_t^T / \sum_{t' \in D_T} \theta_{t'}^T & \text{if } \hat{t} = *, \\ f^t(t, \hat{t}) / \sum_{t' \in \kappa_T((t, \hat{tph}))} f(t', \hat{t}) & \text{if } t \in \kappa_T((\hat{tph}, \hat{t})), \\ 0 & \text{if otherwise.} \end{cases}$$

where the function f^t is a function that returns a score representing how likely it is that tactus interval t follows the previous tactus interval, $\hat{t}$. This function is given by

$$f^t(t, \hat{t}) = e^{-\left(\frac{t-\hat{t}}{2}\right)^2}.$$

In the distribution of T, these scores are normalized to create a probability distribution over tactus intervals given a previous tactus interval. These probabilities are maximal when the tactus interval is equal to the previous tactus interval. Note that the normalization factor automatically ensures that when the tactus interval has not ended, the probability of whatever tactus interval has been generated is one, since the congruent states of the variable are a singleton in that case.

Table 4.2: The initial tactus interval distribution as defined by Temperley (2007).

t	9	10	11	12	13	14	15	16	17	18	19	20	21	22
θ_t^T	.1	.2	.3	.23	.13	.03	.006	.002	.001	.0006	.0002	.0001	.00005	.00005

The parameters of the tactus interval distribution in the initial moment (when $\hat{t} = *$), $\{\theta_t^T\}_{t=9}^{22}$, are shown in Table 4.2. These parameters represent the prior probability of different tactus interval durations.

The probability of a lower-level beat location depends on how far it deviates from its most probable location. The conditional probability distributions of *Db*, *Tb1*, and *Tb2* are shown below.

$$\Pr(db \mid \hat{db}, tph, t, l) = \frac{\theta^B_{\text{DEV}(db,1,t,l)}}{\sum_{b \in \kappa_{Db}((\hat{db},tph,t,l))} \theta^B_{\text{DEV}(db,1,t,l)}},$$

$$\Pr(tb1 \mid \hat{tb1}, tph, t, l) = \frac{\theta^B_{\text{DEV}(tb1,1,t,l)}}{\sum_{b \in \kappa_{Tb1}((\hat{tb1},tph,t,l))} \theta^B_{\text{DEV}(tb1,1,t,l)}},$$

$$\Pr(tb2 \mid \hat{tb2}, tb1, tph, t, l) = \frac{\theta^B_{\text{DEV}(b,bph,t,l)}}{\sum_{b \in \kappa_{Tb2}((\hat{tb2},tb1,tph,t,l))} \theta^B_{\text{DEV}(tb2,2,t,l)}}.$$

Note that the three definitions above are virtually identical except for the constant parameters supplied to DEV, and the arguments to the congruency constraints. These distributions share a set of parameters, $\{\theta_d^B\}_{d=0}^3$, that represent the probabilities of different degrees of beat deviation, d. Temperley defines these as follows: $\theta_0^B = .32, \theta_1^B = .24, \theta_2^B = .08$, and $\theta_3^B = .02$.

The remaining parameters are estimated from an empirical sample of rhythms. The parameters θ^U and θ^L represent respectively the probability that the upper level is triple and that the lower level is triple. The distribution of *Uph* has three parameters, $\theta_0^{Uph}, \theta_1^{Uph}$, and θ_2^{Uph} representing respectively the probabilities that the upper-level phase is zero given that upper level is duple, and the probabilities that the upper-level phase is zero or one given that the upper level is triple.

The distribution of whether an onset occurs at the present moment, $\Pr(n \mid \hat{n}, bs)$ has four parameters: $\{\theta_s^N\}_{s=0}^3$. These represent the probabilities with which notes occurs at different levels of metrical salience, s. As described in Section 4.3.1.4, there are four levels of metrical salience: 0, 1, 2, and 3. Level 0 applies to any moment that does not align with a beat in the three metrical levels defined by the model. Levels 1, 2, and 3 apply to moments that represent respectively lower-, tactus-, and upper-level beats. Temperley estimated the following values for these parameters based on a collection of German folksongs: $\theta_0^N = .01, \theta_1^N = .38, \theta_2^N = .74$, and $\theta_3^N = .95$. The probability distribution of N is shown below.

$$\Pr(n \mid \hat{n}, bs) = \begin{cases} 1 & \text{if } \hat{n} = *, \\ \theta^N_{bs} & \text{if } n = t, \\ (1 - \theta^N_{bs}) & \text{otherwise.} \end{cases}$$

Finally, the distribution of TPH has one parameter, θ^{Tph}, that reflects the probability of whether a rhythm begins on a tactus beat. The original model gives extra weight to analyses in which the first onset aligns with a tactus beat. This is achieved using a special value of *Bs*, namely "first tactus beat", which it assumes only when the current moment corresponds to the first tactus beat. Since we have constrained the first pip to always contain an onset, this situation corresponds to whether *Tph* is or is not equal to zero. The distribution of *Tph* is given by

$$\Pr(tph \mid \hat{t}, \hat{tph}, t) = \begin{cases} \theta^{Tph} & \text{if } tph = 0 \text{ and } \hat{tph} = *, \\ \theta^{Tph}/(t-1) & \text{if } tph \neq 0 \text{ and } \hat{tph} = *, \\ 1 & \text{otherwise.} \end{cases}$$

4.4 Summary

In this chapter, we defined a dynamic Bayesian network model with deterministic constraints based closely on a probabilistic rhythm model proposed by Temperley (2007). To present this model, we used the model-definition framework developed in Chapter 3. Detailed aspects of the model, such as the complex and interacting deterministic constraints that govern the position of the lower-level beats, are made fully explicit in a congruency-constraints-based definition. Details such as these are commonly considered "implementation details" of a model and therefore remain hidden in an algorithmic implementation of the model (if one is provided at all). Although they might seem minute, understanding these details is necessary for completely understanding and replicating a model and its behaviour. The model-definition framework of Chapter 3 could be said to reduce the gap between a published definition of a probabilistic generative model and its algorithmic implementation. A model-definition table completely specifies the deterministic constraints at play in a model in a relatively concise manner.

Our efforts to define the model's deterministic constraints brought to light some nuanced issues: The DBN model represents a temporally incremental model of rhythm perception, while the original model emphasizes a generative perspective. The DBN model considers the first moment to represent the onset of the first note in a rhythm, while the original model generates a metrical grid, the first position of which represents the beginning of the first tactus interval. Arguably, the original model represents the perspective of the performer or composer of

a rhythm, who is aware of the position of the first tactus beat, while the DBN model represents the perspective of the listener, whose experience of the rhythm begins at the moment the first onset occurs.

Finally, we highlighted that the model exhibits independent biases toward a short initial tactus interval, and toward long tactus intervals over the entire rhythm (see Section 4.3.3). In general, such biases are easily introduced in probabilistic models with deterministic constraints. However they can be easy to miss, especially when the definition of a model is not fully formal.

We also demonstrated that models defined in the framework of Chapter 3 are temporally incremental process models. That is, moments are processed one by one, and the joint distribution over all the model's variables must be generated on the basis of locally available information. In the DBN model, this led us to use the tactus-phase variable to keep track of the metrical status of the current moment. These variables, and their deterministic constraints, enable all probabilistic decisions, such as when to generate a new tactus interval or when to generate new positions for the lower-level beats, to be made at the moment that they become relevant.

Chapter 5

Rhythm spaces and two rhythm models

5.1 Introduction

In this chapter, we use the model-definition framework described in Chapter 3 to define adaptations of two different rhythm perception models. One model, which we call the *enculturation model*, is a reformulation of a model proposed by Van der Weij et al. (2017 [Chapter 6]). The other model, which we call the *classical model*, is based on a probabilistic rhythm perception model proposed by Temperley (2007). We presented a dynamic Bayesian network version of this model in Chapter 4. The classical model, in contrast to Temperley's original model, generates sequences of symbolic inter-onset intervals, rather than metrical grids. This change enables us to compare the classical directly to the enculturation model. Algorithmic implementations of the model-definition tables presented in this chapter can be found in Appendix A. This chapter discusses the technical details of these models, while in Chapter 7, where the two models are compared in a cross-cultural experiment, they are discussed at a more conceptual level.

The purpose of this chapter is to present formal definitions of two generative rhythm perception models and to describe how their parameters are estimated from empirical samples of rhythms in a rhythm space. A *rhythm space* is a finite set of rhythms over which a probabilistic generative model of rhythms defines a complete probability distribution. In this chapter, we use the term *rhythm model* to refer to models like the classical and enculturation model that define probability distributions over a rhythm space. The rhythm space of a rhythm model defined as a dynamic Bayesian network with deterministic constraints corresponds to its congruent input sequences of a specific length (see Chapter 3, p. 69).

The classical model and the enculturation model represent different theoretical views of how meter is inferred from rhythms. The classical model, like Temperley's original model, assumes that, when the meter is known, the probability that an onset occurs at a particular point in time depends exclusively on the metrical salience of that point in time. The enculturation model, by contrast, assumes that, when the meter is known, the probability that an onset occurs at a particular point in time depends on the preceding context—that is, the rhythmic pattern preceding the current onset—and the meter in which the preceding context is interpreted. While the classical model is constrained primarily by music theory, the enculturation model is constrained primarily by the empirical rhythm sample from which it learns. Regarding the probability of a rhythm given a meter, the classical model only learns the probabilities with which onsets occur at different levels of metrical salience from empirical samples, whereas the enculturation model learns associations between rhythmic patterns and meters from such samples.

The parameters of the two rhythm models can be estimated from an empirical sample of rhythms that occur in a specific rhythm space. This causes the probability distribution over the rhythm space defined by the rhythm model to approximate the distribution from which the sample was drawn. We can distinguish between empirical and learned probability distributions over the set of rhythms in a rhythm space. The empirical distribution represents the hypothetical distribution underlying an empirical sample. The learned distribution is defined by a model whose parameters have been estimated from an empirical sample. To estimate the degree to which the learned distribution approximates an empirical distribution, we can evaluate the probability that rhythms drawn from the empirical distribution have in a learned distribution. This probability is known as the *model evidence* of a rhythm given a model.

In Chapters 6 and 7, we interpret the parameter estimation procedure as a simulation of the influence of long-term exposure to a certain musical environment on rhythm perception. An empirical rhythm distribution, in this scenario, reflects the average probability with which different rhythms are encountered in that musical environment.

The comparison in Chapter 7 uses an evaluation metric, namely estimated cross-entropy that is based on model evidence. This is the same metric that Temperley (2010) uses to compare different rhythm models. However, to be able to compare the behavior of rhythm models in this way, it is important that these models define probability distributions over the same rhythm space. In a grid-based model, for example, model evidence is affected by the *duration* of a rhythm, since the number of random events described by such a model depends on the duration of a rhythm. In a model that generate sequences of inter-onset intervals, model evidence is affected instead by the *number of events* (i.e., note onsets) in a rhythm, since in such models, the timing of each onset onset is a random event. Ensuring that

Table 5.1: A full enumeration of the eight rhythms in a rhythm space with rhythms of length three and inter-onset interval domain $\{1, 2\}$.

1	1	1	1	**5**	2	1	1
2	1	2	1	**6**	2	2	1
3	1	1	2	**7**	2	1	2
4	1	2	2	**8**	2	2	2

models generate the same rhythm space avoids these issues.

This chapter lays the groundwork for the simulations reported in Chapter 7 in which the behavior of the two models is systematically assessed and compared as a function of the empirical rhythm sample from which their parameters are estimated. This is done by defining a fixed rhythm space in which empirical samples are represented and over which the models learn probability distributions. Rhythm spaces and empirical rhythm samples are discussed in Section 5.2. The models are defined in Sections 5.3 (the classical model) and 5.4 (the enculturation model). These sections describe the congruency constraints of the models and explain how their parameters are estimated from empirical rhythm samples. Finally, Section 5.5 summarizes the concepts developed in this chapter.

5.2 Rhythm spaces

The two rhythm models described in Sections 5.3 and 5.4 define, in each moment, a marginal probability distribution over inter-onset intervals. We sometimes refer to this distribution as a model's *prediction* of the inter-onset interval that is to occur in the current moment. An inter-onset interval is the temporal interval between the moments at which two consecutive notes are played. Inter-onset interval predictions are described, in both models, by a random variable I. We assume for simplicity that there is a finite set of inter-onset intervals that can occur. This set, the *inter-onset interval domain*, corresponds to the domain of I.

A rhythm space is the set of congruent input sequences of length n of a rhythm model. Since both models discussed in this chapter generate the domain of I in each moment, a rhythm space is defined by the sequence length, n, and the inter-onset interval domain, D_I. In Chapters 6 and 7, the inter-onset interval domain is defined as the set of unique inter-onset intervals observed in a set of empirical rhythm samples.

Table 5.1 shows an example of a rhythm space for which $n = 3$ and $D_I = \{1, 2\}$. In general, the set of rhythms in a rhythm space with rhythms of length n corresponds to the Cartesian product of n inter-onset interval domains: ${D_I}^n$.

5.2.1 Rhythms

We interpret inter-onset intervals to be a symbolic representation of time intervals. Specifically, an inter-onset interval denotes an integer multiple of an atomic (indivisible) temporal unit. These units encode a symbolic duration, namely the duration of a *whole note* (denoted in music notation by the symbol $\mathbf{o}$) divided by an integer number $\rho \in \mathbb{N}$, called the *resolution* of the representation. We call these units ρ-*units*. In Chapters 6 and 7 we use a resolution of $\rho = 16$, such that units correspond to the duration of a sixteenth note.

Note durations and ρ-units do not directly correspond to physical time intervals. Instead, they specify the ratios between temporal intervals that must hold approximately in a rhythm that they represent. Since the range of durations with which different musical note values are typically played is constrained, the time intervals represented by ρ-units correspond to a constrained range of time intervals.

5.2.2 Empirical rhythm samples and parameter estimation

The parameters of the models are estimated from *empirical rhythm samples*. In Chapters 6 and 7, we use music corpora to obtain these samples. These corpora contain digital representations of music in formats that capture the same information as can be found in music notation. That information includes note and rest durations, bar lines, and time signatures.

Rhythms in a rhythm sample are represented as a sequence of inter-onset intervals. However, a rhythm sample additionally contains information about the metrical interpretation of each rhythm in the form of a time signature and the location of the first bar line. Time-signature changes are not supported by the rhythm models presented here and do not occur in the samples that we use in Chapters 6 and 7. Metrical interpretation information is used only when we estimate the parameters of a rhythm model.

We represent the time signature and position of the first bar line as follows: The numerator of the time signature is denoted by $num \in \mathbb{N}$, and the denominator by $denom \in \mathbb{N}$, such that $(num, denom) = (4, 4)$ represents 4/4 time and $(num, denom) = (6, 8)$ represents 6/8 time. A *pickup*, also known as an *anacrusis*, describes a situation where the first note in a rhythm does not occur at the beginning of the first bar. We define the *pickup interval*, $pickup \in \mathbb{N}$, as the interval in ρ-units between the beginning of the bar in which the first note onset occurs and the first note onset. [1] For example, if a rhythm notated in 4/4 time begins

[1] Lerdahl and Jackendoff (1983, p. 30) define an anacrusis in the context of grouping structure

a quarter-note before the first bar line, then $pickup = 12$ (assuming $\rho = 16$; a sixteenth-note resolution). That is, the first note is positioned three quarter notes away from the beginning of the bar in which it occurs. While the classical model and the enculturation model represent meter in different ways, their representation of meter can be derived from the metrical interpretation information supplied in a rhythm sample.

An empirical rhythm sample is a multiset (a set in which elements may occur multiple times). Each item in a rhythm sample is a 4-tuple $(r, num, denom, pickup)$, where r is a sequence of inter-onset intervals that occurs in a rhythm space. We use all information in these items—that is, rhythm and meter—for estimating the parameters of the rhythm models. The parameter estimation procedure is sometimes referred to as *training* a model. We use maximum-likelihood estimates of the model parameters: the parameter values that maximize the probability of the empirical sample. More details are provided in the Model parameters subsections of Sections 5.3 and 5.4. When we evaluate (or *test*) a model, for example, by calculating the model evidence of the sequence of inter-onset intervals, we only use the sequence of inter-onset intervals, r, and not the metrical interpretation information.

Both models use the first moment to generate a pickup interval. Since the pickup interval is an aspect of the metrical interpretation it cannot be observed by the model. While the pickup interval is generated in the first moment, both models generate the symbol $*$ as a deterministic state of I (the only observable variable in both models). This is reflected by the congruency constraints of I shown in Tables 5.2 and 5.4. All sequences of inter-onset intervals on which these models are evaluated are defined to begin with this symbol as the first observation. Since the $*$ symbol is generated deterministically, its observation carries no information with regard to metrical interpretation.

Conceptually, the first moment corresponds to the moment at which the first onset in a rhythm occurs. Subsequent moments represent subsequent onsets, each of which creates an inter-onset interval with the previous onset. That is, in the second moment, I describes the time interval between the first and second onset, in the third moment, it describes the time interval between the second and third onset, and so on. When we speak of the onset corresponding to a given moment, we mean the onset to which an inter-onset interval generated in that moment leads. Both the classical model and the enculturation model predict rhythms represented as a sequence of inter-onset intervals prefixed by the $*$ symbol. With

---

as the "span from the beginning of a group to the strongest beat in the group." A pickup interval, as we have defined it, applies only to the first note in a rhythm (grouping structure is outside the scope of this chapter) and instead refers to the interval *from* the strongest beat preceding it to the first note. This means that we can treat pickup interval simply as the phase of the first onset.

a common representation of rhythms in place, we now turn to the definition of the classical model and the enculturation model.

5.3 The classical model

The classical model is based on a probabilistic rhythm perception model described by Temperley (2007). In Chapter 4, we described a dynamic Bayesian network version of this model. The classical model, in contrast to these models, generates distributions of symbolic inter-onset intervals expressed in ρ-units (see Section 5.2) in each moment. Since these inter-onset intervals are a symbolic and score-like representation of rhythms, the classical model omits all aspects related to tempo and timing present in Temperley's original model. That is, the tactus-interval duration cannot change during a rhythm and the positions of the lower-level beats are fixed.

Importantly, the classical model preserves the primary assumptions made by the original model regarding the probability of a rhythm given a meter and the prior probabilities of meter: the prior probability of a meter and pickup interval is determined by a set of probabilistic decisions regarding the structure of a metrical hierarchy and the distribution of inter-onset intervals is based on four parameters that are learned from empirical data. These parameters represent the probabilities with which onsets occur at different levels of metrical salience.

Temperley (2007) uses the Essen folksong collection (Schaffrath & Huron, 1995) to estimate the parameters of the model. Since this corpus does not encode observations of tempo and expressive timing in rhythms, some parameters of the model are derived from musical intuition or prior research. By contrast, all parameters of the classical model can be estimated from empirical rhythm samples, since it does not model tempo and expressive timing.

Below, Section 4.3.1 describes the congruency constraints of the model, and Section 4.3.4 describes the conditional probability distributions of its variables. The latter section also describes how the parameters of these distributions are estimated from rhythm samples.

5.3.1 Deterministic constraints

Table 5.2 defines the random variables, dependency relations, and congruency constraints of the classical model. Below, we briefly describe the congruency constraints of each variable. Note, first, that the definitions of the variables U and L are identical to the corresponding definitions in Chapter 4. That is, the classical model generates the same meters as Temperley's original model.

Table 5.2: The congruency constraints of the classical model.

X_i	PA_i	D_{X_i}	$\kappa_{X_i}(pa_{X_i})$
U	$\{\hat{U}\}$	$\{2,3\}$	$\begin{cases} D_U & \text{if } \hat{u} = *, \\ \{\hat{u}\} & \text{otherwise.} \end{cases}$
L	$\{\hat{L}\}$	$\{2,3\}$	$\begin{cases} D_L & \text{if } \hat{l} = *, \\ \{\hat{l}\} & \text{otherwise.} \end{cases}$
T	$\{\hat{T}\}$	D_T	$\begin{cases} D_T & \text{if } \hat{t} = *, \\ \{\hat{t}\} & \text{otherwise.} \end{cases}$
Uph	$\{\hat{Uph}, U\}$	$\{0,1,2\}$	$\begin{cases} \{n \in \mathbb{N} \mid 0 \leq n < u\} & \text{if } \hat{uph} = *, \\ \{*\} & \text{otherwise.} \end{cases}$
Tph	$\{\hat{Tph}, T\}$	$\{n \in \mathbb{N} \mid 0 \leq n \leq t, t \in D_T\}$	$\begin{cases} \{n \in \mathbb{N} \mid 0 \leq n < t\} & \text{if } \hat{tph} = *, \\ \{*\} & \text{otherwise.} \end{cases}$
P	$\{\hat{P}, I, U, Tph, Uph, T\}$	$\{n \in \mathbb{N} \mid 0 \leq n < u \cdot t, u \in D_U, t \in D_T\}$	$\begin{cases} \{tph + uph \cdot t\} & \text{if } \hat{p} = *, \\ \{(\hat{p} + i) \mod (u \cdot t)\} & \text{otherwise.} \end{cases}$
I	$\{\hat{P}, U, L, T\}$	D_I	$\begin{cases} \{*\} & \text{if } \hat{p} = *, \\ D_I & \text{otherwise.} \end{cases}$

As in Chapter 4, the variable T represents the time interval between tactus beats. However, since the atomic temporal units of the model are symbolic ρ-units, T encodes the duration of the tactus-beat interval that is suggested by the time signature (see Section 5.3.2). In contrast with the original model, the tactus interval does not change during a rhythm. This, in a sense, causes T to be part of the representation of meter. The variables that describe meter, namely U, L, and T, are all defined as *persistent* variables. In Chapter 3, we defined a persistent variable to be a variable that generates multiple congruent states only in one moment and retains its value in subsequent moments. The congruency constraints of U, L, and T, correspondingly, show that they generate their domain in the first moment and a singleton set containing their previous value in subsequent moments.

The variables *Uph* and *Tph* generate a value only in the first moment and are disabled in subsequent moments. Given T, they encode the pickup interval as follows: $pickup = tph + uph \cdot t$. As in Chapter 4, the possible upper-level phases are constrained by the number of tactus beats per bar, U, and the possible tactus phases are constrained by the duration of the tactus interval, *Tph*.

P is a deterministic variable that keeps track of the *phase* (position in the metrical cycle, measured in ρ-units) of the previous onset. Recall that each moment represents a time interval created by an onset with the previous onset. In the first moment, there is no previous onset. Here, the value of P represents the pickup

interval. In subsequent moments, P records, given a previous phase, $\hat{p}$ (beginning with the pickup interval) and an inter-onset interval, i, generated in the current moment, the phase created by moving forward along the metrical cycle by the inter-onset interval: $(\hat{p} + i) \mod (u \cdot t)$.

Temperley's original model and the version described in Chapter 4 predict in each moment whether or not an onset will occur in a small temporal interval represented by that moment. This prediction is described by the observed random variable N. The classical model, by contrast, predicts the duration of an inter-onset interval in each moment. This prediction is described by the observed random variable I. The first moment is, as we mentioned in Section 5.2.2, used to generate a pickup interval. During this moment, I deterministically generates $*$. In subsequent moments, I has a probability distribution over the pre-defined inter-onset interval domain, D_I.

The probability distribution of inter-onset intervals, which is described in the next section, relies on the metrical salience of each ρ-unit occurring between the last and current (second of the two onsets of an inter-onset interval). These metrical-salience values are derived from the phase of the last onset and the meter. This is why the dependencies of I are $\{\hat{P}, U, L, T\}$. Correspondingly, the deterministic variable *Bs* that encodes the metrical salience of a pip in Chapter 4 is redundant in the classical model since pips have been replaced by inter-onset intervals. Since U, L, and T contain all information required to derive the metrical salience of any ρ-unit, the variables that describe the positions of lower-level beats in Chapter 4 (*Db*, *Tb1*, and *Tb2*) also also redundant in the classical model.

5.3.2 Model parameters

The variables U, L, and T have categorical distributions. Maximum-likelihood estimates of the probability of each of their possible values are given by the relative frequency with which those values are observed in an empirical rhythm sample. In order to obtain these estimates, however, the time signatures and pickup intervals provided for each rhythm in an empirical rhythm sample need to be translated into values of U, L, T, and *Uph*. Below, we describe how this could be done.

Time signatures in some cases under-specify the metrical hierarchy. A 3/4 time signature, for example, prescribes that a bar consists of three tactus beats but does not specify whether these tactus beats are subdivided by two or by three. Following conventional use in Western classical music, we will assume such cases to indicate duple subdivision of tactus beats.

The values of U and L can be derived from the numerator of the time signature (represented by a pair $(num, denom)$), following conventional interpretations of time signatures (see London, 2012, p. 17). To determine the value of U, the

Table 5.3: A possible mapping between time signatures and the parameters of the classical model that represent meter. The set of time signatures shown here are the time signatures that occur in empirical rhythm samples used in Chapter 7, where simulation results involving the classical model are reported.

$(num, denom)$	U	L	T
$(2, 4)$	2	4	2
$(3, 4)$	3	4	2
$(4, 4)$	2	4	2
$(6, 8)$	2	6	3

following rules could be used (these cover most time signatures used commonly in Western music): if $num \in \{2, 4, 6, 12\}$, then $U = 2$ and if $num \in \{3, 9\}$, then $U = 3$. A possible set of rules for determining L is: if $num \in \{2, 3, 4\}$ then $L = 2$ and if $num \in \{6, 9, 12\}$, then $L = 3$. The duration of the tactus interval, T, in ρ-units, corresponds to the duration of a bar divided by the value of L: $T = \rho \cdot num/(denom \cdot l)$.

The empirical samples used in simulations in Chapter 7 contain a small set of time signatures that are covered by the rules. For illustration purposes, Table 5.3 shows the values of U, L, and T corresponding to time signatures that occur in rhythm samples in Chapter 7. Since the classical model represents only three levels of metrical hierarchy, and since the tactus level must the second of these, the bar level of 4/4 time cannot be accommodated. Therefore, rhythms in 2/4 and 4/4 time receive the same interpretation.

Once a set of meters that the model should generate has been determined, the domain of T can be set such that it supports these meters. For example, to generate meters corresponding to the time signatures in Table 5.3, assuming $\rho = 16$, the domain of T, D_T must be $\{4, 6\}$ (corresponding to a quarter-note and a dotted quarter-note tactus interval).

The variable *Uph* has a conditional probability distribution that depends on U: $\Pr(uph \mid u)$. The upper-level phase is derived from the pickup interval as follows:

$$Uph = \textsc{floor}(pickup/t),$$

where $\textsc{floor}(r)$ returns the highest integer $\leq r$.

To obtain the maximum-likelihood estimates of the parameters of U, L, T, and *Uph*, each time signature and pickup interval observed in an empirical rhythm sample is transformed into an observation of these variables using the mechanisms described above. Since time-signature changes do not occur in the rhythm samples

used in this thesis, each rhythm corresponds to one observation of a time signature and pickup interval. The maximum-likelihood estimate of the probability that U, L, T, or Uph assumes a certain value corresponds to the number of times that that value is observed in the empirical rhythm sample, divided by the total number of rhythms in the sample.

The probability distribution of Tph has one parameter, p, which represents the probability that a rhythm begins on a tactus beat. This parameter is estimated by the relative frequency with which rhythms begin on a tactus beat in an empirical rhythm sample. That is, the number of times that $pickup \mod t = 0$, divided by the total number of rhythms in the sample. Given p, the distribution of Tph is

$$\Pr(tph \mid \hat{t}, \hat{tph}, t) = \begin{cases} p & \text{if } tph = 0 \wedge \hat{tph} = *, \\ p/(t-1) & \text{if } tph \neq 0 \wedge \hat{tph} = *, \\ 1 & \text{otherwise.} \end{cases}$$

The probability distribution of the observed variable I assigns a probability to each $i \in D_I$. This inter-onset interval distribution is based on four *onset-probability* parameters that are estimated from empirical rhythms: $\{\theta_s^I\}_{s=0}^3$. These parameters represent the probability with which onsets occur at different levels of metrical salience, s, in the empirical rhythm sample. They are the same parameters as those of the distribution of the N variable, $\{\theta_s^N\}_{s=0}^3$, described in Chapter 4. The lowest level of metrical salience, $s = 0$, represents positions not corresponding to a beat in the metrical hierarchy, and levels 1, 2, and 3 represent respectively a lower-, tactus-, and upper-level beat.

To estimate the onset-probability parameters, $\{\theta_s^I\}_{s=0}^3$, each rhythm in a sample is represented as a metrical grid: a grid of subsequent temporal intervals with a symbolic duration of one ρ-unit. Each grid position represents an *silence/onset event*: it either does or does not contain an onset. For example, if we encode onsets as 1 and no onset (silence) as 0, the inter-onset interval pattern $(2, 1, 2)$ corresponds to the following grid: $(1, 0, 1, 1, 0, 1)$. The metrical salience of each grid position is derived from the time signature and pickup interval. The onset-probability parameters for each level of metrical salience, s, are estimated by counting the number of times that an onset occurs at a grid position with metrical salience s and dividing it by the total number of grid points with metrical salience s.

The inter-onset interval distribution is derived from the onset-probability parameters by viewing an inter-onset interval as a set of silence/onset events on a grid. Each inter-onset interval, i, implies that an onset occurs at a grid position exactly i units away from that of the previous onset. That is, it implies the following silence/onset events: first, $i - 1$ silence events occur at the grid positions between the previous onset and the current onset, then a silence/onset event

occurs at the grid position at the end of the inter-onset interval, i. Given the meter (u, l, and t) and the phase of the previous onset ($\hat{p}$), the metrical salience of each of these grid positions can be determined. Given the metrical salience of each grid position, the probability of each silence/onset event is given by the onset-probability parameters, $\{\theta^N_s\}^3_{s=0}$. Below, we describe two functions that determine respectively the metrical salience of each grid position and the joint probability of silence/onset events implied by an inter-onset interval.

The metrical salience of a grid position that is $i \in \mathbb{N}$ units away from the position of the previous onset, given a meter (u, l, and t) and the phase of the previous onset ($\hat{p}$) is determined by the function BSAL: $D_U \times D_L \times D_T \times D_P \times \mathbb{N} \to \{0, 1, 2, 3\}$. This function is defined as follows:

$$\text{BSAL}(u, l, t, \hat{p}, i) = \begin{cases} 3 & \text{if } (\hat{p} + i) \mod (u \cdot t) = 0, \\ 2 & \text{if } (\hat{p} + i) \mod t = 0 \wedge (\hat{p} + i) \neq 0, \\ 1 & \text{if } (\hat{p} + i) \mod (t/l) = 0 \wedge (\hat{p} + i) \mod t \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

The joint probability of the silence/onset events implied by an inter-onset interval is calculated by the function f^i: $D_U \times D_L \times D_T \times D_P \times D_I \to [0, 1)$. This function is defined as follows:

$$f^i(u, l, t, \hat{p}, i) = \theta^I_{\text{BSAL}(u,l,t,\hat{p},i)} \prod_{i'=1}^{i-1} (1 - \theta^I_{\text{BSAL}(u,l,t,\hat{p},i')}).$$

This function represents the probability that Temperley's original model assigns to an inter-onset interval if the duration of pips would correspond to ρ-units and if the model were constrained to not consider tempo and expressive timing aspects of rhythms. Note that the probability of onsets decays exponentially as the inter-onset interval duration increases.

The function f^i represents a probability distribution over inter-onset intervals of any duration. However, we would like to obtain a probability distribution over the finite inter-onset interval domain, D_I. This is achieved by normalizing the probabilities returned by f^i as follows:

$$\Pr(i \mid \hat{i}, \hat{p}, u, l, t) = \frac{f^i(i, \hat{p}, u, t, l)}{\sum_{i' \in D_I} f^i(i, \hat{p}, u, t, l)}.$$

The resulting distribution depends on both the meter and the phase of the previous onset. Figure 5.1 illustrates the inter-onset interval distributions for two different

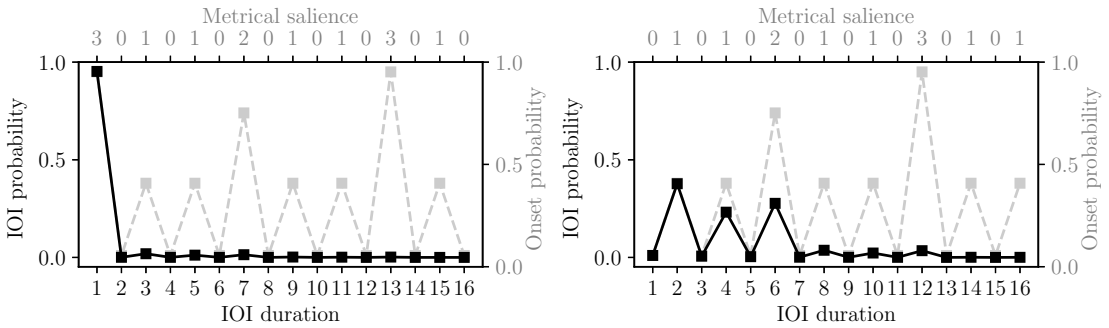


Figure 5.1: Two examples of inter-onset interval (IOI) distributions (the *black* squares) of the classical model generated at different points in the metrical cycle. The figure on the *left* shows the IOI distribution when the phase of the previous onset corresponds to the last position in the measure (i.e., $\Pr(i \mid \hat{P} = 11, U = 2, L = 3, T = 6)$). The figure on the *right* shows the IOI distribution when the phase of the previous onset is the downbeat (i.e., $\Pr(i \mid \hat{P} = 0, U = 2, L = 3, T = 6)$). The onset-probability estimates corresponding to the metrical salience of the grid positions to which each IOI leads are visualized by the *gray* squares. The onset-probability parameter estimates used for this distributions are given in the main text.

values of $\hat{p}$. These distributions are derived from onset-probability parameter estimates provided by Temperley (2007, p. 35) for $\Pr(n \mid bs)$: $\theta^I_0 = .01, \theta^I_1 = .38, \theta^I_2 = .74$, and $\theta^I_3 = .95$. Both distributions are based on a duple compound meter ($U = 2$ and $L = 3$) with a period of 12 ρ-units. The figure on the right shows the distribution of I when the previous phase is 11, the moment just before the downbeat. In this distribution, most probability mass is absorbed by the downbeat position that immediately follows phase 11. When the previous onset occurred on the downbeat, probability mass is spread out more evenly over the next two lower-level beats (inter-onset intervals 2 and 4), and some probability mass is also assigned to the next tactus beat (inter-onset interval 6).

Note, finally, that the only parameters of the distribution in Figure 5.1 that are estimated from empirical data are the four onset-probability parameters, $\{\theta^I_s\}^3_{s=0}$. This means that the amount of information that the classical model can learn about the relation between rhythm and meter from empirical data is limited.

5.4 The enculturation model

The enculturation model is based closely on a model presented by Van der Weij et al. (2017 [Chapter 6]). Compared to the classical model, it can learn significantly more complex associations between rhythm and meter. The model infers meter

from rhythms based on rhythmic patterns that it learned from empirical data. In contrast to the classical model, the enculturation model does not require that onsets in these patterns align with levels of metrical salience in a particular way. In order for the model to infer meter, rhythmic patterns only need to align with the metrical cycle in an approximately consistent within rhythms in a particular style or idiom (as represented by an empirical rhythm sample).

We use the term *metrical category* instead of *meter* to refer to the structure that the enculturation model infers from rhythms. This is to emphasize the following point: a metrical category need not constrain the rhythmic patterns by a recurring pattern of strong and weak beats. Metrical categories may correspond to any continuous cyclical structure that might underlie a rhythm (e.g., *clave*, *tala*, *timeline*, or *usul*). In this thesis, however, time signatures serve as metrical categories.

The key mechanism by which the model learns to associate metrical categories with rhythmic patterns is by representing a rhythm as a sequence of metrical fingerprints. These fingerprints encode how inter-onset intervals in a rhythm align with the metrical cycle. We call these fingerprints *downbeat distances*. A downbeat distance represents an inter-onset interval as the phase of the last onset plus the inter-onset interval (this is described in more detail in Section 5.4.2). Patterns of downbeat distances are learned from an empirical rhythm sample by a sequence model. Each metrical category is associated with its own sequence model describing the sequential statistics of the metrical fingerprints of rhythms associated with that category.

Before we define the congruency constraints of the enculturation model in Section 5.4.2, we introduce notation and terminology that we use to describe variables that represent sequences in Section 5.4.1. We then contextualize some aspects of the model's design as mechanisms that avoid biases that depend on the duration of the metrical cycle in Section 5.4.3. The conditional probability distributions and parameter-estimation methods are described in Section 5.4.4. Finally, Section 5.4.5 compares our reformulation to the original formulation and explains the differences and similarities.

5.4.1 Sequences and accumulator variables

Metrical fingerprints are recorded by a variable whose states correspond to sequences of downbeat distances. A sequence of n symbols, $\boldsymbol{s}$, is an n-tuple $\boldsymbol{s} = (s_0, s_1, \cdots, s_n)$. We consistently use bold-face symbols to refer to variables and states of variables that represent sequences. The elements s_i of $\boldsymbol{s}$, where $0 \leq i < n$, are drawn from a set A called the *alphabet*. The set of all sequences that can be formed by elements from A is denoted by A^*.

Recall that in Chapter 3, we defined $\boldsymbol{xy}$ to denote the concatenation of the tuples $\boldsymbol{x}$ and $\boldsymbol{y}$. As such, $\boldsymbol{x}(s)$ denotes the sequence obtained by appending the symbol s to the sequence $\boldsymbol{x}$. Here, we additionally define an operation FINAL that returns the final symbol of a sequence such that $\text{FINAL}((\cdots, s_n)) = s_n$.

