A Cognitive Information Theory of Music:
A Computational Memetics Approach

Tak-Shing Thomas Chan
Department of Computing
Goldsmiths, University of London
New Cross
London SE14 6NW

Abstract

This thesis offers an account of music cognition based on information theory and memetics. My research strategy is to split the memetic modelling into four layers: Data, Information, Psychology and Application. Multiple cognitive models are proposed for the Information and Psychology layers, and the MDL best-fit models with published human data are selected. Then, for the Psychology layer only, new experiments are conducted to validate the best-fit models.

In the information chapter, an information-theoretic model of musical memory is proposed, along with two competing models. The proposed model exhibited a better fit with human data than the competing models. Higher-level psychological theories are then built on top of this information layer. In the similarity chapter, I proposed three competing models of musical similarity, and conducted a new experiment to validate the best-fit model. In the fitness chapter, I again proposed three competing models of musical fitness, and conducted a new experiment to validate the best-fit model. In both cases, the correlations with human data are statistically significant.

All in all, my research has shown that the memetic strategy is sound, and the modelling results are encouraging. Implications of this research are discussed.

Contents

List of Figures

Acknowledgements

Declaration

I, Tak-Shing Thomas Chan, hereby declare that the work presented in this thesis is entirely my own. Two publications arose from this thesis:

Chan, T.-S. T. and Wiggins, G. A. (2005). A computational memetic approach to music information and aesthetic fitness. In Gervás, P., Veale, T., and Pease, A., editors, *Proceedings of the IJCAI'05 Workshop on Computational Creativity*, pages 105–108. Departamento de Sistemas Informáticos y Programación, Universid Complutense de Madrid, Madrid.

Chan, T.-S. T. and Wiggins, G. A. (2006). More evidence for a computational memetic approach to music information and new interpretations of an aesthetic fitness measure. In Colton, S. and Pease, A., editors, *Proceedings of the ECAI'06 Workshop on Computational Creativity*, pages 13–17. Università di Trento, Trento, Italy.

Signature ....

Date ....

Chapter 1

Introduction

My research hypothesis is that a cognitive modelling approach to music information can account for memetic phenomena (henceforth "a cognitive information hypothesis"). To test this hypothesis, I will first develop a multi-layer strategy for this research (this chapter). I will then construct a novel model of cognitive information, which quantifies the amount of musical information in bits, and will correlate model predictions with human data in two experiments.

1.1 What is Computational Memetics of Music?

Memetics is the study of Darwinian evolution in culture (Hofstadter, 1985), where the *meme* is defined as a unit of cultural transmission and *copying-fidelity* refers to the accuracy of such transmissions (Dawkins, 1976). Chan and Wiggins (2002) coined the term *computational memetics of music*; while Chan and Wiggins were focusing on one specific simulation model, the term *computational memetics of music* is used here to cover the intersection between two existing subfields of memetics:

1. Computational memetics: the study of memetics using computational modelling techniques (Best, 2001);

2. Memetics of music: the application of the memetic framework to musicology, including cognitive musicology (Jan, 2000a).

Researchers in this newly proposed subfield aim to produce theoretical and empirical computational models of musical culture, modelling musical memory (the meme), musical similarity (copying-fidelity), musical value (fitness), and related topics within the unifying biological framework of memetics. As yet, no one seems to know concretely

what a meme is (Hull, 2000). Is it then premature to study the science of memetics? Hull (2000) thinks not:

> [M]emeticists cannot begin to understand what the science of memetics is until they generate some general beliefs about conceptual change and try to test them. These tests are likely to look fairly paltry, but in the early stages of a science, attempts at testing always look fairly paltry [...] I want to urge memeticists to ignore the in-principle objections that have been raised to memetics no matter how cogent they may turn out to be and proceed to develop their theory in the context of attempts to test it. (p. 49)

1.2   Why a Cognitive Information Hypothesis?

How should we measure musical culture quantitatively? What constitutes a unit of cultural evolution? We can answer these questions by borrowing from information theory and cognitive science.

By definition, memes contain information. Indeed, the subtitle of the *Journal of Memetics* is *Evolutionary Models of Information Transmission*. Information theory allows us to break musical memes down to bits. It even allows us to use bioinformatics tools in conjunction with the measures proposed here, if we accept the meme-gene analogy.[1] In other words, when a meme is operationally defined as a unit of information, we have a precise, testable measure of cultural evolution.

Secondly, the musical experience is cognitive and affective, so a theory of music information should reflect this.[2] In other words, music memetics should be a cognitive theory. Note that my cognitive memetics is different from Castelfranchi's (2001) "cognitive memetics", as Castelfranchi is talking about "cognition" in the context of autonomous agents (not necessarily human); he models agents that can decide to accept or reject incoming memes based on "cognitive" rules. While I share with Castelfranchi the idea that memetics should be cognitive, my thesis is at a much lower level of abstraction. Here I follow Neisser (1967) who defines *cognition* as

> [...] all the processes by which the sensory input is transformed, reduced, elaborated, stored, recovered, and used. It is concerned with these processes even when they operate in the absence of relevant stimulation, as in images and hallucinations. Such terms as *sensation*, *perception*, *imagery*, *retention*, *recall*,

---

[1]It is outside the scope of this thesis to discuss the advantages and pitfalls of the meme-gene analogy.
[2]See Chapter 2 for Meyer's view on music and information theory.

problem-solving, and *thinking*, among many others, refer to hypothetical stages or aspects of cognition. (p. 4)

Thirdly, in anthropology, Goodenough (1957) defines culture as cognitive in nature:

[A] society's culture consists of whatever it is one has to know or believe in order to operate in a manner acceptable to its members, and do so in any role that they accept for any one of themselves. Culture, being what people have to learn as distinct from their biological heritage, must consist of the end product of learning: knowledge, in a most general, if relative, sense of the term. (p. 167)

Furthermore, there is what Plotkin (2000) calls "Kitcher's rule": Kitcher (1987) claimed that without psychological foundations there cannot be a natural science of culture (the context was sociobiology). Plotkin (2000) took Kitcher's claim as a self-evident truth in the context of memetics, and I agree with Plotkin.

Another motivation for my cognitive information hypothesis comes from Schneider et al.'s (1986) measure of information for DNA sequences. Using modern terminology,[3] their measure is equal to the sum of mutual information values between the nucleotide sequence $B \in \{A, C, G, T\}$ and each of the binding sites L:

$$R_{\text{sequence}} = \sum_L I(B; L).$$

Binding sites are essentially DNA regions that have been recognised by specific macromolecules such as polymerases and ribosomes (Schneider et al., 1986). Put another way, binding sites correspond with recognisable DNA patterns. As such, the R_{sequence} measure is at least related to recognition if not cognition as its value depends on both the sequence itself and the macromolecular recognisers. The fact that even geneticists have something like cognition in their information measures motivates me to do the same in the context of memetics.

1.3 A Multi-Layer Research Strategy

Here I propose the research strategy used throughout this thesis (Figure 1.1). My strategy is to split the memetic modelling into four layers. The *Data Layer* corresponds to low-level perceptual inputs. Its role is to provide the cultural molecules from which the memetic

---

[3]See Chapter 2 for Shannon's definition of *information rate*, or *mutual information* as it is usually known today.

codes are built. The *Information Layer* corresponds to my cognitive information measure, whose output should ideally be validated by psychophysical experiments. The *Psychology Layer* would include aesthetic fitness, categorisation, familiarity, similarity, and so on, where each component must be validated psychologically. Finally, the *Application Layer* would include, *inter alia*, creative systems, evolutionary musicology and music information retrieval. The advantages of multi-layer modelling is that it allows us to reuse the same *Application* and *Psychology Layers* to describe phenomena across different domains while requiring only changes in the *Information* and *Data Layers*.

Level	Layer	Examples
4	Application	Creative systems and cultural ecology
3	Psychology	Similarity and aesthetic fitness measures
2	Information	Cognitive information measures
1	Data	Computational musical code

Figure 1.1: A multi-layer research strategy

1.4 Scope and Limitations

For this thesis, I will limit my domain to polyphonic Western tonal music and restrict my investigations to the following topics:

1. Modelling musical complexity (cognitive information) and extending this information measure to include joint, conditional and mutual information (Information Layer);

2. Modelling musical similarity based on this cognitive information measure, and testing the model with psychological experiments (Psychology Layer);

3. Modelling musical value (defined here as a subjective measure of psychological affect) based on this cognitive information measure, and testing the model with psychological experiments (Psychology Layer).

As music memetics is not limited to the three topics above, the enumeration and subsequent investigations of other topics (along with assessments of their statistical importance) would be the subject of future research (*e.g.*, recombination, mutation and transmission mechanisms in general). As in many other computational theories of music perception (Lerdahl and Jackendoff, 1983; Large et al., 1995; Narmour, 1999; Temperley and Sleator, 1999), my proposed theory does not account for lyrics, timbre and dynamics, as

these factors are not hierarchical in nature (Lerdahl and Jackendoff, 1983). Furthermore, my theory is unable to account for microtonal music and problems of attention. These are the main limitations of my theory.

1.5   Organisation of The Thesis

This thesis is set out to demonstrate the feasibility of, and to provide evidence for, my cognitive information hypothesis. This thesis consists of six chapters. Chapter 2 reviews relevant literature which forms the basis of this research. Chapter 3 presents the cognitive information theory and its mathematical properties (Information Layer). Within the framework of this information theory, Chapters 4 and 5 present models of musical similarity and musical fitness (Psychology Layer). Chapter 7 summarises the methodologies, results and implications of this thesis, with a short rebuttal to the claim that memetics is a "discredited label" (Edmonds, 2005), followed by prospects of research. The information sheet and consent form used in Chapters 4 and 5 are reproduced in Appendix A. An introduction to computability is presented in Appendix B. Terms not defined in the main text are defined in the Glossary.

Chapter 2

Literature Review

2.1 Introduction

My proposed cognitive model of music information is a synthesis of musicology and information theory. In general, this synthesis is not new and dates back to Pinkerton (1956), or even Birkhoff (1933) if we allow the modern interpretation (Kolmogorov, 1965; Stiny and Gips, 1978; Koshelev, 1998) that order and complexity are measures of information. However, my synthesis is novel because it incorporates physiological, psychological and evolutionary principles into an information model (hitherto lacking in the current literature). Therefore, in this chapter, I will present the background necessary for the construction of my information theory of music.

Meyer (1957) was the first to postulate an explicit link between information theory and music psychology. He first hypothesised that "the psycho-stylistic conditions which give rise to musical meaning, whether affective or intellectual, are the same as those which communicate information" (p. 412). Assuming the central role of expectations in musical experience, he then interpreted musical expectations as internalised probabilities, and thus as equivalent to uncertainties or information.

In this seminal paper, Meyer postulated music as a Markov process with a built-in "systemic uncertainty", countered by the "designed uncertainty" of the composer. According to Meyer, there is a systemic tendency for information to vanish as the music unfolds, and the composer's "designed uncertainty" has the effect of going against this tendency. In relation to musical styles, he called the probabilities from the norm of a style as the "latent expectation" of that style.

Finally, Meyer speculated that "perhaps values as well arise only as the result of the uncertainties involved in making means-end choices" (p. 424). This potentially provides a bridge between a value-neutral information theory and musical value, which is cur-

rently a difficult unresolved problem.

In pondering whether an accurate quantification of musical information is possible at all, he identified two important requirements for the quantification to work (pp. 422–423):

1. "First we must arrive at a more precise and empirically validated account of mental [behaviour] which will make it possible to introduce the more or less invariant probabilities of human mental processes into the calculation of the probabilities involved in the style. This account need not necessarily be statistical itself."

2. "Second, and this is ultimately dependent upon the first, it is necessary to develop a more precise and sensitive understanding of the nature of musical experience."

Subsequent literature searches have not yielded any relevant papers[1] that treat information as a cognitive measurement. There are two ways to proceed: one is to use Shannon information, which requires a precise knowledge of mental events and their probabilities; the other is to propose a new cognitive model of information. As precise models of the brain are beyond our current technological reach, by elimination the only alternative is to propose a new cognitive information theory. However, there is yet another problem: Meyer's "mental behaviour" and "musical experience" are rather vague requirements. Therefore, I will refine them into three major building blocks (in decreasing order of importance):

1. Mental representations: the study of how one encodes knowledge in the mind. Assuming the physicalist[2] position, mental representations can be studied either physiologically or psychologically (or both). In this review, I will report several experimental and theoretical results in the literature, with a primary focus on music representation;

2. Music informatics: the application of data structures and algorithms to music research. This includes general computer science, computer music research, and computational musicology. Music informatics provides the computational tools for modelling musical behaviour;

3. Biomusicology: Wallin (1991) combines the neurophysiological, neuropsychological, and evolutionary aspects of music research into the new field of biomusicology.

---

[1] While there are numerous papers on music and information theories (e.g. Kraehenbuehl and Coons, 1959; Conklin and Witten, 1995), neuroscientific evidence (see Section 2.2.1) suggests that melody-like stimuli are encoded more accurately than random stimuli. Therefore, theories based on transition probabilities are unlikely to account for the perceptual and cognitive constraints in music processing. See also the criticisms of the information-theoretic approach to music by Cohen (1962) and Sharpe (1971), especially the ones based on Chomsky's (1956) observation that no Markov process can include an English grammar.

[2] The physicalist position holds that all mental events are physical.

Since then, this field has been expanded to include comparative musicology and applied biomusicology (Brown et al., 2000).

The rest of this chapter will be devoted to these building blocks, and is organised as follows. I will review mental representations in Section 2.2, and music informatics in Section 2.3. Section 2.4 reviews biomusicology, and finally, Section 2.5 concludes with a summary.

2.2 Mental Representations

Research in mental representations usually fall into two different levels of analysis: the biophysical level or the cognitive level. The first level is concerned with neurophysiology while the latter is concerned with cognitive psychology. Using Marr's terminology (Phillips, 1997), neurophysiology belongs to the hardware implementation level, whereas cognitive psychology belongs to the computational level.

2.2.1 Biophysical Representations

At the biophysical level, researchers aim to model low-level neuronal structures instead of high-level cognitive schemata. Relevant research along this line includes Licklider (1951) and de Cheveigné (1993) on auditory perception, Oja (1982) and Sanger (1989) on mathematical modelling of neural networks, Patel and Balaban (2000) on cortical representations, and the speculation of Narmour (1999) on neuronal representations of melodies.

In auditory perception, Licklider (1951) proposed a model of pitch perception (*i.e.*, mental imagery of pitch), in which the main component is called the "neuronal autocorrelator", hypothesised to exist in the brain, with inputs supplied by the cochlea (a part of the inner ear that translates sounds into nerve impulses). The neuronal autocorrelator is defined as follows (for a single channel of cochlear input):

$$E_{D_K}(t) = \overline{N_A(t)N_A(t - k\Delta\tau)}$$

where E_{D_K} denotes the excitation of the output neuron, the overbar denotes the running average, t denotes time, $N_A(t)$ denotes the state of the input neuron, and $\Delta\tau$ denotes the synaptic delay (propagation delay from one neuron to the next). Licklider (1951) calls it a "duplex theory of pitch perception" because it incorporates frequency-domain analysis (by the cochlea) and time-domain analysis (by the neuronal autocorrelators). The weakness of this model is the assumption of long delay lines, which lacked physiological

evidence (de Cheveigné, 1993). Licklider (1951) also proposed a model of the overall analyser, in which an array of neuronal autocorrelators were used for multiple channels of cochlea inputs (see Figure 2.1). In this figure, x represents the frequency dimension and τ represents the synaptic delay. Together with the time dimension t, the overall system represents an auditory stimulus in three dimensions: t, τ, and x (Licklider, 1951).
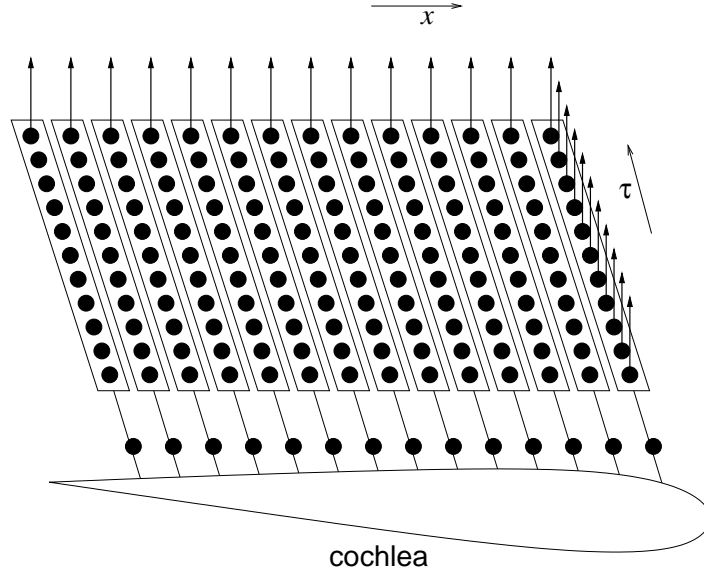
Figure 2.1: Array of neuronal autocorrelators (after Licklider, 1951)

Following Licklider's line of research, de Cheveigné (1993) proposed the neural cancellation model for auditory processing (see Figure 2.2):

$$o(t) = \max(0, i(t) - i(t - T)).$$

It combines the time-domain comb filter in signal processing with the non-negativity constraints in physiology. Non-negativity constraints refer to the fact that the firing rates of neurons cannot be negative. There are two main assumptions for this filter to be physiologically plausible: the existence of long delay lines (same requirement as Licklider's model) and inhibitory synapses. As de Cheveigné (1993) has noted, while the first assumption is still lacking evidence, the second assumption is well accepted in neuroscience. The model was tested on guinea pig auditory-nerve fiber discharge data obtained in response to double vowel stimuli, with success in separating concurrent vowel sounds. In Chapter 3, I will use a modified form of Licklider's (1951) array in which the neuronal autocorrelators are replaced by the neural cancellation filters (de Cheveigné, 1993).
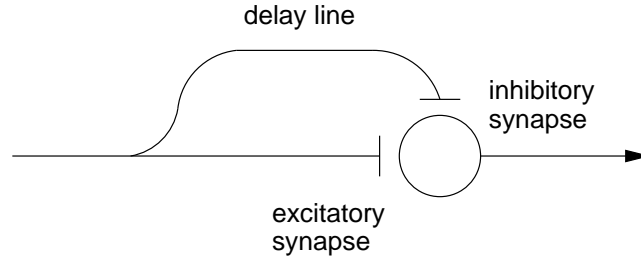
Figure 2.2: Neural "comb" filter (after de Cheveigné, 1993)

The above papers all deal with low-level cochlear inputs, but my proposed research deals with high-level symbolic data. Is it legitimate to adapt the neural cancellation filter for symbolic inputs? It is helpful to view this in light of Lőrincz et al. (2002), who advanced an interesting view of computational neuroscience. They argued that the traditional association of anatomical structure with computational function could be unwarranted. They demonstrated their point by developing a hierarchical neural network model of long-term memory. The model worked well in simulations, but when they mapped the model to the anatomy of the neocortex, functional discrepancies began to appear. They claimed that the discrepancies can be resolved by "questioning the identification of functional and anatomical layers". In other words, it is not *a priori* wrong to reuse the neural cancellation filter to model a different brain function, for it is possible that both functions share the same neural mechanism.

Now I shall turn to the mathematical modelling of neural networks, specifically the well-known works of Oja (1982) and Sanger (1989). Oja (1982) proposed a simplified neuron model that acts as a principal component analyser. Oja's neuron model is:

$$\eta = \sum_{i=1}^{n} \mu_i \xi_i.$$

where $\xi_1, \ldots, \xi_n$ are the inputs, $\mu_1, \ldots, \mu_n$ are the synaptic strengths, and η is the output. His learning equation is based on the normalised Hebbian rule,

$$\mu_i(t+1) = \frac{\mu_i(t) + \gamma \eta(t) \xi_i(t)}{\{\sum_{i=1}^{n} [\mu_i(t) + \gamma \eta(t) \xi_i(t)]^2\}^{1/2}},$$

where γ is positive. Oja (1982) proved that if the input vector $[\xi_1(t) \ldots, \xi_n(t)]^T$ represents a stochastic process, then the neuron would become a principal component analyser as t approaches infinity.

Sanger (1989) went beyond a single neuron and proposed an "optimality principle"

of neural network training by maximising the ability to reconstruct the input data given the network outputs. For a single-layer feedforword network (see Figure 2.3), Sanger proposed the following "Generalised Hebbian Algorithm":

$$C_{ji}(t+1) = C_{ji}(t) + \gamma \left[y_j(t)x_i(t) - \sum_{k=1}^{j} y_j(t)y_k(t)C_{ki}(t) \right]$$

where $C(t)$ is the weight matrix, γ is the learning rate, $x(t)$ is the input vector, and $y(t)$ is the output vector such that $y_j(t) = \sum_{i=1}^{n} C_{ji}(t)x_i(t)$.

$$x_n(t) \qquad C_{m \times n}(t) \qquad y_m(t)$$
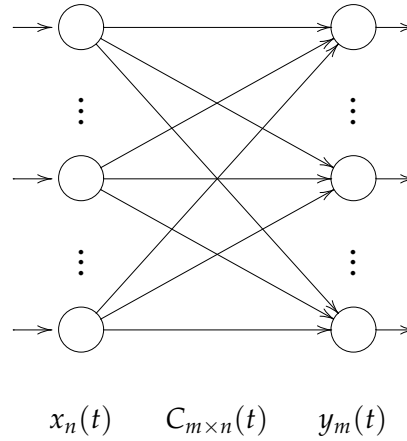
Figure 2.3: A single-layer feedforward neural network

Sanger (1989) proved that his rule would cause the rows of $C(t)$ to converge to the eigenvalues of the correlation matrix $E[xx^T]$, which coincides with the singular value decomposition.