Dynamic Bayesian networks are first-order Markov models, but higher-order Markov models can be accommodated by using sequences as the values of its random variables. We define an *accumulator variable* to be a variable that incrementally generates sequences. In the first moment, an accumulator variable generates either an empty sequence or a sequences consisting of one element from A. In subsequent moments, an accumulator variable generates states obtained by appending an element from the alphabet, A, to its previous state (a sequence of elements from A). The congruency constraints of an accumulator variable is of the following form:

$$\kappa_{\boldsymbol{X}}((\hat{\boldsymbol{x}})) = \begin{cases} \{(s) \mid s \in A\} & \text{if } \hat{\boldsymbol{x}} = *, \\ \{\hat{\boldsymbol{x}}(s) \mid s \in A\}\} & \text{otherwise.} \end{cases}$$

The probability of a state of $\boldsymbol{X}$ given its previous state, $\Pr(\boldsymbol{x} \mid \hat{\boldsymbol{x}})$ corresponds to the probability that a symbol from $s \in A$ follows the *context* $\hat{\boldsymbol{x}}$. The enculturation model uses an accumulator variable to represent rhythmic patterns. Accumulator variables may be modeled by a sequence model that estimates the probability that a symbol $s \in A$ follows a given context.

5.4.2 Deterministic constraints

Table 5.4 defines the deterministic constraints of the enculturation model. We describe these constraints in more detail below.

The persistent variable M represents a metrical category. Each metrical category, m, is associated with a *metrical cycle duration*, denoted by T_m and measured in ρ-units. The metrical category domain, D_M, and the period associated with each $m \in D_M$ are given a priori. If metrical categories correspond to time signatures, which are represented by pairs $(num, denom)$, the duration of their metrical cycle, T_m, is given by $\rho \cdot num/denom$, which corresponds to the duration of a bar.

Like the classical model, the enculturation model has a variable P that represents the phase of the onset to which the current inter-onset interval leads. In all moments but the first, P is deterministic and its value is based on a value of $\boldsymbol{D}$ (a downbeat distance, see below). In the first moment, P generates the possible pickup intervals given a metrical category. The possible pickup intervals represent the possible positions in the metrical cycle at which the first onset may occur. These are constrained by the period of the metrical category, T_m. The pickup

Table 5.4: Model-definition table of the enculturation model.

X_i	PA_i	D_{X_i}	$\kappa_{X_i}(pa_{X_i})$
M	$\{\hat{M}\}$	D_M	$\begin{cases} D_M & \text{if } \hat{m} = *, \\ \{\hat{m}\} & \text{otherwise.} \end{cases}$
P	$\{\hat{P}, M, \boldsymbol{D}\}$	$\bigcup_{m \in D_M} \{n \in \mathbb{N} \mid 0 \leq n < T_m\}$	$\begin{cases} \{n \in \mathbb{N} \mid 0 \leq n < T_m\} & \text{if } \hat{p} = *, \\ \{\textsc{final}(\boldsymbol{d}) \mod T_m\} & \text{otherwise.} \end{cases}$
$\boldsymbol{D}$	$\{\hat{\boldsymbol{D}}, \hat{P}, M\}$	$\{p + i \mid p \in D_P, i \in D_I\}^*$	$\begin{cases} \{()\} & \text{if } \hat{\boldsymbol{d}} = *, \\ \{\hat{\boldsymbol{d}}(\hat{p} + i) \mid i \in D_I\} & \text{otherwise.} \end{cases}$
I	$\{\hat{P}, \boldsymbol{D}\}$	D_I	$\begin{cases} \{*\} & \text{if } \boldsymbol{d} = (), \\ \{\textsc{final}(\boldsymbol{d}) - \hat{p}\} & \text{otherwise.} \end{cases}$

interval corresponds to what Van der Weij et al. (2017 [Chapter 6]) call the "phase" of a metrical interpretation.

$\boldsymbol{D}$ is an accumulator variable whose states represent sequences of metrical fingerprints called downbeat distances. A *downbeat distance* is defined as the phase of the last onset plus the inter-onset interval: $\hat{p} + i$. As such, a downbeat distance indirectly represents the position of an onset in the metrical cycle. Given the last downbeat distance in a state of $\boldsymbol{D}$, the corresponding position in the metrical cycle is the downbeat distance modulo the duration of the metrical cycle.

Consider, for example, the inter-onset interval pattern $(1, 2, 1, 1, 1, 1)$. If the first value of P (the pickup interval) is 5, and $T_m = 6$, the following sequence of downbeat distances would result: $(6, 2, 3, 4, 5, 6)$. If the pickup interval is 0, the downbeat distance pattern $(1, 3, 4, 5, 6, 1)$ emerges. If the pickup interval is 0 and $T_m = 3$, the pattern is $(1, 3, 1, 2, 3, 1)$. Note that downbeat distances correspond generally to phases, but values greater than the duration of the metrical cycle indicate that an inter-onset interval bridges subsequent metrical cycles.

Given a previous phase, $\hat{p}$, and a metrical category, m, each inter-onset interval, $i \in D_I$, that could occur in the current moment will result in a different downbeat distance (namely $\hat{p} + i$). The a priori congruent states of $\boldsymbol{D}$, given $\hat{p}$ and m, and a context, $\hat{\boldsymbol{d}}$ (the sequence of preceding downbeat distances) are obtained by appending each of these possible downbeat distances to the context $\hat{\boldsymbol{d}}$, as can be seen in Table 5.4. While the variable $\boldsymbol{D}$ is not observed directly, its congruent states are constrained by observing I (as explained below). After observing the inter-onset interval, only one downbeat distance state per meter and per previous phase remains congruent.

In the first moment, the value of P represents the pickup interval. Since $\boldsymbol{D}$ generates a set of congruent states for each value of meter, m, and previous

phase, $\hat{ph}$, and since only one downbeat distance per meter and phase remains congruent a posteriori, the a posteriori congruent states of $\boldsymbol{D}$ reflect the possible metrical interpretations of the rhythm. That is, given each meter and each pickup interval, there is exactly one downbeat distance that is consistent with each possible inter-onset interval.

The downbeat-distance variable $\boldsymbol{D}$ is constrained by observations of I, which may be considered a *representation variable* as it defines how inter-onset intervals are represented, given a meter and previous phase, by downbeat distances. Given a downbeat distance, $\boldsymbol{d}$, and a previous phase, $\hat{p}$, I has one congruent state: the inter-onset interval that corresponds to the difference between the previous phase and the downbeat distance (which, recall, is defined as a previous phase plus an inter-onset interval). [2] Therefore, observing the inter-onset interval causes only downbeat distances that are consistent with the observed inter-onset interval to remain congruent.

To summarize: In each moment, the model generates a set of possible downbeat distances. Given a metrical category and a phase, there is a one-to-one correspondence between downbeat distances and inter-onset intervals. The variable $\boldsymbol{D}$ accumulates downbeat distances into sequences, each of which represents an alternative reading of a rhythm that depends on the pickup interval (the first phase that was generated) and the cycle duration of a metrical category. These downbeat-distance sequences are modeled by a sequence model as described Section 5.4.4.

5.4.3 Period-dependent biases

The reason that we chose to represent the metrical status of onsets as downbeat distances, rather than phases, is that the latter representation results in unwanted inferential biases. Since the model predicts sequences of metrical fingerprints, rather than sequences of inter-onset intervals, it must make use of a mapping between inter-onset intervals and metrical fingerprints. Downbeat distances, crucially, map one-to-one to inter-onset intervals (given a metrical interpretation). Phases, on the other hand, may map to multiple inter-onset intervals. The number of inter-onset intervals that are consistent with a given phase depends on the inter-onset interval domain *and* on the duration of the metrical cycle. In other words, the duration of the metrical cycle affects the *granularity* of the phase representation. This period-dependent granularity introduces a bias towards

[2] This mechanism is similar to a mechanism in multiple viewpoint systems that Conklin (1990) calls the *completion* of a viewpoint and which is described by Pearce (2005, p. 114) as an *inverse viewpoint function*. The difference between this mechanism and the model described here is that the downbeat-distance representation depends not only on observed events but also on the value of a hidden (latent) variable, namely meter.

metrical interpretations with short cycles: the phases of these short cycles are more likely to correspond to an observed inter-onset interval. The downbeat-distance representation avoids this bias by ensuring that there is a one-to-one correspondence between downbeat distances and inter-onset intervals.

Potential for another period-dependent bias lies in the distribution of pickup intervals, The number of a priori congruent states of this distribution is equal to the period of the metrical cycle. Therefore, its entropy (inherent uncertainty) is on average higher for metrical categories with long metrical cycles, and pickup intervals of such categories tend to have a lower probability. This would result in a preference for metrical categories with short metrical cycles. In the next section, we describe a correction factor encoded in the distribution of metrical categories that neutralizes this bias.

Note that in the classical model, too, the duration of the tactus interval constrains the number of a priori congruent states of the tactus phase, *Tph*. This causes the model to have a small bias toward meters with short tactus intervals. The effect of this bias is small since the tactus interval is generated only once in the first moment.

5.4.4 Model parameters

In the first moment, a priori congruent states of P correspond to the possible pickup intervals given a meter. In this moment, P, has a uniform distribution, following the definition of Van der Weij et al. (2017 [Chapter 6]). In subsequent moments, the distribution P is deterministic. As such, its probability distribution is given by

$$\Pr(p \mid \hat{p}, \boldsymbol{d}) = \begin{cases} 1/|\kappa_P(\hat{p}, \boldsymbol{d})| & \text{if } p \in \kappa_P(\hat{p}, \boldsymbol{d}), \\ 0 & \text{otherwise.} \end{cases}$$

P has more than one a priori congruent state given each value of its dependencies only in the first moment. Therefore, the above definition ensures that $\Pr(p \mid \hat{p}, \boldsymbol{d})$ is a uniform distribution in the first moment and a deterministic distribution in subsequent moments.

The prior distribution of M, $\Pr(m \mid *)$, represents the a priori probabilities of metrical categories. These probabilities are based on the relative frequency with which a metrical category is observed in a training sample. However, the distribution of M is defined to incorporate a correction factor that neutralizes a bias towards meters with short cycles created by the prior distribution of P (the pickup interval; see Section 5.4.3). Let $\theta_m^M \in [0, 1)$ be the relative frequency of

a metrical category, m, in an empirical sample. The probability distribution of metrical categories is then given by

$$\Pr(m \mid \hat{m}) = \begin{cases} T_m \theta^M_m \,/\, \sum_{m' \in \kappa_M(\hat{m})} T_{m'} \theta^M_{m'} & \text{if } \hat{m} = *, \\ 1/|\kappa_M(\hat{m})| & \text{if } m \in \kappa_M(\hat{m}), \\ 0 & \text{otherwise.} \end{cases}$$

In the above piece-wise definition, the second and third cases apply to non-initial moments. They specify a uniform distribution over congruent states and assign zero probability to incongruent states. Since a priori congruent states of M are a singleton in non-initial moments, the probability $\Pr(m \mid \hat{m})$ is one if m is congruent—that is, if $m = \hat{m}$ (see the congruency constraint in Table 5.4)—and zero otherwise. The correction factor applied in the above equation multiplies the estimated prior probability of a metrical category, θ^M_m, by its period, T_m, to cancel out the effect that cycle duration has on the prior probability of pickup intervals: pickup-interval probabilities of meters with long cycle durations are smaller than those of meters with short cycle durations since the number of possible pickup intervals depends on T_m.

Note that Van der Weij et al. define a joint prior distribution over meter and phase (which in our definition is the pickup interval). It can easily be seen that their definition is identical to the joint distribution of the prior (initial-moment) distributions of meter and pickup interval: $\Pr(m \mid *) \Pr(p \mid *, ())$.

Downbeat-distance sequences are modeled by a sequence model. There is a different sequence model for each metrical category. The parameters of the sequence models are obtained by representing each rhythm in an empirical rhythm sample as a sequence of downbeat distances. These sequences can be derived from inter-onset-interval sequences in empirical training data based on the pickup interval and metrical category of the rhythm as provided in the sample. The resulting downbeat-distance sequences are are grouped by their metrical category and the parameters of a sequence model are estimated from each group of downbeat-distance sequences. How this parameter estimation process works depends on the details of the sequence model.

For a given state of $\boldsymbol{D}$, a sequence model estimates the probability with which a generated downbeat distance, $\textsc{final}(\boldsymbol{d})$ (the last downbeat distance in the sequence), follows a context, $\hat{\boldsymbol{d}}$ (the downbeat distances preceding the last). We denote the estimate of this probability of a sequence model associated with a metrical category, m, as $q^m(\textsc{final}(\boldsymbol{d}))$. The distribution of $\boldsymbol{D}$ is given by

$$\Pr(\boldsymbol{d} \mid \hat{\boldsymbol{d}}, \hat{p}, m) = \frac{q^m(\textsc{final}(\boldsymbol{d}), \hat{\boldsymbol{d}})}{\sum_{\boldsymbol{d}' \in \kappa_{\boldsymbol{D}}(\hat{\boldsymbol{d}}, \hat{p}, m)} q^m(\textsc{final}(\boldsymbol{d}'), \hat{\boldsymbol{d}})}$$

Note that these probabilities are normalized by the total probability that the sequence model, q^m, assigns to all possible downbeat distances that can occur given a meter and a downbeat-distance context, $\hat{\boldsymbol{d}}$. This normalization is necessary for sequence models that generate probability distributions over the entire sequence alphabet. The a priori congruent states of $\boldsymbol{D}$, however, are constrained by the preceding downbeat distance, $\hat{\boldsymbol{d}}$ and may not correspond to the alphabet of a sequence model of $\boldsymbol{D}$.

In Chapters 6 and 7, we use a prediction algorithm called prediction by partial match (PPM) (Cleary & Witten, 1984; Cleary & Teahan, 1997) as a sequence model for downbeat distances. This algorithm implements a variable-order Markov model. It also serves as the sequential prediction mechanisms of the IDyOM modeling framework described by Pearce (2005). Pearce (2005, pp. 79–110) describes several configurations of PPM models. The configurations used in Chapters 6 and 7 use interpolated smoothing, rather than the backoff-smoothing strategy described originally by Cleary and Witten (1984), Cleary and Teahan (1997) and Method C for calculating escape probabilities (Moffat, 1990). Simulations in Chapter 6 use a bounded model where the order of the variable-order Markov models used by the PPM model is constrained to a maximum. Simulations in Chapter 7 use both a bounded model and an unbounded model, in which the maximum order of the variable-order Markov models used by the PPM model is unconstrained.

5.4.5 Connection with original formulation

The differences between the current definition and that of Van der Weij et al. (2017 [Chapter 6]) (referred to below as the *original model*) relate primarily to the manner of presentation. The original model is presented as an extension of IDyOM and is stated in multiple-viewpoint systems terminology. In this chapter, we gave a self-contained definition of the model based on congruency constraints.

Instead of using a *metrical viewpoint*, as in the original model, we represent metrically interpreted rhythms by a latent variable, $\boldsymbol{D}$, that represents sequences of downbeat distances and is constrained by observed inter-onset intervals by means of a representation variable, I. This mechanisms is equivalent to that of metrical viewpoints, but the definition presented here shows how the mechanism is represented by a Bayesian network with deterministic constraints.

At the surface level, the representations of metrical fingerprints are different: we used downbeat distance, while the original model represents onsets, given a metrical interpretation, as a pair consisting of the phase of an onset and the number of bar lines crossed since the last onset. At a deeper level, however, the downbeat-distance representation can be mapped one-to-one to the representation of the original model. The two representations are therefore interchangeable

without affecting the behavior of the model.

In one respect, the behavior of the original model is slightly different from that of the current formulation. The original model is based on rhythms represented by sequences of absolute onset times. In this representation, it is possible that the first note has an onset time not equal to zero. Simultaneously, however, the model has a variable, phase, that represents how a meter aligns with a rhythm. The phase of an interpretation is represents "the interval between the first bar and the time point marked by zero in the encoding of the rhythmic pattern" (Van der Weij et al., 2017 [Chapter 6], p. 4). Since both the phase and the absolute onset time of the first note can be used to encode how a meter aligns with a rhythm, one of them is redundant. While the effect on the model's behavior is marginal, the current formulation avoids this redundancy by representing rhythms as sequences of inter-onset intervals.

5.5 Summary

In this chapter, we introduced the notion of a rhythm space and described a specific rhythm space for rhythms represented by symbolic inter-onset intervals. We then described two probabilistic generative rhythm models—the classical model and the enculturation model—that define probability distributions over this rhythm space. The parameters of the models can be estimated from empirical rhythm samples containing rhythms from a rhythm space and in which the metrical interpretation information of each rhythm is provided. This procedure causes a model to approximate an empirical distributions of rhythms represented by the empirical sample.

The enculturation model is an alternative formulation of the model proposed by Van der Weij et al. (2017 [Chapter 6]). The classical model is based on a model proposed by Temperley (2007) with two modifications: First, the classical model does not model tempo and expressive timing aspects and generates symbolic representations of rhythms. Second, we have derived an inter-onset interval distribution from the original model's distribution over whether or not an onset occurs at a certain position in a metrical grid. This ensures that the model defines probability distributions over the same rhythm space as the enculturation model and enables the classical model and the enculturation model to be compared quantitatively, as we will do in Chapter 7.

In the enculturation model, the probability with which onsets occur at different positions in the metrical cycle of a given meter depends both on the metrical category of the rhythm and the preceding pattern of onsets, represented by metrical fingerprints. In the classical model, the probability that an onset occurs at a position with a given level of metrical salience depends only on the level

of metrical salience. The classical model learns the probabilities with which onsets occur at different levels of metrical salience from empirical rhythm samples. The enculturation model, by contrast, learns the sequential statistics of patterns of metrical fingerprints (metrically interpreted onsets) associated with different meters. Compared to the enculturation model, the classical model may be said to rely primarily on music theory, while the enculturation model relies primarily on patterns learned from empirical data.

The purpose of this chapter was to introduce the classical and enculturation model, to provide detailed and formal definitions of their deterministic and representation constraints, and to show how their parameters are derived from empirical rhythm samples. These concepts are applied in Chapter 7, where we compare the behavior of these models using rhythm samples from different musical idioms.

Chapter 6

A probabilistic model of meter perception*

6.1 Introduction

In a variety of settings, perception appears to be tuned to statistical properties of the environment. It has for example been found that certain properties of neuron receptive fields in early visual processing (Olshausen & Field, 1996) and early auditory processing (Smith & Lewicki, 2006) emerge from information theoretically efficient learning algorithms trained respectively on natural images or sounds. Perception, it has been suggested, is actively shaped by statistical properties of the environment, both on an evolutionary time-scale through gradual adaptation, and on an ontogenetic time scale, through brain plasticity (Clark, 2013).

The perception of meter in music appears to be shaped by cultural differences in musical conventions. Exposure to rhythmically different music has been shown to influence perception from an early age (Hannon & Trehub, 2005a, 2005b), but such shaping possibly continues into adulthood (Creel, 2011, 2012). In the current paper, we hypothesize that considering meter perception from the perspective of *predictive coding* (Clark, 2013; Friston, 2005; Rao & Ballard, 1999) can help to understand how meter perception is shaped by one's environment.

Rhythm is an important component of music traditions all over the world (Savage et al., 2015). When listening to rhythms, onsets in the rhythm are perceived

*This chapter was previously published as van der Weij, B., Pearce, M. T., & Honing, H. (2017). A probabilistic model of meter perception: Simulating enculturation. *Frontiers in Psychology*, 8, 1–18.

relative to a periodic and hierarchically organized framework of beats (Honing, 2013). This mental framework, called meter, is induced in the mind of the listener by the rhythm. The relation between rhythm and meter is complex. For a meter to be perceived, not every beat in the meter needs to coincide with onsets in the rhythm. In many cases, listeners can, through conscious effort, alter their metrical interpretation of a rhythm. At the same time, not every meter is equally easy to hear in every rhythm. Meter, once induced, tends to show a certain resistance to change. Therefore, meter perception is a fundamentally incremental process (Longuet-Higgins & Steedman, 1971): the same rhythmic passage can sound different depending on the meter induced by the rhythm preceding the passage (Honing, 2013).

The organizing structure of meter is commonly described as a hierarchy of pulses, yielding a periodic pattern of metrical accents varying in salience at different points in time. Metrical accent, or metrical salience, is commonly treated as a proxy for temporal expectation, or the probability of an event onset at a particular pulse (Palmer & Krumhansl, 1990). By investigating a corpus of Western classical music, Palmer and Krumhansl (1990) found that the distribution of onsets over different positions relative to the meter reflected theoretical descriptions of metrical hierarchy (Lerdahl & Jackendoff, 1983). Using a goodness-of-fit paradigm, Palmer and Krumhansl (1990) found that temporal *expectations* of North-American listeners also reflect metrical hierarchy, although musicians showed evidence of deeper hierarchical differentiation than nonmusicians. Based on these findings, Palmer and Krumhansl (1990) suggested that composers communicate meter to listeners through the distribution of onsets at different metrical positions. Listeners, in turn, acquire their knowledge about meter through the distribution of onsets over metrical positions in the music they are exposed to.

More recent work has addressed the question of whether hierarchical organization of onset distributions is a general property of rhythmic organization or whether it is specific to Western classical music and related styles. Holzapfel (2015), for instance, found that in traditional Turkish makam music, the distribution of onsets is modulated by the specific *usul*—a type of rhythmic mode, corresponding in some ways to meter—underlying a piece. Furthermore, the distribution of onsets within one usul in Turkish makam music does not always exhibit hierarchical organization. London et al. (2017) found that peaks in onset distributions in a corpus of Malian drumming recordings are not periodically spaced. London et al. (2017) conclude that in makam music and Malian drumming, distributions of onsets do reflect metrical structure, but this structure is not always isochronous or strictly hierarchical.

London et al. (2017) point out that their and Holzapfel's (2015) results question a basic assumption made by many computational models, as well as empirical studies, namely that metrical accent is equivalent to the likelihood of an onset. A more

likely alternative is that metrical expectations are derived from extensive exposure to a musical idiom, by which, beyond distributions of onsets and style-specific, stereotypical rhythmic patterns associated with certain meters are learned.

Consistent with this suggestion, an increasing number of empirical studies show that rhythm perception is affected by enculturation (cf. Morrison & Demorest, 2009). For example, Bulgarian or Macedonian adults are better in detecting metrical violations in meters with a non-isochronous *tactus* level—the level of beat that listeners are most likely to tap along with—(e.g., 5/8 or 7/8) than North-American listeners (Hannon & Trehub, 2005a). This effect appears to be specific to complex meters to which the listeners have been exposed (Hannon et al., 2012).

There have also been a number of observations in the ethnomusicological literature suggesting that individuals from different cultures perceive rhythms differently. For example, during field work in the Bolivian Andes, while studying Easter songs from Northern Potosí, Stobart and Cross (2000) realized that while they had assumed many of the tunes where indisputably anacrustic (i.e., a rhythm starting on an off-beat), the local populations appeared to perceive them as beginning on a downbeat. Another example is provided by rhythms from West-African Sub-Saharan musical cultures, which are characterised by a great deal of metrical ambiguity (Locke, 1982). In particular, many of these rhythms can be interpreted as having a binary or ternary pulse. While individuals from West-African cultures appear to perceive both pulses with equivalent ease, it can take great effort for Western listeners to hear the ternary pulse in some of these rhythms.

The idea that perception, in general, is shaped by statistical properties of the environment is not new (e.g. Barlow, 1961). However, it recently has been developed into a framework which has been argued to bear the promise of providing an overarching theory of perception (Clark, 2013). Under the name of predictive coding (Rao & Ballard, 1999), this framework firmly grounds perception in prediction, based largely on previous sensory experience. In fact, the theory proposes that the brain's primary occupation is to explain sensory input using hierarchical generative models gleaned from previous experience (Clark, 2013). Such models are realized in a hierarchical organization of layers. The lowest layer in the hierarchy represents sensations received directly from the senses. Through feed-forward connections, information travels upward in the hierarchy. Meanwhile, layers higher up in the hierarchy attempt to predict information, propagated by layers below. These predictions are cast to lower layers through feedback connections. Successful prediction cancels out the upward propagation of information. As a result, only *prediction error*, information that higher layers failed to predict, propagates upwards in the hierarchy. Based on prediction error, layers gradually adapt their processing characteristics in a way that minimizes prediction error with respect to layers lower in the hierarchy. By this process of

adaptation, the hierarchy of layers is gradually shaped into a *generative* model of sensations, where layers higher up in the hierarchy track causes in the external world that underlie the received sensations (Friston, 2005). From an information-theoretic point of view, the resulting coding scheme is highly efficient: the more accurate the top-down predictions, the less bottom-up information is left to be processed.

We propose a predictive coding account of meter perception that involves statistical learning of musical rhythms and generation of probabilistic expectations for event timings. Meters are modeled as distinct causes underlying the musical surface. Inferring the underlying meter from rhythm allows the rhythm to be related to rhythms previously heard in that meter, which may help prediction performance. Enculturation is modeled by estimating the parameters of the generative model on a corpus of quantized rhythms annotated with meter. Since the model learns the statistical properties of rhythms through exposure and performs metrical inference based on these, it has the potential to simulate enculturation effects in meter perception.

The paper is organized in six sections. In the remaining part of the current section, Section 6.1.1 develops an account of meter perception based on predictive coding, while Section 6.1.2 discusses relevant work in computational modeling of music perception. Section 6.2 presents the probabilistic model of meter perception in detail, concluding with a set of behaviors we expect the model to exhibit. Section 6.3 presents the methods used in a series of simulations designed to test these behaviors, while Section 6.4 presents the results of the simulations. Section 6.5 discusses the results in the context of the existing literature and includes implications for future research.

6.1.1 Meter perception as predictive coding

The dynamic interaction of top-down and bottom-up processing postulated by predictive coding is reminiscent of dynamic interaction of bottom-up meter-induction and top-down influence exerted by the induced meter, as pointed out by (Vuust & Witek, 2014).

The hypothesis we explore in this paper is that predictive coding can explain how meter perception is influenced by enculturation. To explore the consequences of this idea, we present a probabilistic model of meter perception, based on an empirical Bayes scheme. Empirical Bayes schemes describe how generative systems, such as the generative models posited by predictive coding, are updated by experience (Friston, 2005). We model meters as virtual causes underlying the rhythmic surface: a meter imposes constraints the likelihood of rhythms. A listener commanding an appropriate generative model reflecting this relationship (i.e., how

rhythms are generated from meters), can, when presented only with a rhythmic surface, infer the underlying meter. This process of inferring underlying causes (meters) of experienced sensations (rhythms) involves inverting the generative model of those sensations (which are the end-product of the generative process). We hypothesize that interpreting the rhythm in the context of an inferred meter will reduce the discrepancy between predicted and experienced sensations. In other words, inferring meter makes the rhythm more predictable.

The generative model includes prior expectations, obtained from previous experience, about which metrical categories are likely to occur in general. For example, meters with non-isochronous pulses ("complex" meters) are relatively uncommon in Western-European music, but much more common in music from the Balkans and Eastern Mediterranean region. Listeners from these regions may be more likely to interpret a rhythm in a meter with non-isochronous pulses than listeners from Western Europe. These kind of prior biases might underlie the findings of Hannon and Trehub (2005a) mentioned in the previous section.

Metrical categories favored by prior biases entail expectations regarding the surface structure of rhythms. As bottom-up evidence from the rhythm begins to flow in, these (top-down) expectations are either confirmed or violated. Prediction error results from a violation of the top-down expectations by the incoming evidence. To reduce prediction error, the listener revises their metrical interpretation of the rhythm, which in turn alters the flow of top-down predictions. A predictive coding perspective of meter perception thus posits a dynamic interplay between bottom-up evidence and top-down expectations.

Crucially, both prior biases towards certain meters and the dependencies between meter and the rhythmic surface—which rhythms can be generated by a certain meter—are the result of previous exposure. The generative model in the mind of the listener underlying these representations is carved out by previous experience in predictive processing of rhythmic signals. Since the statistical properties of rhythms vary between styles (e.g., Holzapfel, 2015; London et al., 2017), the processing biases of listeners with significant differences in their exposure to musical styles are likely to vary as well.

6.1.2 Related work

Our approach in some respects resembles other recent probabilistic models, in particular a generative model presented by Temperley (2007). Temperley (2007, pp. 23–48) models meter perception as probabilistic inference on a generative model whose parameters are estimated using a training corpus. Meter is represented as a multileveled hierarchical framework, which the model generates level by level. The probability of onsets depends only on the metrical status of the corresponding onset

time. Temperley (2009) generalizes this model to polyphonic musical structure, and introduces a metrical model that conditions onset probability on whether onsets occur on surrounding metrically stronger beats. This approach introduces some sensitivity to rhythmic context into the model. In later work, Temperley (2010) evaluates this model, the *hierarchical position model*, and compares its performance to other metrical models with varying degrees of complexity. One model, called the first-order metrical position model, was found to perform slightly better than the hierarchical position model, but this increase in performance comes at the cost of a higher number of parameters. Temperley concludes that the hierarchical position model provides the best trade-off between model-complexity and performance.

In a different approach, Holzapfel (2015) employs Bayesian model selection to investigate the relation between usul (a type of rhythmic mode, similar in some ways to meter) and rhythm in Turkish makam music. The representation of metrical structure does not assume hierarchically organization, allowing for arbitrary onset distributions to be learned. Like the models compared by Temperley (2010), this model is not presented explicitly as a meter-finding model, but is used to investigate the statistical properties of a corpus of rhythms.

The approach presented here diverges from these models in that it employs a general purpose probabilistic model of *sequential* temporal expectation based on statistical learning (Pearce, 2005) combined with an integrated process of metrical inference such that expectations are generated given an inferred meter. The sequential model is a variable-order metrical position model. Taking into account preceding context widens the range of statistical properties of rhythmic organization that can be learned by the model. In particular, the model is capable of representing not only the frequency of onsets at various metrical positions, but also the probability of onsets at metrical positions conditioned on the preceding rhythmic sequence. The vastly increased number of parameters of this model introduces a risk of *over-fitting*; models with many parameters may start to fit to noise in their training data, which harms generalization performance. However, we employ sophisticated smoothing techniques that avoid over-fitting (Pearce & Wiggins, 2004). Furthermore, we to some extent safe-guard against over-fitting by evaluating our model using cross-validation.

6.2 The probabilistic model

In this section and the sections that follow, we use the words metrical category and metrical interpretation in a specific sense. *Metrical categories*, denoted by m, represent different metrical frameworks in which rhythms can be interpreted. Metrical categories correspond directly to time signatures taken from scores. Each

metrical category has an associated *period*, denoted by T_m. The period is encoded as a discrete number representing the duration of one bar of m in basic quantized units of time (see Section 6.2.1). The *phase* parameter, ψ, encodes how a metrical category aligns with the rhythmic surface. More precisely, ψ encodes the time-interval between the downbeat of the first bar and the time point marked by zero in the encoding of the rhythmic pattern. Together, a metrical category and phase form a *metrical interpretation*.

The approach described below deals not with real audio signals. Instead, the musical surface is represented as a sequence of events. Each event corresponds to a note, as it might be found in a musical score. The nth event in a sequence is denoted by e_n. A sequence of events, starting at event n and ending at event m is denoted by $\mathbf{e}_n^m$. Section 6.2.1 provides more details the representation of rhythmic patterns.

Predictive coding postulates internal generative models reflecting the causal structure of the external world. In analogy to this, we model meter perception as the inversion of a generative model of rhythms. Enculturation through exposure to rhythms is modeled by deriving the parameters of the generative model from a corpus of rhythms annotated with metrical interpretation. During listening, the metrical category underlying a given rhythm is generally not known to the listener. Instead, it has to be inferred from rhythmic surface, which is assumed to result from the generative model. The likelihood of a metrical interpretation given an observed rhythm (i.e., a sequence of events) can be inferred from the generative model through the application of Bayes' formula, as shown in Equation 6.1.

$$\overbrace{p(m, \psi \mid \mathbf{e}_0^n)}^{\text{posterior}} = \frac{\overbrace{p(\mathbf{e}_0^n \mid m, \psi)}^{\text{likelihood}} \overbrace{p(m, \psi)}^{\text{prior}}}{\underbrace{p(\mathbf{e}_0^n)}_{\text{evidence}}}. \tag{6.1}$$

Two factors play a role in calculating the likelihood of a metrical category: The *a priori* likelihood of the metrical category itself, operationalized here as the metrical category's conventionality. In Equation 6.1, this distribution is labeled *prior*. The other factor is the likelihood of the rhythmic pattern given a certain metrical structure. In Equation 6.1, this function is labeled *likelihood*. The distribution over metrical interpretations inferred from the observed events is called the *posterior* distribution. The factor labeled *evidence* in Equation 6.1 is a constant with respect to metrical interpretation. It ensures that the distribution sums to unity.

The proposed generative model is illustrated in Figure 6.1. To generate a rhythm, a metrical category is first generated from a distribution, $p(m)$, reflecting the prior likelihood of metrical categories. Next, a phase is sampled from a uniform distribution over a range of discrete phases allowed in m. From a model associated

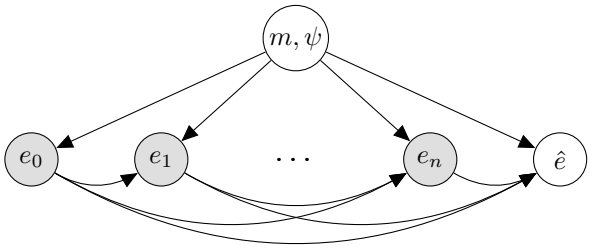


Figure 6.1: Conditional dependency relations assumed by the model between its probabilistic variables visualized as a graphical model Bishop (2006). Shaded nodes represent observed variables, unshaded nodes represent unobserved, *hidden* variables. Each node in the graph is associated with a discrete probability distribution. If one or more arrows terminate at a node, its associated probability distribution is conditioned on the node(s) that the arrows originate from. Nodes labeled e_n represent musical events indexed by n. The hidden variable at the top represents a metrical interpretation. The hidden variable labeled $\hat{e}$ represents a predicted subsequent event.

with the selected metrical category, events are then generated in an incremental fashion. As can be seen in Figure 6.1, the likelihood of an event is conditioned on underlying metrical category and preceding events.

Equation 6.1 can be expanded into the incremental and recursive equation shown in Equation 6.2. This equation expresses the posterior distribution given all events as proportional to the product of the likelihood of the last event, e_n and the posterior given all but the last event, $\mathbf{e}_0^{n-1}$. Inferring the posterior incrementally after each event by refining the posterior that resulted from the previous events can be interpreted intuitively as the listener integrating the (bottom-up) information provided by each event into their (top-down) beliefs about the underlying metrical category. Note that the evidence normalization constant has been omitted for clarity.

$$\overbrace{p(m, \psi \mid \mathbf{e}_0^n)}^{\text{per-event posterior}} \propto \begin{cases} \overbrace{p(e_n \mid m, \psi, \mathbf{e}_0^{n-1})}^{\text{per-event likelihood}} \overbrace{p(m, \psi \mid \mathbf{e}_0^{n-1})}^{\text{updated prior}} & \text{if } n > 0, \\ p(e_n \mid m, \psi) p(m, \psi) & \text{else.} \end{cases} \tag{6.2}$$

To infer the posterior distribution over metrical interpretations, Equation 6.2 is evaluated for a set of possible metrical interpretations. This set is constrained to include only metrical categories that occur in the model's training data. The number of different phases considered per metrical category depends on the period of the category, T_m.

To evaluate Equation 6.2, two probability distributions need to be approximated:

the prior distribution over metrical interpretations, $p(m, \psi)$, and the likelihood function $p(\mathbf{e}_0^n \mid m, \psi)$. We discuss both in the following paragraphs.

First, we consider estimating the prior, which uses supervised learning from a corpus of rhythms labelled with metrical category. The parameters of the distribution defining the *a priori* likelihood of metrical categories (not phases), $p(m)$, are set to their maximum likelihood estimate, namely the relative frequency of occurrence of a metrical category in the empirical training data.

$$p(m) = \frac{N_m}{N}, \tag{6.3}$$

where N_m is the number of times m was observed in the training data and N is the total number of training examples (rhythms) in the training data.

The prior distribution over metrical interpretations (i.e., the joint distribution over phase and metrical category) is defined as follows:

$$p(m, \psi) = \frac{p(m)}{\sum^{m'} T_{m'} p(m')}. \tag{6.4}$$

Each metrical interpretation is assigned a probability proportional to the probability of its category. This definition entails a reweighing of metrical categories to compensate for the duration of their periods; it prevents meters with long periods (many possible phases) from being at a disadvantage due to the uniform spreading out of their probability over a large number of phases.

Second, we consider estimating the likelihood. Calculating the likelihood of an observed rhythm given a hypothesized metrical interpretation involves two steps: First, the rhythm under consideration is *interpreted* in a hypothesized metrical interpretation specified by m and ψ. Interpretation is operationalized in the present model as converting the events in the rhythm into a sequence of symbols encoding the position of each event relative to the beginning of the bar in which it occurs under the currently considered metrical interpretation. The details of this conversion are discussed in Section 6.2.3. Second, the likelihood of the resulting sequence of symbols is estimated using an unsupervised probabilistic model trained on metrically interpreted rhythms in the training corpus annotated with the same metrical category, m. The likelihood that a rhythm is generated by given metrical interpretation thus becomes the likelihood of the sequence of symbols resulting from metrically interpreting the event onset times in the rhythm. The likelihood of the metrically interpreted rhythm, in turn, is determined on the basis of a corpus of rhythms belonging to the same metrical category.

Equation 6.2 decomposes the likelihood function into the product of the per-event likelihoods, i.e., the likelihood of each (metrically interpreted) event given the

sequence of preceding (metrically interpreted) events. In the present work, IDyOM (Pearce, 2005) is used to approximate the per-event likelihood function.

IDyOM is a flexible modeling framework based on variable-order Markov modeling combined with a multiple-viewpoint system for music prediction (Conklin & Witten, 1995). It was designed for modeling dynamically changing auditory expectations, based on long-term and short-term statistical learning, which evolve as a piece of music unfolds. Empirical research has demonstrated that IDyOM accurately simulates listeners' predictive processing of melody in many perceptual tasks involving pitch expectation (Pearce, 2005; Pearce, Müllensiefen, & Wiggins, 2010; Omigie, Pearce, & Stewart, 2012; Omigie et al., 2013), uncertainty (Hansen & Pearce, 2014), segmentation (Pearce et al., 2010) and emotional response (Egermann et al., 2013; Gingras et al., 2016).

Section 6.2.3 describes how our model is implemented on top of IDyOM. While the present model does not make use of the full range of modeling opportunities that the multiple-viewpoint approach has to offer, presenting the model as an extension of IDyOM highlights the continuity between the two probabilistic modeling approaches.

Aspects of multiple viewpoint systems and IDyOM relevant to the present model are introduced in Section 6.2.1 and Section 6.2.2. Our treatment of this topic is far from complete; for a complete overview, we refer the reader to Conklin and Witten (1995) and Pearce (2005).

6.2.1 Representation of rhythmic patterns

Multiple viewpoint systems represent the musical surface as a sequence of multi-dimensional datapoints encoding basic attributes of musical events, such as pitch, onset time and duration. These basic attributes of events are accessed through *viewpoints*. A viewpoint maps sequences of events, rather than individual events, to an element of its corresponding *type*, τ. The set of all possible elements of a type τ is called the *alphabet* of τ and is denoted by $[\tau]$. A viewpoint function may be undefined for some sequences of events. The inter-onset-interval viewpoint, for example, is undefined for the sequence $\mathbf{e}_0^0$, which consists of only a single event, e_0. Hence, a viewpoint is defined by a partial function that maps sequences of events to elements of a type

$$\Psi_\tau : \zeta^* \rightharpoonup [\tau],$$

where the symbol ζ^* denotes the set of all possible sequences of events.

A distinction between two types of viewpoints is made. A *basic viewpoint* simply

returns one of the basic attributes of the last event in the sequence to which it was applied (i.e., a projection function). The alphabet of a basic viewpoint is determined by the set of values of its corresponding attribute observed in the training corpus (see Section 6.2.2). A *derived viewpoint* derives more abstract attributes from one or more basic attributes of one or more basic events. Its alphabet can be derived from the alphabets of the basic viewpoints that the viewpoint is derived of. The inter-onset-interval viewpoint and metrical viewpoints introduced in Section 6.2.3 are examples of derived viewpoints. For derived viewpoints, multiple different sequences of events may map to the same element.

The function Φ_τ returns the sequence of viewpoint elements of type τ obtained by applying the viewpoint function Ψ_τ incrementally to to all prefixes of the sequence in order of increasing length:

$$\Phi_\tau(\mathbf{e}_0^n) = \begin{cases} \Phi_\tau(\mathbf{e}_0^{n-1})\Psi_\tau(\mathbf{e}_0^n) & \text{if } \Psi_\tau(\mathbf{e}_0^n) \neq \perp, \\ \Phi_\tau(\mathbf{e}_0^{n-1}) & \text{else,} \end{cases}$$

where $\perp$ is a symbol indicating that the viewpoint is undefined for the given sequence of events.

The model introduced here makes use of a single basic viewpoint, namely `on`, returning the onset attribute of the last event in a sequence, and a set of derived metrical viewpoints. The alphabet of onset, [`on`], contains natural numbers that encode the temporal position of a note as an integer-multiple of basic quantized units. To obtain a finite, meaningful alphabet for `on`, the onset alphabet is constructed online by adding the set of inter-onset intervals encountered in the training data to the onset of the previous event.

6.2.2 Predicting musical events

Predicting sequences of musical events in IDyOM requires specifying a set of viewpoints, $\tau_0, \tau_1, \cdots, \tau_n$, on which to base predictions. A predictive model is associated with each of these viewpoints. Each predictive model is trained on the set of symbol sequence obtained by applying the associated viewpoint function Φ_τ to all event sequences in the training corpus. To approximate the predictive distribution for a future event, $p(\hat{e} \mid \mathbf{e}_0^n)$, given a sequence of preceding events $\mathbf{e}_0^n$, the function Φ_τ is applied, once for each of the specified viewpoints, to $\mathbf{e}_0^n$ to obtain a set of sequences of viewpoint elements.

The per-viewpoint predictions, $p_\tau(\Psi_\tau(\hat{e}) \mid \Phi_\tau(\mathbf{e}_0^n))$ are then combined into a single event prediction, using a mechanism that involves a weighted geometric mean. Some subtleties are involved in converting the predictive distributions to a single domain so that they can be combined (see Pearce, 2005, pp. 111–128). These need

not concern us, as the model proposed here only uses a single viewpoint to predict a single attribute of the event representation (although it could be extended in the future to include use multiple viewpoints).

IDyOM thus reduces the challenge of estimating $p(\hat{e} \mid \mathbf{e}_0^n)$ to the parallel prediction of symbol sequences by estimating $p_\tau(\Psi_\tau(\hat{e}) \mid \Phi_\tau(\mathbf{e}_0^n))$ for each viewpoint $\tau_0, \tau_1, \cdots, \tau_n$. The (domain-general) method employed by IDyOM for predicting symbol sequences is based on a data-compression scheme called prediction by partial matching (PPM) introduced by Cleary and Witten (1984). Pearce and Wiggins (2004) provide an overview of various modifications and improvements to the original PPM scheme that have been proposed over the years, and compare their performance using an information-theoretic performance measure (see Section 6.2.4). IDyOM implements multiple prediction schemes and furthermore allows predictions to be based on two separate models: a long-term model trained on a corpus of training data and a short-term model trained, online, on only the current sequence of events. In our simulations, we use only a long-term model (see Pearce, Conklin, & Wiggins, 2005), employing a PPM* scheme using method C (Moffat, 1990) for calculating escape probabilities and adapted to use interpolated smoothing—the configuration Pearce and Wiggins (2004) found to yield the best results for a long-term model. A parameter called model order-bound parameter limits the amount of previous events taken into account in the predicting the next event, $\hat{e}$: An order-bound of b means that it is assumed that $p(\hat{e} \mid \mathbf{e}_0^n) \approx p(\hat{e} \mid \mathbf{e}_{n-b}^n)$. While Pearce and Wiggins (2004) found that an unbounded model order worked best, the present paper presents results for varying model order-bounds of up to four.

6.2.3 Metrical viewpoints, metrical models, and metrical inference

The per-event likelihood function in Equation 6.2 is a predictive distribution that, based on events observed so far and a hypothesized metrical interpretation, specified by m and ψ, predicts the next event. This relies on interpreting the sequence of events in the given metrical interpretation and estimating the likelihood of the resulting sequence of symbols given a predictive model of such sequences in the provided metrical category. Interpretation of a rhythm in a specific metrical interpretation is achieved in IDyOM through the introduction of a set of *metrical viewpoints*. Metrical viewpoints transform a sequence of absolute onset times into a sequence of symbols that depend on the metrical interpretation implemented by the viewpoint.

The general form of a metrical viewpoint $\tau_{m,\psi}$ is

$$\Psi_{\tau_{m,\psi}}(\mathbf{e}_0^n) = f(m, \psi, \mathbf{e}_0^n),$$

where f is a function that implements the metrical interpretation given a phase and metrical category.

The present model uses a simple metrical interpretation function that returns the *metrical position* of an onset. This function makes few assumptions about the structural organization of meter, and can accommodate complex, non-isochronous meters. The metrical position of an onset is defined as its position relative to the period and phase of an interpretation. The general definition of the resulting metrical position viewpoint, `mp`, is given below

$$\Psi_{\mathtt{mp}_{m,\psi}}(\mathbf{e}_0^n) = (\Psi_{\mathtt{on}}(\mathbf{e}_0^n) - \psi) \mod T_m,$$

where the viewpoint `on` is a basic viewpoint that returns the onset of the last event in a sequence of events.

One metrical viewpoint is created for each metrical interpretation considered by the model by instantiating m and ψ to a specific value.

The alphabet of the `mp` viewpoint is given by

$$[\mathtt{mp}_{m,\psi}] = \{0, 1, \cdots, T_m - 1\}.$$

Using metrical viewpoints, metrical inference can be implemented on top of the standard IDyOM machinery, with one important caveat: the predictive model of a metrical viewpoint, $\tau_{m,\psi}$ is trained only on those sequences in the training data that have been annotated with metrical category m. Hence, the predictability of a metrically interpreted rhythm depends only on rhythms previously observed in the corresponding metrical category.

One further subtlety needs to be addressed to complete the model. Note that the per-viewpoint predictive distributions mentioned in Section 6.2.2 are defined over a viewpoint's alphabet $[\tau]$. In order to predict the onset of the next event this alphabet needs to be mapped back to the alphabet of the onset viewpoint, [`on`]. However, any metrical position in [`mp`] theoretically corresponds to an infinite number of periodically spaced onset times. To be able to generate predictions for *specific* onset times, and for metrical inference to work correctly, it is necessary that the alphabet of a metrical viewpoint maps to unique onset times. This can be achieved by *linking* the metrical position viewpoint to another metrical viewpoint, which encodes the distance in bars between the last event and the predicted event.

The equation below defines the bar distance viewpoint, `bd` in terms of an intermediate metrical viewpoint, `bn` (bar number), which calculates the number of bars elapsed between time zero and the onset of the last event.

$$\Psi_{\mathtt{bd}_{m,\psi}}(\mathbf{e}_0^n) = \Psi_{\mathtt{bn}_{m,\psi}}(\mathbf{e}_0^n) - \Psi_{\mathtt{bn}_{m,\psi}}(\mathbf{e}_0^{n-1}),$$

where metrical viewpoint `bn` is defined as

$$\Psi_{\mathtt{bn}_{m,\psi}}(\mathbf{e}_0^n) = \text{integer}\left(\frac{(\Psi_{\mathtt{on}}(\mathbf{e}_0^n) - \psi)}{T_m}\right).$$

A linked viewpoint is a special case of a derived viewpoint composed of a number of constituent viewpoints. The elements of linked viewpoints are tuples containing the values of the constituent viewpoints. A linked viewpoint composed of $\tau_1, \cdots, \tau_n$ is denoted by $\tau_1 \otimes \cdots \otimes \tau_n$, its alphabet is given by the Cartesian product of the constituent viewpoints' alphabets: $[\tau_1] \times \cdots \times [\tau_n]$.

The linked metrical viewpoint used in our simulations is denoted by $\mathtt{mp} \otimes \mathtt{bd}$, and encodes metrical position and distance in bars between the last event. Elements in the alphabet of this viewpoint have a one-to-one correspondence to elements in $[\mathtt{on}]$.

To summarize: metrical viewpoints and separate predictive models per metrical category enable using IDyOM to estimate the per-event likelihood function in Equation 6.2. In this model, the likelihood of a metrical interpretation m depends on the predictability of the sequence of symbols that results from interpreting the rhythm in that metrical interpretation. This predictability in turn depends on the set of rhythms previously observed in m.

6.2.4 Expectation and information content

We have focussed our discussion so far on the issue of inferring a posterior distribution over metrical interpretations. In order to calculate prediction error, it is necessary to derive the predictive distribution over future note onsets given a preceding rhythmic context and an inferred meter.

To estimate prediction error, we look at the amount of information communicated by each observation. Although it is sometimes referred to as cross-entropy (e.g. Manning & Schütze, 1999), we call this quantity the *information content* (MacKay, 2003) of an event. Information content is defined as the negative logarithm of the likelihood of observing the next event given the predictive distribution conditioned on the sequence of events observed so far:

$$h(\hat{e} \mid \mathbf{e}_0^n) = -\log_2 p(\hat{e} \mid \mathbf{e}_0^n). \tag{6.5}$$

In an information-theoretic sense, this quantity is equivalent to prediction error. An unlikely (unexpected) event results in a high prediction error, signaled by high information content. Conversely, a likely event results in a low prediction error, signaled by low information content.

The predictive distribution corresponds to the probability distribution associated with the hidden variable labeled $\hat{e}$ in the graphical model in Figure 6.1. This distribution is obtained from the generative model by marginalizing out meter and phase from the posterior distribution inferred from the preceding events:

$$p(\hat{e} \mid \mathbf{e}_0^n) = \sum^{m} \sum^{\psi} p(\hat{e} \mid m, \psi, \mathbf{e}_0^n) p(m, \psi \mid \mathbf{e}_0^n), \tag{6.6}$$

where the summation over meters sums over all metrical categories considered by the model, $m \in M$, and the summation over phases sums over all possible phase of category m, $\psi \in \{0, 1, \cdots, T_m - 1\}$.

Equation 6.6 shows that the prediction of the onset of the next event is subject to top-down influence from the distribution over metrical interpretations inferred from bottom-up information from the events observed so far.

6.2.5 Hypotheses

We expect an accurate computational model of human meter perception to show certain patterns of behavior. First, we expect it to be able to infer meters that agree with the time signatures in notated scores (Longuet-Higgins & Lee, 1982; Temperley, 2004). Second, we argued that the metrical knowledge, acquired by listeners through exposure to a musical idiom, is characterized not only by the distribution of onsets over metrical positions, but also by the probabilistic properties of how rhythms in particular meters sequentially unfold. Thus, we expect that a model that can learn such properties will lead to increased performance in finding time signatures notated in scores compared to a similar model that does not learn these properties. Third, we argued above that categorizing rhythms into metrical categories can plausibly be regarded as a strategy to reduce prediction error for those rhythms. Therefore, we expect that our model will show better performance in predicting the timing of musical events than a comparable model that is agnostic of meter. Fourth, we expect that our model will simulate enculturation by showing sensitivity to the statistical properties of the rhythms it was trained on. A model trained on rhythms with similar statistical properties as the rhythms it is evaluated on will perform better than a model that was trained

on rhythms with different statistical properties. If the statistical properties of rhythms originating from two cultures with different cultural practices regarding rhythm are sufficiently different, we expect that a model trained on rhythms from the same culture as the rhythms it is evaluated on will outperform a model trained on rhythms from a culture with different rhythmic practices. We evaluate these expectations in Sections 6.3 and 6.4.

6.3 Methods

6.3.1 Resolution of onset time and phase

For reasons of computational efficiency, the resolution the phase parameter of metrical interpretations is restricted to sixteenth notes. This means that, for example, in the 3/4 category twelve different phases are possible (since the duration of one 3/4 bar is twelve sixteenth notes). Since all onset times in rhythms used in this study encode distance from the beginning of the first bar in the annotated meter the correct phase of a rhythm can be represented under any phase resolution. The representation of rhythms in a phase of zero does not influence the evaluation: as far as the model is concerned, all phases are initially equally likely, since the prior distribution over phase is uniform. The presence of 32th notes and 16th-note triplets in the training data requires that onset times are represented as integer multiples of symbolic units corresponding to 96th notes.

6.3.2 Training data

Except for one artificially constructed test set, the datasets used in our simulations are all derived from the Essen folksong collection (Schaffrath & Huron, 1995). The Essen folksong collection is a corpus consisting of monophonic transcriptions of folksongs, originating from various geographical regions across the globe. The majority of the folksongs in this dataset originate from regions in Germany and China. We use a version of the Essen folksong collection encoded in humdrum format, which we obtained from http://kernscores.stanford.edu.

Folksongs without an annotated time signature, or with multiple time signatures are filtered out. The simulations described below use different subsets of this filtered version of the Essen folksong collection.

6.3.3 Classification performance and the influence of preceding context

The first expectation formulated in Section 6.2.5 concerns the model's ability to infer meters that agree with time signatures notated in scores. To evaluate this, classification performance is measured using ten-fold cross validation on a dataset of German folksongs. In a cross validation scheme, the model is trained and evaluated ten times on different partitions of the dataset into a training set and a test set. Reported classification scores are based on the average classification score over all ten partitions.

The second expectation we formulated is that models exploiting sequential probabilistic properties will perform better in this task than a similar model that does not exploit such properties. To evaluate this, we measure classification performance of five different models configured with order-bounds ranging from zero to four using cross validation. The order-bound parameter (see Section 6.2.2) allows us to vary the degree to which the model can learn sequential probabilistic properties of rhythms, interpolating between a model that can only learn distributions of onsets over metrical positions (order-bound zero) and a model that predicts the subsequent metrical position based on the metrical positions of the last four events (order-bound four).

The result of performing inference on the generative model—inferring meter from a rhythm—is not a single classification, but a posterior probability distribution over metrical interpretations. To determine in which meter the model interprets a rhythm, an additional inferential step is required. All classification scores reported in this paper are based on the interpretations with the highest posterior probability after observing the entire rhythm. An interpretation is considered correct if its phase and category agree with the annotated time signature.

For these simulations, we used rhythms extracted from 4966 German folksongs in the Essen folksong collection. This set is constructed by selecting all melodies with an "ARE" record (area of origin; Huron, 1999) indicating a region of Germany from the Essen folksong collection, subject to the constraints described in Section 6.3.2. Figure 6.2 shows the distribution of meters in the resulting dataset. The most frequently appearing time signatures in this set are 4/4, 2/4, 3/4, and 6/8.

6.3.4 Does metrical inference reduce prediction error?

The third expectation we formulated is that a model using inferred meter to predict the onsets of musical events will outperform comparable models that do not use metrical inference. To assess whether metrical inference increases predictive performance we compare the model an IDyOM model that predicts

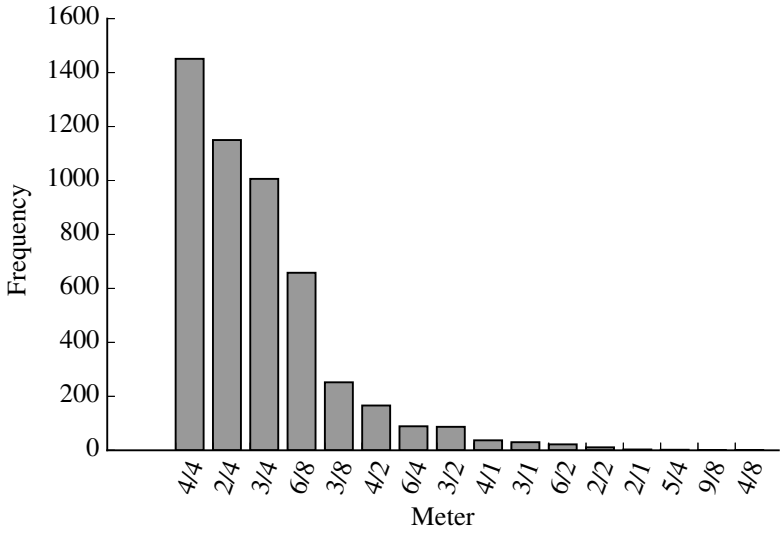


Figure 6.2: Histogram showing the of the distribution of meters in the dataset of 4966 German folksongs from the Essen folksong collection.

event onset time without inferring meter. Prediction performance is measured by looking at average information content (see Section 6.2.4), which represents the discrepancy between predicted and observed events.

This IDyOM model is configured to use a single viewpoint, encoding inter-onset intervals between subsequent events, to predict onset time. Inter-onset interval is defined as the difference between the onset time of the final and penultimate event. Both models are trained and evaluated on the same dataset using cross-validation, and the input of both models consists only of onset times encoded in the event representation.

The results are reported, as before, for order-bounds varying from zero to four. The values represent average information content over cross validation folds.

6.3.5 Simulating enculturation

The fourth expectation concerning the model's behavior we formulated is that it should show sensitivity to the statistical properties of its training data. To investigate this, two types of statistical aspects of training data that affect the model's behavior in different ways are distinguished. The first aspect is the distribution of metrical categories in the training rhythms. This distribution is directly reflected in the prior distribution, encoding a priori likelihood of different metrical categories. The effect of the prior distribution on the model's behavior can be seen as *inferential biases*. The second aspect concerns the sequential

structure of the training rhythms themselves. This aspect includes the distribution of onsets over different metrical positions, but also the typical unfolding of rhythms interpreted in a specific meter and the presence of stereotypical rhythmic patterns.

These two aspects of training data may influence the encountered prediction error on novel rhythms as well as the metrical category in which rhythms are interpreted. To investigate the effect of inferential biases, we focus on consequences of inferential biases for metrical interpretation. In the investigation of the statistical properties of rhythms themselves we focus on the effects of training data on prediction error.

The simulations described below are all conducted using an order-bound of four, since the cross validation results indicate that, out of the considered order-bounds, four works best (see Section 6.4).

6.3.5.1 Inferential biases

A high prevalence of certain metrical categories in the music to which a listener has been exposed to previously may lead to inferential biases: a tendency to interpret rhythms in the pervasive category. In probabilistic terms, this is a sensible behavior: in the presence of uncertainty, it is optimal to tend towards categories with a high *a priori* likelihood of occurring. Such likelihoods are represented in the prior distribution over metrical categories. Inferential biases are top-down in the sense that they are independent of the particular rhythm encountered by the model. Once the model begins to process a rhythm, the prior distribution is updated by bottom-up evidence from the rhythm. Inferential biases can alternatively be understood as changing the initial state of meter induction. Meters favored by the prior distribution require less evidence from rhythmic events to gain a high posterior likelihood. In cases where a rhythm is ambiguous (i.e., provides evidence for two or more metrical categories), inferential biases towards either category can be decisive in the model's interpretation.

To we investigate the effect of inferential biases, we train two models on a subset of the German folksongs described in Section 6.3.3 containing 658 2/4 (a simple duple meter), 658 3/4 (a simple triple meter) and 658 6/8 (a compound duple meter) training examples. We bias the prior distribution of one model to favor 3/4 interpretations while the other model is biased to favor 6/8 interpretations.

In this simulation the prior distribution is not estimated empirically using the relative frequency of metrical categories in the training data. Instead, the parameters of the prior distribution are manually set to the values shown in Table 6.1. The rationale behind this choice is that if we would manipulate the prior distribution by altering the number of training rhythms in a metrical category, the number of training examples from which the model predictive model of that category is

Table 6.1: Prior probabilities of metrical categories used for simulating inferential biases.

Category	3/4 biased	6/8 biased
2/4	4/9	4/9
3/4	4/9	1/9
6/8	1/9	4/9

learned would be affected, which introduces performance differences that cannot be attributed solely to the prior distribution.

The consequences of the biased prior distribution are investigated using an artificially constructed test set. To construct this set, first, a set of rhythmic patterns is constructed by generating all possible patterns within the following constraints: the total duration of a pattern is exactly twelve sixteenth notes, none of the patterns begin with a rest and the minimum inter-onset interval is a sixteenth note. The resulting set consists of 2^{11} rhythmic patterns: each pattern begins with an onset and each sixteenth-note time point between the second and twelfth sixteenth-note can contain an event onset. Because twelve sixteenth notes is exactly the duration of one 3/4 or 6/8 bar, this set contains all rhythms with a minimum interval of a sixteenth note that fit in one bar of a 3/4 or 6/8 meter. To construct the final test set, each of these patterns is repeated four times. The repetition allows the model more time to converge on a single interpretation.

Both models are used to infer meter for each rhythm in the test set. Note that while three different categories, 2/4, 3/4, and 6/8, are considered, the quadruple repetition of patterns with a duration of twelve sixteenth notes may favor 3/4 and 6/8 interpretations. Since this potential bias is a property of the test set on which both models are evaluated, it does not cause problems for the evaluation of the effect of inferential biases.

We expect that inferential biases will increase the number of rhythms interpreted in the category corresponding to the bias. Due to the juxtaposition of 3/4 and 6/8 inferential biases, and the bar-level period-correspondence between these two meters, we expect to find the greatest degree of disagreement in interpretation of rhythms in the test set between the 3/4 and 6/8 categories: the 3/4 biased model will likely interpret rhythms classified by the 6/8 biased model as 6/8 in 3/4 and vice versa.

It seems plausible that 3/4 and 6/8 inferential biases will lead to some disagreement about the 2/4 category. An inferential bias may lead a model to interpret rhythms classified by the other model as 2/4 in the category corresponding to its bias. At the tactus level, 2/4 and 3/4 exhibit structural similarities: by convention, 2/4 and 3/4 both imply simple meters, where beats are subdivided into two smaller

units. The 6/8 time signature, on the other hand, implies a compound meter. These (music-theoretic) similarities between 2/4 and 3/4 may lead the 3/4 biased model to interpret more rhythms, interpreted in 2/4 by the 6/8 biased model, according to its bias than the 6/8 biased model will out of the rhythms interpreted in 2/4 by the 3/4 biased model. It is worth noting that 2/4 and 6/8 have a different structural similarity at the level above the tactus: they are both duple meters. However, the duration of beat in 2/4 and 6/8, in our quantized input representation, is different, preventing this similarity from playing a role in our model.

The set of rhythms interpreted differently by both models likely consists of rhythms that do not strongly imply one specific interpretation. We expect such rhythms to be either ambiguous, or metrically over- or under-determined (London, 2012, pp. 75–76). Because we define a classification as the interpretation with the maximum posterior probability, the model always produces an interpretation of a rhythm, even if evidence from the rhythm is weak or conflicting. Therefore, some of the rhythms about which the models disagree may be metrically vague, i.e., not strongly suggesting any interpretation.

6.3.5.2 Cultural distance between Chinese and German rhythms

In two simulations, we investigate how the model responds to being trained on folksongs originating from China or Germany. Music from these two areas might be different enough to lead to differences in rhythmic processing between enculturated individuals. By training the model on a dataset of Chinese and German folksongs, we can simulate how, according to the model, exposure to these stylistically different sets of rhythms affects perception.

To this end, we use two dataset sets: containing folksongs originating respectively from Germany and China. The German dataset is the same one that is used for the cross validation simulations described in Section 6.3.3. The dataset of Chinese folksongs is constructed in the same way as the German dataset, namely by selecting all folksongs from the Essen folksong collection whose "ARE" reference record (Huron, 1999) indicated a region in China and after first filtering out folksongs with zero or more than one annotated time signatures.

We run simulations in two separate conditions. In both conditions, two models are trained: one on a Chinese training set, and one on a German training set. Both of these models are subsequently evaluated on a separate Chinese and German test set consisting of rhythms that do not occur in the training data. In contrast to the simulation described above, we estimate the prior distribution in its normal way (see Equation 6.3 and 6.4).

The number of rhythms of each metrical category used in the test and train sets

in the first and second condition are shown in Table 6.2.

In the first condition (see the columns under "identical" in Table 6.2), we control for the effect of the prior distribution and use identical distributions of metrical categories in the training data of both models. This allows us to attribute observed effects to differences in the statistical properties of rhythms, ruling out effects of differences in the number of training examples or the differences in prior distributions. Meters considered in the simulation need to be well represented in both datasets. In the German and Chinese dataset that we have available, this constraint leaves 2/4, 3/4, and 3/8 as suitable categories. Despite this reduction, the number of rhythms in meters other than 2/4 in the Chinese dataset remains rather small.

Due to the small number of rhythms in meters other than 2/4 in the Chinese dataset, it is not possible to use a uniform distribution of meters in the test sets for this condition. Instead, we only include rhythms in 2/4 in the German and Chinese test set.

In the second condition (see the columns under "empirical" in Table 6.2), we allow the prior distribution to influence results and use empirical distributions of metrical categories in the training data of both models. By empirical, we mean that the relative frequencies of meters in the test and training sets that we used are equal to those observed in the Essen folksong collection. Both training sets contained in total an equal number of training examples.

Rhythms in the test sets for this condition are distributed to the same proportions as in the corresponding training sets. The Chinese test set predominantly contains rhythms annotated in 2/4 while the German test set also contains substantial numbers of rhythms in 3/4 and 4/4.

We expect that, on the Chinese and the German test sets, the model trained and tested on culturally similar music will exhibit lower average information content and higher classification performance than the model trained on culturally different music. We expect to see this pattern of results both for the identical, as well as for the empirical distribution of meters in the training data.

6.4 Results

6.4.1 Classification performance and preceding context

Figure 6.3a shows the average number of correct interpretations found by our model at order-bounds ranging from zero to four. The averages are obtained by first averaging all per-event information contents (see Section 6.2.4) in the test set

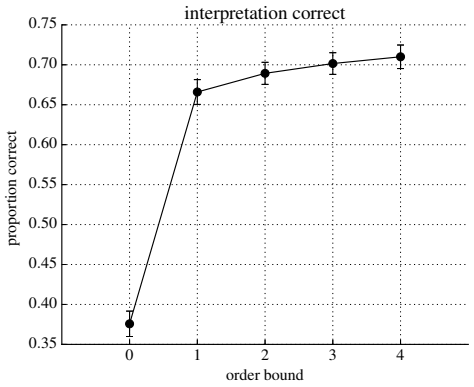
Table 6.2: Number of rhythms in different metrical categories in training and test sets in two different conditions (*identical* or *empirical*) used in the simulation of enculturation. The *upper* table shows the counts for the identical condition and the *lower* table shows counts for the empirical condition.

Distribution of meters		Identical			
Country of origin		Germany		China	
Dataset		Training	Test	Training	Test
	2/4	950	200	950	200
	4/4	132	0	132	0
Meter	3/4	35	0	35	0
	3/8	19	0	19	0
Total count		1136	200	1136	200

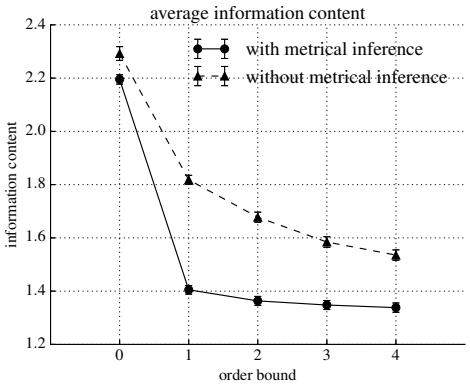
Distribution of meters		Empirical			
Country of origin		Germany		China	
Dataset		Training	Test	Training	Test
	2/4	339	60	1009	178
	4/4	427	75	90	16
Meter	3/4	296	52	24	4
	3/8	74	13	13	2
Total count		1136	200	1136	200

of one cross validation fold, and subsequently over all cross-validation folds. The standard deviations are calculated over the averages per cross validation fold. At order-bound zero, the model interprets rhythms in agreement with annotated the time signatures in, on average 38%, of the cases. At order-bound one, classification performance increases sharply to, on average, 67% of the rhythms in agreement with the annotated time signature. Increasing order-bound further yields modest improvements. At order-bound four, the highest we tested, on average, 71% the rhythms were interpreted in agreement with the annotated time signature.

Variability in performance between different partitions of the data in a training and test set is low, as the small error bars in Figure 6.3a show.



(a) Proportions of correctly classified interpretations.



(b) Average information contents for the model (with metrical inference) compared to IDyOM without metrical inference.

Figure 6.3: Classification performance and average information content for five different models varying in order-bound, evaluated using ten-fold cross-validation. Markers represent values obtained by averaging over the ten folds. Error bars represent one standard deviation above and below the average.

6.4.2 Metrical inference and prediction error

Figure 6.3b shows prediction performance in terms of average per-event information content of rhythms under IDyOM (without metrical inference) and our extended version of IDyOM (with metrical inference). Both models were tested at order-bounds ranging from zero to four.

The results shows that, in general, information content decreases as order-bound increases for both the IDyOM model (without metrical inference) and our model (with metrical inference). The results also show that for all tested order-bounds, the average information content is lower our model (with metrical inference): for example 2.19 compared to 2.29 for order-bound zero and 1.34 compared to 1.54 at order-bound four.

6.4.3 Simulating enculturation

6.4.3.1 Inferential biases

The results obtained from contrasting two models with manually manippulated prior distributions on an artificially generated test set are summarized in Table 6.3.

Table 6.3: A contingency table showing the number of time-signature classifications by a 3/4 biased model and a 6/8 biased model.

		3/4 Biased			
		6/8	3/4	2/4	All
6/8 Biased	6/8	471	83	40	594
	3/4	0	395	0	395
	2/4	0	54	1005	1059
	All	471	532	1045	2048

The results shows that both models interpret approximately half of all rhythms in 2/4. The rightmost column in bold shows that the 6/8 biased model interprets more rhythms in 6/8 than in 3/4, while the bottom row in bold shows that the 3/4 biased model interprets more rhythms in 3/4 than in 6/8.

The numbers on the diagonal show that both models agree on the vast majority of interpretations. Both models agree on the interpretation of rhythms that are classified *despite* an inferential bias as 3/4 or 6/8: None of the rhythms that the 3/4 biased model interprets as 6/8 are interpreted differently by the 6/8 biased model. Similarly, none of the rhythms that the 6/8 biased model interprets as 3/4 are classified differently by the 3/4 biased model.

The numbers off the diagonal show that the greatest degree of disagreement occurs between the 6/8 and 3/4 categories, but there is also substantial disagreement between 2/4 and 3/4 and 2/4 and 6/8.

There are two categories of rhythms sensitive to inferential biases: The first category consists of 83 rhythms that the 6/8 biased model interprets in 6/8 while the 3/4 biased model interprets them in 3/4. The second category consists of rhythms that one model interprets in 2/4 while the other model interprets them in the category its biased towards. The 6/8 biased model interprets 40 rhythms in 6/8 that the 3/4 biased model interprets in 2/4. Out of the rhythms classified by the 6/8 biased model as 2/4, the 3/4 biased model interprets slightly more rhythms in agreement with its bias (namely 54), than the 6/8 biased model does out of the rhythms classified by the 3/4 biased model as 2/4 (namely 40).

6.4.3.2 Cultural distance between Chinese and German rhythms

Table 6.4 shows average information content and classification performance obtained in the simulations of enculturation with German or Chinese folksongs. Results from two conditions are reported: one in which the German and Chinese training sets have an identical distribution of metrical categories and one in which

Table 6.4: Average information content and classification performance of models trained and evaluated on test sets with rhythms from Germany and China. Results are reported for two different conditions. One in which training sets contain *identical* distributions of metrical categories, and one in which training sets contain *empirical* distributions of metrical categories.

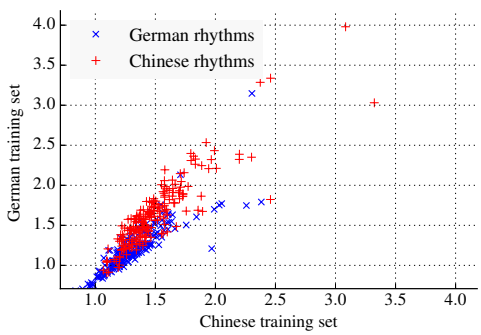
		Test set			
		Identical priors		Empirical priors	
	Training set	German	Chinese	German	Chinese
Information content	German	1.21	1.63	1.34	1.72
	Chinese	1.32	1.49	1.70	1.49
Classification	German	0.84	0.80	0.73	0.72
	Chinese	0.59	0.77	0.47	0.75

they have empirical distributions of metrical categories.

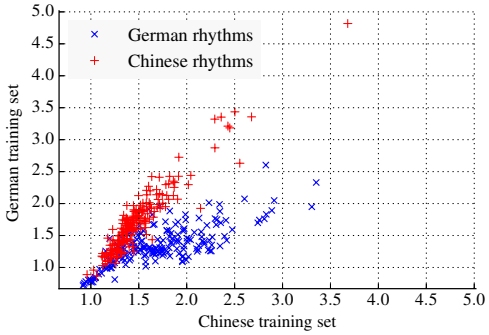
In both conditions the results can be said to show effects of enculturation: The average information content for models evaluated on rhythms from the same country as the rhythms in their training data (culturally familiar) is lower than for models trained on rhythms from the other country (culturally unfamiliar). Classification performance shows a similar pattern: in most cases, classification performance is better for models evaluated on culturally familiar rhythms. However, in the identical prior condition, classification performance of the German model on the Chinese test set was slightly higher than of the Chinese model. Furthermore, in the identical prior condition, the average information content of the Chinese model is lower when evaluated on the German test set compared to the Chinese test set.

For both models and in both conditions, but most notably in the identical priors condition, information content of rhythms in the Chinese test set was slightly higher than that of rhythms in the German test set.

Figure 6.4a and 6.4b project the rhythms from both test tests onto a two-dimensional plane. The coordinates of each rhythm are determined by the average information content of events in the rhythm under the Chinese model (x-axis) and German model (y-axis). Under this projection, rhythms from the two cultures form clusters that are to some degree spatially separated. The degree of separation is stronger in the empirical prior condition (Figure 6.4b). For both conditions, average information content of events in a single test set is highly correlated between both models (see Table 6.5).



(a) Results for the training and test sets with fixed distributions of meters.



(b) Results for the training and test sets with empirical distributions of meters.

Figure 6.4: Scatter plots of the average information content of rhythms for the Chinese and German models.

Table 6.5: Pearson product-moment correlation coefficients between average information content per rhythm under the German and the Chinese model, showing the degree to which information-content assigned to the same rhythms by both models is related.

		Test set	
		German	Chinese
Prior	Fixed	0.74	0.94
	Empirical	0.86	0.89

6.5 Discussion

A predictive coding view of perception entails that perception depends on generative models in the mind of the perceiver that are tuned by statistical properties of the environment, both through evolutionary adaptation and sensory experience, to predict sensations. We hypothesized that effects of enculturation on the perception of meter can be understood in terms of predictive coding. To explore the consequences of this idea, we presented a probabilistic model of meter perception for which predictive coding served as the conceptual basis. The underlying hypothesis is that meter perception is the result of a strategy, based on statistical learning, probabilistic prediction and inference, for increasing predictive accuracy in processing of temporal events in music.