This correlational aspect of learning has also been verified physiologically by *in vitro* hippocampal slice recordings. Using a pair of strong and weak stimuli, Stanton and Sejnowski (1989) showed that an increase in synaptic strength (long-term potentiation) is elicited when the stimuli are applied in phase, while a decrease in synaptic strength (long-term depression) is elicited when the stimuli are applied out of phase. Stanton and Sejnowski then concluded that the mechanisms of associative long-term potentiation and depression can compute and store the covariance matrix of the inputs in the hippocampus. This provides further evidence for Sanger's "optimality principle".

On cortical representations, Patel and Balaban (2000) provided neuroscientific evidence that parts of the brain are tracking the pitch contour of tone sequences where the tracking accuracy is proportional to musical predictability. Their participants listened

to twenty-eight one-minute tone sequences of varying degrees of predictability, chosen from random sequences (hardest to predict), $1/f$ sequences (the second hardest), $1/f^2$ sequences (easier), and scales (the easiest). The MEG signals were simultaneously recorded by a 148-channel whole-head biomagnetometer. They found that in all cases the reconstructed MEG phase spectrum bore a significant resemblance to the stimuli pattern. For all participants, the correlation coefficients between the input sequences and the MEG increased in the order of: *random* $< 1/f < 1/f^2 <$ *scales*, meaning that their participants are worst at tracking random music and best at tracking scales. This unequal tracking accuracy motivates me to propose, in Chapter 3, an alternative information theory of music that is not directly based on the probabilities of the tone sequences.

Finally, Narmour (1999) speculated that in a neuronal representation of melodies in the brain, each neuron should be *level-topic* (store hierarchical function) and *tonotopic* (store melodic pitch), and perhaps *chronotopic* as well (store manifest duration). The connections between neurons would then store the learned path of expectations. Narmour's speculation implies that hierarchical levels are as important as pitches and durations. This point will be taken into consideration in Chapter 3.

In summary, this subsection reviews biophysical representations. The concepts specifically relevant to my proposed research are: Licklider's (1951) array, de Cheveigné's (1993) neural cancellation filter, the link from neural networks to the "optimality principle" (Sanger, 1989), the experimental result that melody-like music is encoded more faithfully than random music (Patel and Balaban, 2000), and the importance of hierarchical levels in music representations (Narmour, 1999). Next, I review cognitive representations.

2.2.2 Cognitive Representations

At the cognitive level, representations refer to higher-level cognitive schemata, which are usually functional rather than biophysical. In this line of research, relevant work includes Atkinson and Shiffrin (1968) on short-term memory; Levitin (1994) and Levitin and Cook (1996) on absolute memory; Lerdahl and Jackendoff (1983) and Temperley and Sleator (1999) on well-formedness and preference rules; and Large et al. (1995) on reduced memory representations.

Atkinson and Shiffrin (1968) proposed an influential model of human memory (reproduced in Figure 2.4) where it has three components: sensory register (SR), short-term store (STS) and long-term store (LTS). The short-term store can hold information for about thirty seconds without rehearsal (Atkinson and Shiffrin, 1968). Another theory, at least in music psychology, is that the STS has a duration of 3–5 seconds on average, occasionally

up to 10–12 seconds depending on the complexity of stimuli (Snyder, 2000). The "brain-clock" theory (Poppel, 1989) goes even further and suggests that musical memory are segmented into three-second units. For the purpose of this thesis, I will not attempt to resolve this discrepency, but instead I will follow Atkinson and Shiffrin's original thirty-second limit.

$$\boxed{\text{Environmental Input}} \longrightarrow \boxed{\text{SR}} \longrightarrow \boxed{\text{STS}} \longleftrightarrow \boxed{\text{LTS}}$$
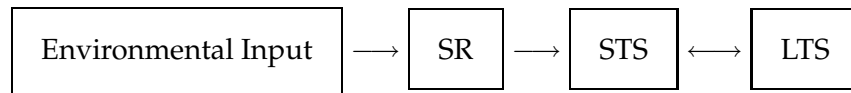
Figure 2.4: Block diagram of Atkinson and Shiffrin's (1968) model

In the quest of relative/absolute memory for music, Levitin (1994) and Levitin and Cook (1996) obtained evidence that long-term auditory memory is absolute, with respect to both pitch and tempo. Levitin (1994) asked forty-six participants to sing two popular songs from memory. Each song constitutes a trial. All of them reported that they had not heard their selected song in the past seventy-two hours. Three participants withdrew from the experiment after the first trial. Of the forty-three participants who completed both trials, 12% of the participants got the correct pitch on both trials, and a further 32% were within two semitones of the correct pitch on both trials. This result suggests that pitch memory is absolute. In Levitin and Cook (1996), the same dataset was re-analysed for tempo. For both trials combined, 72% of the participants were within 8% of the correct tempo. This result suggests that memory for tempo is also absolute (Levitin and Cook, 1996). This absolute nature of tempo memory prompts me to use an absolute notion of time in the proposed theory.

As regards well-formedness and preference rules, Lerdahl and Jackendoff (1983) were the first to model music psychology using such rules. In their well-known *Generative Theory of Tonal Music* (GTTM), Lerdahl and Jackendoff (1983) first made two important idealisations: that the listeners are experienced in Western tonal music; and that there exists a final state of understanding (*i.e.*, a cognitive representation of the music). These are two important idealisations that I will adopt wholesale into my proposed information model.

Secondly, Lerdahl and Jackendoff (1983) limited their investigations to four hierarchical parts of listener's musical intuitions: grouping structure (segmentation of music into sections and smaller units), metrical structure (multiple levels of strong and weak beats), time-span reduction (a tree structure showing relative importance of the notes), and prolongational reduction (showing tension and relaxation). They achieved this goal by proposing three types of rules (Lerdahl and Jackendoff, 1983):

1. Well-formedness rules: these constrain the space of possible musical structures. This rules correspond to generative grammar in linguistic theory;

2. Preference rules: these rules correspond to experienced listeners' preferred interpretation of a piece. Although preference rules do not correspond to any part of Chomskian linguistics, they are necessary because musical intuitions can often be ambiguous;

3. Transformational rules: these include the grouping overlap, grouping elision and metrical deletion rules, which cannot be modelled using the well-formedness rules. These rules do not play a major role in GTTM.

Temperley and Sleator (1999) provided a preference-rule approach of meter modelling based on GTTM. Temperley and Sleator (1999) found that GTTM had problems with rubato performances due to the regularity well-formedness rule, so they relaxed this into a regularity preference-rule (prefer evenly spaced beats), avoiding the rigidity of the well-formedness rule. Their approach also consists of two other preference rules adapted from GTTM: event rule (prefer beats that aligns with event onsets) and length rule (prefer beats that aligns with onsets of longer notes). The actual search procedure is based on dynamic programming with a score table, with columns representing quantised time (35ms) and rows representing beat intervals. They have implemented this in their `meter` program, which I will use in this research.

Finally, Large et al. (1995) proposed a reduced memory model of music as a neural network that performs lossy compression on input melodies. The model was validated experimentally: six skilled pianists were asked to improvise ten variations each on three children's melodies, and the variations thus produced correlated significantly with the reconstruction errors predicted by the neural network (Large et al., 1995). This result suggests that the human brain might use a form of compression in storing melodies.

In summary, this subsection reviews cognitive representations of music. The concepts specifically relevant to my proposed research are: short-term memory (Atkinson and Shiffrin, 1968), absolute musical memory (Levitin, 1994; Levitin and Cook, 1996), Temperley and Sleator's (1999) `meter` program, and a compression-based model of musical memory (Large et al., 1995).

2.3 Music Informatics

As I am proposing a new information theory of music, it is logical that a review of music informatics is in order. In this section, I will mainly review information theories and time

series analysis as applied to the mathematical modelling of music.[3]

2.3.1 Information Theories

Shannon's (1948) paper is the seminal paper in information theory. Shannon is interested in the problem of message transmission from an information source to its destination, where semantic aspects are not relevant. The basis of his information theory is Hartley's information measure,

$$H(N) = \log_2 N,$$

where N is the number of possible messages. Shannon noticed that the number of possible messages increases exponentially with time. If the communication system is governed by stochastic processes, then by using this statistical structure one could reduce the required capacity of the transmission channel.

Shannon began by assuming that every sequence generated by the same information source has the same statistical structure. Then he defined the well-known information measure called the entropy,

$$H(x) = -\sum_{i=1}^{n} p_i \log p_i,$$

for the set of probabilities $\{p_1, \ldots, p_n\}$ that characterises x (the information source). Furthermore, Shannon defined the joint entropy

$$H(x, y) = -\sum_{i,j} p(i, j) \log p(i, j)$$

and the conditional entropy

$$H_x(y) = -\sum_{i,j} p(i, j) \log p_i(j),$$

corresponding to joint and conditional probabilities of events. In modern notation, $H_x(y)$ is usually written as $H(y|x)$. He then proved the following theorems (for proofs see Shannon, 1948):

Theorem 2.1 (Shannon). *$H(x, y) \leq H(x) + H(y)$ with equality iff x and y are independent.*

Theorem 2.2 (Shannon). *$H(x, y) = H(x) + H_x(y) = H(y) + H_y(x)$.*

---

[3]Concerned readers might object to my omission of *music and connectionism*. In my view, music and connectionism belongs elsewhere since their nature is psychological rather than mathematical (*cf.* my citation of Large et al. (1995) in Section 2.2.2). More to the point, my proposed model is at the computational (mathematical) level, not the algorithmic (connectionist) level.