A set of expectations concerning the model's behavior was derived based on: the relevance of the model as a cognitive model of meter perception, theoretical proposals about the relation between rhythm and meter, the model's ability to

reduce prediction error, and finally the model's potential to simulate enculturation. To investigate the degree to which the model meets these expectations, we ran a series of simulations. The results show that the model can infer metrical structure from rhythms, and that this ability improves when statistical properties of the succession of onsets in the metrical context are taken into account. A comparison with a similar model that does not use metrical inference demonstrates that metrical inference reduces prediction error in predicting the timing of musical events. Finally the results show hypothesised patterns of enculturation when models are trained on corpora varying, both naturally and artificially, in terms of distribution of metres and rhythmic properties.

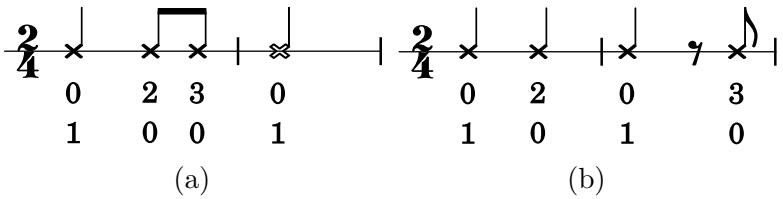
The following sections discuss the simulation results in detail.

6.5.1 Meter classification and preceding context

A model of meter perception can reasonably be expected to interpret a simple rhythm in a meter that agrees with the time signature that an educated listener would use when transcribing that rhythm. The used rhythms were taken from folksongs in the Essen folksong collection (Schaffrath & Huron, 1995). Despite its possible relevance to determining the time signature, melodic information was disregarded. This limitation notwithstanding, cross-validation results indicate that the model generally infers interpretations that agree both in category and phase with annotated time signatures. The best performing model configuration interprets rhythms in a time signature and phase that agrees with annotations in the Essen folksong collection in 71% of the cases. These classifications were selected by the model out of a large pool of alternatives. Summing the number of possible phases per considered metrical category (see Section 6.2) yields 320 possible metrical interpretations. Many of these categories occur very infrequently in the training data, resulting in a low *a priori* likelihood for these categories. If we limit interpretations to the four most frequently occurring metrical categories—4/4, 2/4, 3/4, and 6/8—the number of interpretation options reduces to 48.

By varying the model's order-bound (the amount of preceding events that inform the prediction of the next event, see Section 6.2.2), we investigated to what degree learning statistical properties of the succession of metrical positions in rhythms improved the model's performance.

Increasing the order-bound from zero to one yields the most significant improvement in classification performance. This finding is consistent with results obtained by Temperley (2010) in a comparison of six onset-prediction models. Some of these models were metrical, which means they made use of provided (rather than probabilistically inferred) metrical information. Temperley (2010) found that out of the compared models, the two metrical and context-sensitive models, namely



(a) (b)

Figure 6.5: Two rhythms that result in different orderings of the same set of mp ⊗ bd viewpoint elements. The number-pairs below the notes are the values of mp ⊗ bd. The top number represents the value of the bd (bar-distance) viewpoint, the bottom number represents the value of the mp (metrical position; expressed in multiples of an eighth note duration) viewpoint.

the first-order metrical duration model and hierarchical position model, yielded the lowest cross-entropy (information content) score.

The performance increase between order bound zero and one is unsurprising. In a zeroth-order model, events in a rhythm are conditionally independent given a meter. If the meter is known, the probability of the next event only depends on its metrical status and is independent of preceding events. [1] In a zeroth-order model, a rhythm is a "bag of notes": the order in which notes occur is irrelevant to the final outcome. However, note-order bears consequences for the metrical interpretation of a rhythm, as illustrated in Figure 6.5. The rhythm in Figure 6.5a is structurally different from the rhythm in Figure 6.5b, yet under a zeroth-order model using mp ⊗ bd metrical viewpoints (see Section 6.2.3) these rhythms are indistinguishable.

The results show that classification and prediction performance, increases further when order-bound is increased to four. Since this improvement is relatively modest, it remains to be seen to what extent probabilistic information about the succession of multiple events facilitates metrical inference. Perhaps the effect of order-bound would be more pronounced for music styles with more complex rhythms than the folksongs used here.

6.5.2 Metrical inference reduces prediction error

We proposed that meter perception may result from predictive coding: interpreting onsets in a rhythm as the result of a generative model with different periodic categories (meters), that are inferred from the pattern of onsets itself, may facilitate prediction of future onsets. Interpreting a rhythm in a metrical framework allows a listener to relate the observed events to patterns they observed previously. A

[1] The bd viewpoint used in our simulations indirectly introduces minor context dependency: if its value zero it means that the current note is the first note in the bar.

computational probabilistic model that infers meter to predict the timing of events, such as the one presented here, should therefore encounter a lower prediction error in empirical rhythms compared to a similar model that does not infer meter.

To evaluate this, we compared prediction performance of the presented model to an IDyOM model that predicts the event onset times without using metrical inference. This comparison seems natural because the presented model implements metrical inference directly on top of IDyOM as explained in Section 6.2.

Simulations show that the meter inferring model reduces prediction error compared to IDyOM (without metrical inference) under all tested order-bounds. These results support the suggestion that inferring meter may improve temporal prediction of events in rhythms.

6.5.3 Simulating enculturation

The goals of the simulations concerning enculturation were to investigate how our model's behavior is shaped by the statistical properties of rhythms in its training data, and to investigate the extent to which these statistical properties can be exploited to improve the prediction and metrical interpretation of stylistically similar rhythms. We first explored the consequences of inferential biases on an artificially constructed set of potentially ambiguous rhythms. Then, we studied the effect of statistical properties of sets of rhythms on metrical inference. The results show that when tested on Chinese rhythms, models trained on rhythms of Chinese folksongs show better prediction performance than models trained on German folksongs. The converse was true when the models were tested on German folksongs.

This simulation of enculturation should be seen as a proof-of-concept: Patterns of quantized onset times annotated with meter are a limited representation of the rich variety of musical and non-musical experiences that may shape listeners' perception of meter. In the musical domain, timbre, polyphony, expressive timing and dynamics are some examples of aspects not considered by our approach that all could plausibly form part of the experiences that shape meter perception. Nevertheless, it is possible that monophonic corpora of rhythms from different cultures can predict some enculturation effects. The methodology presented here is an illustration of how such predictions could be made.

6.5.3.1 Inferential biases

Inferential biases were introduced into the model by directly manipulating the prior distribution, to avoid differences in the amount of training examples per metrical category, which would influence the results.

We contrasted two models: one with a 6/8 inferential bias, another with a 3/4 inferential bias. The models were evaluated on an artificially constructed test set of rhythms with the potential for ambiguity between 3/4 and 6/8. These test rhythms were not annotated, as we intended find the set of rhythms for which inferential biases could swing the model's interpretation.

The results show that inferential biases affected the distribution of interpretations over metrical categories in ways that we expected: Each model interpreted more rhythms in the category corresponding to its bias than the other model. Both models agreed on the interpretation of the majority rhythms. These rhythms contained enough evidence towards a particular interpretation to override the model's inferential bias. As we expected on music theoretic grounds, the 3/4 biased model swung the interpretation of slightly more rhythms, interpreted in 2/4 by the 6/8 biased model, to a 3/4 interpretation than the 6/8 model did out of the set of rhythms interpreted in 2/4 by the 3/4 biased model.

Eight rhythms interpreted were interpreted in 3/4 without pick up by the 3/4 biased model and in 6/8 without pick up by the 6/8 biased model. These rhythms are shown, by way of example, in Figure 6.6, along with metrical grids contrasting a simple 3/4 interpretation with a compound 6/8 interpretation. That the interpretation of these rhythms could be influenced depending on inferential bias of the model suggests that they are ambiguous (e.g., 6.6vi), and/or metrically underdetermined (e.g., 6.6i and 6.6iv), or metrically vague, i.e., not strongly suggesting any interpretation (e.g., 6.6viii).

6.5.3.2 Cultural distance between Chinese and German rhythms

In general agreement with the hypotheses presented in Section 6.2.5, the results in Table 6.4 show that models evaluated on a test set with rhythms from the same country as the rhythms they were trained on exhibit a lower average per-event information content. The classification scores for models trained on culturally familiar rhythms were also higher compared to models trained on culturally unfamiliar rhythms, except on the Chinese test set in the identical prior scenario. It could be that rhythms in the Chinese portion of the Essen folksong collection (Schaffrath & Huron, 1995) were less consistently annotated, but further investigation is necessary to determine whether this is the case. The pattern of results suggests that the statistical properties of Chinese and German rhythms are different, and that these differences can be exploited to optimize prediction and metrical inference on rhythms from one of the countries.

In a recent study comparing recognition memory in North American listeners on Turkish classical music and Western art music, (Demorest, Morrison, Nguyen, & Bodnar, 2016) found that rhythmic properties of music did not contribute to an

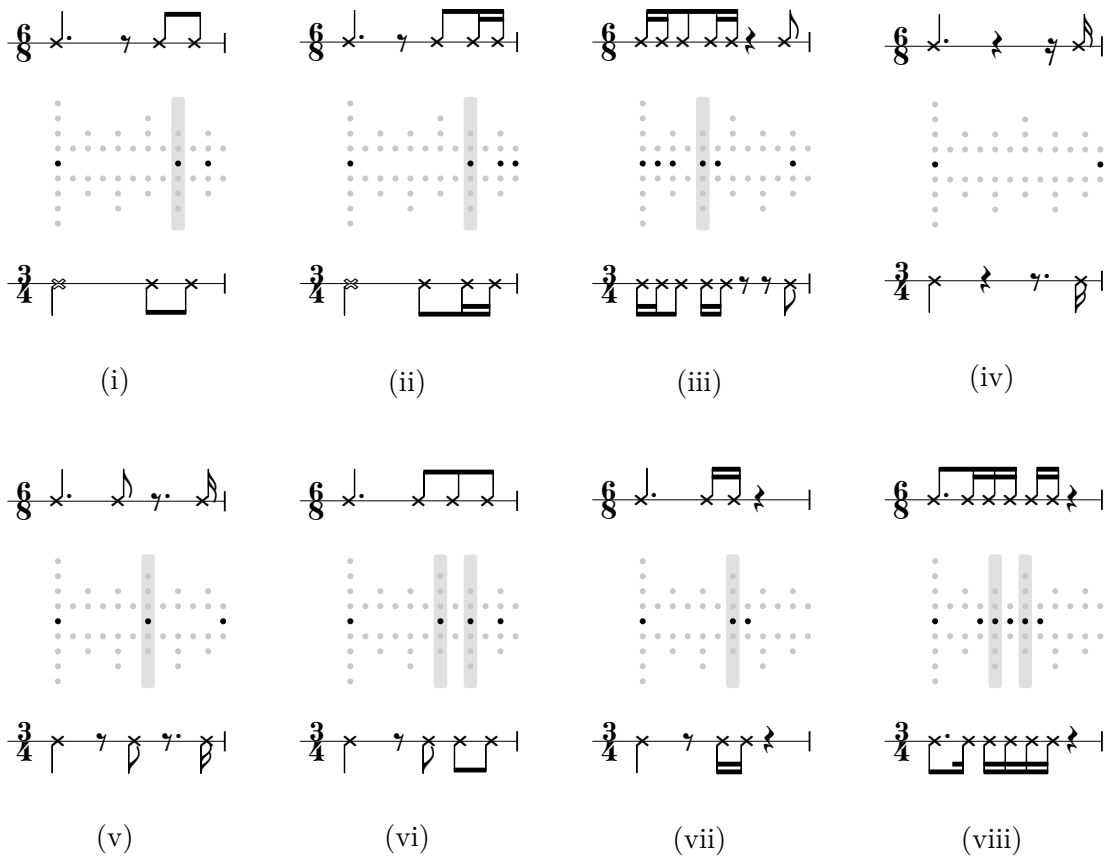


Figure 6.6: The full set of rhythms interpreted as non-anacrustic 3/4 rhythm by a model with a 3/4 inferential bias and a non-anacrustic 6/8 rhythm by a model with a 6/8 inferential bias. Each rhythm is shown in dot notation between a compound 6/8 metrical grid (above) and a 3/4 simple metrical grid (below). The grey bars highlight onsets that fall on beats with different theoretical saliencies in both interpretations. Score transcriptions of the rhythms in 6/8 and 3/4 are shown above and below the grids.

enculturation effect on memory performance. At a first sight, these results seem surprising in the light of earlier studies that did find effects of enculturation related to rhythmic organization of music (Hannon & Trehub, 2005a, 2005b; Hannon et al., 2012). However, it is possible that while rhythms are capable of eliciting effects of enculturation, such rhythms did not occur in the stimuli used by Demorest et al. (2016). Demorest et al. (2016) used a small set of stimuli that were not specifically selected to contain rhythms likely to elicit an effect of enculturation. The methodology applied in this simulation of enculturation is an example of how probabilistic models of rhythm perception can be employed to predict which rhythms are likely to elicit an effect of enculturation.

We hypothesized that fine-tuning of perception to the statistical properties of

musical rhythms in one's environment in a way that leads to a reduction of prediction-error in rhythms typical of one's environment leads to differences in the processing of meter. This idea is closely related to the notion of *cultural distance*—the degree to which pitch relations in a musical excerpt resemble the pitch relations typical to music from one's own culture—introduced recently by Demorest and Morrison (2016). The *cultural distance hypothesis* (Demorest & Morrison, 2016) states that cultural distance is predictive of various culturally dependent responses such as preference, tension, expectation, and memory. This hypothesis is supported by a series of studies where cultural distance of stimulus material was found to affect memory performance (for an extensive overview, see Morrison & Demorest, 2009). Demorest and Morrison (2016) propose that cultural distance could be measured using probabilistic models of melodic expectancy, such as IDyOM, that learn the statistical properties of music from a particular culture. Music that is culturally distant from the music such a model is trained on should be predicted less effectively than culturally familiar music. As such, in the context of a cross-cultural study, average information content—the degree to which observed events deviate from one's expectations—can be seen as an operational definition of cultural distance.

The model presented here can supplement predictions about *melodic* cultural distance as provided by existing probabilistic models, with predictions about *rhythmic* cultural distance. Cultural distance, as predicted by our probabilistic model, can then be read directly from Figures 6.4a and 6.4b. If the probabilistic aspects of rhythm learned by the presented model correspond to those implicitly learned by human listeners, then, according to the cultural distance hypothesis, rhythms in the top-right part of Figure 6.4a and 6.4b should be more difficult to remember for German listeners while rhythms in the bottom-right part of Figure 6.4a and 6.4b should be more difficult to remember for Chinese listeners.

Other culturally dependent responses mentioned by Demorest and Morrison (2016) such as, expectation, preference, and tension can be potentially linked to information content as well. Regarding expectation, information content is a direct consequence of predictive failure and has been shown to account well for human pitch expectations (Pearce, 2005; Hansen & Pearce, 2014). Regarding preference, perceived groove and experienced pleasure have been hypothesized to depend on the right balance between predictability and unpredictability (Witek et al., 2014). Furthermore, influential proposals have postulated close ties between expectation and both emotional responses to music (Huron, 2006) and musical meaning (Meyer, 1957). Regarding tension, melodic expectation has recently been linked to expressive performance, which in turn was linked to perceived tension (Gingras et al., 2016).

6.5.4 General discussion

While it is commonly assumed that the metrical accent of a beat, as derived from formal hierarchical descriptions of meter (Lerdahl & Jackendoff, 1983), is proportional to the probability of onsets at those beats, recent findings by Holzapfel (2015) and London et al. (2017) challenge this view. London et al. (2017) suggested that onset frequency need not be correlated with metrical accent for effective communication of meter. Instead, they argue, it is the recurrence and stability of rhythmic figures in the context of specific meters that may play a key role in the relation between rhythm and meter.

The results we presented show that models which take into account the preceding context of musical events, thus possessing the potential to learn the typical unfolding of multiple characteristic rhythmic patterns under different meters, are generally better at predicting rhythms and reconstructing annotated meters from note onsets alone. These findings, we would argue, provide further support for the idea the relationship between rhythm and meter is not only characterised by the distribution of note onsets, but also by characteristic rhythms and statistical properties of succession of interval between events.

The model we presented learns a generative model of rhythms from an annotated corpus. The supervised aspect of this approach challenges the cognitive plausibility of our model. Humans develop a feel for meter in their own culture without someone explicitly informing them about the "right" metrical interpretation. Nevertheless, situated exposure to rhythm almost always happens within a context containing an abundance of multi-sensory information related to the rhythmic practice. Within the music itself, other instruments, expressive timing and dynamics may provide strong metrical cues. In the environment, being rocked to music as an infant, participating in dancing or observing other people dance all contribute to the multi-sensory context by which rhythm perception is shaped. While not entirely putting concerns related to the supervised aspect of our approach to rest, metrical annotations in our training data can potentially be seen as capturing some of the information communicated in situated exposure to rhythms.

We have only considered event onset times in the present study while other musical aspects such as melodic repetition are known influence the perception of meter as well (Hannon, Snyder, Eerola, & Krumhansl, 2004). A full account of meter perception should take these aspects into account. Our model could be a good starting point for such an account: due to the implementation of the model in IDyOM, it is possible to link metrical viewpoints with melodic viewpoints and incorporate melodic aspects into the generative model.

Another limitation of the current model is its relatively simple representation of metrical structure. Time signatures fall short in capturing the structural complexity of perceived meter. The model treats metrical categories as independent

generative models and structural similarities between meters remain unexploited. The model is limited in its interpretation of rhythms into metrical categories by the categories observed in training data. In future work, we will seek to address these limitations by extending the model's representation of metrical structure.

The model introduced here represents an extension of previous work in probabilistic modelling of music (Conklin & Witten, 1995; Pearce, 2005). It is worth pointing out that the predictive mechanisms on which the model presented here is based, are domain independent (Pearce & Wiggins, 2004). The PPM* sequence prediction methods we employ can be applied to any domain that can be represented as structured sequences of symbols. Indeed, they were originally proposed in the field of text compression, but have proven to be useful in cognitive models of melodic expectation as well (Pearce, 2005; Pearce & Wiggins, 2012).

In summary, we have presented a computational probabilistic model meter perception, grounded in a predictive coding perspective of perception. The model has the potential to simulate musical expectations resulting from the perception of meter, shaped by previous exposure. The results show that the model can interpret simple rhythms in meters that agree with annotated time signatures and that it generates the hypothesized effects of enculturation. Simulations such as the ones presented here, can be used to generate theoretical predictions for cross-cultural studies of rhythm perception. Future research will determine the extent to which the learning processes implemented by our model capture aspects of those at work in human listeners.

Chapter 7

Statistical affordances for meter in makam and Western rhythms

7.1 Introduction

It is commonly claimed that the perception and appreciation of music are affected by internalized statistical patterns characteristic of musical styles that listeners are familiar with (Meyer, 1957; Huron, 2006; Temperley, 2007; Longuet-Higgins, 1979; Pearce, 2018). Cross-cultural studies have shown that the musical environment can indeed significantly influence music perception (for reviews, see Stevens, 2012; Patel & Demorest, 2013). These claims and findings have fueled interest in probabilistic generative approaches to modeling music perception (Temperley, 2007; Pearce, 2005). Such models can be used to model how music perception may adapt to statistical patterns in music (Temperley, 2007; Pearce, 2018; Morrison et al., 2019).

Using probabilistic generative models to simulate how perception is shaped by statistical patterns in the environment is consistent with predictive processing theories of perception and cognition. These theories posit that perception and perceptual learning (E. J. Gibson, 1963) both are symptoms of prediction-error minimization (respectively on fast and slow timescales) in a probabilistic generative model of sensations (Clark, 2013). Because of this unification, the shaping of perception by statistical patterns in the environment fits naturally in a predictive processing perspective on perception and cognition. This study adopts such a perspective by viewing generative models of rhythm perception as theories of the predictive models that listeners employ while perceiving a musical rhythm. If predictive processing theories are accurate, then the shaping of rhythm perception

by statistical patterns in the musical environment can be described maximizing the probability of rhythms encountered by a generative model of rhythms over time, based on rhythms encountered previously.

There is empirical and quantitative evidence that rhythm perception is influenced by previous activities and experiences (Hannon & Trehub, 2005b; Soley & Hannon, 2010; Hannon et al., 2012; Cameron, Bentley, & Grahn, 2015; Jacoby & McDermott, 2017; Polak et al., 2018). However, along which dimensions rhythm perception is shaped, whether statistical patterns in rhythms play a role in bringing this shaping about, and if so, which patterns are most important remain interesting questions. Both statistical patterns present in the musical environment and the statistical learning mechanisms of listeners are relevant to these questions.

In the current study, we approach these questions by considering the effect that the musical environment, the learning mechanisms of listeners, and the stylistic properties of a rhythm may have on the ability to perceive meter in a rhythm. These three factors give rise to the concept of statistical affordances for meter which we describe in Section 7.2.2. We represent musical environments by three empirical samples of rhythms (see Section 7.2.1), two of which contain rhythms from German and Dutch folk melodies, and one of which contains rhythms from Turkish makam music. Learning mechanisms are modeled using two probabilistic generative models of rhythm perception. One, which we call the *classical model* Section 7.3.2), is based closely on a probabilistic rhythm model described by Temperley (2007), and one, which we call the *enculturation model* (Section 7.3.4), is based on a model described by Van der Weij et al. (2017 [Chapter 6]). The classical model is consistent with what we call classical theories of meter (Section 7.3.1). The enculturation model is consistent with theories of meter that posit a greater influence of the musical environment on rhythm perception than classical theories do (Section 7.3.3). We compare three variants of the enculturation model that are constrained to various degrees in the length of statistical patterns they are sensitive to. Both models are described in detail in Chapter 5. All models and their variants used in this study can be ordered in a hierarchy of sensitivity to statistical patterns, described in Section 7.3.5.

After describing the concepts of musical environments and statistical affordances for meter in Section 7.2 and the models in Section 7.3, we formulate a set of research questions in Section 7.4 and explain how we investigate these questions in a set of three experiments. The first two experiments involve model simulations of listeners with long-term exposure to a certain musical environment. Here, we assess how different models of listeners are shaped by different musical environments and how this affects the ability of these models to predict inter-onset intervals, based on inferences about the underlying meter, in rhythms drawn from either the same or a different musical environment. The general methodology applied in the first two experiments is explained in Section 7.5. In the third experiment, we

consider the statistical properties of the rhythm samples directly and compare two different representations of metrical context of onsets in a rhythm: metrical salience, and phase (position in the metrical cycle).

7.2 General concepts

7.2.1 Musical environments

Rhythm perception takes place in a context of rich and multi-modal sensations, commonly accompanied by forms of movement, such as dancing, or active participation in the music-making (Trehub et al., 2015). While these aspects plausibly play a role in the shaping of rhythm perception, this study focuses only on the possible effects of passive *exposure* to rhythms. Therefore, we describe the musical environment conceptually as a probability distribution that describes the probability with which a listener encounters different rhythms. This listener-specific distribution depends on aspects like the listener's social and cultural context, as well as their actions and preferences.

However, instead of attempting to obtain such listener-specific distributions, we focus on coarse-grained distributions of rhythms that may conceivably represent a certain musical environment. To represent these distributions, we use music corpora containing relatively large amounts of music, categorized by style or geographical region of origin. In the current study, we draw independent samples from three different corpora: one containing Dutch folk melodies, one containing German folk melodies, and one containing Turkish makam music. These samples represent three musical environments, two of which (the German and Dutch folk melodies) may be expected to be similar, and one of which may be expected to be different from the other two (the Turkish makam music). The samples can be classified as belonging to two different musical idioms: the folk melodies as Western tonal music, and the Turkish music as Middle-Eastern makam music (see Section 7.5.6).

Two further limitations apply: we consider only the rhythms created by the timing of note onsets in monophonic melodies, specifically the intervals between note onsets, represented by a discrete, symbolic, and score-like representation. We therefore refrain from making statements about the influence of expressive timing (notes that are timed slightly early or late, for example for expressive reasons) or tempo (such as rubato, the natural speeding up and slowing down of rhythms for expressive reasons). Motivations for these simplifications are partly practical—they reflect the format in which empirical datasets representing music from different musical idioms are available—and partly theoretical—they help to narrow the scope of our study.

7.2.2 Statistical affordances for meter

Meter is a perceived recurring pattern of strong and weak accents that is a prerequisite for moving to a rhythm and playing music in synchrony. It has been described in a multitude of ways by different authors. Some consider it an abstract mental phenomenon (Longuet-Higgins & Lee, 1984; Lerdahl & Jackendoff, 1983), emphasizing its perceptual nature. Others view it as coupled oscillation (Large & Kolen, 1994; McAuley, 1995), emphasizing its dynamic, in-time character. Yet others describe it as skilled active behavior (London, 2012), emphasizing the role of experience and training. These perspectives were discussed in more detail in Chapter 2 of this thesis.

Traits of listeners, such as the ability to perceive meter, that enable them to participate in musical activities have been considered from an evolutionary perspective (Honing, Ten Cate, Peretz, & Trehub, 2015). However, these traits are also shaped by the diverse cultural environments (Trehub et al., 2015) in which listeners are embedded. Furthermore, due to the musical diversity of different cultural environments, rhythms prevalent in one cultural environment may offer more (or different) opportunities to perceive meter to listeners embedded in that environment than to other listeners (Stobart & Cross, 2000; Hannon & Trehub, 2005a, 2005b). To emphasize that the ability to perceive meter depends on the listener, the style of a rhythm, and the musical environment in which the listener is embedded, we introduce the concept of *statistical affordances for meter*.

An *affordance* may be described as a relation between an animal and a property of a situation that affords a certain behavior (Chemero, 2003). For example, Chemero (2003) notes that the affordance of "eating" is provided by an apple only to animals capable of eating and digesting apples. Affordances can be perceived: an animal capable of eating and digesting apples can perceive the edibility of apples. Similarly, we might say that the affordance to entrain metrically to a rhythm may be perceived by listeners with the cognitive capacities required for detecting certain kinds of regularity in rhythms. In this view, meter is a perceived affordance that can be characterized as a relation between properties of listeners and properties of rhythms.

However, since the cognitive capacities of listeners appear to be shaped by patterns and regularities in their musical environment, the musical environment and the learning mechanisms of listeners also play a role in whether the affordance of entraining metrically to a rhythm is available. A *statistical affordance for meter* is available when internalized rhythmic patterns and regularities enable a listener to perceive meter in a given rhythm. Whether a statistical affordance for meter is available depends on three factors: (1) the characteristics of a rhythm (e.g., its style or idiom), (2) the statistical patterns in the musical environment that have shaped the listener, and (3) how listeners are shaped by these patterns

and regularities—that is, their learning mechanisms, or sensitivity to statistical patterns. In summary, statistical affordances for meter can be perceived in rhythms belonging to a certain style or idiom by *enculturated* listeners.

For example, in Western classical music, the frequency with which onsets occur in different positions in the metrical cycle can serve as a cue for meter (Temperley, 2007, 2010; Palmer & Krumhansl, 1990). This cue for meter is available to listeners who internalize these frequencies through long-term exposure to Western music. Such listeners may perceive statistical affordances for meter in Western classical music courtesy of their sensitivity to this statistical pattern and their long-term exposure to Western classical music.

7.3 Models

7.3.1 Classical theories of meter

Classical theories of meter describe meter as a multileveled hierarchy of beats (Lerdahl & Jackendoff, 1983), or as trees generated by grammars associated with different meters that recursively subdivide metrical intervals (Longuet-Higgins & Lee, 1984). These hierarchies create a recurring pattern of beats with alternating levels of metrical salience (also referred to as beat strength, such that metrically salient beats are strong beats). The metrical salience of a beat is determined by the number of beats that align with its onset (Lerdahl & Jackendoff, 1983), or highest metrical level that the beat initiates (Longuet-Higgins & Lee, 1984). For example, the bar-level beats of a 6/8 meter are separated by the duration of six eighth notes, which is subdivided by two into a new metrical level with a period of three eighth notes. Finally, this level is subdivided to produce the lowest metrical level containing three beats per mid-level beat, each of which is separated by the duration of one eighth note. A 3/4 meter has the same inter-beat interval at the bar level as a 6/8 meter but entails a different hierarchy: the bar-level is subdivided first by three and then by two.

Classical theories of meter are consistent in their requirement that in order to establish the perception of a metrical hierarchy, onsets and accents must reinforce this hierarchy by accentuating metrically strong beats: Lerdahl and Jackendoff (1983) say that metrical interpretations in which strong beats at each metrical level are stressed (by events in the rhythm) are to be preferred. Longuet-Higgins and Lee (1984) defined meter as a grammar and rhythm as the structures generated by that grammar. Meters that minimize syncopation by ensuring that if onsets occur on metrically weak beats, they are followed by onsets on metrically strong beats, are preferred according to this theory. These requirements leave relatively little room for idiomatic or stylistic influences on the structure of rhythms.

7.3.2 The classical model

The classical model is an adaptation of a rhythm perception model proposed by Temperley (2007). It posits the same assumptions concerning the way rhythm is constrained by meter as Temperley's model but omits aspects related to tempo and timing. An adaptation of Temperley's model that does incorporate these aspects and is very close to Temperley's original was described in Chapter 4. Meter, in these models, is represented by a multileveled hierarchy of beats, subject to well-formedness constraints outlined by Lerdahl and Jackendoff (1983). Based on the meter, the metrical salience of each position in which an onset can occur can be calculated. The classical model assumes that the probability of whether a note onset occurs at any of these positions is explained fully, if the meter were known, by the metrical salience of that position. These onset probabilities are described by a set of parameters that Temperley calls the *note-beat profile*, as they describe the probability of a note occurring at a particular beat. The note-beat profile is estimated from empirical rhythm samples and thus reflects the statistical patterns that listeners internalize from prior exposure to rhythms according to the classical model.

While the model does not *a priori* require that onsets are most likely to occur on metrically salient positions, the model is primarily compatible with classical theories of meter: the only way in which it can infer meter from a rhythm is if onsets align with different levels of metrical salience in a consistent, context-independent way. Temperley argues that the alignment of onsets with strong beats is required to clearly establish meter in a listener's mind. Styles in which rhythmic onsets align less strongly with metrically salient beats, for example by means of syncopation, create more metrical ambiguity, which, Temperley argues, must be compensated for by stricter adherence to tempo (tolerance for tempo changes is a parameter of Temperley's original model but is not included in the classical model since tempo and expressive timing aspects are omitted).

Its small number of parameters endows Temperley's model, and the classical model, with the desirable quality of making narrow predictions about rhythm perception. However, it simultaneously constrains the model's ability to adapt to diverse rhythm distributions. Although the model does have the ability to learn some statistical patterns from rhythms, it may be characterized as conservative with regard to the plasticity of rhythm perception. This causes the structure of rhythms that afford listeners (as simulated by the model) to perceive meter to be strongly constrained by the hierarchical structure of meter, and less by the statistical patterns in the rhythms to which listeners have been previously exposed.

7.3.3 Alternative theories of meter

The frequency with which onsets occur at different positions in the metrical cycle in a given empirical sample of rhythms can be made visible by histograms depicting these frequencies for each metrical position. Palmer and Krumhansl (1990) constructed such histograms using music samples from different styles and periods of Western classical music and found the results to be constrained more by the hierarchical structure of meter, as described by classical theories of meter, than by the style or period of a rhythm. They concluded from this that patterns of metrical salience serve as reliable cues for meter, that listeners may learn patterns of metrical salience from extensive exposure to music but also that the probabilities with which onsets occur at different positions in the metrical cycle are constrained primarily by the metrical salience patterns predicted by classical theories of meter, despite having investigated only Western classical music.

Simultaneously, classical theories of meter were inspired primarily by the study and analysis of Western classical music, and were developed by authors who, as they themselves readily acknowledge (Lerdahl & Jackendoff, 1983; Longuet-Higgins, 1979), are primarily familiar with this musical idiom. The extent to which the constraints that these theories impose on meter perception generalize to rhythms and listeners familiar with styles beyond Western classical music has been questioned. In particular, the requirement that a rhythm must, in order to reliably establish a meter, place onsets and accents primarily on metrically salient positions has been argued not to hold universally.

Iyer (1998, p. 44), for example, argued that "one should not regard the global musical preponderance of 'syncopation' (off-beat accents) as a vast set of exceptions to the 'normal' accentual rules of meter but rather as convincing counterexamples to such proposed accentual rules." The normal accentual rules for meter here refer to the principles posited by classical theories of meter, according to which syncopation is a deviation from the norm and according to which the pattern of strong and weak beats prescribes where onsets should (predominantly) occur.

An alternative view is that there are ways complementary to reinforcing metrically salient beats with rhythmic accents by which meter can be established. These complementary ways may rely on learned associations between rhythmic patterns and meters, as has been suggested by London (2004, 2012). As long as certain rhythmic patterns occur consistently in a certain orientation to a metrical cycle, listeners with the appropriate experience and training may be able to perceive meter in rhythms that evoke these learned associations. Syncopations, in this view, are not deviations from a norm but a property of how onsets in a rhythm align with the metrical accents of a meter. Crucially, in this view metrical salience is not seen as identical to expectancy but as a phenomenological aspect of the perception of meter which may or may not coincide with expectations for events. This view

demands a certain skillfulness of listeners for them to perceive the (culturally appropriate) meter which involves sensitivity to more complex statistical patterns in rhythms than only the frequency with which onsets occur at different levels of metrical salience. [1]

There is empirical evidence suggesting that meter can be established when onsets do not occur predominantly on metrically strong beats. London et al. (2017) analyzed recordings of three Malian drum ensemble pieces. They found that the pieces, importantly, do suggest a metrical cycle and a regular beat, yielding a metrical organization comparable to a 12/8 time signature. However, they also found that onsets in these rhythms occur more frequently in off-beat positions than in on-beat positions (apart from the downbeat). These findings are at odds with the predictions of classical theories of meter, and also at odds with the theory that the probability of onsets in a rhythm depends primarily on metrical salience (Palmer & Krumhansl, 1990; Temperley, 2007). Based on these results, London et al. (2017) propose that metrical entrainment can be supported by statistical associations between rhythmic patterns and metrical orientations, which are learned through practice and experience.

Furthermore, Holzapfel (2015) presented evidence suggesting that metrical categories can be inferred on the basis of drum-stroke patterns associated with different rhythmic modes. Holzapfel investigated rhythms derived from a corpus of makam music (which is also used in this study) and used drum-stroke patterns associated with different *usuls* (rhythmic modes) to classify the usul of a rhythm, represented by the way they onsets in the rhythm are distributed over different positions in the metrical cycle. Holzapfel found that this could be done successfully, showing that the frequency with which onsets occur in different metrical positions in makam music is influenced by the usul pattern, in addition to meter.

7.3.4 The enculturation model

The enculturation model is based on a model proposed by Van der Weij et al. (2017 [Chapter 6]). A detailed and technical description of this model is given in Chapter 5. The model learns associations between *metrical categories* and rhythmic patterns. What is regarded as a metrical category by the enculturation model is to some extent arbitrary. In principle clave patterns, timelines or usuls could all serve as metrical categories. The only requirement of meter is that it can

[1] It should be noted that (purely perceptual) statistical learning is not the only way in which enculturated meter perception can be established. Agawu (2006, p. 18), for example, noted the importance of the integration of rhythm in dance in different African communities to arrive at a "culturally sanctioned" understanding of rhythm. Such cultural customs may give rise to patterns in symbolic representations of rhythms that are detectible by statistical learning mechanisms.

be associated with a metrical cycle. In the current study, time signatures are used as metrical categories, and the duration of metrical cycles corresponds to a bar.

A *metrical interpretation* of a rhythm is defined by a metrical category and by a representation of how the metrical cycle aligns with the sequence of inter-onset intervals (that is, where the first onset occurs in the metrical cycle). Given a metrical interpretation, a sequence of inter-onset intervals can be represented as a sequence of *downbeat distances*, which is a representation that incorporates the position of the second of the two onsets in an inter-onset interval relative to the metrical cycle. The way in which inter-onset intervals are represented, given a metrical interpretation, thus depends on the duration of the metrical cycle and the way in which the metrical cycle aligns with the rhythm. Downbeat distances can therefore be seen as *metrical fingerprints* that are left by inter-onset intervals given a metrical interpretation.

Since the enculturation model learns associations between metrical categories and patterns of onsets, it is sensitive to more complex statistical patterns and supports more nuanced associations between rhythm and meter than the classical model. Patterns of metrical fingerprints are learned by a sequence model that describes probability distributions of the possible metrical fingerprints to follow a sequence of preceding metrical fingerprints, given an underlying meter. This supports the learning of complex statistical associations between metrical categories and rhythmic patterns. It can be likened to how sequence models of melodies can expose differences between statistical properties of musical idioms (Pearce, 2018).

The length of the patterns to which the enculturation model is sensitive can be controlled. In the current study, we test variants of the model sensitive to patterns of up to two (*short* patterns), five (*long* patterns), or an *unbounded* number of metrical fingerprints. While a pair of subsequent note onsets create nothing but an inter-onset interval, two metrical fingerprints can arguably be called a rhythmic *pattern* since they represent not just an interval but encode locations of the onset in a metrical cycle (their metrical context).

7.3.5 Hierarchy of sensitivity

The classical model and the three variants of the enculturation model can be ordered in terms of the complexity of statistical patterns to which they are sensitive, creating a *hierarchy of sensitivity*. Models sensitive to simple statistical patterns are placed at lower levels than models sensitive complex statistical patterns. Crucially, a model at any particular level is also sensitive to the statistical patterns that models at lower levels are sensitive to.

Table 7.1 shows the classical model and three variants of the enculturation model ordered in this manner. The table also shows the kind of representation of metrical

Table 7.1: A hierarchy of sensitivity to statistical patterns in musical environments. The classical model and the three variables of the enculturation model are arranged hierarchically by the complexity of statistical patterns to which they are sensitive.

Level	Model	Metrical context	Maximum pattern length
0	Classical	Metrical salience	N/A
1	Enculturation	Downbeat distance	Two (short)
2			Five (long)
3			Unbounded

context used by the model and the length of rhythmic patterns that it is sensitive to. The classical model, which is sensitive only to the note-beat profile of a sample of rhythms, is placed at level zero of the hierarchy. The enculturation model sensitive to rhythmic patterns of two metrical fingerprints (downbeat distances) occupies level one. This model can learn the note-beat profiles from a sample of rhythms but is additionally sensitive to the metrical fingerprints of two subsequent notes. Above this model, at levels two and three, are variants of the enculturation model sensitive to patterns of five metrical and an unbounded number of fingerprints.

7.4 Research questions

We investigate how the learning mechanisms of a listener (as simulated by a model) and their musical environment (the *source* sample) affect the availability of statistical affordances for meter (as measured by cross-entropy reduction) in rhythms drawn from either the same or a different musical environment (the *target* sample) by comparing both between models at successive levels of the hierarchy of sensitivity, between *within-sample*, *within-idiom* and *across-idiom* simulations, and between different musical environments.

The source sample refers to the rhythm sample from which a model learns and the target sample refers to the rhythm sample on which a model is evaluated. In within-sample and within-idiom simulations, the idiom of the source and target samples is the same while in across-idiom simulations the source and target samples reflect different musical idioms. These concepts are explained in more detail in Sections 7.5.1 and 7.5.2. How we measure the availability of statistical affordances for meter as cross-entropy reduction is explained in Section 7.5.3.

In the experiments, we assess three kinds of effects:

1. In within-sample simulations, we compare models at successive levels of the hierarchy of sensitivity to investigate whether sensitivity to more complex

statistical patterns increases the availability of statistical affordances for meter. This assesses the effect of model.

2. In within-sample simulations, we test whether patterns to which a model is sensitive make affordances for meter available to the same degree in different musical environments. This assesses the interaction between musical environment and model.
3. In within-sample, within-idiom, and across-idiom simulations, we test whether patterns learned from one musical environment make affordances for meter available in rhythms from similar and dissimilar musical environments. This assess the interaction between simulation type and model.

In Experiment 1, we investigate whether there are statistical affordances for meter that rely on sensitivity to long or unbounded-length patterns by investigating effects 1 and 2. This experiment involves only the enculturation model, for which we can vary the maximum length of patterns to which it is sensitive. In Experiment 2, we investigate two questions: do statistical affordances for meter rely on similar or different statistical patterns in different musical idioms and do statistical affordances for meter in different musical environments require different kinds of sensitivity to statistical patterns? Here, we investigate effects 2 and 3 involving the classical model and the enculturation model.

If we find in Experiment 1 that there are statistical affordances for meter available exclusively to models sensitive to long patterns or patterns of unbounded length, that suggests that such patterns can in principle serve as cues for meter. This would mean that there are other statistical patterns, besides the frequency with which onsets occur at different positions in the metrical cycle (Palmer & Krumhansl, 1990), or at different levels of metrical salience (Temperley, 2007), that can support metrical inference, at least for listeners with sufficient prior exposure to rhythms prevalent in a certain musical environment.

Classical theories of meter do not strictly rule out the existence of such patterns but do hold that these patterns are not necessary for meter perception. That the distribution of onsets across different positions can be used to infer the meter, however, is posited to be a necessary condition. Therefore, if classical theories are accurate and if rhythms from the studied musical environments are equally metrically unambiguous, the classical model should not perform disproportionally worse in across-idiom simulations in Experiment 2 compared to the enculturation model.

If, on the other hand, the alignment of onsets with metrically strong beats is a cue for meter especially in rhythms from the Western musical idiom, then in Experiment 2 the classical model may perform better in within-sample and within-idiom simulations involving Western rhythms than in simulations involving the makam target sample compared to the enculturation model. Furthermore, if the perception of meter relies partly on internalized statistical patterns specific

to different musical environments, the classical model may be expected to show a smaller performance difference between within- and across-idiom simulations in Experiment 2 than the enculturation model, which is able to exploit these idiom-specific statistical patterns.

In terms of the constraints that meter perception poses on rhythms in the musical environment, classical theories posit that rhythms are constrained relatively strongly by their meter, since a rhythm must reinforce the metrical hierarchy by accentuating metrically strong beats. Alternatively, if musical environments can influence rhythm perception, then listeners with enculturated rhythm perception in turn influence the musical environment, resulting in complex dynamics of cultural transmission (Ravignani, Thompson, Grossi, Delgado, & Kirby, 2018; Thompson et al., 2016). Through these dynamics, different musical environments may emerge in which metrical entrainment relies partly on culture-specific statistical patterns. In Experiment 2, we would then expect to find statistical affordances for meter that rely on exposure to idiom-specific statistical patterns. That is, we would expect to find that statistical patterns learned from a source sample make statistical affordances available in within-sample and within-idiom simulations but not in across-idiom simulations.

Finally, if in Experiment 2 we find performance differences between the classical model and the enculturation model sensitive to short patterns, these differences may result from the enculturation model's sensitivity to statistical patterns of *multiple* (two) metrical fingerprints, or only from the more detailed representation of metrical context as downbeat distances rather than metrical salience. In order to investigate which of these properties is likely to be responsible for any observed performance differences, we investigate the samples directly in Experiment 3 and compare the degree to which their statistical properties can be described by metrical salience representations or by phase representations. These experiments are comparable to those performed by Temperley (2010) to investigate which statistical principles best describe "common-practice rhythm". The phase representation is more fine-grained than metrical salience and similar to the metrical fingerprints used by the enculturation model. If we find that the phase representation is significantly better at describing the statistical properties of the different samples than metrical salience, that suggests that performance differences between the classical model and the enculturation model are attributable to their representation of metrical context while finding few differences between the two representations would suggest that the enculturation model's sensitivity to patterns of multiple metrical fingerprints is more important.

7.5 General methodology

7.5.1 Model training and rhythm spaces

Probabilistic generative models define probability distributions of what they are designed to model, which in our case is rhythms. The models can be made to approximate specific distributions by estimating their parameters from empirical samples of rhythms drawn from these distributions. The process of parameter estimation is referred to as *training* a model on an empirical sample.

When training the models used in this study, information about the metrical interpretation of each rhythm is provided. After training, a rhythm model defines a probability distribution of rhythms that is learned from the training sample. The probability of each rhythm (without information about metrical interpretation) according to this distribution is called its *model evidence*. The model training strategy that we use maximizes the total model evidence of the training sample. Equivalently, the procedure causes the model to maximize the model evidence of rhythms it expects to encounter, thereby minimizing prediction error. The actual model evidence of encountered rhythms depends on how well the training sample represents the probability with which rhythms are encountered in a given musical environment, and on how well the model learns the relevant statistical patterns in the rhythms.

In Chapter 5, we described the concept of *rhythm spaces*: a finite set of rhythms, defined by an inter-onset interval domain and a sequence length, over which a probability distribution can be defined by a generative model. We described how the parameters of the classical model and the enculturation model can be derived from a sample of rhythms, which consists of rhythms that occur in the rhythm space. In the present study, we use a symbolic rhythm space consisting of all rhythms that can be represented by a sequence of twenty-nine inter-onset intervals (this choice is motivated in Appendix B), using a resolution of sixteenth notes. The inter-onset interval domain is based on the unique inter-onset intervals observed in empirical data. In the current study, the inter-onset interval domain is $\{1, 2, 3, 4, 5, 6, 7, 8, 10, 12, 14, 16, 18, 20, 24\}$. We use three music corpora (see Section 7.5.6) to derive samples from different empirical probability distributions of this rhythm space and use the techniques described in Chapter 5 to train the models.

7.5.2 Within- and across-idiom simulations

Different empirical samples are used to represent different musical environments. We call the sample on which a model is trained, which represents long-term

exposure to a specific musical environment, the *source sample*. The trained model defines a probability distribution of rhythms that is the model's approximation of the distribution underlying the source sample. The sample on which a model is evaluated is called the *target sample*.

The source sample may be drawn from the same dataset as the target sample, or from a different dataset. We distinguish between three simulation types, depending on the datasets from which the source and target sample have been drawn. When drawn from the same dataset, we speak of *within-sample* simulations, when drawn from different datasets representing the same musical idiom, we speak of *within-idiom* simulations, and when drawn from different datasets representing different musical idioms, we speak of *across-idiom* simulations.

7.5.3 Measuring statistical affordances for meter as predictive success

The goal of the evaluation procedure is to measure how closely a probability distribution of rhythms learned from a source sample by a given model resembles the empirical distribution underlying the target sample. We measure this by estimating the *cross-entropy* of distribution learned from the source sample relative to the target sample. This may be understood intuitively as measuring the overall "surprisingness" of the target sample to the model trained on the source sample. Cross-entropy-based evaluation methods have been used previously in probabilistic approaches to music cognition (Temperley, 2010; Pearce & Wiggins, 2004; Conklin & Witten, 1995), and are popular for evaluating statistical models of language (Jurafsky & Martin, 2000).

Cross-entropy estimation involves generating predictions for each inter-onset interval in each rhythm, based on the preceding inter-onset intervals, and evaluating the accuracy of the prediction (see Section 7.5.5). These predictions rely not on the surface pattern of inter-onset intervals but on the meter that has been inferred from these inter-onset intervals. Predictive success thus measures the degree to which the model successfully uses metrical inference to predict the unfolding of a sequence of inter-onset intervals. We therefore interpret this measure to reflect the degree to which statistical affordances for meter are available in rhythms from a particular target sample, to a particular model, trained on a particular source sample.

There are two main motivations for using cross-entropy as a measure for the availability of statistical affordances for meter instead of something that directly considers the specific meters inferred by the models. First, we consider meter to be part of the generative model that is brought to bear by a listener to predict the temporal unfolding of rhythms (Van der Weij et al., 2017 [Chapter 6]). Cross-

entropy is a measure of predictive success, which is what listeners, according to predictive processing accounts of cognition, optimize for. Second, especially in across-idiom simulations where models are evaluated on "unfamiliar" rhythms, it may be that the metrical interpretation inferred by the model is not exactly the same as the metrical interpretation observed in the corpus but that it still enables the model to successfully predict inter-onset intervals. By using cross-entropy, we avoid the need to label one metrical interpretation as the "correct" interpretation and also avoid problems in judging how appropriate a metrical interpretation is given the "correct" interpretation (see Temperley, 2004).

An important caveat of using cross-entropy is that its lower bound is the entropy of the target distribution. The entropy of a probability distribution of rhythms may be seen as the inherent complexity of these rhythms. While this complexity cannot be measured directly, cross-entropy results are sensitive to it. Therefore, whenever we draw comparisons between different source or target distributions, we do consider the performance differences between these samples only in relation to those of another model. That is, when comparing different target or source samples (including different simulation types such as within- and across-idiom), we only interpret interactions between the effects of model and the target or source sample on cross-entropy.

7.5.4 Metrical inference versus sequential prediction

A caveat of the enculturation model's use of sequence models is that these models have a very large number of parameters. A sequence model can, in theory, and given enough training data, approximate any probability distribution of sequences. [2] However, how readily a distribution of sequences is learned by a sequence model—that is, the amount of training data required to obtain a reliable estimate of its parameters—depends on whether the sequence representation describes properties to which the sequence-distribution is variant. For example, the distribution of pitch sequences in melodies is somewhat invariant to absolute pitch of the first note, therefore pitch intervals or pitch classes are examples of representations that effectively capture relevant properties of pitch sequences. Another aspect that affects the efficiency with which distributions can be learned is on which latent variables sequences are assumed to depend. In a melody, for example, the distribution of pitch-class sequences is likely to depend strongly on a melody's

[2] An upper bound on the number of parameters of a PPM model, which underlies sequential modeling the enculturation model, is the number of unique sequences of lengths up to the length of the sequence. The number of unique contexts of length n, given an alphabet Σ (the set of symbols that occur in the sequence), is $|\Sigma|^n$, which, in many cases, is an astronomically large number. The upper bound on the number of parameters of a PPM model for sequences up to length l is $\sum_{n=0}^{l}|\Sigma|^n$.

tonal center. A model that conditions such sequences on their underlying key is therefore likely to do well in approximating distributions of melodies.

In light of the above caveat, in Experiment 1, we compare the prediction performance of the enculturation model with that of another model that we call the *IOI model*. The IOI model predicts inter-onset interval sequences directly, instead of generatively by assuming an underlying meter. This comparison enables us to assess whether the enculturation model is able to learn distributions of rhythms more readily than the IOI model due to its inferring of meter.

7.5.5 Evaluation measures

We primarily use estimated cross-entropy as an evaluation measure, but in Experiment 1 we additionally verify whether, in within-sample evaluations, lower cross-entropy corresponds to a greater agreement between the meter inferred by the enculturation model and the meter by measuring metrical interpretation performance.

7.5.5.1 Metrical interpretation performance

We define metrical interpretation performance as the posterior probability that the enculturation model assigns to the metrical interpretation of a rhythm that agrees with the one observed in the corpus after having processed the last inter-onset interval. Metrical interpretation performance is not considered in Experiment 2, where we compare the enculturation model to the classical model. The classical model estimates the parameters of variables that determine the probability of pickup intervals from the training data, while the enculturation model assumes a uniform distribution (see Chapter 5). There is significantly less variety in pickup intervals in the Turkish sample compared to the Dutch and German samples, which confounds the interpretation of metrical interpretation performance of the classical model and the enculturation model across different source samples.

7.5.5.2 Cross-entropy

Cross-entropy is proportional to the dissimilarity between two probability distributions—for example, a learned distribution of rhythms and an empirical distribution. Let R_t be a random variable describing the distribution of rhythms underlying a target sample, and R_s a random variable describing the distribution of rhythms underlying a source sample. In general, let a model's estimate of the distribution of a random variable X be described by the variable $\tilde{X}$. The model

evidence of a rhythm is then given by $P(\tilde{R}_s = r)$. The cross-entropy of $\tilde{R}_s$ with respect to R is given by

$$H(R_s, \tilde{R}_t) = -\sum_{r \in D_{R_t}} P(R_t = r) \log_2 P(\tilde{R}_s = r). \tag{7.1}$$

The quantity $-\log_2 P(\tilde{R}_s = r)$ corresponds to the information content, measured in bits, carried by the observation of rhythm r to an observer represented by the model. The cross-entropy is the average amount of information the observer is expected to receive by encountering rhythms with probabilities described by the actual distribution of rhythms. This actual distribution is unknown, but we do have access to samples drawn from this distribution. Let the multiset S_t represent a sample from the target distribution. If we assume that the stochastic process that generates rhythms is stationary and ergodic (Cover & Thomas, 2006), then the cross-entropy is approximated by

$$\tilde{H}(S_t, \tilde{R}_s) = -\frac{1}{|S_t|} \sum_{r \in S_t} \log_2 P(\tilde{R}_s = r), \tag{7.2}$$

which corresponds to the average per-rhythm information content. We refer to $\tilde{H}(S_t, \tilde{R}_s)$ as the estimated cross-entropy of a model with respect to a target sample.

Estimated cross-entropy is minimized if the target distribution and the model's estimate of the source distribution of rhythms are identical. The entropy of a rhythm distribution represents the irreducible uncertainty about which rhythm will be encountered next, when rhythms are drawn from the distribution. Cross-entropy is maximized if the model's estimate of the true distribution is a uniform distribution: a distribution that assigns equal probability to each possible outcome. Thus, while cross-entropy is proportional to dissimilarity of two distributions, its lower bound depends on the entropy of the target distribution.

In within-sample simulations, $\tilde{H}(S_t, \tilde{R}_s)$ estimates the cross-entropy of the model's estimate of the source with respect to the source distribution. In across-sample simulations, it estimates the cross-entropy the model's estimate of the source distribution with respect to the target distribution. When the source and target distributions are similar, cross-entropy estimates in within- and across-sample simulations should be similar.

In our simulation results, we report the estimated per-inter-onset-interval entropy, which corresponds to $\tilde{H}(S_t, \tilde{R}_s)/29$ (where 29 is the fixed length of a rhythm). This measures the mean cross-entropy per observation of I (an inter-onset interval). The upper bound on this cross-entropy is the entropy of a uniform distribution of

I. Since the size of the inter-onset interval alphabet used in our simulations is 15, the upper bound estimated cross-entropy is $-\log_2 \frac{1}{15} \approx 3.91$.

7.5.5.3 Within- and across-sample cross-validation

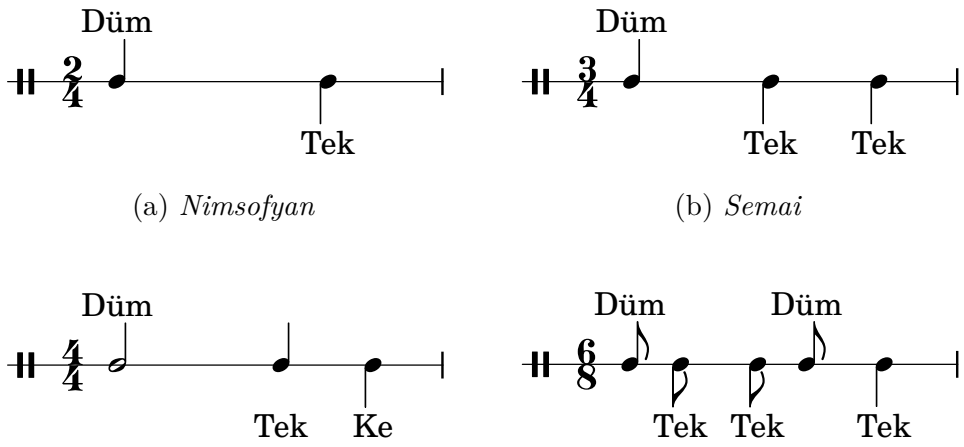
In order to estimate cross-entropy and metrical inference performance, a testing sample that is independent of the training sample is required. If the source and target sample are independent, the target sample can be used for this. However, in within-sample simulations, the source and target sample are identical. Splitting the sample into two non-overlapping independent samples would yield either a training sample that is too small to reliably estimate model parameters from or a testing sample too small to obtain a good estimate of performance. Therefore, we employ a generalization of a technique known as *ten-fold cross-validation* to estimate model performance. The generalization enables cross-validation to be applied to different source and target samples, ensuring that simulations results are comparable between within- and across-sample simulations.

To perform ten-fold cross-validation, the source and target samples are randomly split into ten approximately equal-sized samples, called folds. Ten pairs of training and testing samples are constructed by combining nine folds from the source sample into a training sample and using one fold from the target sample for testing. When the source and target sample are identical, the fold used for testing is the remaining fold that is not used in the training sample. A model is trained and tested ten times, each time using a different pair of testing and training samples. When the source and target samples are identical, this procedure is identical to ten-fold cross-validation. This method enables us to iteratively test a model on all available data, while using most of the data for training in each iteration.

The samples contain rhythms in different meters and we expect meter to play a significant role in the performance of models. We therefore constrain the relative number of rhythms in each meter in training and test sample to be the same as in the whole sample, a technique known as *stratified cross-validation*. Finally, it should be kept in mind that cross-validation inevitably under-estimates variance due to the sampling of training data because the training samples are overlapping and therefore not independent (Dietterich, 1998).

7.5.6 Materials

We derived three rhythm samples from Turkish makam music, and German and Dutch folk songs. Dutch and German folk songs are treated here as Western tonal music, and part of the Western musical idiom, while the Turkish makam



(a) *Nimsofyan* (b) *Semai*

(c) *Sofyan* (d) *Yürüksemai*

Figure 7.1: Skeletal rhythms associated with each usul in our sample of Turkish music. The text above and below the rhythms denotes the type of drum stroke. Left-hand strokes are shown below the rhythm, while right-hand strokes are shown above. "Düm" strokes are associated with the highest intensity. The time signatures correspond to the time signatures in which pieces using these usuls are notated in the SymbTr corpus.

music is taken to represent a different musical idiom. The German folk songs were sourced from the Essen folksong collection (Schaffrath & Huron, 1995), which contains a large number of folk songs from Germany in addition to folk songs from a variety of geographical regions. The Dutch folk songs were sourced from the Meertens Tunes Collection, which has been compiled by the Meertens Institute (Van Kranenburg, Bruin, De Grijp, & Wiering, 2014). Turkish makam music was sourced from the SymbTr dataset, which contains some pieces that are considered art music as well as some considered folk music (Karaosmanoğlu, 2012).

Regarding rhythmic organization, a relevant theoretical difference between Western and makam music is that makam pieces are categorized by rhythmic modes called *usuls*. These rhythmic modes are characterized by a rhythmic pattern of two types of drum strokes: "düm" strokes, which produce a deep and resonant sound and are considered most important in establishing the rhythmic mode, and "tek" strokes (Marcus, 2001). These usuls, and the time signatures in which they are notated in the SymbTr corpus, are shown in Figure 7.1. Usul patterns have been shown to influence the distributions of onsets at different positions in the metrical cycle (Holzapfel, 2015). Rhythms of Western music, by contrast, can be expected to be constrained strongly by the patterns of metrical salience associated with their time signature (Palmer & Krumhansl, 1990; Temperley, 2010).

Details on how we derived datasets of German, Turkish, and Dutch rhythms from

Table 7.2: The number of rhythms that occur in each dataset for different meters, sorted in descending order by the minimum number of rhythms in each meter across datasets. Only the top-8 meters are shown. Rows corresponding to meters selected to be included in our samples are colored gray.

Meter	Dutch	German	Turkish	Minimum count
4/4	990	1323	247	247
3/4	379	879	107	107
6/8	982	603	97	97
2/4	289	945	80	80
6/4	69	82	149	69
3/8	54	216	7	7
4/8	9	4	3	3
9/8	37	1	308	1

these corpora are provided in appendix B. Importantly, the samples that we use are subject to three constraints: First, the total number of rhythms in each sample is identical. This ensures that performance differences across samples cannot be attributed to differences in sample size. Second, each sample contains rhythms in the same four meters, and the number of rhythms in each meter is the same in each sample. This ensures that performance differences across samples cannot be attributed to the presence of different meters. Third, within the Turkish samples, rhythms in the same meter must also be associated with the same usul. In the SymbTr corpus, some time signatures are associated with multiple usuls. Holzapfel (2015) has shown that usul patterns influence represent statistical affordances for meter. The third constraint rules out that this influence plays a role in our results.

Ensuring that the number of rhythms in each meter is the same *within* a sample is a precaution. The enculturation model estimates an independent sequence model from the rhythms in each meter. The sequence modeling algorithm involves certain heuristics that make the effect that the size of the training sample has on its performance unpredictable. This could introduce biases if the number of rhythms used to estimate the parameters of each sequence model is not constant.

These constraints impose strong limitations on the size of the empirical samples that we can use: the constraint that the number of rhythms in each meter must be balanced means that the maximum number of rhythms in each meter is constrained by the smallest number of rhythms that is available in each of the involved meters in any of the datasets. Since we are interested in the effects of metrical inference, we want to ensure that testing and training data contain rhythms in a variety of meters. We therefore look for a set of meters for which each dataset contains a reasonable number of rhythms in those meters.

Table 7.3: The number of rhythms observed for each combination of time signature and usul in the Turkish dataset. Rows corresponding to usuls selected to be included our samples are colored gray.

Meter	Usul	Count
4/4	Sofyan	247
	Yürüksemai	7
3/4	Semai	107
6/8	Yürüksemai II	81
	Âzeri Yürüksemai	9
2/4	Nimsofyan	79
	Yürüksofyan	1

Table 7.2 shows the number of rhythms in each dataset per meter, sorted by the minimum number of examples of rhythms in that meter in either of the datasets. This table only shows meters for which each dataset contains at least one rhythm in that meter. In order to ensure that a sufficient number of rhythms each meter can be used for training, we select the meters in the top four rows of this table (highlighted in bold) for constructing the samples. Table 7.3 shows the different usuls observed in the Turkish rhythms for the selected meters. The usuls selected for inclusion in the samples are highlighted in bold in Table 7.3.

The resulting selection represents a reasonable amount of metrical variety: two binary simple meters (2/4 and 4/4), one binary compound meter (6/8), and one ternary simple meter (3/4). Each sample contains seventy-nine rhythms in each of the four meters, and contains 316 rhythms in total, corresponding to 9164 inter-onset intervals. To construct these samples, seventy-nine rhythms in each meter are drawn at random and without replacement from each dataset.

7.6 Experiment 1

7.6.1 Methods

The three variants of the enculturation model and the IOI model (with short, long, and unbounded maximum pattern lengths) are evaluated in within-sample simulations using the Turkish, German, and Dutch samples. We are not specifically interested in differences between the Dutch and German samples and we merge the results of these simulations into the results for Western rhythms (the distinction becomes relevant in Experiment 2, where we compare within-sample to within-idiom and across-idiom simulations). The Turkish rhythms are referred to as the

makam rhythms.

The sequence models used by the enculturation model and the IOI model are variable-order Markov models that are also used by the IDyOM modeling framework (see Pearce, 2005, pp. 79–110). The maximum pattern length to which the sequence models are sensitive is controlled by considering two versions of the sequence modeling algorithm used by these models: a *bounded* and an *unbounded* version. The order of a Markov model indicates the number of notes immediately preceding the current moment on which the probability distribution of inter-onset intervals is conditioned: e.g., a first-order Markov model is sensitive to two notes. The bounded sequence model has a parameter called *order bound* $n \in \mathbb{N}$, which constrains the maximum order of the Markov models that it combines. The unbounded version uses Markov models of any order up to the length of a rhythm. We consider two unbounded models with order bounds of one and four (maximum pattern lengths two and five), and one unbounded model.

While Van der Weij et al. (2017 [Chapter 6]) also consider an order bound of zero, we do not consider this option here. The probability distribution of a metrical fingerprint of an inter-onset interval is strongly constrained by the metrical fingerprint of the previous inter-onset-interval. No sensitivity to sequential patterns corresponds to assuming independence between these two events, which would cause a significant drop in performance (as can be seen in Van der Weij et al.'s results) that cannot only be attributed to the decreased sensitivity to sequential patterns.

7.6.2 Results

Figure 7.2a shows estimated cross-entropy as a function of the maximum pattern length and which model (enculturation or IOI) is used. Figure 7.2b shows estimated cross-entropy as a function of the training (source) sample and which model is used. We analyzed the results by performing a three-way factorial analysis of variance with estimated cross-entropy as the dependent variable and source sample (Western or makam), maximum pattern length (short, long, or unbounded) and model (enculturation or IOI) as the independent variables.

We find that the cross-entropy, averaged across models and source samples, is lower for models sensitive to long patterns ($M = 1.52$) rather than short patterns ($M = 1.63$), $PRE = 0.36$, $F(1, 169) = 94.5$, $p < 0.001$ and marginally lower for models sensitive to patterns of unbounded length ($M = 1.47$) rather than long, $PRE = 0.08$, $F(1, 169) = 14.1$, $p < 0.001$. Cross-entropy, averaged across models and levels of sensitivity, is higher when using the makam source sample ($M = 1.66$) than when using the Western samples ($M = 1.48$). We find that which model (enculturation or IOI) is used interacts with whether the Western or makam

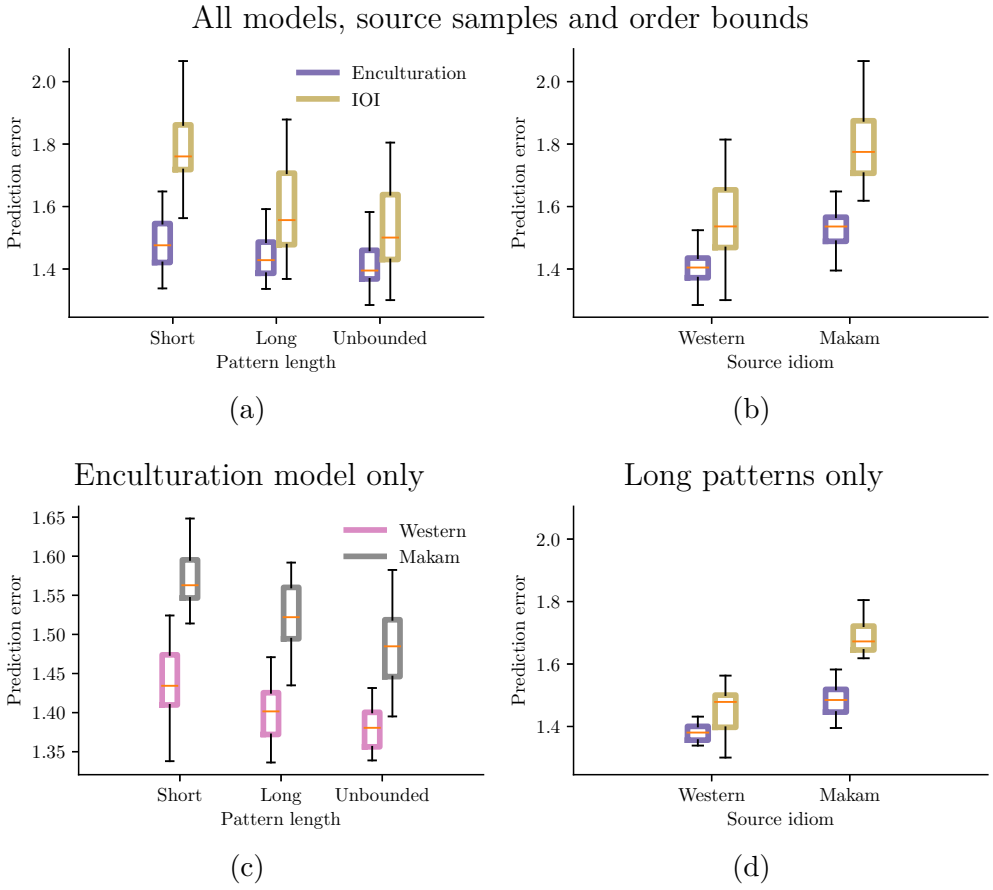


Figure 7.2: Estimated cross-entropy of the enculturation model and the IOI model in within-sample simulations with Western or makam source samples. Figure 7.2a shows cross-entropy estimates averaged across source samples for models sensitive to patterns of different maximum lengths. Figure 7.2b shows cross-entropy estimates averaged across maximum pattern lengths for the Western or makam source samples. Figure 7.2c shows the results of different maximum pattern lengths of the enculturation model in isolation. Figure 7.2d shows the result of the IOI and enculturation model sensitive to patterns of unbounded length for Western or makam source samples.

source sample is used $PRE = 0.18$, $F(1, 169) = 36.9$, $p < 0.001$ and whether the maximum pattern length is short or long, $PRE = 0.19$, $F(1, 169) = 38.2$, $p < 0.001$: When the maximum pattern length is long, the difference between the average performance of the enculturation model ($M = 1.44$) and the IOI model ($M = 1.59$) is smaller than when it is short (where the cross-entropy is 1.48 using the enculturation model and 1.78 using the IOI model).

To investigate whether sensitivity to longer patterns results in lower cross-entropy for the enculturation model, we analyze its results in isolation. These results are shown in Figure 7.2c. To analyze them, we performed a two-way factorial analysis of variance with cross-entropy as the dependent variable and source sample and maximum pattern length as independent variables. We find that the enculturation model achieves lower cross-entropy when the maximum pattern length is long ($M = 1.44$) rather than short ($M = 1.48$), $PRE = 0.12$, $F(1, 85) = 11.1$, $p = 0.001$. When an unbounded maximum pattern length is used ($M = 1.42$), cross-entropy is marginally lower than when it is long, $PRE = 0.06$, $F(1, 85) = 5.1$, $p = 0.027$. Again, cross-entropy is on average higher in the makam source sample ($M = 1.53$) than in the Western source sample ($M = 1.41$). We find no interactions between whether the source sample is Western or makam and the maximum pattern length is short or long, $PRE = 0.00$, $F(1, 85) = 0.3$, $p = 0.603$, or whether it is long or unbounded, $PRE = 0.01$, $F(1, 85) = 0.5$, $p = 0.487$.

Since it looks like the IOI model performs almost as well as the enculturation model when they are sensitive to patterns of unbounded length, we analyze these results in isolation. Figure 7.2d shows the results of both models with unbounded maximum pattern lengths. We performed a two-way factorial analysis of variance with cross-entropy as the dependent variable and source sample and model as the independent variables. We again find an interaction between whether the source sample is Western or makam and which model is used, $PRE = 0.23$, $F(1, 57) = 17.1$, $p < 0.001$. When the source sample is Western, the average estimated cross-entropy of the enculturation model is 1.38, versus 1.45 for the IOI model, but when the source sample is makam, the cross-entropy of the IOI model ($M = 1.69$) increases more than that of the enculturation model ($M = 1.49$). An analysis of the simple effects shows that the enculturation model performs better than the IOI model both in case of the Western rhythms, $PRE = 0.24$, $F(1, 39) = 11.8$, $p = 0.001$, and in case of the makam rhythm, $PRE = 0.76$, $F(1, 19) = 55.5$, $p < 0.001$.

Finally, we assess whether lower cross-entropy corresponds to a higher posterior probability of the metrical interpretation observed in the corpus. Figure 7.3 shows the average posterior probability of the meters observed in the corpus as predicted by the enculturation model as a function of its source sample and the maximum pattern length. We find that the posterior probability of the observed metrical interpretation is higher when maximum pattern length is unbounded ($M = 0.64$)

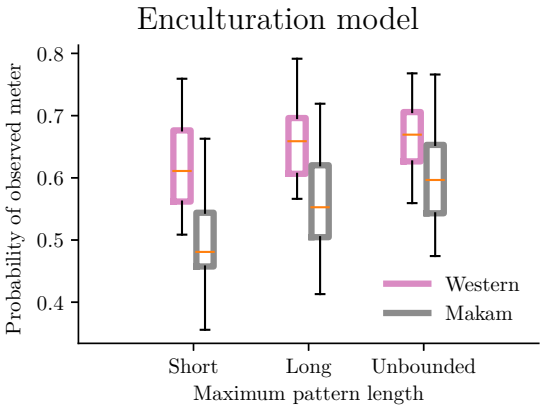


Figure 7.3: A comparison of the effects of the source sample and maximum pattern length on the average posterior probabilities of the meter observed in the corpora for each rhythm.

rather than short ($M = 0.58$), $PRE = 0.15$, $F(1, 85) = 14.5$, $p < 0.001$, but we do not find that the model sensitive to patterns of unbounded length performs better than the model sensitive to long patterns, $PRE = 0.02$, $F(1, 85) = 1.6$, $p = 0.205$. The enculturation model assigns less posterior probability to the observed metrical interpretation ($M = 0.55$) in simulations involving the makam source sample than in simulations involving the Western source sample ($M = 0.65$), $PRE = 0.30$, $F(1, 85) = 36.8$, $p < 0.001$. We observe no interactions between whether the source sample is makam or Western and whether the maximum pattern length is short or long, $PRE = 0.01$, $F(1, 85) = 0.5$, $p = 0.470$, or whether it is long or unbounded, $PRE = 0.01$, $F(1, 85) = 0.7$, $p = 0.396$.

7.6.3 Discussion

The results suggest that statistical affordances for meter relying on sensitivity to patterns of two to five metrical fingerprints are available in both makam and Western rhythms: the cross-entropy of the enculturation model sensitive to patterns of length five is lower than that of the enculturation model sensitive patterns of length two. Patterns longer than five metrical fingerprints also occasionally improve prediction of the timing of onsets: the enculturation model sensitive to patterns of unbounded length performed marginally better than the model sensitive to patterns of five. This suggests that besides the statistical cues for meter posited by classical theories of meter (Lerdahl & Jackendoff, 1983; Palmer & Krumhansl, 1990; Temperley, 2007), both Western and makam rhythms contain more complex statistical patterns that can provide cues for meter.

Sensitivity to patterns of more than two metrical fingerprints also helps the enculturation model predict the time signatures observed in the score representations

of the melodies from which rhythms are derived. Here, the effect of pattern length was less pronounced and we found no evidence that sensitivity to patterns beyond five metrical fingerprints improves metrical interpretation. This suggests that lower cross entropy indeed is associated with enhanced metrical interpretation performance, although it also seems possible for cross-entropy to improve when metrical interpretation does not.

The results do not clearly indicate that sensitivity to long- or unbounded-length patterns is useful to different degrees in makam or Western rhythms: neither in the case of cross-entropy, nor in the case of metrical interpretation performance does the source sample interact with the maximum pattern length.

The results reveal a general pattern where for all models, cross-entropy is higher with respect to the makam rhythms. This suggests that the hypothetical probability distribution from which these rhythms are drawn has a higher entropy than the distribution of the Western rhythms. A possible explanation for this may be the musical styles from which the rhythms in both samples are derived: The makam sample consists of both folk and Turkish art music (Karaosmanoğlu, 2012), while the Western samples consist of folk song melodies, which arguably exhibit relatively simple melodic and rhythmic structure.

Finally, the results show that the enculturation model is better at predicting inter-onset intervals in rhythms than a sequence model of inter-onset intervals. This shows that inferring meter allows the timing of onsets in a rhythm to be predicted more accurately. However, for Western rhythms, the difference in cross-entropy of the IOI and enculturation model is relatively small when the maximum pattern length is long or unbounded, while for makam rhythms, the difference remains substantial for these pattern lengths. Apparently, when the IOI model is sensitive to sufficiently long patterns, it predicts the inter-onset intervals in Western rhythms almost as well as the enculturation model, without inferring the underlying meter. When sensitive to shorter patterns, the enculturation model has a more significant advantage over the IOI model in predicting inter-onset intervals.

Earlier findings of Van der Weij et al. (2017 [Chapter 6]) suggested that models sensitive to longer patterns are better able to learn probability distributions and better able to predict the time signatures of rhythms as observed in the Essen folksong collection. The present findings replicate these earlier findings using the refined enculturation model and generalize them to the three different samples used in this study.