Finally, Shannon defined the information rate, also known as mutual information $I(x;y)$, as $R = H(x) - H_y(x)$, where $H(x)$ represents the information source and $H_y(x)$ represents the equivocation or noise entropy, which characterises the ambiguity due to transmission noise. This definition, plus the two theorems above, are ubiquitous in the literature; I will refer to them collectively as the *three laws of Shannon*, and will attempt to prove these laws for my new information theory as well. A version of his information theory for continuous probability distributions is also given by Shannon (1948):

$$H(x) = - \int_{-\infty}^{\infty} p(x) \log p(x) dx,$$

where the joint and conditional entropies are

$$H(x,y) = - \iint p(x,y) \log p(x,y) dx dy$$

and

$$H_x(y) = - \iint p(x,y) \log p_x(y) dx dy.$$

Kolmogorov (1968) noted that the Shannon definition of entropy

> [...] used probability concepts, and thus does not pertain to individual values, but to random values, *i.e.*, to probability distributions within a group of values [...] By far, not all applications of information theory fit rationally into such an interpretation of its basic concepts. I believe that the need for attaching definite meaning to the expressions $H(x|y)$ and $[I(x;y)]$, in the case of individual values x and y that are not viewed as a result of random tests with a definite law of distribution, was realized long ago by many who dealt with information theory. (p. 662)

In a related paper, Kolmogorov (1965) proposed an algorithmic approach to the quantitative definition of information. He believed that the algorithmic approach would give rise to a correct definition of "hereditary information", for instance the amount of information required for the reproduction of a cockroach (Kolmogorov, 1965). His approach is based on the "quantity of information conveyed by an individual object x about an individual object y". Kolmogorov noted that while this can be done in the probabilistic approach:

$$I(x;y) = \iint P_{xy} \log_2 \frac{P_{xy}}{P_x P_y} dx dy,$$

it is not always meaningful in practice as $I(x;y)$ depends on the complexity of the schemes used to describe the objects. Furthermore, the characteristics of objects might

not be random variables. Accordingly, Kolmogorov defined the relative complexity of an object y given x as the length of the minimal program p that outputs y given x,

$$K_A(y|x) = \min_{A(p,x)=y} l(p)$$

where A is the asymptotically optimal programming method such that for any other programming methods $\varphi(p,x)$ we have the inequality

$$K_A(y|x) \leq \min_{\varphi(p,x)=y} l(p) + C_\varphi$$

where the constant C_φ depends only on φ. Kolmogorov then defined the complexity of y as

$$K_A(y) = K_A(y|1)$$

and the "quantity of information conveyed by x about y" as

$$I_A(x:y) = K_A(y) - K_A(y|x).$$

Note that in modern literature (e.g. Bennett et al., 1998), the subscript A is often dropped, and the optimal programming method is tacitly assumed. Interestingly, up to a logarithmic term, Kolmogorov complexity obeys the three laws of Shannon as well (Kolmogorov, 1965; Hammer et al., 2000).

A closely related measure is called the Levin complexity. While Kolmogorov complexity does not consider the running time, Levin complexity does. Levin complexity is defined as (Koshelev, 1998; Levin, 1973):

$$a(x) = \min_{A(p)=x} \left\{ t(p) \cdot 2^{l(p)} \right\}$$

where $t(p)$ is the running time of the program p and A is the asymptotically optimal programming method as described in the previous paragraph. While the concepts of Kolmogorov and Levin complexities will not be directly used in my thesis (since they are uncomputable; see Appendix B), they are nonetheless essential for understanding two other sections in this literature survey (specifically the ones dealing with similarity and algorithmic aesthetics). Furthermore, Kolmogorov's argument for a probability-free information theory will apply with equal force to my proposed information theory (Chapter 3).

In summary, this subsection covers information theories. In Chapter 3, I will develop my proposed information theory by building on Kolmogorov (1965), while incorporating

existing knowledge about mental representations (Section 2.2), and relating it to the three laws of Shannon. The next section deals with time series analysis.

2.3.2 Time Series Analysis

A distinctive approach to the analysis of symbolic musical data is time series analysis. One of the pioneering papers, Dirst and Weigend (1994), noted that themes in fugues are traditionally subjected to symmetry transformations (see Figure 2.5). Motivated by this musicological fact, they devised three representation schemes for four-part fugues:

- The x-representation: x_t is a four-dimensional vector denoting the pitch values of the four voices at time t (in semiquavers).

- The difference representation: $d_t = x_t - x_{t-1}$, representing the four pitch intervals at time t.

- The run length representation: each note is denoted by (p, l), where p denotes the pitch number and l denotes its length. Dirst and Weigend remarked that this scheme does not preserve vertical alignment.

Musical term	Operation
Transposition	$x \leftarrow x + c$ (translation)
Retrograde	$t \leftarrow -t$ (time reversal)
Inversion	$d \leftarrow -d$ (pitch reflection)
Diminution	$t \leftarrow 2t$
Augmentation	$t \leftarrow 0.5t$

Figure 2.5: Symmetry transformations (after Dirst and Weigend, 1994)

Given these time series representations, Dirst and Weigend proposed several ways to analyse the horizontal (melody), vertical (harmony), and higher order structure of music. For horizontal analysis, they proposed to use Markov models and Fourier transforms on the pitch, intervals, or length time series. For vertical analysis, they proposed to use standard connectionist techniques on the x-representation. For higher order structure, the authors have presented a theme finding algorithm based on the difference representation and clustering techniques. Finally, for the modelling of expectations, they proposed to use a neural network model that predicts the next event, whereby a large prediction error is interpreted as a violation of the musical expectation.[4] What are noteworthy here

---

[4]However, nothing was said about empirical validation.

are their repeated mentions of the difference representation, which I will also be using in Chapter 3.

Boon and Decroly (1995) showed that symbolic pitch data can be quantised into a time series so that dynamical systems theory can be used for music analysis (for multi-part music, each part is analysed separately). They used phase space dimensions and spectral analysis to identify global dynamics in a corpus, and a novel entropy measure for local dynamics.

The phase portrait for an n-part composition can be constructed by plotting the n-dimensional pitch trajectories over time in an n-dimensional phase space (Boon and Decroly, 1995). The phase space dimensions (D_f) can be obtained from the log-log plot of $\mathcal{N}(\lambda)$ against λ, where $\mathcal{N}(\lambda)$ is calculated by dividing the phase space into small boxes of size λ and counting the total number of occupied boxes in that space, and D_f is the slope of the plot:

$$\log \mathcal{N}(\lambda) = -D_f \log \lambda.$$

D_f is found to be in the range of $0.94 \leq D_f \leq 1.86$ for the 23 pieces in their corpus. Secondly, the slopes of the log-log power spectra $S(f) \sim 1/f^\nu$ showed that for musical pieces, ν is in the range of $1.79 \leq \nu \leq 1.97$ (Boon and Decroly, 1995).

The most interesting aspect is their new entropy measure. Empirically, Boon and Decroly discovered that conditional entropy of the note s_{i+1} given the previous note s_i "did not reflect consistent significance" for musical sequences. They identified a possible cause as the lack of reference to tonality, then defined a new entropy measure by dividing the note distributions into two sets, P for notes belonging to a reference scale (*e.g.*, A major), and Q for notes outside the scale. This new entropy measure is defined as:

$$H_0' = \frac{H(\gamma P) + H(\delta Q)}{\log N},$$

where N is the total number of notes, $\gamma + \delta = 1$ and $\delta > \gamma$. The last inequality means that non-tonal notes are more surprising and therefore are assigned a higher entropy. Generalising to first order transitions, they used:

$$H_1' = \sum_S H'(S)\mathcal{P}(S)$$

where

$$S \;=\; \text{MIDI pitch number,}$$
$$\theta \;=\; \text{the reference scale,}$$

$$
\begin{aligned}
\nu(S) &= \text{total number of occurrences of pitch } S, \\
H'(S) &= -\frac{\sum_{s \in \theta} \gamma P(s|S) \log(\gamma P(s|S)) + \sum_{s \notin \theta} \delta P(s|S) \log(\delta P(s|S))}{\log \nu(S)}, \\
\mathcal{P}(S) &= \begin{cases} \gamma \nu(S)/N, & \text{if } S \in \theta, \\ \delta \nu(S)/N, & \text{otherwise.} \end{cases}
\end{aligned}
$$

They systematically plotted all measured quantities against each other. Their widely scattered results suggest that there is no evidence of correlations between global and local dynamics (with the notable exception that D_f might be related to H'_1). This result suggests that low-order Markov models (even when tweaked to accommodate tonality) is unable to capture all information within the piece, which corrobates my motivation for a new information theory.

In summary, this subsection deals with time series analysis of symbolic musical data. When viewed in light of biophysical relevance, the difference representation (Dirst and Weigend, 1994) bears a close resemblance to the neural cancellation filter (de Cheveigné, 1993). Therefore, with biological realism in mind, the difference representation (Dirst and Weigend, 1994) will be used in Chapter 3.

2.4 Biomusicology

Biomusicology (see p. 16) is of particular relevance to my thesis. The very first evolutionary tree of music was produced by Lomax's (1980) cantometrics project (Brown et al., 2000), but Lomax did not use the term *meme*, nor did this term exist when the work was carried out in the 1960's. Lomax (1980) and colleagues quantified musical culture by what they called a cantometric profile (consisting of thirty-seven variables, such as nasality, tempo and melodic range, each of them within a 3–6 point scale). After collecting 148 cantometric profiles from all over the world, Lomax (1980) performed a multifactor analysis on the collected profiles and discovered ten major regional factors: Siberian, Circum-Pacific, Nuclear America, African Gatherer, Early Agriculture, Proto-Melanesian, Oceanic, Old High Culture, Central Asian, and West Europe. By subjecting these ten factors to a further clustering procedure, Lomax and colleagues were able to create an evolutionary tree of folk song styles. This tree[5] is reproduced in Figure 2.6 (only primary bonds are reproduced here). Given this tree, Lomax predicted that all world songs have two evolutionary parents: Siberia and African Gatherer. However, the correctness of Lo-

---

[5]This tree should be read from the left to right, with the leftmost nodes (Siberia and African Gatherer) representing the roots. Evolutionary age (causality) is assumed to be positively rank-correlated with the level of socio-economic development, shown on the horizontal axis (Lomax, 1980). Strictly speaking, this is not a tree but a network due to the reticulations.