7.7 Experiment 2

7.7.1 Methods

The results of Experiment 1 suggested that sensitivity to statistical patterns of multiple subsequent metrical fingerprints makes statistical affordances for meter available to the enculturation model in both Western and makam rhythms. Since the difference in cross-entropy between sensitivity to patterns of up to five or an unbounded number of metrical fingerprints is marginal, we only include two variants of the enculturation model in Experiment 2: one sensitive to short (maximum length two) patterns and one sensitive to patterns of unbounded length.

The applied methodology is the same as in the previous experiment, except that we now evaluate each model on all target samples, rather than only on a target sample drawn from the same distribution as the source sample. This results in three sets of within-sample simulations (using German, Dutch and Turkish training samples), two sets of within-idiom simulations (where the source–target pairs are German–Dutch or Dutch–German) and four sets of across-idiom simulations (a German or Dutch source sample with a Turkish target sample, or a Turkish target sample with a German or Dutch source sample). In this experiment, the distinction between Dutch and German samples is relevant as we compare within-sample to within-idiom results in order to distinguish between potential sample-specific or idiom-specific statistical affordances for meter.

7.7.2 Results

Figures 7.4a, 7.4b, and 7.4c show average per-note cross-entropy (surprisal) results of individual rhythms (from all cross-validation folds and target samples) obtained with the enculturation model sensitive to patterns of unbounded length. Figures 7.4d, 7.4e, and 7.4f show the same results obtained with the classical model. The figures show two dimensional projections of a three-dimensional space. The x, y, and z axes of this space represent average per-note information content (surprisal) of a rhythm for a model trained on the Dutch, German, or Turkish source sample. As such, the coordinates of each rhythm in this space represent the degree to which its inter-onset intervals are predicted accurately by models trained on different source samples. The figures show two-dimensional projections of this space on planes spanned by different pairs of training samples.

If a rhythm is located on the diagonal of any of these figures, the cross-entropy of models trained on either source sample with respect to that rhythm is similar. In the results of the enculturation model, Turkish rhythms extend outward from the diagonal when one of the axes represents a Turkish model (Figures 7.4b and

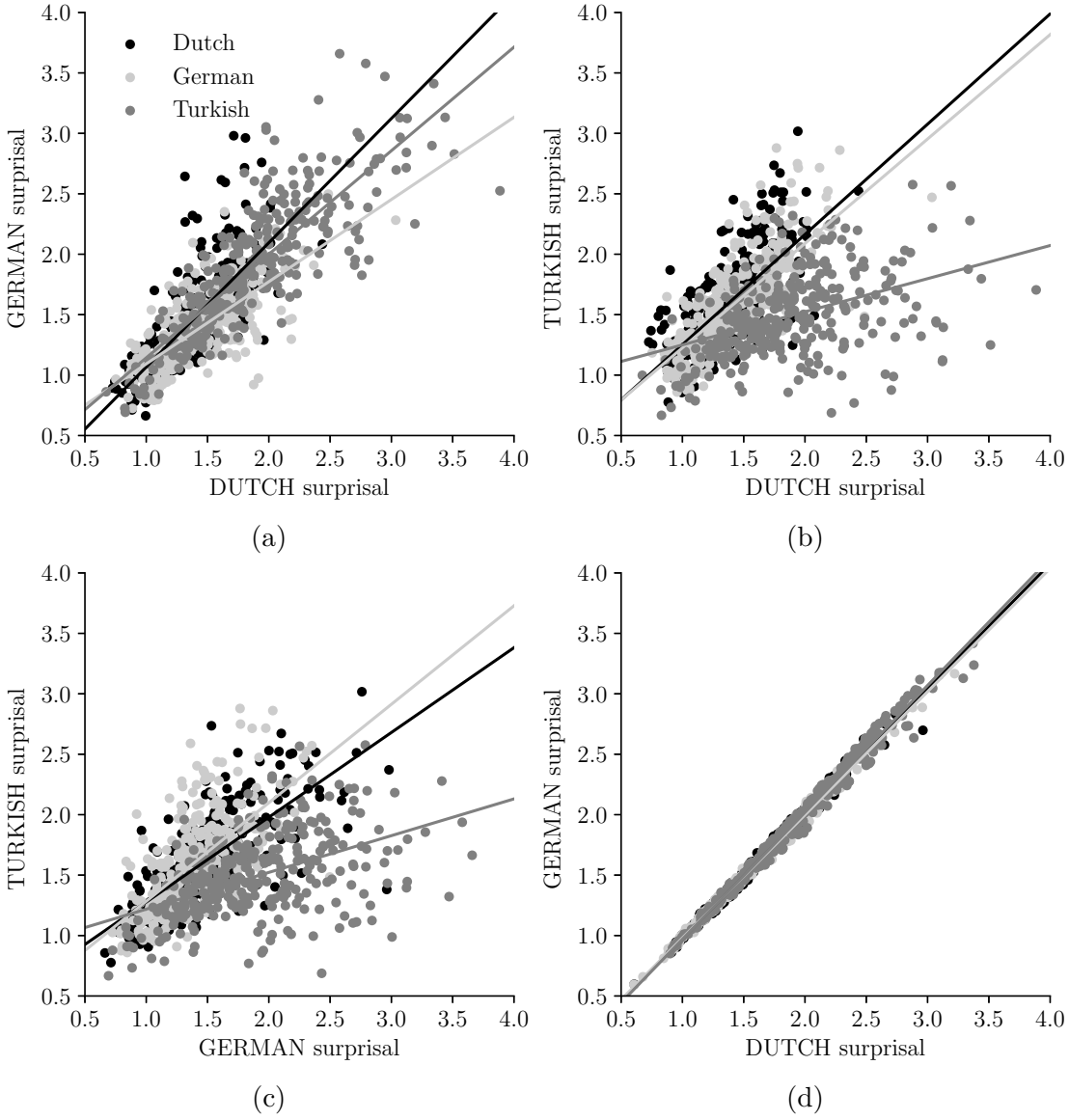


Figure 7.4: *(Figure continues on the next page.)* Two-dimensional projections of a three-dimensional space containing all rhythms in all target samples.

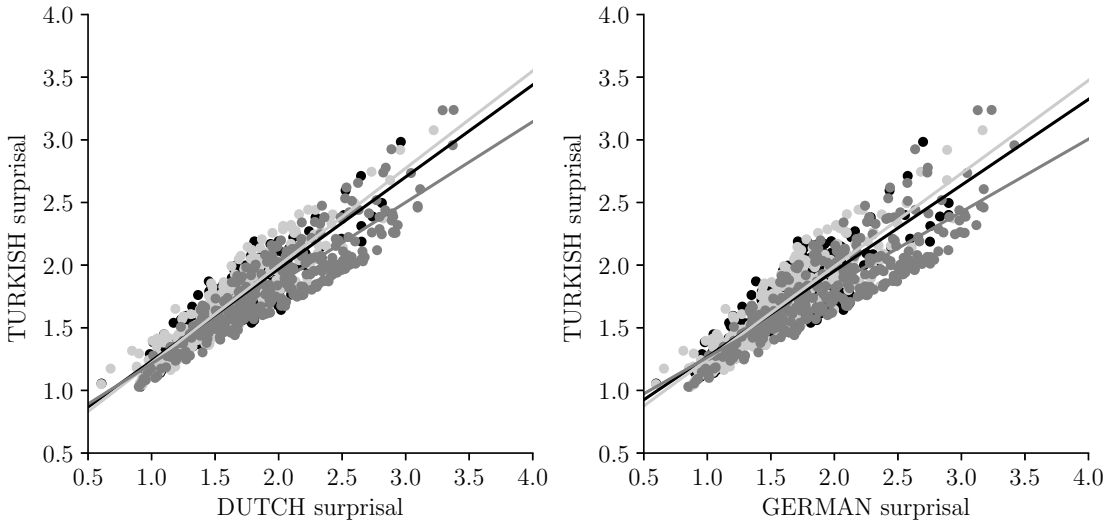


Figure 7.4: *(Continued).* Two-dimensional projections of a three-dimensional space containing all rhythms in all target samples. Each rhythm's coordinates correspond to the average per-inter-onset-interval surprisal (negative logarithm, base two, of the probability of each inter-onset interval) of models trained on different source samples with respect to the rhythm. The three figures in the top row (7.4a, 7.4b, and 7.4c) visualize the results of the enculturation model sensitive to patterns of unbounded length and the three figures in the bottom row show the results of the classical model (7.4d, 7.4e, and 7.4f). The lines represent first-degree polynomials fitted to the rhythms in each target sample.

7.4c). In Figure 7.4a, the rhythms center predominantly around the diagonal (although there are some clear exceptions), indicating that the Dutch and German predict all rhythms to similar degrees. Compared to the results of the classical model (Figures 7.4d, 7.4e, and 7.4f), the enculturation model clearly shows greater differentiation between models, as indicated by the rhythms spreading farther from the diagonal than in the results of the classical model.

We first analyze the results separately, per source sample, in order to compare the results of within-sample to within-idiom simulations and to investigate whether the model used interacts with the simulation type, which would indicate the presence of sample- or idiom-specific affordances for meter. For the Dutch and German source samples, we performed two-way factorial analyses of variance with estimated cross-entropy as the dependent variable, and model (classical, or enculturation sensitive to short or long patterns, or patterns of unbounded length), and simulation type (within-sample, within-idiom, and across-idiom) as the independent variables. For example, in the analysis of the German source sample, the within-idiom group would represent the Dutch rhythms, and the across-idiom group would represent the Turkish rhythms. For the Turkish source sample, the only simulation types are within- and across-idiom since we have no other sample representing Middle-Eastern makam music.

In the German source sample results (Figure 7.5a), averaged across all models, cross-entropy is slightly lower in within-sample simulations ($M = 1.46$) than in within-idiom simulations ($M = 1.51$), $PRE = 0.07$, $F(1, 82) = 6.2$, $p = 0.015$. In the Dutch source sample results (Figure 7.5b), we find a small and not statistically significant difference in cross-entropy between within-sample ($M = 1.48$) and within-idiom simulations ($M = 1.50$), $PRE = 0.03$, $F(1, 82) = 2.3$, $p = 0.137$. In neither the Dutch source sample nor the German source sample results do we find that there is an interaction between whether the simulation type is within-sample or within-idiom and any of the contrasts between between models. The differences between the within-sample and within-idiom simulations are small and only significant in the German source sample results. This is consistent with the expectation that patterns useful for inferring meter are consistent within musical idioms. For further analyses, we merged the results of these two simulation types into a within-idiom group.

Furthermore, in none of the per-source-sample analyses do we find that whether the enculturation model is sensitive to short patterns or patterns of unbounded length interacts with the simulation type (within-sample, within-idiom, or across-idiom). Therefore, we exclude the enculturation model sensitive to patterns of unbounded length from consideration in further analyses, enabling us to focus on the contrast whether simulation type (within- or across-idiom) interacts with model (classical or enculturation with a short maximum pattern length).

We first analyze the results of the Western source samples (Dutch and German)

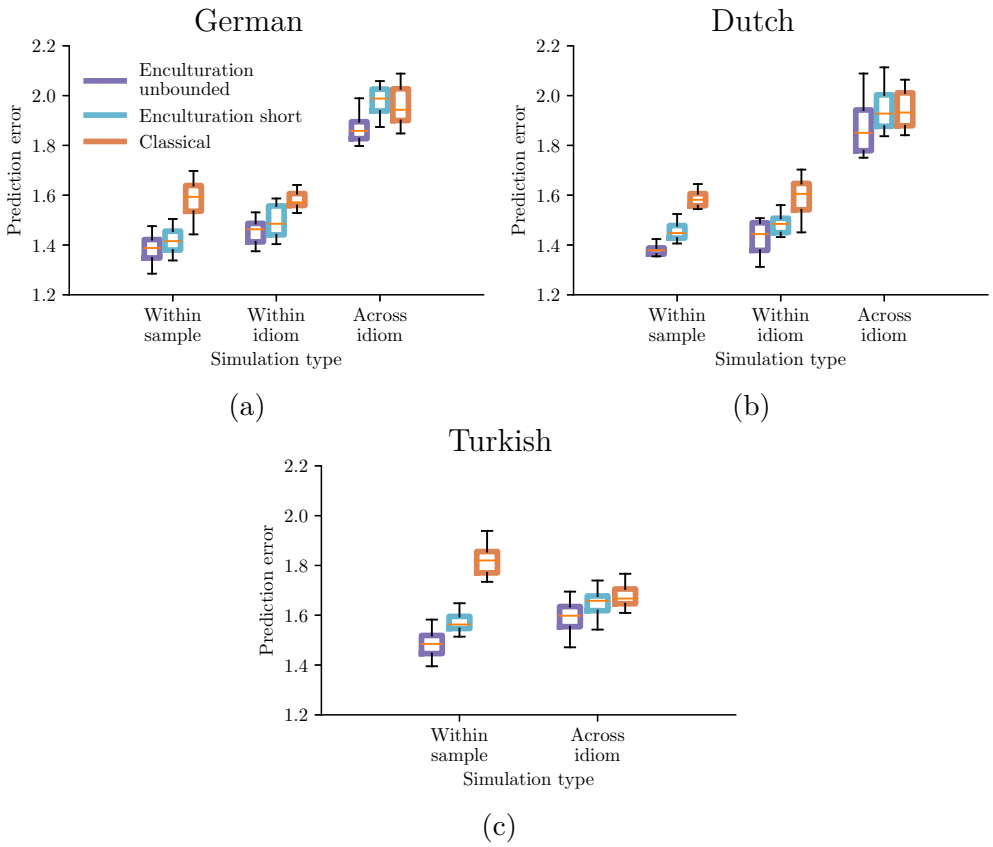


Figure 7.5: The results of Experiment 2. The estimated cross-entropy of the two variants of the enculturation model and the classical model on different target samples. The results are shown separately for each source sample. The results using the German source sample are shown in Figure 7.5a, those using the Dutch source sample in Figure 7.5b, and those using the Turkish source sample in Figure 7.5c.

and the makam source sample separately. For each, we perform a two-way factorial analysis of variance with cross-entropy as the dependent variable and simulation type (within- or across-idiom) and model (enculturation or classical) as independent variables.

In the Western source sample results (Figure 7.6a), we find that, averaged across all models, cross-entropy is lower in within-idiom simulations ($M = 1.53$) than in across-idiom simulations ($M = 1.96$), $PRE = 0.88$, $F(1, 117) = 873.2$, $p < 0.001$. However, we observe an interaction between model (enculturation or classical) and simulation type (within- or across-idiom), $PRE = 0.17$, $F(1, 117) = 23.9$, $p < 0.001$: in within-idiom simulations, the enculturation model performs better ($M = 1.46$) than the classical model ($M = 1.59$), while in across-idiom simulations, the performance of the enculturation model ($M = 1.97$) is similar to that of the

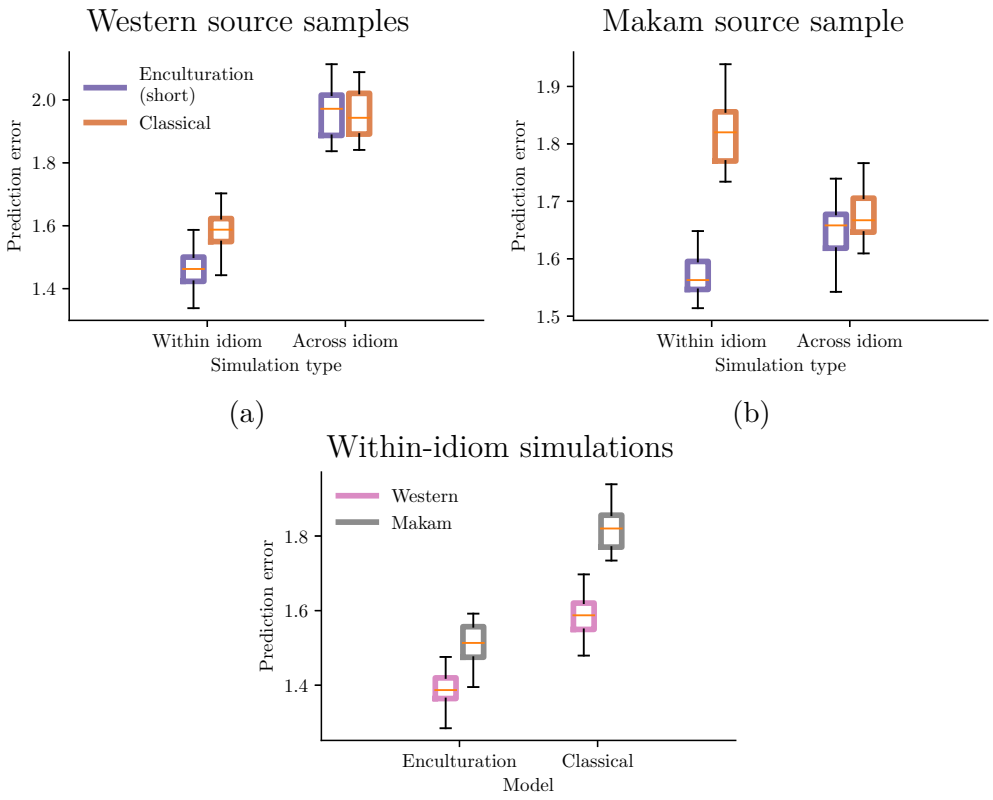


Figure 7.6: The re-grouped results of Experiment 2, with the German and Dutch source samples merged into a group of Wester source samples, and excluding results of the enculturation sensitive to long patterns. The two figures on top show the results using the Western (German and Dutch) source samples (Figure 7.6a) and makam (Turkish) source sample (Figure 7.6b). Below, Figure 7.6c shows the results of within-idiom simulations comparing the Turkish and Western source samples and the classical model and the enculturation model.

classical model ($M = 1.96$).

In the makam source sample results (Figure 7.6b), we again find an interaction between model and simulation type, $PRE = 0.46$, $F(1, 57) = 47.7$, $p < 0.001$. This time, however, the performance of the classical model is worse in within-idiom simulations ($M = 1.82$) than in across-idiom simulations ($M = 1.68$). The performance of the enculturation model is slightly better in within-idiom ($M = 1.57$) than in across-idiom simulations ($M = 1.65$). An analysis of the simple effect shows that this difference is significant, $PRE = 0.32$, $F(1, 179) = 13.2$, $p = 0.001$ (Figure 7.6b).

Analyses of other simple effects show that in within-idiom simulations, the enculturation model obtains lower cross-entropy than the classical model, both when

using the Western, $PRE = 0.47$, $F(1, 179) = 69.4$, $p < 0.001$ (Figure 7.6a), or the makam source sample, $PRE = 0.11$, $F(1, 117) = 14.4$, $p < 0.001$ (Figure 7.6b). In across-idiom simulations, however, the differences between the cross-entropy of the classical and enculturation model are no longer statistically significant, both when using the Western source sample, $PRE = 0.01$, $F(1, 179) = 0.3$, $p = 0.572$ (Figure 7.6a), and when using the makam source sample, $PRE = 0.07$, $F(1, 179) = 2.8$, $p = 0.104$ (Figure 7.6b).

The cross-entropy of both the classical model and the enculturation model is higher for the makam rhythms compared to the Western rhythms. To investigate whether this difference is greater for one of the models, we analyze the within-sample results in isolation. Figure 7.6c shows the cross-entropy results for the two source samples as a function of which model is used. A two-way factorial analysis of variance with cross-entropy as the dependent variable and model (classical or enculturation) and source sample (Western or makam) as the independent variables reveals an interaction between the source sample (Western or makam) and model (classical or enculturation), $PRE = 0.18$, $F(1, 87) = 19.4$, $p < 0.001$. Using the Western source samples, the difference in performance between the classical and enculturation model is smaller ($M = 1.39$ for the enculturation model and $M = 1.59$ for the classical model), than when using the makam source sample ($M = 1.51$ for the enculturation model and $M = 1.82$ for the classical model).

7.7.3 Discussion

Visual inspection of the results of the enculturation model and the classical model in Figure 7.4 illustrates the inter-dependence between patterns in the environment and learning mechanisms: The enculturation model is sensitive to complex patterns in rhythms and reveals a gradient of degrees to which rhythms conform to the statistical properties learned from different samples. A gradient which to some extent separates the Turkish rhythms from the German and Dutch rhythms because these rhythms can be predicted more accurately by models trained on a Turkish source sample. The classical model is significantly less sensitive to statistical patterns in the musical environment and less clearly distinguishes German and Dutch rhythms from Turkish rhythms.

Complementing the results of Experiment 1, Experiment 2 shows that in within-sample simulations, more statistical affordances for meter are available to the enculturation model (sensitive to short patterns) in German and Dutch rhythms than to the classical model.

As in Experiment 1, the cross-entropy of both models is higher on the makam rhythms than on the Western rhythms. However, in Experiment 2 we observed that the cross-entropy of the classical model increases more sharply than that

of the enculturation model (sensitive to patterns of two metrical fingerprints) when evaluated on makam rhythms compared to Western rhythms: the classical model is able to learn fewer statistical patterns in makam rhythms that are useful for inferring meter than the enculturation model. This suggests that while the makam rhythms do contain statistical patterns that experienced listeners can use to perceive meter, these patterns rely, more than is the case for the Western rhythms, on statistical patterns detected by the enculturation model but not by the classical model. In fact, when trained on Turkish rhythms, cross-entropy of the classical model is *higher* in within-idiom simulations compared to across-idiom simulations (with German or Dutch target samples).

The results show that both makam and Western rhythms contain idiom-specific statistical patterns that can serve as cues for meter. These cues rely on statistical patterns that can be detected by the enculturation model sensitive to patterns of two metrical fingerprints but cannot be detected by the classical model. The difference in cross-entropy between this enculturation model and the classical model in within-idiom simulations largely disappears in across-idiom simulations. Furthermore, the patterns appear to be idiom-, rather than sample-specific: cross-entropy of all models was comparable in within-sample and within-idiom simulations.

Although we found in Experiment 1 that patterns longer than two metrical fingerprints can serve as cues for meter, the results of Experiment 2 do not provide evidence these patterns longer than two metrical fingerprints are idiom-specific: We did not find that whether enculturation model is sensitive to patterns of unbounded length or not interacts with whether the simulation is within- or across-idiom.

It is possible that sensitivity to longer patterns has consequences for *which* metrical interpretations are inferred for a given target sample by models trained on different source samples. The cross-entropy measure represents the extent to which inter-onset intervals in a rhythm can be predicted based on meter inferred from the rhythm but is not informative about the actual inferred metrical interpretations on which these predictions are based. Investigating which meters are inferred by in with- and across-idiom simulations remains a topic for future work to address.

7.8 Experiment 3

Experiment 2 showed that the enculturation model sensitive to patterns of two metrical fingerprints learns idiom-specific patterns that provide affordances for meter in within-idiom simulations and that statistical affordances for meter are less available to the classical model than the enculturation model in the case of makam rhythms. However, these results do not reveal whether the statistical

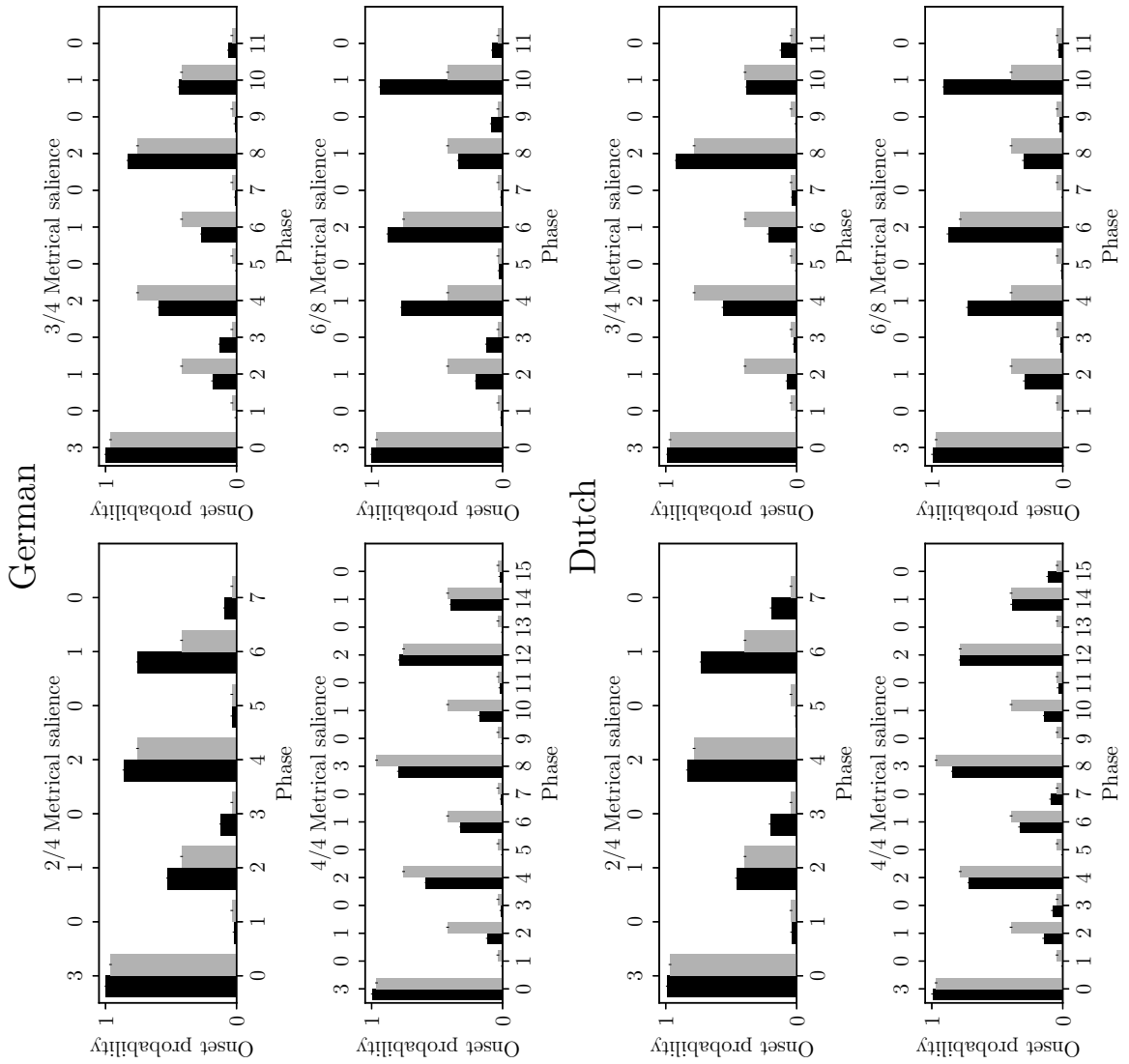
patterns learned by the enculturation model rely on the model's representation of metrical fingerprints, or to its sensitivity to patterns of two of such fingerprints.

In Experiment 3, we investigate whether the representation of metrical context phase could plausibly account for the observed differences between the classical model and the enculturation model. We investigate two representations: the representation of metrical salience used by the classical model and *phase*: the position of an onset in the metrical cycle. Although the enculturation model's metrical fingerprints, downbeat distances, are marginally more informative than phase, they encode the same information.

Figure 7.7 shows the relative frequencies with which onsets occur at different phases (the black bars) and at the level of metrical salience associated with these phases (the gray bars) for each sample and each meter. Here we use the four levels of metrical salience posited by the classical model: three, two, one, and zero, indicating respectively a bar-level downbeat, a tactus beat, a beat immediately below the tactus level and any other beat (see Chapter 5 for details). For example, the height of the black bar at phase two of the 2/4 meter in the Dutch sample represents the proportion of times at which second sixteenth-note position in a 2/4 bar contains an onset out of the total number of times that this position occurs in the Dutch sample. The corresponding level of metrical salience is one, and the relative frequency of onsets at this level of metrical salience corresponds to the proportion of times that an onset occurs at this level of metrical salience (in any meter) in the Dutch sample. The proportion of times that an onset occurs at the four different levels of metrical salience in a given sample specifies the note-beat profile that is used by the classical model. Note that we are assuming a grid-based representation of rhythms here, in which rhythms are represented by a grid of sixteenth-note durations during which an onset either does or does not occur.

It is apparent from Figure 7.7 that the overall pattern of the probabilities with which onsets fall on metrical positions corresponding to different levels of metrical salience are similar in German, Dutch, and Turkish rhythms. However, the onset probability distributions of the Turkish rhythms exhibit more uncertainty, especially in the case of the 2/4 and 4/4 rhythms: the probabilities are less close to one and zero. This pattern was also noted by Holzapfel (2015), who suggested that meter in Turkish makam rhythms may be less stratified.

We consider two simple models that describe the probability that a note occurs at each grid point used to represent a rhythm of which the meter is given. That is, the position of each grid point in the metrical cycle—its phase—is known. The probability that a note occurs at a grid point is described by the random variable N, with the possible values $\{0, 1\}$, where 1 indicates the occurrence of an onset and 0 indicates no onset at a grid point. The *phase model* describes the probability that a note occurs at a given phase, p, in the metrical cycle of a meter, m: $P(n \mid p, m)$. These probabilities are estimated by the black bars in Figure 7.7.


German
Dutch
2/4 Metrical salience
3/4 Metrical salience
4/4 Metrical salience
6/8 Metrical salience
Onset probability
Phase

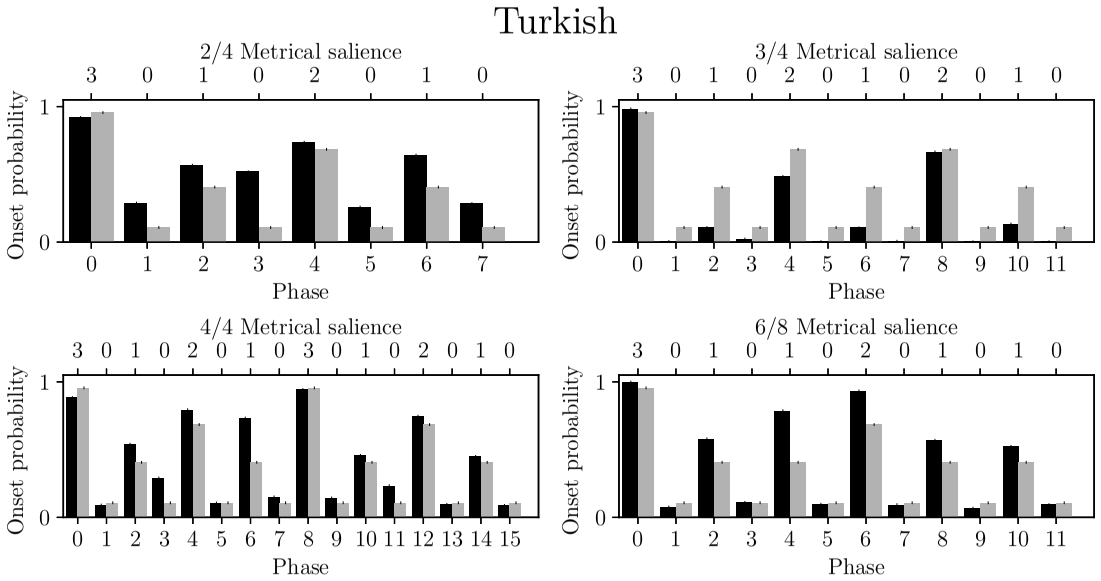


Figure 7.7: Relative frequencies with which onsets occur at different positions in the bar (black bars) and at different levels of metrical salience (gray bars). The gray bars are based on the relative frequency of onsets at four levels of metrical salience posited by the classical model in the entire sample. The black bars are based on the relative frequency of onsets in rhythms in the corresponding meter.

The *salience model* describes the probability that an onset occurs at a given level of metrical salience, s: $P(n \mid s)$. [3] These probabilities are estimated by the gray bars in Figure 7.7.

Since in a grid representation of a rhythm, every phase occurs an approximately equal number of times, we can calculate the cross-entropy of these models for different target samples. We denote the probability with which notes occur at different phases in different meters in a target sample, t, by $P(n^t \mid p, m)$. The cross-entropy of the phase model with respect to a target sample t is given by

$$-\frac{1}{2\,T_m}\sum_{p=0}^{T_m-1}\sum_{n=0}^{1} P(n^t \mid p, m)\log_2 P(n \mid p, m),$$

where T_m is the period of a meter m in sixteenth notes. For example, if $m = 4/4$, then $T_m = 16$. Note that the above equation corresponds to the expected value of information content of each observation of a note, given that notes occur at different phases, p, of a meter , m, with probability $P(n^t \mid p, m)$.

The cross-entropy of the salience model with respect to a target sample t is given by

$$-\frac{1}{2\,T_m}\sum_{p=0}^{T_m-1}\sum_{n=0}^{1} P(n^t \mid p, m)\log_2 P(n \mid \mathit{sal}(m, p)),$$

where *sal* is a function that returns the metrical salience of a given phase, p, and meter, m. For example, if $m = 2/4$ and $p = 4$, then $\mathit{sal}(2/4, 4) = 2$ (a tactus beat). Similarly, $\mathit{sal}(2/4, 0) = 3$ (a bar-level downbeat), and $\mathit{sal}(2/4, 1) = 0$ (not a beat).

Table 7.4 shows the cross-entropies obtained by applying the above formulas to the phase and salience distributions estimated from different source samples with respect to different target samples (the numbers shown in the Table are averaged across the four meters). Note that the phase and salience models do not infer meter, like the models in Experiment 1 and 2. Nevertheless, the numbers in Table 7.4 provide an indication of the degree to which the occurrence of onsets at grid points can be predicted by the phase model and the salience model given that the meter is known.

Overall, cross-entropies of the phase and salience models in Table 7.4 can be seen to describe the same pattern as the results of the classical model: Using Western source samples, the cross-entropies are lower when the target sample is also Western compared to when it is Turkish. When using the Turkish source

[3] The phase model is identical to the fine-grained position model and the salience model to the metrical position model described by Temperley (2010).

Table 7.4: The average cross-entropies of onset distributions estimated from each source sample with respect to each target sample. The onset probabilities are conditioned on different representations of metrical context: either their position in the metrical cycle (phase) or their metrical salience.

		Target					
		Phase			Metrical salience		
		Dutch	German	Turkish	Dutch	German	Turkish
Source	Dutch	0.429	0.454	0.805	0.508	0.504	0.680
	German	0.456	0.427	0.839	0.510	0.502	0.687
	Turkish	0.564	0.582	0.601	0.532	0.530	0.650

sample, the cross-entropies are *higher* when the target sample is also Turkish compared to when it is Dutch or German. The same pattern was observed for the classical model in Experiment two in within-idiom and across-idiom simulations. There, too, the estimated cross-entropy was higher in within-idiom simulations than in across-idiom simulations when the source sample was Turkish. The cross-entropy of the enculturation model, on the other hand, was lower in within- versus across-idiom simulations using the Turkish source sample.

This suggests that the different results of the enculturation model cannot be attributed primarily to the more detailed representation of metrical context but derive, at least in part, from its sensitivity to patterns of two metrical fingerprints. However, it remains an opportunity for future work to investigate this issue more directly, for example by using models in which the representation of metrical context and maximum pattern length can be varied independently. The design of the enculturation model did not allow us to reduce the maximum length of patterns to which it is sensitive to one.

7.9 General discussion

Classical theories of meter propose that its perception relies on schematic patterns of metrical salience that are inferred from a rhythm and influence the subsequent perception of the rhythm. These perceived patterns arise, according to classical theories, from the alignment of hierarchically organized periodicities (Lerdahl & Jackendoff, 1983), or from a generative grammar underlying a rhythm (Longuet-Higgins & Lee, 1984). How, precisely, meter is inferred from a rhythm is described by these classical theories in different ways, but they both agree that onsets and accents must predominantly fall on metrically strong beats, either to minimize syncopation (Longuet-Higgins & Lee, 1984) or to increase congruency between

rhythm and meter (Lerdahl & Jackendoff, 1983).

It has been suggested that listeners learn schematic patterns of metrical salience patterns from extended exposure to music, while the schemas themselves have been argued to be constrained primarily by the hierarchical structure of meter, and not the style or period, of a rhythm (Palmer & Krumhansl, 1990). Similarly, Temperley (2007) proposes that listeners are sensitive to the frequency with which onsets occur at different levels of metrical salience (the *note-beat profile*). Both of these theories, which are largely consistent with classical theories of meter, view metrical salience to be strongly linked to onset expectancy. It has been argued, however, that this relation holds primarily in Western classical musical. Iyer (1998) and London et al. (2017), for example, cite musics from West-Africa and from the African diaspora as counterexamples.

In this study, we investigated statistical affordances for meter that might be available to enculturated listeners in makam and Western rhythms. A statistical affordance for meter is an opportunity for an enculturated listener to perceive meter in a rhythm from a specific musical idiom. The availability of such affordances depends on three factors: (1) the statistical characteristics of the rhythm, (2) the statistical patterns to which a listener (or model) is sensitive and (3) the musical environment that reflects the long-term previous exposure of a listener (or model). That is, if a statistical affordance for meter is available, a particular rhythmic pattern serves as a *cue for meter* to an enculturated listener that has internalized certain statistical patterns in a musical idiom.

We investigated whether in addition to the frequency with which onsets occur at different levels of metrical salience there are other statistical patterns that make statistical affordances for meter available. We also compared the effectiveness with which different kinds of sensitivity to statistical patterns make statistical affordances for meter available in rhythms from different musical idioms. Finally, we investigated whether rhythms from different musical idioms contain statistical patterns that make affordances for meter available specifically in that idiom.

The results suggest, first, that rhythms in both Western folk melodies and Turkish makam music contain statistical patterns more complex than note-beat profiles that can serve as cues for meter to enculturated listeners. Second, both makam and Western rhythms appear to contain idiom-specific patterns that can serve as cues for meter. Third, we found that note-beat profiles are less effective as statistical cues for meter in makam rhythms than statistical patterns in the metrical fingerprints (see Section 7.3.4) of up to two subsequent inter-onset intervals.

We found that sensitivity to patterns of between two and five metrical fingerprints of subsequent note onsets makes statistical affordances for meter available in rhythms from the same musical environment as the musical environment representing the model's long-term exposure. That is, statistical affordances for meter relying on

such relatively long patterns are available in both Western and makam rhythms to listeners familiar with the corresponding musical idioms.

Furthermore, we found that while all models we tested are worse at predicting inter-onset intervals in the makam rhythms compared the Western rhythms, the performance of the classical model, which is sensitive only to note-beat profiles, decreased to a greater extent than that of the enculturation model. This suggests that while note-beat profiles are less useful for inferring meter in makam rhythms, makam rhythms do contain other statistical patterns that can serve as cues for meter. The results of Experiment 3 suggested that it is the sensitivity of the enculturation model to patterns of two metrical fingerprints, in addition to its representation of metrical context, that is responsible for these results.

Finally, we found that the enculturation model learns idiom-specific statistical patterns in both makam and Western rhythms, which are not learned by the classical model. This suggests that statistical affordances for meter are available in makam and Western rhythm to listeners with long-term familiarity with the respective musical idioms. In other words, enculturated listeners may have more rhythmic patterns at their disposal by which they can infer meter when hearing a rhythm in a musical idiom that they are familiar than when hearing a rhythm in a musical idiom that they are not familiar with. Simultaneously, however, we did not find evidence that these idiom-specific patterns consist of more than two metrical fingerprints, even though we found that these longer patterns can serve as cues for meter in Western and makam rhythms.

Our results may be compared to the predictions of the cultural distance hypothesis (Demorest & Morrison, 2016; Morrison et al., 2019), which states that the degree to which the statistical properties of music from different cultures are similar predicts the ability of listeners from those cultures to process music from the other culture. The evidence that we found for idiom-specific statistical patterns in rhythms is consistent with this hypothesis.

However, we note that when quantifying cultural distance using cross-entropy, one must be aware that cross-entropy is not a symmetric metric, and therefore not strictly a distance: the cross-entropy of a model that has learned the statistical properties of culture X on the music from culture Y may be different from the cross-entropy of culture Y with respect to culture X. For example, our results suggested that the entropy of the distribution of the makam rhythms in our sample is higher than that of the rhythms of Western folk melodies. Especially when a musical idiom is reduced to a low-dimensional representation of only one of its facets, such as monophonic melodies, or patterns of inter-onset intervals, such effects are likely to play a role since the complexity of different musical idioms may reside in different representations.

Across-idiom simulation results reported by Pearce (2018) for Chinese and Western

and by Morrison et al. (2019) for Chinese, Western and makam melodies suggest that the statistical differences between melodies from different cultures are more pronounced than is the case for rhythms in our samples, judging from visual comparison of the results in Figure 7.4 of this chapter and, for example, Figure 4 of Pearce (2018). Similarly, our analyses suggested that, as far as the Western and Turkish melodies from which our samples were derived, rhythms contain only a modest amount of idiom-specific structure. This may explain why Demorest et al. (2016) found that rhythm did not contribute to an enculturation effect observed in an earlier study (Demorest, Morrison, Münir, & Jungbluth, 2008). If the differences in the distributions of rhythms typical of different musical cultures are more subtle than the differences in distributions of melodies, then it may be comparatively likely that a rhythm sampled from either culture has a low cultural distance. The approach that we proposed for finding idiom-specific statistical affordances for meter using probabilistic generative models can be used to construct stimuli that are specifically predicted by models to elicit differences based on the previous cultural exposure of listeners. Such stimuli could be used in a cross-cultural experiment, similar to the experiment of Demorest et al. (2008), to investigate whether idiom-specific cues for meter can predict cultural differences in rhythm perception.

Our findings can also be related to a recent debate concerning the degree to which so-called beat-based timing and memory-based (or duration-based) timing contribute to expectations regarding the timing of auditory events on short timescales (Bouwer et al., 2020). Beat-based timing refers to the prediction of auditory events based on inferred regularity and is generally interpreted in terms of the entrainment of attention based on mechanisms of coupled oscillation (Large & Jones, 1999). Memory-based timing refers to expectations derived from predictable patterns (Bouwer et al., 2020). *Predictable*, in the context of these experiments, means predictable from the stimulus itself rather than from internalized stylistic aspects of music. Bouwer and Honing (2015) and Bouwer et al. (2020) interpret memory-based to rely on mechanisms posited by predictive processing theories of perception. These studies have found evidence that both beat- and memory-based expectations contribute to auditory expectations of the timing of events (Bouwer & Honing, 2015; Bouwer et al., 2020). The present findings appear consistent with these observations: while rhythms from different musical idioms contain some idiom-specific patterns, these differences are modest, suggesting that rhythms are constrained partly by principles shared between different idioms.

The approach pursued in this chapter suggests several topics that could be investigated in future work. For example, it would be interested to perform a more fine-grained comparison between the classical model and the enculturation model in which representation of metrical context (metrical salience or phase) and sensitivity to patterns (none or short) are varied independently. Furthermore, while our results did not provide evidence for idiom specific affordances for meter

relying on patterns of more than two metrical fingerprints, we did not consider the actual meters inferred by the tested models. It could be that the length of patterns to which a model is sensitive does influence the meters in which models interpret rhythms in within- and across-idiom simulations. As more empirical data becomes available from which rhythm samples can be drawn, it would be interesting to extend the study to include samples that represent musical idioms more broadly by including for example both folk and art music. Finally, it would be interesting to investigate whether the affordances for meter available to real listeners rely on statistical patterns identified in this study. To this end, cross-cultural studies are crucial since they enable the testing of hypotheses about the effects that statistical patterns in the environment may have had on perception.

The materials used in this study—score-like representations of the inter-onset interval patterns in rhythms—are in many ways impoverished representations of music. It seems plausible that expressive timing, tempo and tempo changes, dynamics, timbre, note durations (articulation), and the broader embodied context in which rhythm perception occurs can have a significant influence on enculturated rhythm perception and the differentiation of musical idioms. It is interesting, however, that even in this constrained representation, statistical patterns that are specific to different musical idioms can be observed. Accordingly, we think that computational models of music perception and cognition can play a valuable role in identifying the similarities and differences between musical idioms found in music across the world (Honing & Bouwer, 2019; Savage et al., 2015; Mehr et al., 2019).

Chapter 8

Discussion and conclusion

Predictive processing is an exciting and relatively new framework in which perception and cognition can be described and modeled. As reviewed in Chapter 2, rhythm and meter perception show considerable flexibility and plasticity under the influence of the previous experience, practice, and training of culturally embedded listeners. I have argued that the existing offering of computational models of rhythm perception does not sufficiently address these aspects. Predictive processing suggests a modeling framework, namely probabilistic generative models, that is particularly suitable for describing and modeling such flexibility and plasticity. This thesis takes one step toward an account of rhythm perception that considers the effects of experience, practice, and training. In Chapter 6, I propose a probabilistic generative model intended to simulate the effects of long-term *exposure* to rhythms in a musical environment on meter perception. Compared to earlier probabilistic generative approaches, this places significantly more emphasis on learning from musical patterns and regularities in the environment, and less emphasis on music theory.

Chapters 3, 4, and 5 were concerned with the technical and formal details of the modeling approach pursued in this thesis. In Chapter 3, I proposed a framework in which dynamic Bayesian network models with deterministic constraints can be defined. The framework supports the definition of a variety of probabilistic generative models of music perception that operate on abstract and symbolic representations. These formal definitions demand a high level of precision and explicitness, which results in compact definitions of the, sometimes complex and interacting, deterministic constraints of such models. Such definitions transparently reveal the structure and assumptions of probabilistic cognitive models.

Furthermore, the model definitions can be translated, with relatively little effort, into implementations that can be used in computer simulations. This may facilitate the sharing of modeling work and make it easier for other researchers to reproduce simulation results and to extend and build upon modeling work defined in the framework. The framework itself has been implemented as a cognitive modeling toolkit, which was used to generate the simulation results presented in Chapter 7. [1] In Appendix A, the model definitions are stated in Common Lisp, demonstrating the close mapping between model definition tables and their functional implementations in this framework.

The advantages of the toolkit are comparable to those of the IDyOM modeling framework (Pearce, 2005), which enables researchers to specify statistical models of sequences based on multiple derived representations of these sequences. Compared to IDyOM, however, the framework is more general and supports the specification of models that infer latent underlying structure (key or meter). On the other hand, it also is more low-level and requires a greater degree of programming ability on the part of the researcher using it. Dynamic Bayesian network models defined in this framework are directly comparable to models of statistical learning of multiple representations of melodies, as can be defined in IDyOM (Pearce, 2005, 2018): they incrementally predict events in a sequence, from which quantifications of expectedness and uncertainty can be derived.

The modeling framework arose from my efforts to generalize IDyOM and the multiple viewpoint systems (Conklin & Witten, 1995) on which it is based, to support inferring structure such as key and meter from sequences of events. The work presented in Chapter 6, which uses the terminology of multiple viewpoint systems to define the novel model presented in this thesis, is a result of that approach. It turned out, however, that the resulting framework could be formulated in a more general way, and the current modeling framework is the result of this.

There are many opportunities for future work to build and improve upon the models and their algorithmic implementations presented in this thesis. For example, the temporal prediction capabilities of the present model could be used in tandem with IDyOM's melodic prediction capabilities to yield a richer characterization of the dynamic interplay between listener expectations and musical progression. Furthermore, the meter perception model presented in this thesis could be extended to simultaneously infer other latent structure, such as key signatures. The probabilistic approach naturally accommodates inferring multiple kinds of latent structure simultaneously.

Chapters 4 and 5 provide practical examples of the usage of the framework

[1] The author intends to make this implementation and that of the models used in the present research available as free (as in freedom) and open-source software as soon as possible at https://osf.io/z4389/.

of Chapter 3. Chapter 4 defines an adaptation of Temperley's (2007) rhythm perception model that can evaluate a rhythm incrementally in time. The original formulation of this model emphasizes a top-down perspective in which a metrical grid is generated first and a rhythm is generated on top of it. A few minor modifications were required to enable the model to be evaluated in a temporally incremental fashion, similar to the other models presented in this thesis that are of an incremental nature (i.e., process models). Furthermore, by defining the model as a dynamic Bayesian network with deterministic constraints, some of its interacting deterministic constraints are revealed more explicitly than in the model's original definition.

Chapter 5 defines two generative models of meter perception and describes the technical details of a methodology that allows them to be compared systematically. There are two key steps to this methodology: first, the definition of a rhythm space over which both models define complete probability distributions, and, second, training the models on empirical samples of rhythms in this space. Training a model (estimating its parameters from an empirical sample) causes it to approximate an empirical distribution over the rhythms in the rhythm space. As such, these two steps allow us to assess how well different models approximate different empirical distributions of rhythms.

The emphasis on modeling the flexibility and plasticity of rhythm perception led us to adopt a cross-cultural approach. That is, we evaluate models not only on rhythms of the same style, repertoire, or idiom as that of the rhythms from which they have learned but also on rhythms from a different style, idiom or repertoire than those from which they have learned. The importance of a cross-cultural approach for investigating perception has been recognized for some time (Huron, 2006; Patel & Demorest, 2013), and appears recently to have gained a new impulse and urgency (London et al., 2017; Jacoby et al., 2019; Mehr et al., 2019; Jacoby et al., 2020). The role that probabilistic modeling approaches can play in generating theoretical predictions about the effect of enculturation in perception has recently been highlighted too (Pearce, 2018; Morrison et al., 2019). The work in this thesis concerns precisely the kind of models that could be used in this setting.

The cross-cultural approach is elaborated most fully in Chapter 7, where I consider the influence of three factors on the availability of statistical affordances for meter (whether statistical patterns in rhythms enable a model of an enculturated listener to infer meter): (1) the musical environment that represents long-term exposure, (2) the statistical patterns to which a model is sensitive, and (3) the musical environment in which the model is evaluated. I investigate this using empirical samples of rhythms from two different musical idioms: Western folk melodies and Turkish makam melodies.

The results suggest, first, that *rhythmic pattern matters*: Western and Turkish rhythms contain patterns that can serve as cues for meter consisting of patterns

of multiple events. Such cues are more complex than the frequency with which onsets occur at different levels of metrical salience (Palmer & Krumhansl, 1990; Temperley, 2007, 2010). This supports claims by London (2004, 2012, p. 68) and London et al. (2017) that musical styles contain characteristic rhythmic patterns that facilitate beat and meter perception in listeners familiar with these patterns. Second, as has been previously been argued by Iyer (1998), it appears that *metrical salience is most useful in Western rhythms*: the frequency with which onsets occur at different levels of metrical salience is significantly less useful for inferring meter in Turkish rhythms compared to Western rhythms. Third, *cultural familiarity matters*: some of the patterns that can serve as cues for meter appear to be idiom-specific. Listeners exposed to rhythms that rely on idiom-specific cues for meter may more clearly perceive the meter if they are familiar with the relevant musical idiom. Simultaneously however, idiom-specific patterns in rhythms appear to be short and to provide only a modest contribution to the availability of statistical affordances for meter overall. The patterns useful for inferring meter in Western and makam rhythms thus appear to an extent to be shared. This is consistent with studies suggesting that rhythms worldwide share common patterns, most prominently simple integer ratios between the durations of temporal intervals (Savage et al., 2015; Jacoby & McDermott, 2017; Mehr et al., 2019). Furthermore, neurobiological studies have suggested that different kinds of auditory expectations can be distinguished: beat-based expectations, which rely on inferred regularity, and memory-based expectations, which rely on predictable patterns (Bouwer et al., 2020). While "predictable" in this work refers to the context of a single stimulus, and not to internalized patterns from long-term exposure, the results are consistent with the suggestion that expectations regarding the timing of events in a rhythm are driven both by patterns shared between cultures and patterns specific to different cultures.

The above findings concern only patterns that matter for *models* of meter perception. Whether they matter for listeners too is at the moment an open question. The computational models on which these results are based can be used to investigate this. It would, for example, be possible to quantify the extent to which individual rhythms rely on idiom-specific cues. Such quantifications could be used in a cross-cultural experiment involving listeners with different histories of musical experience to construct rhythmic stimuli containing idiom-specific patterns that these listeners are or are not expected to be familiar with.

The methodology used in Chapter 7 is closely related to the *cultural distance hypothesis* proposed by Demorest and Morrison (2016) and Morrison et al. (2019). Demorest and Morrison (2016, p. 189) define cultural distance as "the degree to which the musics of any two cultures differ in the statistical patterns of pitch and rhythm" and propose that it can be quantified by statistical models of melodies. The hypothesis states that cultural distance between music from different cultures predicts the efficacy with which familiar with either culture process music of the

other culture. The approach pursued in Chapter 7 adds another dimension to this concept by showing that the way in which the perceptual and perceptual learning abilities of listeners are modeled plays a role in the quantification cultural distance. Furthermore, I have cautioned (in the General discussion section of Chapter 7, beginning at page 187) that one should be mindful of the information-theoretic properties of measures used to quantify cultural distance: estimates of cross-entropy (the average information content of events in a composition) have a lower bound that corresponds to the entropy (the inherent uncertainty, or loosely, complexity) of the distribution from which a sample of music is drawn. In practical terms, there may be differences in the complexity of stimulus materials from different cultures that confound observed differences between the performance of (human or modeled) test subjects on these materials. A similar point has been made by Cameron et al. (2015), who emphasized that in a cross-cultural study, only statistical interactions between the cultural background of participants and the cultural origin of stimulus materials should be interpreted.

Predictive-processing accounts of rhythm perception have been previously proposed by Vuust et al. (2014), Vuust and Witek (2014) and Vuust, Dietz, Witek, and Kringelbach (2018). These accounts, however, are based on a conceptual analysis and description of the predictive processing theory. As the introduction of this thesis states, a compelling argument for the use of computational or mathematical models to express theories is that they demand a high level of precision and explicitness and that they can be used to work out the precise consequences of a theory. This is valuable especially when a theory proposes interacting and path-dependent processes. For example, Vuust and Witek (2014) observe that a predictive processing account of rhythm perception suggests that prior experience plays an important role in rhythm perception. However, regarding this conclusion, it might seem invoking the predictive processing theory does not offer much additional insight. On the other hand, a generative model such as the one described in this thesis provides the opportunity to *simulate* the effect of prior experience, using empirical samples of rhythms, and to generate precise predictions of the consequences for rhythm perception. These predictions are based partly on the predictive processing theory and partly on the particular implementation of the generative model.

It could be argued that one of the main strengths of the predictive processing theory is that it suggests a specific way in which perception and cognition can be modeled, namely using probabilistic generative models. Vuust et al. (2018) suggest that syncopations cause prediction error and that prediction error can therefore be calculated using Longuet-Higgins and Lee's (1984) model of syncopation. Simultaneously, however, Vuust et al. (2014) suggest that syncopations may "become predicted at the higher levels" (p. 349). This illustrates that the meaning of "prediction error" depends crucially on the generative model and how prediction errors emerge in this model as a result of an organism's interaction with the

environment. The ideas and findings of Vuust and Witek (2014) and Vuust et al. (2018) are interesting but could be made more powerful when combined with concrete descriptions of a generative model such as the one described in this thesis. Their theory of rhythmic incongruity (Vuust et al., 2018), for example, is based on notions like prediction error and "metrical uncertainty", both of which can be precisely quantified by a generative model of rhythms, making it possible to actually test whether the predictions of the theory agree with musical intuitions or are borne out in experiments.

The above considerations show that probabilistic generative models of rhythm perception can play a relevant role in contemporary theorizing about rhythm perception. The unified approach of using dynamic Bayesian network models that generate the same type of observations (rhythms) based on different underlying generative models, as demonstrated in Chapter 7, makes it possible not only to quantify concepts like syncopation, rhythmic incongruity, and metrical uncertainty but to quantify them as a function of different histories of previous exposure to music and different theories of rhythm perception. This makes it possible to investigate many new questions: Which rhythms are predicted to be metrically ambiguous according to a generative model based on traditional theories of meter? At which points in the same rhythm do the strongest violations of expectation occur for listeners with different histories of previous exposure? Can we find rhythms that contain salient syncopations that are nevertheless strongly predicted? Addressing such questions depends crucially on formal descriptions of generative models that can be implemented and tested in computer simulations.

In summary, I have argued that existing models of rhythm and meter perception do not sufficiently account for the effects of culturally embedded experience, practice, and training on rhythm perception. I have proposed a model that learns from patterns and regularities in datasets of rhythms and presented a set of tools and methodologies for designing, describing, and evaluating probabilistic models of music perception in a cross-cultural context. The model represents one step toward greater consideration in computational modeling of rhythm perception of the environment in which perception is shaped and fine-tuned. I believe that the direction in which this step takes us is both fruitful and worth pursuing further, and I hope that the tools, techniques, and results presented in this thesis will promote this.

Appendix A

Common Lisp model implementations

Below, implementations of the enculturation model and the classical model, as defined in Chapter 5, are shown. These implementations are complete specifications the two models save for a definition of the conditional probability distributions.

The definitions make use of a Common Lisp implementation of the framework described in Chapter 3, developed by the author. It is the author's current intention to make this framework, tentatively named "jackdaw" available as free (as in freedom) software as soon as possible.

Models are defined by a macro `DEFMODEL`, which extends `DEFCLASS`. The general form of a model definition is `(DEFMODEL NAME SUPER-CLASSES SLOTS VARIABLES DISTRIBUTIONS)`, where `SLOTS` is a set of parameters of the congruency constraints, `VARIABLES` is a list of variable definitions, and `DISTRIBUTIONS` defines the probability distribution of each variable that does not have a uniform distribution (which is the default). An example of a parameter to the congruency constraints used by the classical model and the enculturation model is the inter-onset interval domain, stored in the `ioi-domain` class slot in the definitions below.

Each variable definition has the following form: `(VARIABLE-NAME DEPENDENCIES CONGRUENCY-CONSTRAINT)`. `DEPENDENCIES` is a the subset of the variable's dependencies that are relevant to the congruency constraints. In this list, horizontal dependencies are marked by the `^` prefix, such that `^X` refers to a horizontal dependency on `X`. `CONGRUENCY-CONSTRAINT` defines the congruency constraint. Congruency constraints are defined as an anonymous function that is applied to the variables listed in `DEPENDENCIES`. The values of these dependencies are referenced in the congruency constraint by prefixing their name with the `$` symbol (this is to ensure the reserved symbol `T` can also be used as a variable name). For

example, `$x` refers to a value of `X` and `$^x` refers to the previous value of `X`. For the sake of consistency with the notation conventions used in Chapter 3, I use capital letters to define variable names and lowercase letters to reference their values.

Probability distribution definitions in `DISTRIBUTIONS` are of the form `(VARIABLE-NAME DEPENDENCIES DISTRIBUTION)`. `DEPENDENCIES` is the set of variables required to assign a probability to each congruent state. `DISTRIBUTION` is a reference to an implementation of the probability distribution, which is not shown here.

In Chapter 3, there is no distinction between the dependencies of a variable's probability distribution and of its congruency constraint. In the implementation, however, they may be different sets. This is because it may be that not all variables on which a probability distribution depends are relevant to the congruency constraint and vice versa. An example of this is the `D` variable of the enculturation model: its probability distribution depends on the metrical category while its congruency constraint depends on the previous downbeat distance, `d`, and the previous phase, `^p`. The actual dependencies of the variable are given by the union of probability-distribution dependencies and congruency-constraint dependencies.

A.1 Enculturation model

The code snippet below shows an implementation of the enculturation model.

```
(defmodel enculturation (generative-model)
  ((ioi-domain :initarg :ioi-domain :reader ioi-domain)
   (meter-domain :initarg :meter-domain :reader meter-domain)
   (training? :initarg :training? :accessor training?
              :initform nil))
  ;; Congruency constraints
  ((M (^m)
      (if (eq $^m '*)
      (meter-domain model)
      (list $^m)))
   (D (^d ^p)
      (if (eq $^d '*)
      (list '(*))
      (loop for i in (ioi-domain model)
         collect (cons (+ $^p i) $^d))))
   (P (^p m d)
      (if (eq $^p *)
      (list (mod (car $d) (car $m)))
      (loop for p below (car $m)
         collect p)))
   (I (d ^p)
    (if (eq $d '(*))
        (list '*)
        (list (- (car $d) $^p)))))
  ;; Probability distributions
  ((D (m) (accumulator-model))
   (M () (categorical)))
  :required-fields (ioi-domain))
```

The names of the variables are the same as those used in the model-definition table in Chapter 5. The above definition is intentionally verbose in order to explicitly highlight the continuity with the model-definition table. For example, the variable definition of `M` (M) shows that its congruency constraint has one dependency, `^m` ($\hat{m}$). If `^m` equals the symbol $*$, which is the deterministic state of each variable in the initialization model, it generates the meter domain (a set of possible metrical categories). Otherwise, it generates its previous value.

The variable definitions can compressed by introducing some *macros* which define different types of generic behavior. In particular, the macros `PERSISTENT`, `RECURSIVE`, and `DETERMINISTIC` are used to define persistent, recursive, and

deterministic variables. Their definitions are shown below.

```
(defmacro recursive (constraint initialization-constraint)
  `(if (eq $^self '*) ,initialization-constraint ,constraint))

(defmacro persistent (constraint)
  `(recursive (list $^self) ,constraint))

(defmacro deterministic (congruent-value)
  `(list ,congruent-value))
```

Here, `^self` is a special variable name that always refers to the variable itself.

The `DETERMINISTIC` macro does not save any typing, but is intended to enhance readability.

Using these macros, the definition of the enculturation model can be compressed into the definition below.

```
(defmodel enculturation (generative-model)
  ((ioi-domain :initarg :ioi-domain :reader ioi-domain)
   (meter-domain :initarg :meter-domain :reader meter-domain)
   (training? :initarg :training? :accessor training?
              :initform nil))
  ;; Congruency constraints
  ((M (^m) (persistent (meter-domain model)))
   (D (^d ^p)
      (recursive (loop for i in (ioi-domain model)
            collect (cons (+ $^p i) $^d))
         (deterministic '(*))))
   (P (^p m d)
      (recursive (loop for phase below (car $m)
            collect phase)
         (list (mod (car $d) (car $m)))))
   (I (d ^p)
    (recursive (list (- (car $d) $^p))
           (list '*))))
  ;; Probability distributions
  ((D (m) (accumulator-model))
   (M () (categorical)))
  :required-fields (ioi-domain))
```

A.2 Classical model

The definition of the classical model, using the macros defined above, is shown below.

```
(defmodel classical (generative-model)
  ((tactus-intervals :initarg :tactus-intervals
                     :reader tactus-intervals)
   (ioi-domain :initarg :ioi-domain :reader ioi-domain))
  ((U (^u) (persistent '(2 3)))
   (L (^l) (persistent '(2 3)))
   (T (^t) (persistent (tactus-intervals model)))
   (UPH (^uph u) (persistent (loop for uph below $u collect uph)))
   (TPH (^tph t) (persistent (loop for tph below $t collect tph)))
   (P (^p i u t uph tph)
       (recursive (deterministic (mod (+ $^p $i) (* $u $t)))
          (deterministic (+ $tph (* $uph $t)))))
   (I (^i ^p)
    (recursive (ioi-domain model)
           (deterministic '*))))
  ((U () (bernouilli :symbols '(3 2)))
   (L () (bernouilli :symbols '(3 2)))
   (T () (categorical))
   (UPH (u) (bar-phase))
   (TPH (t) (tactus-phase))
   (I (^p t u l) (classical-ioi))))
```

Appendix B

Empirical rhythm space samples

This appendix describes how the rhythm datasets used in Chapter 7 were derived from music corpora. The German rhythms were derived from the Essen Folksong Collection (Schaffrath & Huron, 1995), Dutch rhythms from the Meertens Tunes Collection (Van Kranenburg et al., 2014), and Turkish rhythms from the SymbTr corpus (Karaosmanoğlu, 2012).

We used all melodies from the SymbTr corpus and Meertens Tunes Collections. From the Essen folksong collection, we used only melodies that originate from Germany (based on the `ARE` reference record [Huron, 1999]).

The melodies were first converted from the representations used by the corpora into a common tabular format. We then segmented each rhythm into passages during which the time signature does not change. This yields three rhythm *datasets*. From these dataset, empirical *samples* of a rhythm space were created. These two steps are described below.

B.1 Dataset construction

The SymbTr corpus represents melodies in a format called `mu2` and a text-based format consisting of tab-separated data derived from the `mu2` format, described by Karaosmanoğlu (2012). Because we could not find a precise definition the `mu2` format, we derived our samples from the text-based format. We used a custom script to extract the onset information from the text files and to create a tabular representation that records the onset time of each note in each melody. Melodies in the Meertens Tunes collection and the Essen folksong collection are provided

in the **krn representation of the Humdrum syntax (Huron, 1999). We used a custom script (based on IDyOM's **krn parser) to convert each melody into the same tabular representation format. The resulting rhythms were then segmented into passages during which the time signature does not change. Any unmetered passages were discarded.

B.2 Empirical rhythm space samples

We chose a rhythm length of 29 for the rhythm space to make optimal use of the available data within the constraints imposed by the study. Longer rhythm lengths would have resulted in having to discard more rhythms because they are not long enough, but a shorter rhythm length means more notes are discarded because they are truncated. The value of 29 approximately optimizes the total number of events that end up in the dataset.

Sixteenth notes were chosen as a resolution because most rhythms in the three corpora can be accommodated by this resolution. Furthermore, the computational requirements of the models, in particular the enculturation model, are affected by this resolution. Using sixteenth-notes, these computational requirements remain acceptable.

To obtain a set of rhythms that satisy the length and resolution constraints, we first discarded all rhythms with less than 29 inter-onset intervals, which would have resulted in a larger sample. From the resulting rhythms, we discarded all rhythms that contain inter-onset intervals that cannot be expressed by integer multiples of sixteenth notes. Each rhythm was then truncated to the first 29 inter-onset-intervals. The resulting set of rhythms is treated as a sample of an empirical distribution over rhythms in the rhythm space.

We opted not to split each rhythm into segments of 29 inter-onset intervals. This choice was made to rule out a "beginning effect" or an "ending effect": It could be that being near the beginning or near the ending of a melody influences the statistical properties of rhythms.

Data-driven studies involving rhythms sometimes choose to discard only the events that cannot be accommodated (e.g. Holzapfel, 2015; Temperley, 2010), rather than the entire rhythm as we have done. Our choice to discard the entire rhythm is motivated by our interest in the sequential statistics of rhythmic patterns: removing events in the middle of a rhythm may disrupt these statistics.

Bibliography

Agawu, K. (1995). The invention of "African rhythm". *Journal of the American Musicological Society*, *48*(3), 380–395.

Agawu, K. (2006). Structural analysis or cultural analysis? competing perspectives on the "standard pattern" of West African rhythm. *Journal of the American Musicological Society*, *59*(1), 1–46.

Allen, M. & Friston, K. J. (2018). From cognitivism to autopoiesis: Towards a computational framework for the embodied mind. *Synthese*, *195*(6), 2459–2482.

Anderson, M. L. (2003). Embodied cognition: A field guide. *Artificial Intelligence*, *149*(1), 91–130.

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. In *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Bouwer, F. L., Burgoyne, J. A., Odijk, D., Honing, H., & Grahn, J. A. (2018). What makes a rhythm complex? The influence of musical training and accent type on beat perception. *PLOS One*, *13*(1), 1–26.

Bouwer, F. L. & Honing, H. (2015). Temporal attending and prediction influence the perception of metrical rhythm: Evidence from reaction times and erps. *Frontiers in Psychology*, *6*, 1–14.

Bouwer, F. L., Honing, H., & Slagter, H. A. (2020). Beat-based and memory-based temporal expectations in rhythm: Similar perceptual effects, different underlying mechanisms. *Journal of Cognitive Neuroscience*, *32*(7), 1221–1241.

Brooks, R. A. (1991a). *Intelligence without reason*. Massachusetts Institute of Technology. AI Memo 1293.

Brooks, R. A. (1991b). Intelligence without representation. *Artificial Intelligence*, *47*(1–3), 139–159.

Bundy, A. (1990). What kind of field is AI? In D. Partridge & Y. Wilks (Eds.), *The foundations of artificial intelligence: A sourcebook* (pp. 215–222). Cambridge: Cambridge University Press.

Cameron, D. J., Bentley, J., & Grahn, J. A. (2015). Cross-cultural influences on rhythm processing: Reproduction, discrimination, and beat tapping. *Frontiers in Psychology*, *6*(366), 1–11.

Carroll, J. & Long, D. (1989). *Theory of finite automata*. Englewood Cliffs: Prentice Hall.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Where next? *TRENDS in Cognitive Sciences*, *10*(7), 292–293.

Chemero, A. (2003). An outline of a theory of affordances. *Ecological Psychology*, *15*(2), 181–195.

Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, MA: MIT Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences*, *36*(3), 181–253.

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford: Oxford University Press.

Clarke, E. F. (1987). Categorical rhythm perception: An ecological perspective. In A. Gabrielsson (Ed.), *Action and perception in rhythm and music* (pp. 19–33). Stockholm: Royal Swedish Academy of Music.

Clarke, E. F. (1989). The perception of expressive timing in music. *Psychological Research*, *51*(1), 2–9.

Clarke, E. F. (1999). Rhythm and timing in music. In D. Deutsch (Ed.), *The psychology of music* (2nd ed., Chap. 13, pp. 473–500). New York: Academic Press.

Clayton, M. (2000). *Time in Indian music*. Oxford: Oxford University Press.

Cleary, J. G. & Teahan, W. J. (1997). Unbounded length contexts for PPM. *The Computer Journal*, *40*(2/3), 67–75.

Cleary, J. G. & Witten, I. H. (1984). Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications*, *32*(4), 396–402.

Conklin, D. (1990). *Prediction and Entropy of Music* (Master's thesis, University of Calgary).

Conklin, D. & Witten, I. H. (1995). Multiple viewpoint systems for music prediction. *Journal of New Music Research*, *24*(1), 51–73.

Cooper, G. & Meyer, L. B. (1960). *The rhythmic structure of music*. Chicago: The University of Chicago Press.

Cover, T. M. & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Hoboken: John Wiley & Sons.

Creel, S. C. (2011). Specific previous experience affects perception of harmony and meter. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(5), 1512–1526.

Creel, S. C. (2012). Similarity-based restoration of metrical information: Different listening experiences result in different perceptual inferences. *Cognitive Psychology*, *65*(2), 321–351.

Dechter, R. & Larkin, D. (2001). Hybrid processing of beliefs and constraints. In J. S. B. Breese & D. Koller (Eds.), *Proceedings of the 17th conference in uncertainty in artificial intelligence* (pp. 112–119). San Francisco: Morgan Kaufmann.

Demorest, S. M. & Morrison, S. J. (2016). Quantifying culture: The cultural distance hypothesis of melodic expectancy. In J. Y. Chiao, S.-C. Li, R. Seligman, & R. Turner (Eds.), *The Oxford handbook of cultural neuroscience* (Chap. 12, pp. 183–194). Oxford: Oxford University Press.

Demorest, S. M., Morrison, S. J., Münir, N. B., & Jungbluth, D. (2008). Lost in translation: An enculturation effect in music memory performance. *Music Perception*, *25*(3), 213–223.

Demorest, S. M., Morrison, S. J., Nguyen, V. Q., & Bodnar, E. N. (2016). The influence of contextual cues on cultural bias in music memory. *Music Perception*, *33*(5), 590–600.

Desain, P. & Honing, H. (1992). *Music, mind and machine*. Amsterdam: Thesis Publishers.

Desain, P. & Honing, H. (1994). Advanced issues in beat induction modeling: Syncopation, tempo and timing. In *Proceedings of the 1994 international computer music conference* (pp. 92–94). San Francisco: International Computer Music Association.

Desain, P. & Honing, H. (1999). Computational models of beat induction: The rule-based approach. *Journal of New Music Research*, *28*(1), 29–42.

Desain, P. & Honing, H. (2003). The formation of rhythmic categories and metric priming. *Perception*, *32*(3), 341–365.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895–1923.

Eck, D., Gasser, M., & Port, R. (2000). Dynamics and embodiment in beat induction. In P. Desain & L. Windsor (Eds.), *Rhythm perception and production* (pp. 157–170). Lisse: Swets & Zeitlinger.

Egermann, H., Pearce, M. T., Wiggins, G. A., & McAdams, S. (2013). Probabilistic models of expectation violation predict psychophysiological emotional responses to live concert music. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(3), 533–553.

Fitch, W. T. & Rosenfeld, A. J. (2007). Perception and production of syncopated rhythms. *Music Perception*, *25*(1), 43–58.

Friston, K. J. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *360*(1456), 815–836.

van Gelder, T. (1995). What might cognition be, if not computation? *The Journal of Philosophy*, *92*(7), 345–381.

Gibson, E. J. (1963). Perceptual learning. *Annual Review of Psychology*, *14*, 29–56.

Gibson, J. J. (1979). *An ecological approach to visual perception*. Boston, MA: Houghton-Mifflin.

Gingras, B., Pearce, M. T., Goodchild, M., Dean, R. T., Wiggins, G. A., & McAdams, S. (2016). Linking melodic expectation to expressive performance timing and perceived musical tension. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(4), 594–609.

Gordon, A. D., Henzinger, T. A., Nori, A. V., & Rajamani, S. K. (2014). Probabilistic programming. In *FOSE 2014: Future of software engineering proceedings* (pp. 167–181). FOSE 2014. New York: Association for Computing Machinery.

Gouyon, F. & Dixon, S. (2005). A review of automatic rhythm description systems. *Computer Music Journal*, *29*(1), 34–54.

Grahn, J. A. & Brett, M. (2007). Rhythm and beat perception in motor areas of the brain. *Journal of Cognitive Neuroscience*, *19*(5), 893–906.

Hannon, E. E., Snyder, J. S., Eerola, T., & Krumhansl, C. L. (2004). The role of melodic and temporal cues in perceiving musical meter. *30*(5), 956–974.

Hannon, E. E., Soley, G., & Ullal, S. (2012). Familiarity overrides complexity in rhythm perception: A cross-cultural comparison of American and Turkish listeners. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(3), 543–548.

Hannon, E. E. & Trehub, S. E. (2005a). Metrical categories in infancy and adulthood. *Psychological Science*, *16*(1), 48–55.

Hannon, E. E. & Trehub, S. E. (2005b). Tuning in to musical rhythms: Infants learn more readily than adults. *Proceedings of the National Academy of Sciences*, *102*(35), 12639–12643.

Hansen, N. C. (2010). The legacy of Lerdahl and Jackendoff's A Generative Theory of Tonal Music: Bridging a significant event in the history of music theory and recent developments in cognitive music research. *Danish Yearbook of Musicology*, *38*, 33–55.

Hansen, N. C. & Pearce, M. T. (2014). Predictive uncertainty in auditory sequence processing. *Frontiers in Psychology*, *5*, 1–17.

Hoel, P. G., Port, S. C., & Stone, C. J. (1971). *Introduction to probability theory*. Boston: Houghton Mifflin.

Holzapfel, A. (2015). Relation between surface rhythm and rhythmic modes in Turkish makam music. *Journal of New Music Research*, *44*(1), 25–38.

Honing, H. (2006). Computational modeling of music cognition: A case study on model selection. *Music Perception*, *23*(5), 365–376.

Honing, H. (2009). *Musical cognition*. New Brunswick: Transaction Publishers.

Honing, H. (2013). Structure and interpretation of rhythm in music. In D. Deutsch (Ed.), *The psychology of music* (3rd ed., pp. 369–404). London: Academic Press.

Honing, H. & Bouwer, F. L. (2019). Rhythm. In J. Rentfrow & . Levitin D (Eds.), *Foundations of music psychology: Theory and research* (pp. 33–70). Cambridge, MA: MIT Press.