max's tree has been questioned: Brown et al. (2000) claimed to have an independent (but unpublished) cluster analysis on Lomax's raw cantometric data that contradicts Lomax's prediction. As their tree is not published (nor do they say anything about their methodology), I have no basis for judgement here—I can only say that we need more data before making any such claims.
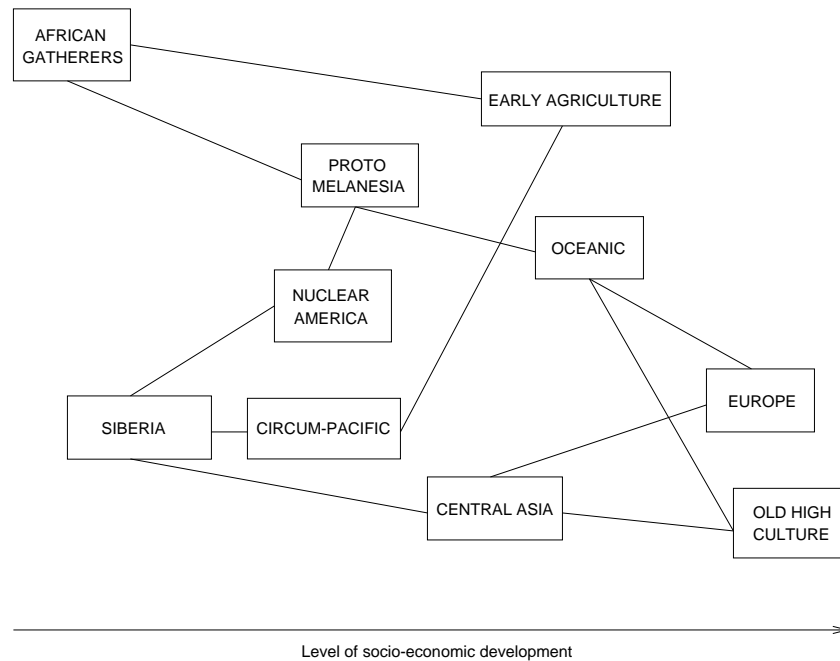
Figure 2.6: Clustering of song styles (after Lomax, 1980)

The first scientific study of musical memes (with the word *meme* explicitly mentioned) was carried out by Lynch and Baker (1994). They began by defining the "song meme" as a sequence of "syllable types", where the syllable types were determined *post hoc* by visual inspection of discontinuities in the recorded spectrograms of chaffinch songs. The sequences were then analysed using population biology methods. Lynch and Baker (1994) found that the levels of cultural differentiation among chaffinch populations can be explained by high mutation rates (memetic drift) and low migration rates (memetic isolation).

While Lomax (1980) and Lynch and Baker (1994) might rightfully be called the fathers of music memetics, Jan (2000a,b) was the first to apply the memetic paradigm to Western tonal music and link it to music theory and psychology. Taking a more theoretical standpoint, Jan (2000a) distinguished between the *phemotype* and the *memotype*, which correspond to memetic behaviours and artefacts (*e.g.*, scores and recordings) and their engendering neural structures (*e.g.*, mental representations), respectively. He also

requires *coequality*, by which he means the segmentation of music into discrete, comparable units (analogous to the DNA code). Jan (2000a) identified two dimensions of memetic hierarchies, *cultural* and *structural*: cultural hierarchies consist of intraopus style, idioms, dialects, rules and laws; whereas structural hierarchies can be mapped to Narmour's hierarchical style structures, defined as "[themes] that listeners implicatively map from the top down onto incoming foreground variations" (Narmour, 1999, p. 444). Jan's account is a good starting point for future work, but as it stands there is not much beyond a vague mapping from musicological terms to memetic ones (thus unfalsifiable). I will address this inadequacy by furnishing an explicit and in principle falsifiable definition of memetic information in Chapter 3.

Jan (2000a) also linked aesthetics to memetics: "[the cultural fitness] of a [musical meme] is an index of its intrinsic appeal to the environment of a brain, which is circumscribed both by innate perceptual and cognitive attributes, *and by the receptivity to incursion of the complement of memes already encoded therein*" (Jan, 2000a, my emphasis). This cultural fitness has not been explicitly modelled before, and I will investigate this further in Chapter 5.

In summary, this subsection reviews music memetics. Of particular relevance is the link between memetics and music psychology (Jan, 2000a). While Lomax (1980) and Lynch and Baker (1994) did pioneering work on music memetics, Jan (2000a) was the first to link memetics to music psychology as well as aesthetics. By putting Jan (2000a) together with (Meyer, 1957, reviewed in Section 2.1), one could unearth an indirect link between memetics and information theory—ultimately, both seek to model music psychology and aesthetics. I believe that this link addresses an important caveat regarding memetic information (Hull, 2000):

> The solar system, an enclosed gas, and a molecule of table salt all contain information. So does a molecule of DNA. It is a double helix. The bonds that run along the 'backbones' of this molecule do not rupture as easily as those holding the base pairs together. Hence, the molecule can zip and unzip with great facility. However, another sort of information is also contained in a molecule of DNA—in the sequence of its base pairs. As far as I know, none of the current analyses of evidence can distinguish between these two sorts of information, and until they do, memetics is in real trouble. (p. 59)

By looking at memes from an information-processing perspective, we can analyse the second sort of information quantitatively.

2.5 Summary

Measures of music information have been around for a while, but their physiological, psychological and evolutionary validity have been hitherto lacking. Based on the previous work of Meyer (1957), I have identified three major building blocks for a biologically plausible quantification of music information: mental representations, music informatics and biomusicology. In mental representations, I have reviewed biophysical and cognitive representations, providing the most important building block for my proposed information theory. In music informatics, I have reviewed information theory and time series analysis, providing the computational techniques for my proposed research. Finally, in biomusicology, I have reviewed the memetics of music, providing a biological basis for my proposed measure. All in all, this chapter provided a strong foundation for my thesis.

Chapter 3

Cognitive Information

3.1 Introduction

This chapter proposes a novel *cognitive information* theory of music. The heart of this theory consists of a non-Shannon information measure for symbolic musical time series such as those extracted from MIDI files. The aim of this theory is to model the amount of short-term memory required to store pieces of music in the human brain. This theory provides a unifying information-theoretic and simultaneously perceptually-motivated framework for music complexity, music similarity, and aesthetics of music. In this chapter, I will further motivate my cognitive information theory, detail my research methodology for this and the next two chapters, and propose and test my new information measures.

3.2 Motivation

Recall from Chapter 2 the two requirements for the construction of an accurate information theory of music (Meyer, 1957):

1. "First we must arrive at a more precise and empirically validated account of mental behavior which will make it possible to introduce the more or less invariant probabilities of human mental processes into the calculation of the probabilities involved in the style. This account need not necessarily be statistical itself" (p. 422–423).

2. "Second, and this is ultimately dependent upon the first, it is necessary to develop a more precise and sensitive understanding of the nature of musical experience" (p. 423).

   In Chapter 2, I have recast these requirements into three main building blocks (for a biologically plausible information theory of music): mental representations, music infor-

matics and biomusicology. I propose that this kind of information modelling—although difficult from a pure information-theoretic point-of-view—is actually possible if we actively incorporate knowledge from other fields, especially theoretical neuroscience and music psychology. This approach is built on Meyer's suggestion that the above two conditions are necessary and sufficient. Even if they are not, my model could still be a partial cognitive model for a clearly defined subset of tonal music, such as fugues, on a model-fitting basis. The definition of such subsets, if such a need arises, would be future work.

The modelling of brain activities would give us a sound basis for further investigation into computational aesthetics and, as I will argue, a sound basis for music memetics as well. The following is a brief rationale for my new information theory.

1. To a first approximation, cognitive information corresponds to perceived complexity, manifested in self-reported responses of musical complexity (*cf.* Conley, 1981; Shmulevich and Povel, 2000, reviewed below). For Lerdahl (1988, p. 255), "complexity refers not to musical surfaces but to the richness of their (unconscious) derivation by the listener". A similar view is expounded by Toop (1993). For Toop, complexity is "essentially a subjective, perceptual phenomenon [...] something [that the listeners sense unreflectingly] as richness" (p. 48). There seems to be a consensus that complexity is a perceived phenomena and not a physical one. This corroborates my cognitive information hypothesis (see Chapter 1).

2. Recall from Chapter 2 that musical probabilities do not match brain probabilities, because neuroscientific experiments have shown that melody-like music is encoded more faithfully in the brain than random music (Patel and Balaban, 2000). Currently, all existing Shannon-based models of music employ the probabilities of the musical text, therefore they are unlikely to reflect the probabilities of mental processes as per Meyer's requirements. A notable exception is Conklin and Witten's (1995) multiple viewpoint system, which combines multiple probabilistic models of music (called *viewpoints*), thus making it possible to recreate the nonlinearities observed by Patel and Balaban (2000); however, the correlation between combined viewpoints and brain probabilities is not currently known.

3. Instead of Shannon information theory (as specified by Meyer), I will instead propose a new, compression-based information theory that does not involve probabilities. I believe that this change does not detract from Meyer's arguments, because Meyer explicitly said that "this account need not necessarily be statistical itself". Furthermore, Large et al. (1995) empirically validated a compression-based model of musical memory, providing indirect support for my compression-based

approach.

4. My model addresses Meyer's first requirement by having a short-term memory component derived from biologically plausible building blocks known as delay lines and cancellation. These building blocks have previously been hypothesised to exist in the human auditory system by different researchers (Licklider, 1951; de Cheveigné, 1993), although they have not yet been located experimentally. In this thesis, I will simply assume that they exist, on the basis that an explanatory theory derived from this assumption lends weight to it.

5. I will adopt the optimality principle of neural processing (Sanger, 1989). Specifically, I assume that the auditory short-term memory performs fairly-optimal compression. This is further motivated by an experimentally validated, compression-based model of musical memory (Large et al., 1995).

6. Towards fulfilling Meyer's second requirement, I use the preprocessor of Temperley and Sleator (1999) as my front-end, which results in a stream of three-tuples of the form ⟨onset, pitch, metrical level⟩ from a MIDI file. This representation is essentially the same as the neuronal speculation of Narmour (1999), with a slightly different interpretation of "chronotopic" (Narmour used durations, I used onsets). Evidence for beat-tracking mechanisms is well-documented in music psychology (Lerdahl and Jackendoff, 1983); furthermore, absolute pitch and tempo mechanisms have been documented by Levitin (1994) and Levitin and Cook (1996). Moreover, Temperley and Sleator (1999) claimed that their beat-finding algorithm is robust even for performance rubati.

7. Finally, at Marr's computational level (Phillips, 1997), I hypothesise that the way I put together the above components reflects the cognitive constraints on music processing.

In conclusion, both of Meyer's requirements are addressed by my proposed model, explained in detail below. Therefore, in his terms at least, I have proposed a potentially accurate quantification of music in information-theoretic terms, that could in principle lead to a quantitative measure of musical value.

Of course, the validity of my premises and assumptions cannot be proven in pure deductive logic, given their falsifiable nature. The best we can get is empirical evidence (*i.e.*, testing the model predictions, see Subsections 4.4.2 and 5.3.4), and even that cannot be done in full in the scope of this thesis. Due to inherent time constraints, many validation tasks have to be left for future work.