Honing, H., ten Cate, C., Peretz, I., & Trehub, S. E. (2015). Without it no music: Cognition, biology and evolution of musicality. *Philosophical Transactions of the Royal Society B*, *370*, 1–8.

Honing, H. & de Haas, W. B. (2008). Swing once more: Relating timing and tempo in expert jazz drumming. *Music Perception*, *25*(5), 471–476.

Hünefeldt, T. & Brunetti, R. (2004). Artificial intelligence as "theoretical psychology": Christopher Longuet-Higgins' contribution to cognitive science. *Cognitive Processing*, *5*(3), 137–139.

Huron, D. B. (1999). *Music research using humdrum: A user's guide*. Ohio State University.

Huron, D. B. (2006). *Sweet anticipation: Music and the psychology of expectation*. Cambridge, MA: MIT Press.

Huron, D. B. (2008). Lost in music. *Nature*, *453*(7194), 456–457.

Iyer, V. (1998). *Microstructures of feel, macrostructures of sound: Embodied cognition in West African and African-American musics* (Doctoral dissertation, University of California, Berkeley).

Jacoby, N., Margulis, E. H., Clayton, M., Hannon, E., Honing, H., Iversen, J., ... Wald-Fuhrmann, M. (2020). Cross-cultural work in music cognition: Challenges, insights, and recommendations. *Music Perception*, *37*(3), 185–195.

Jacoby, N. & McDermott, J. H. (2017). Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. *Current Biology*, *27*(3), 359–370.

Jacoby, N., Undurraga, E. A., McPherson, M. J., Valdés, J., Ossandón, T., & McDermott, J. H. (2019). Universal and non-universal features of musical pitch perception revealed by singing. *Current Biology*, *29*(19), 3229–3243.

Jones, M. R. & Boltz, M. (1989). Dynamic attending and responses to time. *Psychological Review*, *96*(3), 459–491.

Jurafsky, D. & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Englewood Cliffs: Prentice Hall.

Karaosmanoğlu, M. K. (2012). A Turkish makam music symbolic database for music information retrieval: SymbTr. In *Proceedings of the 13th international society for music information retrieval* (pp. 223–228).

Knill, D. C. & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, *27*(12), 712–719.

van Kranenburg, P., Bruin, M., de Grijp, L. P., & Wiering, F. (2014). The Meertens Tune Collections. *Meertens Online Reports*, *2014*(1), 1–17.

Kvifte, T. (2007). On the perception of meter. *Ethnomusicology*, *51*(1), 64–84.

Large, E. W. (2008). Resonating to musical rhythm: Theory and experiment. In S. Grondin (Ed.), *Psychology of time* (pp. 189–231). Bingley, UK: Emerald Publishing Group.

Large, E. W. (2010a). A dynamical systems approach to musical tonality. In R. Huys & V. K. Jirsa (Eds.), *Nonlinear dynamics in human behavior* (pp. 193–211). Berlin: Springer.

Large, E. W. (2010b). Neurodynamics of music. In M. R. Jones, R. R. Fay, & A. N. Popper (Eds.), *Music perception* (Vol. 36, pp. 201–231). Springer Handbook of Auditory Research. New York: Springer.

Large, E. W., Herrera, J. A., & Velasco, M. J. (2015). Neural networks for beat perception in musical rhythm. *Frontiers in Systems Neuroscience*, *11*, 1–14.

Large, E. W. & Jones, M. R. (1999). The dynamics of attending: How people track time-varying events. *Psychological Review*, *106*(1), 119–159.

Large, E. W. & Kolen, J. F. (1994). Resonance and the perception of musical meter. *Connection Science*, *6*(1), 177–208.

Large, E. W. & Palmer, C. (2002). Perceiving temporal regularity in music. *Cognitive Science*, *26*(1), 1–37.

Large, E. W. & Snyder, J. S. (2009). Pulse and meter as neural resonance. *Annals of the New York Academy of Sciences*, *1169*(1), 46–57.

Lee, C. S. (1991). The perception of metrical structure: Experimental evidence and a model. In P. Howell, R. West, & I. Cross (Eds.), *Representing musical structure* (pp. 59–127). London: Academic Press.

Lerdahl, F. & Jackendoff, R. (1983). *A generative theory of tonal music*. Cambridge, MA: MIT Press.

Locke, D. (1982). Principles of offbeat timing and cross-rhythm in southern Eve dance drumming. *Ethnomusicology*, *26*(2), 217–246.

London, J. (1993). Loud rests and other strange metric phenomena or, meter as heard. *Music Theory Online*, (2).

London, J. (1995). Some examples of complex meters and their implications for models of metric perception. *Music Perception*, *13*(1), 59–77.

London, J. (2004). *Hearing in time: Psychological aspects of musical meter*. New York: Oxford University Press.

London, J. (2012). *Hearing in time* (2nd ed.). New York: Oxford University Press.

London, J., Polak, R., & Jacoby, N. (2017). Rhythm histograms and musical meter: A corpus study of Malian percussion music. *Psychonomic Bulletin & Review*, *24*(2), 474–480.

Longuet-Higgins, H. C. (1973). Comments on the Lighthill report. In *Artificial intelligence: A paper symposium*. London: Science Research Council. (Republished in Longuet-Higgins (1987), pp. 45–46)

Longuet-Higgins, H. C. (1976). Perception of melodies. *Nature*, *263*, 646–653.

Longuet-Higgins, H. C. (1978). The perception of music. *Interdisciplinary Science Reviews*, *3*(2), 148–156.

Longuet-Higgins, H. C. (1979). The perception of music. *Proceedings of the Royal Society of London B*, *205*(1160), 307–322.

Longuet-Higgins, H. C. (1981). Artificial intelligence—a new theoretical psychology? *Cognition*, *10*(1–3), 197–200. (Republished in Longuet-Higgins (1987), pp. 30–39)

Longuet-Higgins, H. C. (1987). *Mental processes*. Cambridge, MA: MIT Press.

Longuet-Higgins, H. C. & Lee, C. S. (1982). The perception of musical rhythms. *Perception*, *11*(2), 115–128.

Longuet-Higgins, H. C. & Lee, C. S. (1984). The rhythmic interpretation of monophonic music. *Music Perception*, *1*(4), 424–441.

Longuet-Higgins, H. C. & Steedman, M. J. (1971). On interpreting Bach. In B. Meltzer & D. Michie (Eds.), *Machine intelligence* (Vol. 6, pp. 221–241). Edinburgh: Edinburgh University Press.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge: Cambridge University Press.

Manning, C. D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Marcus, S. (2001). Rhythmic modes in Middle-Eastern music. In V. Danielson, D. Reynolds, & S. Marcus (Eds.), *Garland encyclopedia of world music, volume 6: Middle East* (pp. 89–92). London: Routledge.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: W. H. Freeman.

Mateescu, R. & Dechter, R. (2008). Mixed deterministic and probabilistic networks. *Annals of Mathematics and Artificial Intelligence*, *54*(1), 3–51.

McAuley, J. D. (1993). *Learning to perceive and produce rhythmic patterns in an artificial neural network*. Computer Science Department, Indiana University.

McAuley, J. D. (1994). Finding metrical structure in time. In M. C. Mozer, D. S. Touretzky, & P. Smolensky (Eds.), *Proceedings of the 1993 connectionist models summer school* (pp. 219–227). New York: Psychology Press.

McAuley, J. D. (1995). *Perception of time as phase: Toward an adaptive-oscillator model of rhythmic pattern processing* (Doctoral dissertation, Indiana University).

Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., ... Glowacki, L. (2019). Universality and diversity in human song. *Science*, *366*(6468).

Meyer, L. B. (1957). Meaning in music and information theory. *The Journal of Aesthetics and Art Criticism*, *15*(4), 412–424.

Milch, B. (2006). *BLOG: Probabilistic models with unknown objects* (Doctoral dissertation, University of California, Berkeley).

Moffat, A. (1990). Implementing the PPM data compression scheme. *IEEE Transactions on Communications*, *38*(11), 1917–1921.

Morrison, S. J. & Demorest, S. M. (2009). Cultural constraints on music perception and cognition. In J. Y. Chiao (Ed.), *Cultural neuroscience: Cultural influences on brain function* (Vol. 178, pp. 67–77). Elsevier.

Morrison, S. J., Demorest, S. M., & Pearce, M. T. (2019). Cultural distance: A computational approach to exploring cultural influences on music cognition. In M. H. Thaut & D. A. Hodges (Eds.), *The Oxford handbook of music and the brain* (pp. 42–56). New York: Oxford University Press.

Murphy, K. P. (2002). *Dynamic Bayesian networks: Representation, inference and learning* (Doctoral dissertation, University of California, Berkeley).

Newell, A. & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, *19*(3), 113–126.

Olshausen, B. A. & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*, 607–609.

Omigie, D., Pearce, M. T., & Stewart, L. (2012). Tracking of pitch probabilities in congenital amusia. *Neuropsychologia*, *50*(7), 1483–1493.

Omigie, D., Pearce, M. T., Williamson, V. J., & Stewart, L. (2013). Electrophysiological correlates of melodic processing in congenital amusia. *Neuropsychologia*, *51*(9), 1749–1762.

Palmer, C. & Krumhansl, C. L. (1990). Mental representations for musical meter. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(4), 728–741.

Patel, A. D. & Demorest, S. M. (2013). Comparative music cognition: Cross-species and cross-cultural studies. In D. Deutsch (Ed.), *The psychology of music* (3rd ed., pp. 647–681). London: Academic Press.

Pearce, M. T. (2005). *The construction and evaluation of statistical models of melodic structure in music perception and composition* (Doctoral dissertation, City University, London).

Pearce, M. T. (2018). Statistical learning and probabilistic prediction in music cognition: Mechanisms of stylistic enculturation. *Annals of the New York Academy of Sciences*, *1423*(1), 378–395.

Pearce, M. T., Conklin, D., & Wiggins, G. A. (2005). Methods for combining statistical models of music. In U. K. Wiil (Ed.), *Computer music modeling and retrieval* (pp. 295–312). Berlin, Heidelberg: Springer.

Pearce, M. T., Müllensiefen, D., & Wiggins, G. A. (2010). The role of expectation and probabilistic learning in auditory boundary perception: A model comparison. *Perception*, *39*(10), 1367–1391.

Pearce, M. T. & Wiggins, G. A. (2004). Improved methods for statistical modelling of monophonic music. *Journal of New Music Research*, *33*(4), 367–385.

Pearce, M. T. & Wiggins, G. A. (2012). Auditory expectation: The information dynamics of music perception and cognition. *Topics in Cognitive Science*, *4*(4), 625–652.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. New York: Cambridge University Press.

Polak, R., Jacoby, N., Fischinger, T., Goldberg, D., Holzapfel, A., & London, J. (2018). Rhythmic prototypes across cultures: A comparative study of tapping synchronization. *Music Perception*, *36*(1), 1–23.

Polak, R., London, J., & Jacoby, N. (2016). Both isochronous and non-isochronous metrical subdivision afford precise and stable ensemble entrainment: A corpus study of Malian jembe drumming. *Frontiers in Neuroscience*, *10*, 1–11.

Povel, D.-J. & Essens, P. (1985). Perception of temporal patterns. *Music Perception*, *2*(4), 411–440.

Rao, R. P. N. & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Ravignani, A., Thompson, B., Grossi, T., Delgado, T., & Kirby, S. (2018). Evolving building blocks of rhythm: How human cognition creates music via cultural transmission. *Annals of the New York Academy of Sciences*, *1423*(1), 176–187.

Repp, B. H. (1995). Expressive timing in Schumann's "Träumerei:" An analysis of performances by graduate student pianists. *The Journal of the Acoustical Society of America*, *98*(5), 2413–2427.

Repp, B. H. (2005). Sensorimotor synchronization: A review of the tapping literature. *Psychonomic Bulletin & Review*, *12*(6), 969–992.

Repp, B. H. & Su, Y.-H. (2013). Sensorimotor synchronization: A review of recent research (2006–2012). *Psychonomic Bulletin & Review*, *20*(3), 403–452.

Russel, S. & Norvig, P. (2003). *Artificial intelligence: A modern approach* (2nd ed.). Upper Saddle River: Pearson Education.

Russell, S. (2015). Unifying logic and probability. *Communications of the ACM*, *58*(7), 88–97.

Sadakata, M., Desain, P., & Honing, H. (2006). The Bayesian way to relate rhythm perception and production. *Music Perception*, *23*(3), 269–288.

Savage, P. E., Brown, S., Sakai, E., & Currie, T. E. (2015). Statistical universals reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences*, *112*(29), 8987–8992.

Schaffrath, H. & Huron, D. B. (1995). The essen folksong collection in the humdrum kern format. Retrieved 2018, from https://kern.humdrum.org/cgi-bin/browse?l=/essen

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Shmulevich, I. & Povel, D.-J. (2000). Measures of temporal pattern complexity. *Journal of New Music Research*, *29*(1), 61–69.

Shove, P. & Repp, B. H. (1995). Musical motion and performance: Theoretical and empirical perspectives. In J. Rink (Ed.), *The practice of performance: Studies*

in musical interpretation (pp. 55–83). Cambridge: Cambridge University Press.

Simoncelli, E. P. & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*, 1193–1216.

Smith, E. C. & Lewicki, M. S. (2006). Efficient auditory coding. *Nature*, *439*, 978–982.

Soley, G. & Hannon, E. E. (2010). Infants prefer the musical meter of their own culture: A cross-cultural comparison. *Developmental Psychology*, *46*(1), 286–292.

Song, C., Simpson, A. J. R., Harte, C. A., Pearce, M. T., & Sandler, M. B. (2013). Syncopation and the score. *PLOS One*, *8*(9), 1–7.

Steedman, M. J. (1977). The perception of musical rhythm and metre. *Perception*, *6*(5), 555–569.

Stevens, C. J. (2012). Music perception and cognition: A review of recent cross-cultural research. *Topics in Cognitive Science*, *4*(4), 653–667.

Stobart, H. & Cross, I. (2000). The Andean anacrusis? rhythmic structure and perception in easter songs of northern Potosí, Bolivia. *British Journal of Ethnomusicology*, *9*(2), 63–92.

Teki, S., Grube, M., & Griffiths, T. D. (2012). A unified model of time perception accounts for duration-based and beat-based timing mechanisms. *Frontiers in Integrative Neuroscience*, *5*, 1–7.

Temperley, D. (2000). Meter and grouping in African music: A view from music theory. *Ethnomusicology*, *44*(1), 65–96.

Temperley, D. (2001). *The cognition of basic musical structures*. Cambridge, MA: MIT Press.

Temperley, D. (2004). An evaluation system for metrical models. *Computer Music Journal*, *28*(3), 28–44.

Temperley, D. (2007). *Music and probability*. Cambridge, MA: MIT Press.

Temperley, D. (2009). A unified probabilistic model for polyphonic music analysis. *Journal of New Music Research*, *38*(1), 3–18.

Temperley, D. (2010). Modeling common-practice rhythm. *Music Perception*, *27*(5), 355–376.

Temperley, D. (2013). Computational models of music cognition. In D. Deutsch (Ed.), *The psychology of music* (3rd ed., pp. 327–368). London: Academic Press.

Temperley, D. & Sleator, D. (1999). Modeling meter and harmony: A preference-rule approach. *Computer Music Journal*, *23*(1), 10–27.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*, 1279–1285.

Thompson, B., Kirby, S., & Smith, K. (2016). Culture shapes the evolution of cognition. *Proceedings of the National Academy of Sciences*, *113*(16), 4530–4535.

Tichko, P. & Large, E. W. (2019). Modeling infants' perceptual narrowing to musical rhythms: Neural oscillation and Hebbian plasticity. *Annals of the New York Academy of Sciences*, *1453*(1), 125–139.

Todd, N. P. M. & Lee, C. S. (2015). The sensory-motor theory of rhythm and beat induction 20 years on: A new synthesis and future perspectives. *Frontiers in Human Neuroscience*, *9*, 1–25.

Trehub, S. E., Becker, J., & Morley, I. (2015). Cross-cultural perspectives on music and musicality. *Philosophical Transactions of the Royal Society B*, *370*, 1–9.

Velasco, M. J. & Large, E. W. (2011). Pulse detection in syncopated rhythms using neural oscillators. In *Proceedings of the 12th international society for music information retrieval* (pp. 185–190).

Vuust, P., Dietz, M. J., Witek, M., & Kringelbach, M. L. (2018). Now you hear it: A predictive coding model for understanding rhythmic incongruity. *Annals of the New York Academy of Sciences*, *1423*(1), 19–29.

Vuust, P., Gebauer, L. K., & Witek, M. A. G. (2014). Neural underpinnings of music: The polyrhythmic brain. In H. Merchant & V. de Lafuente (Eds.), *Neurobiology of interval timing* (pp. 339–356). New York: Springer.

Vuust, P. & Witek, M. A. G. (2014). Rhythmic complexity and predictive coding: A novel approach to modeling rhythm and meter perception in music. *Frontiers in Psychology*, *5*, 1–14.

van der Weij, B., Pearce, M. T., & Honing, H. (2017). A probabilistic model of meter perception: Simulating enculturation. *Frontiers in Psychology*, *8*, 1–18.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*(6), 598–604.

Wiese, W. & Metzinger, T. (2017). Vanilla PP for philosophers: A primer on predictive processing. In W. Wiese & T. Metzinger (Eds.), *PPP - philosophy and predictive processing* (pp. 1–18). Frankfurt am Main.

Wilson, A. D. & Golonka, S. (2013). Embodied cognition is not what you think it is. *Frontiers in Psychology*, *4*, 1–13.

Witek, M. A. G., Clarke, E. F., Wallentin, M., Kringelbach, M. L., & Vuust, P. (2014). Syncopation, body-movement and pleasure in groove music. *PLOS One*, *9*(4), 1–12.

Witek, M. A. G., Liu, J., Kuubertzie, J., Yankyera, A. P., Adzei, S., & Vuust, P. (2020). A critical cross-cultural study of sensorimotor and groove responses to syncopation among Ghanaian and American university students and staff. *Music Perception: An Interdisciplinary Journal*, *37*(4), 278–297.

Samenvatting

Ervaren Luisteraars: De invloed van langetermijnblootstelling aan muziek op ritmeperceptie

In dit proefschrift onderzoeken we ritmeperceptie met behulp van computationele modelleertechnieken en ontwikkelen we gereedschappen en technieken om probabilistische generatieve modellen van ritmeperceptie te definiëren en evalueren. We beargumenteren dat eerdere computationele modellen van ritmeperceptie onvoldoende verklaren hoe deze vorm van waarneming is gevormd door eerdere ervaringen, oefening en training van cultureel ingebedde luisteraars. Op basis van *predictive processing*-theorieën van waarneming stellen we een probabilistisch generatief model van metrumwaarneming voor, dat, vergeleken met eerdere modellen, voor een groter deel leert van patronen en regelmatigheden in datasets van ritmes (die de muzikale omgeving representeren). De uitkomst van dit leerproces simuleert de effecten van de langetermijnblootstelling aan ritmes die luisteraars ondergaan in hun muzikale omgeving.

De nadruk die het proefschrift legt op patronen en regelmatigheden in muzikale ritmes in de omgeving van de luisteraar leidt ons ertoe het model in een inter- en intra-culturele context te evalueren. We onderzoeken hoe een variatie in de gevoeligheid voor statistische patronen, de stijl of culturele oorsprong van ritmes waaruit het model leert en die van de ritmes waarop het wordt geëvalueerd bijdragen aan de prestaties van het model. Dat wil zeggen dat het model, samen met een alternatief model dat traditionele Westerse theorieën van metrumperceptie vertegenwoordigt, wordt geëvalueerd op cultureel bekende ritmes (ritmes in dezelfde stijl als de ritmes waarvan het model heeft geleerd) en cultureel onbekende ritmes (ritmes in een andere stijl dan de ritmes waarvan het model heeft geleerd). Op deze manier onderzoeken we of er variatie is tussen stijlen in de patronen en

regelmatigheden die nuttig zijn voor het afleiden van metrum en onderzoeken we de hoeveelheid en aard van de gevoeligheid voor statistische patronen die nodig is om deze variatie te kunnen detecteren.

Concreet gezien onderzoeken we empirische selecties van ritmes van Westerse volksliederen en Turkse makammuziek. We stellen vast dat zowel Westerse als Turkse ritmes regelmatigheden bevatten die een luisteraar die bekend is met deze regelmatigheden kan gebruiken om een metrum af te leiden uit individuele ritmes. Deze regelmatigheden omvatten patronen van meerdere ritmische gebeurtenissen en zijn complexer dan de schematische patronen van verwachting, die geassocieerd worden met traditionele theorieën van metrumperceptie. Verder ontdekken we dat sommige patronen alleen voorkomen in Westerse muziek en andere alleen in Turkse makammuziek. De resultaten suggereren echter ook dat makam en Westerse ritmes een groot deel van de patronen en regelmatigheden die het afleiden van metrum faciliteren delen.

Verder presenteert dit proefschrift een kader voor het ontwerp en de implementatie van discrete dynamische Bayesiaanse netwerkmodellen met deterministische beperkingen. Dit kader maakt het mogelijk om formele, bondige en expliciete definities van zulke modellen te geven die eenvoudig vertaald kunnen worden naar werkende en uitvoerbare implementaties. Het kader heeft tot doel de transparantie en reproduceerbaarheid van modelleeronderzoek te verbeteren en het voor andere onderzoekers gemakkelijker te maken om verder te bouwen op bestaand modelleerwerk dat gebruik maakt van dit kader. Het kader is geschikt om theorieën gebaseerd op *predictive processing* te definiëren. Zulke cognitieve modellen, zo ook de modellen besproken in dit proefschrift, kunnen worden gezien als sequentiële muziekvoorspellingsmodellen en kunnen worden gebruikt om muzikale verwachtingen en onzekerheid te modelleren, die zich samen met de muziek dynamisch in de tijd ontwikkelen.

De opbouw van dit proefschrift is als volgt: Hoofdstuk 2 beschrijft een verscheidenheid aan eerdere aanpakken van het modelleren van ritmeperceptie en geeft een overzicht van de eerder aangetoonde manieren waarop ritmeperceptie door eerdere ervaringen, training en oefening wordt beïnvloed. Hoofdstuk 3 ontwikkelt de technische details van het modelleerkader. Hoofdstuk 4 illustreert het gebruik van dit kader door een aanpassing van een eerder voorgesteld generatief model van ritme- en metrumperceptie te presenteren. Hoofdstuk 5 beschrijft de technische details van twee modellen die verderop in het proefschrift worden onderworpen aan empirische evaluaties. Hoofdstuk 6 definieërt en motiveert het primaire model van dit proefschrift (Hoofdstuk 5 bevat een technischer beschrijving van dit model). Tot slot beschrijft Hoofdstuk 7 een inter- en intraculturele studie waarin de modellen en methodologieën ontwikkeld in Hoofdstuk 5 worden toegepast op empirische selecties van Turkse makammuziek en Westerse volksliederen.

Summary

Experienced Listeners: Modeling the influence of long-term musical exposure on rhythm perception

This thesis investigates rhythm perception using computational modeling techniques and develops a set of tools and techniques for the definition and evaluation of probabilistic generative models of music perception. We argue that previously proposed computational models of rhythm perception insufficiently account for how perception has been shaped by culturally embedded listeners' prior experience, practice, and training. Motivated by predictive processing theories of perception, we propose a probabilistic generative model of meter perception which, compared with previous models, to a greater extent learns from patterns and regularities in datasets of rhythms (representing a musical environment). The outcome of this learning process simulates the effects of the long-term exposure that listeners receive to rhythms in their musical environment.

The emphasis on patterns and regularities in musical rhythms in the environment leads us to evaluate the model in a cross-cultural context. We investigate how varying degrees of sensitivity to statistical patterns, the style or cultural origin of rhythms from which the model learns, and that of rhythms on which it is evaluated factor into the model's performance. That is, the model, together with an alternative model representing traditional Western theories of meter perception is evaluated on culturally familiar rhythms (of the same style as the rhythms it has learned from) and culturally unfamiliar rhythms (of a different style than the rhythms it has learned from). In this way, we investigate whether there is between-style variety in the patterns and regularities useful for inferring meter, and the amount and type of sensitivity to statistical patterns necessary for detecting this variation.

Concretely, we investigate empirical samples containing rhythms of Western folksongs and Turkish makam music. We find that the Western as well as the Turkish rhythms contain regularities that allow a listener familiar with these regularities to infer meter from individual rhythms. These regularities involve patterns of multiple rhythmic events and are more complex than schematic patterns of expectation associated with traditional theories of meter perception. Furthermore, we find that some of these patterns occur only in Western or only in makam rhythms. However, the results also suggest that the patterns and regularities by which meter can be inferred are to a significant extent shared between makam and Western rhythms.

Additionally, this thesis presents a framework for the design and implementation of discrete dynamic Bayesian network models with deterministic constraints. The framework enables formal, concise, and explicit definitions of such models that can straightforwardly be translated into functional and executable implementations. The framework aims to enhance the transparency and reproducibility of modeling research and to make it easier for other researchers to build further on modeling work that uses the framework. The framework is suitable for defining predictive-processing based theories of music perception. Such cognitive models, including the models discussed in this thesis, can be seen as sequential music prediction models and can be used to model musical expectancy and uncertainty evolving dynamically over time, in lockstep with the temporal progression of music.

Concerning the structure of this thesis, Chapter 2 surveys a variety of previously pursued approaches to modeling rhythm perception and reviews ways in which rhythm perception has been found to be influenced by prior experience, training, and practice. Chapter 3 develops the technical details of the modeling framework. Chapter 4 demonstrates its use by presenting an adaptation of a previously proposed generative model of rhythm and meter perception. Chapter 5 provides technical definitions of two models that are subjected to empirical evaluations later in the thesis. Chapter 6 defines and motivates the main model proposed in this thesis (Chapter 5 contains a more technical description of this model). Finally, Chapter 7 presents cross-cultural evaluations in which the models and the methodology developed in Chapter 5 are applied to empirical samples of Turkish makam music and Western folksongs.

Titles in the ILLC Dissertation Series:

ILLC DS-2009-01: **Jakub Szymanik**
Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language

ILLC DS-2009-02: **Hartmut Fitz**
Neural Syntax

ILLC DS-2009-03: **Brian Thomas Semmes**
A Game for the Borel Functions

ILLC DS-2009-04: **Sara L. Uckelman**
Modalities in Medieval Logic

ILLC DS-2009-05: **Andreas Witzel**
Knowledge and Games: Theory and Implementation

ILLC DS-2009-06: **Chantal Bax**
Subjectivity after Wittgenstein. Wittgenstein's embodied and embedded subject and the debate about the death of man.

ILLC DS-2009-07: **Kata Balogh**
Theme with Variations. A Context-based Analysis of Focus

ILLC DS-2009-08: **Tomohiro Hoshi**
Epistemic Dynamics and Protocol Information

ILLC DS-2009-09: **Olivia Ladinig**
Temporal expectations and their violations

ILLC DS-2009-10: **Tikitu de Jager**
"Now that you mention it, I wonder...": Awareness, Attention, Assumption

ILLC DS-2009-11: **Michael Franke**
Signal to Act: Game Theory in Pragmatics

ILLC DS-2009-12: **Joel Uckelman**
More Than the Sum of Its Parts: Compact Preference Representation Over Combinatorial Domains

ILLC DS-2009-13: **Stefan Bold**
Cardinals as Ultrapowers. A Canonical Measure Analysis under the Axiom of Determinacy.

ILLC DS-2010-01: **Reut Tsarfaty**
Relational-Realizational Parsing

ILLC DS-2010-02: **Jonathan Zvesper**
Playing with Information

ILLC DS-2010-03: **Cédric Dégremont**
The Temporal Mind. Observations on the logic of belief change in interactive systems

ILLC DS-2010-04: **Daisuke Ikegami**
Games in Set Theory and Logic

ILLC DS-2010-05: **Jarmo Kontinen**
Coherence and Complexity in Fragments of Dependence Logic

ILLC DS-2010-06: **Yanjing Wang**
Epistemic Modelling and Protocol Dynamics

ILLC DS-2010-07: **Marc Staudacher**
Use theories of meaning between conventions and social norms

ILLC DS-2010-08: **Amélie Gheerbrant**
Fixed-Point Logics on Trees

ILLC DS-2010-09: **Gaëlle Fontaine**
Modal Fixpoint Logic: Some Model Theoretic Questions

ILLC DS-2010-10: **Jacob Vosmaer**
Logic, Algebra and Topology. Investigations into canonical extensions, duality theory and point-free topology.

ILLC DS-2010-11: **Nina Gierasimczuk**
Knowing One's Limits. Logical Analysis of Inductive Inference

ILLC DS-2010-12: **Martin Mose Bentzen**
Stit, Iit, and Deontic Logic for Action Types

ILLC DS-2011-01: **Wouter M. Koolen**
Combining Strategies Efficiently: High-Quality Decisions from Conflicting Advice

ILLC DS-2011-02: **Fernando Raymundo Velazquez-Quesada**
Small steps in dynamics of information

ILLC DS-2011-03: **Marijn Koolen**
The Meaning of Structure: the Value of Link Evidence for Information Retrieval

ILLC DS-2011-04: **Junte Zhang**
System Evaluation of Archival Description and Access

ILLC DS-2011-05: **Lauri Keskinen**
Characterizing All Models in Infinite Cardinalities

ILLC DS-2011-06: **Rianne Kaptein**

Effective Focused Retrieval by Exploiting Query Context and Document Structure

ILLC DS-2011-07: **Jop Briët**
Grothendieck Inequalities, Nonlocal Games and Optimization

ILLC DS-2011-08: **Stefan Minica**
Dynamic Logic of Questions

ILLC DS-2011-09: **Raul Andres Leal**
Modalities Through the Looking Glass: A study on coalgebraic modal logic and their applications

ILLC DS-2011-10: **Lena Kurzen**
Complexity in Interaction

ILLC DS-2011-11: **Gideon Borensztajn**
The neural basis of structure in language

ILLC DS-2012-01: **Federico Sangati**
Decomposing and Regenerating Syntactic Trees

ILLC DS-2012-02: **Markos Mylonakis**
Learning the Latent Structure of Translation

ILLC DS-2012-03: **Edgar José Andrade Lotero**
Models of Language: Towards a practice-based account of information in natural language

ILLC DS-2012-04: **Yurii Khomskii**
Regularity Properties and Definability in the Real Number Continuum: idealized forcing, polarized partitions, Hausdorff gaps and mad families in the projective hierarchy.

ILLC DS-2012-05: **David García Soriano**
Query-Efficient Computation in Property Testing and Learning Theory

ILLC DS-2012-06: **Dimitris Gakis**
Contextual Metaphilosophy - The Case of Wittgenstein

ILLC DS-2012-07: **Pietro Galliani**
The Dynamics of Imperfect Information

ILLC DS-2012-08: **Umberto Grandi**
Binary Aggregation with Integrity Constraints

ILLC DS-2012-09: **Wesley Halcrow Holliday**
Knowing What Follows: Epistemic Closure and Epistemic Logic

ILLC DS-2012-10: **Jeremy Meyers**
Locations, Bodies, and Sets: A model theoretic investigation into nominalistic mereologies

ILLC DS-2012-11: **Floor Sietsma**
Logics of Communication and Knowledge

ILLC DS-2012-12: **Joris Dormans**
Engineering emergence: applied theory for game design

ILLC DS-2013-01: **Simon Pauw**
Size Matters: Grounding Quantifiers in Spatial Perception

ILLC DS-2013-02: **Virginie Fiutek**
Playing with Knowledge and Belief

ILLC DS-2013-03: **Giannicola Scarpa**
Quantum entanglement in non-local games, graph parameters and zero-error information theory

ILLC DS-2014-01: **Machiel Keestra**
Sculpting the Space of Actions. Explaining Human Action by Integrating Intentions and Mechanisms

ILLC DS-2014-02: **Thomas Icard**
The Algorithmic Mind: A Study of Inference in Action

ILLC DS-2014-03: **Harald A. Bastiaanse**
Very, Many, Small, Penguins

ILLC DS-2014-04: **Ben Rodenhäuser**
A Matter of Trust: Dynamic Attitudes in Epistemic Logic

ILLC DS-2015-01: **María Inés Crespo**
Affecting Meaning. Subjectivity and evaluativity in gradable adjectives.

ILLC DS-2015-02: **Mathias Winther Madsen**
The Kid, the Clerk, and the Gambler - Critical Studies in Statistics and Cognitive Science

ILLC DS-2015-03: **Shengyang Zhong**
Orthogonality and Quantum Geometry: Towards a Relational Reconstruction of Quantum Theory

ILLC DS-2015-04: **Sumit Sourabh**
Correspondence and Canonicity in Non-Classical Logic

ILLC DS-2015-05: **Facundo Carreiro**
Fragments of Fixpoint Logics: Automata and Expressiveness

ILLC DS-2016-01: **Ivano A. Ciardelli**
Questions in Logic

ILLC DS-2016-02: **Zoé Christoff**
Dynamic Logics of Networks: Information Flow and the Spread of Opinion

ILLC DS-2016-03: **Fleur Leonie Bouwer**
What do we need to hear a beat? The influence of attention, musical abilities, and accents on the perception of metrical rhythm

ILLC DS-2016-04: **Johannes Marti**
Interpreting Linguistic Behavior with Possible World Models

ILLC DS-2016-05: **Phong Lê**
Learning Vector Representations for Sentences - The Recursive Deep Learning Approach

ILLC DS-2016-06: **Gideon Maillette de Buy Wenniger**
Aligning the Foundations of Hierarchical Statistical Machine Translation

ILLC DS-2016-07: **Andreas van Cranenburgh**
Rich Statistical Parsing and Literary Language

ILLC DS-2016-08: **Florian Speelman**
Position-based Quantum Cryptography and Catalytic Computation

ILLC DS-2016-09: **Teresa Piovesan**
Quantum entanglement: insights via graph parameters and conic optimization

ILLC DS-2016-10: **Paula Henk**
Nonstandard Provability for Peano Arithmetic. A Modal Perspective

ILLC DS-2017-01: **Paolo Galeazzi**
Play Without Regret

ILLC DS-2017-02: **Riccardo Pinosio**
The Logic of Kant's Temporal Continuum

ILLC DS-2017-03: **Matthijs Westera**
Exhaustivity and intonation: a unified theory

ILLC DS-2017-04: **Giovanni Cinà**
Categories for the working modal logician

ILLC DS-2017-05: **Shane Noah Steinert-Threlkeld**
Communication and Computation: New Questions About Compositionality

ILLC DS-2017-06: **Peter Hawke**
The Problem of Epistemic Relevance

ILLC DS-2017-07: **Aybüke Özgün**
Evidence in Epistemic Logic: A Topological Perspective

ILLC DS-2017-08: **Raquel Garrido Alhama**
Computational Modelling of Artificial Language Learning: Retention, Recognition & Recurrence

ILLC DS-2017-09: **Miloš Stanojević**
Permutation Forests for Modeling Word Order in Machine Translation

ILLC DS-2018-01: **Berit Janssen**
Retained or Lost in Transmission? Analyzing and Predicting Stability in Dutch Folk Songs

ILLC DS-2018-02: **Hugo Huurdeman**
Supporting the Complex Dynamics of the Information Seeking Process

ILLC DS-2018-03: **Corina Koolen**
Reading beyond the female: The relationship between perception of author gender and literary quality

ILLC DS-2018-04: **Jelle Bruineberg**
Anticipating Affordances: Intentionality in self-organizing brain-body-environment systems

ILLC DS-2018-05: **Joachim Daiber**
Typologically Robust Statistical Machine Translation: Understanding and Exploiting Differences and Similarities Between Languages in Machine Translation

ILLC DS-2018-06: **Thomas Brochhagen**
Signaling under Uncertainty

ILLC DS-2018-07: **Julian Schlöder**
Assertion and Rejection

ILLC DS-2018-08: **Srinivasan Arunachalam**
Quantum Algorithms and Learning Theory

ILLC DS-2018-09: **Hugo de Holanda Cunha Nobrega**
Games for functions: Baire classes, Weihrauch degrees, transfinite computations, and ranks

ILLC DS-2018-10: **Chenwei Shi**
Reason to Believe

ILLC DS-2018-11: **Malvin Gattinger**
New Directions in Model Checking Dynamic Epistemic Logic

ILLC DS-2018-12: **Julia Ilin**
Filtration Revisited: Lattices of Stable Non-Classical Logics

ILLC DS-2018-13: **Jeroen Zuiddam**
Algebraic complexity, asymptotic spectra and entanglement polytopes

ILLC DS-2019-01: **Carlos Vaquero**
What Makes A Performer Unique? Idiosyncrasies and commonalities in expressive music performance

ILLC DS-2019-02: **Jort Bergfeld**
Quantum logics for expressing and proving the correctness of quantum programs

ILLC DS-2019-03: **András Gilyén**
Quantum Singular Value Transformation & Its Algorithmic Applications

ILLC DS-2019-04: **Lorenzo Galeotti**
The theory of the generalised real numbers and other topics in logic

ILLC DS-2019-05: **Nadine Theiler**
Taking a unified perspective: Resolutions and highlighting in the semantics of attitudes and particles

ILLC DS-2019-06: **Peter T.S. van der Gulik**
Considerations in Evolutionary Biochemistry

ILLC DS-2019-07: **Frederik Möllerström Lauridsen**
Cuts and Completions: Algebraic aspects of structural proof theory

ILLC DS-2020-01: **Mostafa Dehghani**
Learning with Imperfect Supervision for Language Understanding

ILLC DS-2020-02: **Koen Groenland**
Quantum protocols for few-qubit devices

ILLC DS-2020-03: **Jouke Witteveen**
Parameterized Analysis of Complexity

ILLC DS-2020-04: **Joran van Apeldoorn**
A Quantum View on Convex Optimization

ILLC DS-2020-05: **Tom Bannink**
Quantum and stochastic processes

Institute for Logic, Language and Computation