3.3 Methodology

For this and the next two chapters, the *meta-gMDL$_+$* methodology (proposed in this section) will be used throughout. This methodology combines minimum description length and meta-analysis, detailed below. The main idea is to combine description lengths from different studies and then select the hypothesis that minimises the combined code length.

3.3.1 Minimum Description Length

The *minimum description length* (MDL) principle states that the best data model is the one that minimises the code length for the observed data (Rissanen, 1978; Barron and Rissanen, 1998). MDL is better than traditional goodness-of-fit measures (such as variance accounted for) because MDL penalises complex models, thus preventing overfitting (Pitt and Myung, 2002; Grünwald, 2005). Since psychological data are usually on an interval scale, a good way of comparing them with computational models would be linear regression (specifically, one-tailed correlation model where negative correlations are ignored).

The non-negativity constraint is motivated by the existence of models that output the correct magnitude but with the wrong sign. Consider the following datasets: $(x_{data}, y_{data}) = \{(0,0), (1,1), (2,0)\}$ and $(x_{model}, y_{model}) = \{(0,1), (1,0), (2,1)\}$. Quadratic fits are shown in Figure 3.1. With a two-tailed correlation, the model accounts for 100% of the variances in the data, which is not right in a model selection context because the model is exactly the opposite of what it should be. With a one-tailed correlation, however, the variance accounted for is 0%, which is exactly what we wanted.

(a) Data   (b) Model
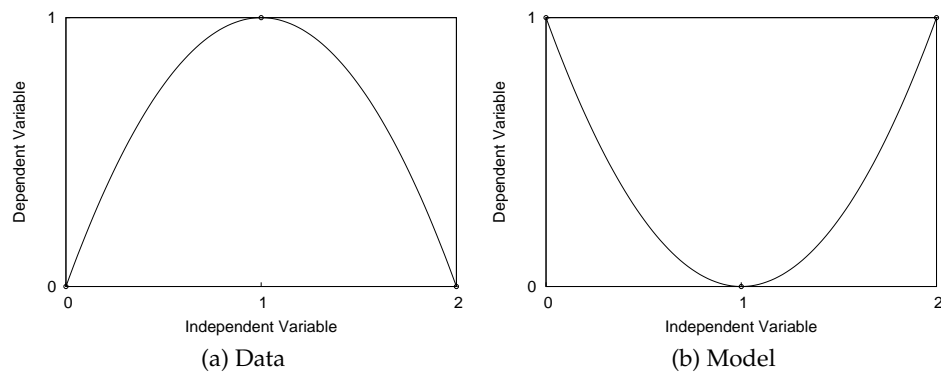
Figure 3.1: Toy Data and Model

The gMDL criterion (Hansen and Yu, 2001) is an appropriate measure of code length for regression problems like two-tailed correlation. Although there are many other forms

of MDL criteria, I chose gMDL because it has a simple closed-form expression that is amenable to my one-tailedness modification.

Consider a non-negative, one-tailed correlational model with n pair of data points,

$$y_i \;=\; R_+ x_i + \epsilon_i$$

where $1 \le i \le n$, x and y are standardised vectors (zero mean and unit variance) such that $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2 = n - 1$, and the errors ϵ_i are normally distributed with zero mean and an unknown variance. The vectors are standardised so that we are comparing apples with apples, and the choice of $n - 1$ is such that the unbiased estimate of population variance becomes unity. Here R_+ denotes the product moment correlation coefficient with negative correlations replaced by zeros,

$$R_+ \;=\; \max\left(0, \frac{\sum_{i=1}^n x_i y_i}{n - 1}\right).$$

In other words, if the correlation between x and y is negative, we ignore x and model y as noise. The residual sum of squares is

$$
\begin{aligned}
\text{RSS} \;&=\; \sum_{i=1}^n \epsilon_i^2 \\
&=\; \sum_{i=1}^n (y_i - R_+ x_i)^2 \\
&=\; \sum_{i=1}^n (y_i^2 - 2 y_i R_+ x_i + R_+^2 x_i^2) \\
&=\; \sum_{i=1}^n y_i^2 - 2 R_+ \sum_{i=1}^n x_i y_i + R_+^2 \sum_{i=1}^n x_i^2 \\
&=\; (n - 1) - 2 R_+^2 (n - 1) + R_+^2 (n - 1) \\
&=\; (n - 1) - R_+^2 (n - 1) \\
&=\; (n - 1)(1 - R_+^2).
\end{aligned}
$$

The gMDL criterion is then given by Hansen and Yu (2001):

$$\text{gMDL}_+ \;=\; \begin{cases} \frac{n}{2} \log \frac{\text{RSS}}{n-k} + \frac{k}{2} \log \frac{(\sum_{i=1}^n y_i^2 - \text{RSS})/k}{\text{RSS}/(n-k)} + \log n, & \text{if } R_+^2 \ge k/n \\[2mm] \frac{n}{2} \log \left(\frac{\sum_{i=1}^n y_i^2}{n}\right) + \frac{1}{2} \log n & \text{otherwise,} \end{cases}$$

where k denotes the number of predictors (only one here, which is R_+).[1] After some

---

[1] I write gMDL$_+$ here because I use R_+ instead of the usual correlation coefficient R.

algebra, this becomes

$$\text{gMDL}_+ \;=\; \begin{cases} \frac{n}{2}\log(1-R_+^2) + \frac{1}{2}\log\frac{(n-1)R_+^2}{1-R_+^2} + \log n, & \text{if } R_+^2 \geq 1/n \\ \frac{n}{2}\log\left(\frac{n-1}{n}\right) + \frac{1}{2}\log n & \text{otherwise.} \end{cases}$$

3.3.2 Meta-Analysis

Glass (1976) coined the term *meta-analysis* to mean the integration of findings through statistical analysis of a large collection of studies. For correlational studies, Glass (1977) recommended that one could use either the average of correlation coefficients, or that the coefficients be "squared, averaged, and the square root taken". Instead of looking at the correlation coefficient, which is a poor measure of model selection (Pitt and Myung, 2002), I propose to add the gMDL code lengths together to obtain a combined code length.[2] The hypothesis with the shortest combined code length will then be chosen as the best explanation. I call this meta-gMDL$_+$ selection.

3.4 Review of Non-Standard Information Models

3.4.1 The Theory of Cilibrasi et al. (2004)

In this chapter, I will make use of two non-standard information models, reviewed below. First I will review the information theory proposed implicitly in Cilibrasi et al. (2004). I write "implicitly" because the explicit one is the uncomputable Kolmogorov complexity (approximated by `bzip2`, with a preprocessing step to ensure fair comparisons of MIDI files).[3] But I believe that any preprocessing that goes beyond lossless data extraction counts as a deliberate modification to the compressor. In fact I am prepared to claim that the effectiveness of their information theory comes entirely from preprocessing alone. This is an important philosophical difference between Cilibrasi et al. (2004) and my work. Their preprocessing step goes like this (Cilibrasi et al., 2004):

> The preprocessor extracts only MIDI Note-On and Note-Off events. These events were then converted to a player-piano style representation, with time quantized in 0.05-sec intervals. All instrument indicators, MIDI control signals, and tempo variations were ignored. For each track in the MIDI file, we

---

[2] In meta-analyses, people often use Cronbach's α (interrater consistency) as weights, such that studies with all participants agreeing with each other will have more weights than those without. My problem is that I do not have the α values for most of the studies in this and the other two chapters, so instead I assume that all studies are equally consistent.

[3] And secondly because they do not use this information measure other than for the purpose of calculating the information distance, the subject of the next chapter.

calculate two quantities: an average volume and a modal note. ("Modal" is used here in a statistical sense, not in a musical sense.) The average volume is calculated by averaging the MIDI Note-On velocity of all notes in the track. The modal note is defined to be the note pitch that sounds most often in that track. If this is not unique, then the lowest such note is chosen. The modal note is used as a key-invariant reference point from which to represent all notes. It is denoted by 0, higher notes are denoted by positive numbers, and lower notes are denoted by negative numbers. A value of 1 indicates a half step above the modal note, and a value of -2 indicates a whole step below the modal note.

The modal note is written as the first byte of each track. For each track, we iterate through each 0.05-sec time sample in order, producing a single signed 8-bit value as output for each currently sounding note (ordered from lowest to highest). Two special values are reserved to represent the end of a time step and the end of a track. The tracks are sorted according to decreasing average volume and then output in succession. (p. 58)

This contradicts with the following claims, made in the same paper:

1. "We do not look for similarity in specific features know to be relevant for classifying music; instead we apply a general mathematical theory of similarilty" (p. 49);

2. "We want to stress again that our method does not rely on any music-theoretical knowledge or analysis but only on general-purpose compression techniques" (p. 62).

Both claims are false, because they made use of specific features such as modal notes, intervals, and average volume. Such features rely on music-theoretical knowledge (such as the relative importance of intervals). Actually, even Rudi Cilibrasi himself admitted that "the preprocessing step is crucial to the success of the method" (personal communication, 23 October 2006). Therefore, there is no clear evidence that the Kolmogorov complexity framework played a role here in the "success" of this measure, despite published claims to the contrary. I would argue that the preprocessor is really a part of their information measure.

3.4.2 T-Complexity Theory

Now I will review the theory of T-complexity (Titchener, 2000), which will be used in my short-term memory model (proposed in the next section). Let $A = \{a_1, a_2, \ldots, a_n\}$ be a

finite alphabet of symbols. By convention (Hopcroft et al., 2000, p. 113), A^+ denotes "one or more" symbols taken from the alphabet A. The T-complexity of a string $x \in A^+$ is "a measure of the *effort* required" to produce x and is derived as follows (Titchener, 2000). First we decompose x into a series of patterns $p_i \in A^+$ such that

$$x = p_t^{k_t} p_{t-1}^{k_{t-1}} p_{t-2}^{k_{t-2}} \ldots p_1^{k_1} a,$$

where $a \in A$, subject to the constraint

$$p_i = p_{i-1}^{m_{i-1,i-1}} p_{i-2}^{m_{i-1,i-2}} \ldots p_1^{m_{i-1,1}} A,$$

where $0 \le m_{j-1,i} \le k_i$. The T-complexity of x is then defined as:

$$C_T(x) = \sum_i \log_2(k_i + 1).$$

The decomposition step can be done in $O(n \log n)$ time using the fast T-decomposition algorithm (Yang and Speidel, 2005). Note that the decomposition itself is not unique, but it has been proven that all possible decompositions of x give the same value of $C_T(x)$ (Titchener, 2000; Yang and Speidel, 2005). The T-complexity measure is selected for my short-term memory model mainly because of its $O(n \log n)$ speed.

Example 3.1. *Prove that $(ME)^3(M)^1E$ is a valid decomposition of MEMEMEME, and calculates its T-complexity.*

Proof. For $(ME)^3(M)^1E$, we have:

1. $A = \{M, E\}$;

2. $p_1 = M, k_1 = 1$. Here $p_1 \in A$;

3. $p_2 = ME, k_2 = 3$. Here $p_2 = (p_1)^1E$ where $E \in A$ and $0 \le 1 \le k_2$.

Therefore the constraint is satisfied and $(ME)^3(M)^1E$ is a valid decomposition. Its T-complexity is $\sum_i \log_2(k_i + 1) = \log_2(1 + 1) + \log_2(3 + 1) = 3$ bits. $\square$

3.5   A Model of Musical Memory

My cognitive information model is based on Atkinson and Shiffrin's (1968) memory model.[4] A block diagram of my information model (with context also shown) can be

---

[4]Note that this is just one possible realisation of the Atkinson-Shiffrin model; I do not claim that it is the best.

found in Figure 3.2. Here OPM stands for onset, pitch and metrical level (the input), TDNN stands for time-delay neural network (the sensory register), STM stands for short-term memory and LTM stands for long-term memory. The double bars in this diagram represent my proposed information measure quantifying the memory usage in the STM.
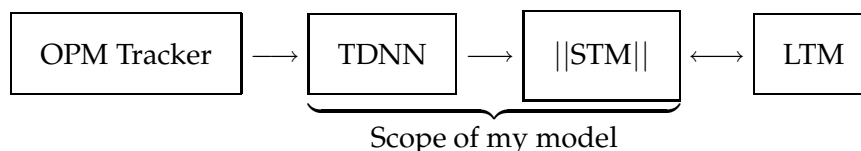
Figure 3.2: Block diagram of my information model

3.5.1 Assumptions

Firstly, I assume the existence of separate STM and LTM mechanisms (Atkinson and Shiffrin, 1968). This is a standard simplifying assumption. Secondly, both Narmour (1999) and Temperley and Sleator (1999) require a beat tracker for their respective cognitive theories to work. This leads to my assumption that metrical level is a crucial building block that should not be omitted from the inputs. Finally, I assume that the neural cancellation filter mechanism is applicable to higher cognitive functions as well. This is justified by the argument of Lőrincz et al. (2002) on the separation of structure and function (see Chapter 2, p. 17).

3.5.2 Inputs

My model takes an array of three-tuples of the form ⟨onset, pitch, metrical level⟩ from a beat-tracked MIDI file, which I will call the OPM representation. Onsets can be in any linear time format (seconds, milliseconds and so on). Pitches are specified in MIDI pitch units. Metrical level starts at 0 (least accent) and goes up to 4 (heaviest accent). This representation is isomorphic to those of Narmour (1999) and Temperley and Sleator (1999). For example, the first bar of J. S. Bach's Invention No. 1 (Figure 3.3) along with its metrical structure (marked by the x's) are converted to the OPM representation in Figure 3.4.

In the rest of this thesis, the input file will be a real-time performance MIDI file, post-processed by the `meter` program (Temperley and Sleator, 1999). However, note that the meter-finding program itself is outside the scope of my model; the OPM input can be prepared by any other meter-finding program as long as it returns the onsets, pitches
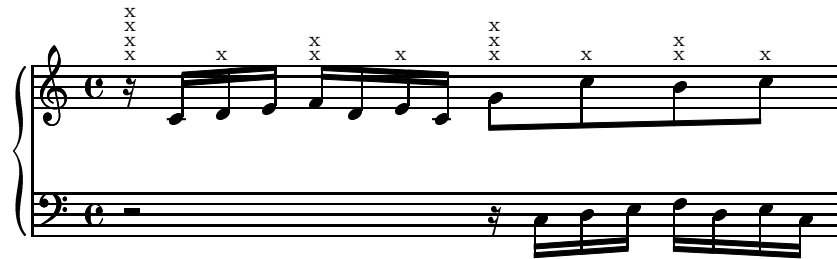
Figure 3.3: First bar of J. S. Bach's Invention No. 1 in C Major (BWV 772)

Onset	Pitch	Level
1	48	0
2	50	1
3	52	0
4	53	2
5	50	0
6	52	1
7	48	0
8	55	3
9	36	0
10	38	1
10	60	1
11	40	0
12	41	2
12	59	2
13	38	0
14	40	1
14	60	1
15	36	0

Figure 3.4: OPM representation of the first bar of Bach's Invention No. 1 (BWV 772)

and metrical levels in the manner prescribed above.

3.5.3 Outputs

Time-Delay Neural Network (TDNN)

The sensory register should have been modelled by the connections between neurons that Narmour (1999) mentions, using a spreading activation model, except that Narmour does not give enough information for a concrete implementation of it. So instead I use the neural cancellation filter (de Cheveigné, 1993), which can be configured to compute all pairwise differences between all nodes. The advantage of this filter is that it takes

into account the notion of intervals (differences) which are crucial to music perception. The calculation of differences also resonates with Bateson's (1973) ecological view that information is a "difference which makes a difference".

In the model below, the time-delay mechanisms are modelled by a *neural cancellation matrix M*, which I define as the output of an array (Licklider, 1951) of neural cancellation filters (de Cheveigné, 1993):

$$m_{ij} = \max(0, x_i - x_{i-j}),$$

for $1 \leq i \leq n$ and $1 \leq j < i$, where $x_1, \dots, x_n$ is a real time series. Intuitively, neural cancellation filters are comb filters with non-negativity constraints (de Cheveigné, 1993). Unfortunately, this matrix is ill-defined because m_{ij} is undefined whenever $i \leq j$. Furthermore, a one-dimensional (real) representation cannot capture the OPM format naturally. So I modify it to:

$$m_{ij} = \max(0, x_i - x_j)$$

where $x_1, \dots, x_n$ is a three-dimensional time series in the OPM format, for $1 \leq i, j \leq n$, where the max operator here operates component-wise. For example, $\max(0, \langle -4, -1, 2 \rangle)$ returns $\langle 0, 0, 2 \rangle$ and $\max(0, \langle 10, -5, 10 \rangle)$ returns $\langle 10, 0, 10 \rangle$.

Short-Term Memory (STM)

Assuming that the STM is performing compression (*cf.* Sanger, 1989; Large et al., 1995), I quantify the compressibility of music with T-complexity (Titchener, 2000). Specifically, I define *cognitive information $H(x)$* as the square root of the T-complexity of the above neural cancellation matrix, which models the STM usage of a meme. This definition is the centre of my cognitive information theory.[5] The square root is used because the neural cancellation filter sends n source elements into n^2 destination slots.

As an example, the first three notes of the aforementioned Bach invention (BWV 772) correspond to the following three-dimensional time series:

$$x = \begin{bmatrix} \langle 1, 48, 0 \rangle \\ \langle 2, 50, 1 \rangle \\ \langle 3, 52, 0 \rangle \end{bmatrix},$$

---

[5]In this thesis, matrices are stored in row-major order and tuple elements are stored left to right, with no zero padding, using 32-bit signed integers in the big-endian format. The symbol alphabet for the calculation of T-complexity is 8-bit. But these are not important details, as Titchener (2000) has demonstrated that T-complexity is empirically a monotonic increasing function of Shannon entropy (for long strings), so we can expect the T-complexity to stay approximately the same after a conversion to another storage format.

for which the modified neural cancellation matrix can be calculated as follows:

$$
M = \begin{bmatrix} \max(0, x_1 - x_1) & \max(0, x_1 - x_2) & \max(0, x_1 - x_3) \\ \max(0, x_2 - x_1) & \max(0, x_2 - x_2) & \max(0, x_2 - x_3) \\ \max(0, x_3 - x_1) & \max(0, x_3 - x_2) & \max(0, x_3 - x_3) \end{bmatrix}
$$

$$
= \begin{bmatrix} \max(0, \langle 0,0,0 \rangle) & \max(0, \langle -1,-2,-1 \rangle) & \max(0, \langle -2,-4,0 \rangle) \\ \max(0, \langle 1,2,1 \rangle) & \max(0, \langle 0,0,0 \rangle) & \max(0, \langle -1,-2,1 \rangle) \\ \max(0, \langle 2,4,0 \rangle) & \max(0, \langle 1,2,-1 \rangle) & \max(0, \langle 0,0,0 \rangle) \end{bmatrix}
$$

$$
= \begin{bmatrix} \langle 0,0,0 \rangle & \langle 0,0,0 \rangle & \langle 0,0,0 \rangle \\ \langle 1,2,1 \rangle & \langle 0,0,0 \rangle & \langle 0,0,1 \rangle \\ \langle 2,4,0 \rangle & \langle 1,2,0 \rangle & \langle 0,0,0 \rangle \end{bmatrix}.
$$

A hexadecimal dump of the above matrix is shown in Figure 3.5.

```
00000000   00 00 00 00   00 00 00 00   00 00 00 00   00 00 00 00
00000010   00 00 00 00   00 00 00 00   00 00 00 00   00 00 00 00
00000020   00 00 00 00   00 00 00 01   00 00 00 02   00 00 00 01
00000030   00 00 00 00   00 00 00 00   00 00 00 00   00 00 00 00
00000040   00 00 00 00   00 00 00 01   00 00 00 02   00 00 00 04
00000050   00 00 00 00   00 00 00 01   00 00 00 02   00 00 00 00
00000060   00 00 00 00   00 00 00 00   00 00 00 00
```

Figure 3.5: Hexadecimal dump of M (first column denotes address)

The reader can verify that the following decomposition (with each symbol underlined) satisfies the constraints set forth in the definition of T-complexity:

$$
(\underline{00}\,\underline{00}
$$
$$
\underline{00}\,\underline{00}\,\underline{00}\,\underline{00}\,\underline{00})^1(\underline{00}\,\underline{00}\,\underline{00}\,\underline{00}\,\underline{00}\,\underline{00}\,\underline{00}\,\underline{01}\,\underline{00}\,\underline{00}\,\underline{00}\,\underline{02}\,\underline{00}\,\underline{00}\,\underline{00}\,\underline{01})^1(\underline{00}\,\underline{00}\,\underline{00}\,\underline{00}\,\underline{00}
$$
$$
\underline{00}\,\underline{00}\,\underline{00}\,\underline{00}\,\underline{00}\,\underline{00}\,\underline{00}\,\underline{00}\,\underline{00})^1(\underline{00}\,\underline{00}\,\underline{00}\,\underline{00}\,\underline{00}\,\underline{00}\,\underline{00}\,\underline{01}\,\underline{00}\,\underline{00}\,\underline{00}\,\underline{02}\,\underline{00}\,\underline{00}\,\underline{00}
$$
$$
\underline{04})^1(\underline{00}\,\underline{00}\,\underline{00}\,\underline{00}\,\underline{00}\,\underline{00}\,\underline{00}\,\underline{01})^1(\underline{00}\,\underline{00}\,\underline{00}\,\underline{02})^1(\underline{00})^{15}\underline{00}.
$$

Its T-complexity is $\sum_i \log_2(k_i + 1) = \log_2(1+1) + \log_2(1+1) + \log_2(1+1) + \log_2(1+1) + \log_2(1+1) + \log_2(1+1) + \log_2(15+1) = 10$. Therefore $H(x)$ is $\sqrt{10} \approx 3.2$ bits.

Long-Term Memory (LTM)

The long-term memory was not modelled here, but it ought to be investigated in future work. For the purpose of this thesis, I will simply assume that the LTM consists of a concatenation of a corpus of pieces, which characterises a particular listener who has learnt the pieces.

3.6 Extended Theory for Two Musical Objects

So far, $H(x)$ has no notion of jointness and conditionedness of two musical objects. To extend the theory for two musical objects, I borrow the pairing operator $\langle x, y \rangle$ from Kolmogorov complexity theory, which returns the concatenation of x and y (Bennett et al., 1998). Armed with this pairing operator, I then define the *joint cognitive information* of x and y as $H(\langle x, y \rangle)$. Theoretically, this joint measure models memory usage learning both x and y together assuming that they have not been learned before. Next, I define *cognitive independence*, which characterises the condition $H(\langle x, y \rangle) \geq H(x) + H(y)$. When x and y are cognitively independent, compressing the concatenated inputs would yield a equal or larger result compared with concatenating both compressed inputs. In other words, the two inputs have so little in common that there could be no memory savings when they are compressed together.

Finally, I define *conditional cognitive information*, which models memory usage transferring song x to a brain containing y, as

$$H(x|y) = \min\{H(x), H(\langle x, y \rangle) - H(y)\},$$

and the *mutual cognitive information* as

$$I(x;y) = \max\{0, H(x) + H(y) - H(\langle x, y \rangle)\}.$$

It follows that if x and y are cognitively independent, then the mutual cognitive information will be zero by definition. This information can also be interpreted as a model of memory savings in learning x and y together.

3.6.1 Conformance to The Three Laws of Shannon

Kolmogorov (1965) observed that the Shannon inequalities are also valid for Kolmogorov complexity (up to a logarithmic term). Hammer et al. (2000) went further and stated that these inequalities are valid for ranks of finite subsets of linear spaces as well. Given this apparent universality (applicable to three completely different information models), I will now prove that these inequalities are valid for my information measures too.

Theorem 3.1. $H(x, y) \leq H(x) + H(y)$ *with equality iff x and y are cognitively independent.*

Proof. By definition, we have $H(x, y) = \min\{H(x) + H(y), H(\langle x, y \rangle)\} \leq H(x) + H(y)$. Again, by definition, $H(\langle x, y \rangle) \geq H(x) + H(y)$ iff x and y are cognitively independent. Therefore, $H(x, y) \leq H(x) + H(y)$ with equality iff x and y are cognitively independent. $\square$

Theorem 3.2. $H(x, y) = H(x) + H(y|x) = H(y) + H(x|y) + O(1)$.

Proof. To prove the first equality, we note that $H(x|y) = \min\{H(x), H(\langle x, y \rangle) - H(y)\} = \min\{H(x) + H(y), H(\langle x, y \rangle)\} - H(y) = H(x, y) - H(y)$. For the second equality, we invoke the symmetry of algorithmic information (Li et al., 2003) and we have $H(\langle x, y \rangle) = H(\langle y, x \rangle) + O(1)$. Therefore the second equality holds. $\square$

Theorem 3.3. $I(x; y) = H(x) - H(x|y)$.

Proof. We have $I(x; y) = \max\{0, H(x) + H(y) - H(\langle x, y \rangle)\} = H(x) + \max\{-H(x), H(y) - H(\langle x, y \rangle)\} = H(x) - \min\{H(x), H(\langle x, y \rangle) - H(y)\} = H(x) - H(x|y)$. $\square$

These theorems will be used in Chapter 4 to establish the measure-theoretic properties of my cognitive information measures.

3.7 Psychological Experiments

In this section, I summarise three psychological experiments on perceived complexity done by other researchers, which I will replicate computationally here in this chapter.

All three experiments used stimuli that sounded realistic and therefore should not suffer from the criticism of being ecologically invalid.

The first experiment was done by Shmulevich and Povel (2000). In their experiment, they asked 25 participants to listen to 35 rhythmic patterns and to rate the complexity of each. The patterns and mean ratings are reproduced in Figure 3.6. Shmulevich and Povel (2000) noted that a compression-based measure of complexity (the Lempel-Ziv measure) could only account for 2.25% of the variances in the human data ($r = 0.15$), whereas their Povel-Shmulevich measure could account for 56.25% ($r = 0.75$). They attributed this to two factors: first, that the Lempel-Ziv measure is unsuitable for short sequences; and second, that the Povel-Shmulevich measure is an empirically-tested perceptual model and therefore it is likely to do better than the Lempel-Ziv measure.

The second experiment was done by Conley (1981). In this seminal work on the perception of complexity in art music, Conley specified 10 predictors of musical complexity and set out to correlate them with human judgements of musical complexity. In her experiments, sixteen Beethoven Eroica Variations (played by Sviatoslav Richter, Angel S-40183) were used as stimuli (see Figure 3.7). These stimuli[6] were chosen because of their ecological validity (Conley, 1981). The effect of musical training is controlled for by dividing the participants into Graduate, Sophomore and Non-major groups (Conley, 1981).

---

[6]Var. 1 is actually the main theme, despite Conley's misleading label.

Pattern	Complexity	Pattern	Complexity
|||||..||.|.|...	1.56	||..||.||||.|...	2.64
|||.|.|||..||...	2.12	||..||.|||.||...	3.24
.|||.|||..||...	2.08	|||||.||.|..|...	3.08
|.|.|||||..|...	1.88	||||.|..|||.|...	3.04
|..||.|.|||||...	1.80	|||..||.|||.|...	3.04
|||.|||.||..|...	2.44	.|||..|.||||...	2.56
|.||||.||..|...	2.20	|.|..||||.|||...	2.56
||..||||||.|.|...	2.56	||||.|.|..|||...	2.84
||...|.|||.|||...	3.00	||.|||.|..|||...	3.60
.|||.|||..|...	2.04	||.|..|||.|||...	2.68
|||.||..||.||...	2.76	.||||.|..|||...	3.28
||.||||.|..||...	2.72	|..|||||.||.|...	3.08
|.||.||||..|...	3.00	||||.|||..|.|...	3.52
||..||.||.|||...	3.16	||||..||.||.|...	3.60
|..|||.|||.||...	2.04	||.||||..||.|...	3.04
||.||||.|..|...	2.88	||.|..|||||.|...	2.88
||.|||.|||..|...	2.60	|.|..|||.||||...	3.08
||.|||..||.||...	2.60		

Figure 3.6: Human data from Shmulevich and Povel (2000). Here a bar represents a tone in middle C, a dot represent a rest, and the events are spaced 200ms apart

Her main results are reproduced in Figure 3.7. For this particular experiment, musical training had an effect and that the best predictor of complexity for all three groups is the rate of rhythmic activity, accounting for more than 70% of the variances in each group (Conley, 1981). At the end of her paper, Conley was careful to warn us that her results might not necessarily generalise to any other settings.

The third experiment was conducted by Heyduk (1975). Heyduk composed four original, thematically similar piano pieces in increasing complexity ($A < B < C < D$) by manipulating chord structure and syncopation. He then asked his participants to rate the complexity of each of the pieces after two listenings. The mean complexity ratings can be seen in Figure 3.8. The ratings exhibited a monotonically increasing relationship against compositional complexity (Heyduk, 1975).

3.8 Computational Replications

Following my proposed methodology in this chapter, I will compare the following three models by using meta-gMDL$_+$ selection:

- The model proposed in Section 3.5;

Var.	Title	Graduate	Sophomore	Non-major
1	Introduzione col Basso del Tema	-1.65	-1.51	-1.17
2	A due	-0.57	-0.99	-1.07
3	A tre	-0.90	-1.09	-0.95
4	A quattro	0.12	0.14	0.05
5	Tema	-0.64	-0.48	-0.22
6	Variation I	0.00	0.40	0.63
7	Variation II	1.12	1.89	1.39
8	Variation III	0.88	0.62	0.29
9	Variation IV	-0.24	0.11	0.04
10	Variation V	-0.08	-0.78	-0.53
11	Variation VIII	0.23	0.09	0.15
12	Variation IX	0.20	0.45	0.65
13	Variation X	1.49	0.93	0.62
14	Variation XI	-0.29	-0.39	-0.07
15	Variation XII	0.69	0.64	0.60
16	Variation XIII	-0.36	-0.02	-0.39

Figure 3.7: Stimuli and mean standardised complexity data from Conley (1981)

Piece	Complexity
A	3.0
B	5.4
C	8.0
D	9.8

Figure 3.8: Complexity data, read from the graph in Heyduk (1975)

- A reduced model, which is the same as above but with the neural cancellation filter removed; in other words, the T-complexity is taken of the input sequence itself, and the final square root is not taken;

- The baseline model using Cilibrasi et al.'s (2004) preprocessor with `bzip2` (see Subsection 3.4.1), which is the current state-of-the-art in algorithmic music information.

In examining the internal validity of my information measure, I noted that there are two potential objections: the short sequence objection and the polyphonic objection. I will answer each of them by computational replications. Finally, I will combine the findings using meta-gMDL$_+$ selection.

3.8.1 Experiment 1

As reviewed above, Shmulevich and Povel (2000) found that a compression-based measure of complexity (Lempel-Ziv) could only account for 2.25% of the variances in human judgements of rhythmic complexity. This raises the question whether my compression-based information measure would perform as poorly for short sequences. The following replication of Shmulevich and Povel (2000) will provide evidence against the short sequence objection.

Method

Materials The stimuli consisted of all thirty-five rhythmic patterns from Shmulevich and Povel's (2000) experiment (as shown in Figure 3.6).

Procedure Stimuli were entered into a computer and then converted into the OPM format. The cognitive information $H(\cdot)$ of all 35 OPM files were then calculated. The correlation between the values thus calculated and the human judgements in Figure 3.6 are then reported, along with its statistical significance and gMDL$_+$ (see Subsection 3.3.1).

Results

Results are shown in Figure 3.9. There is a statistically significant correlation between my proposed model and human data ($p < .05$). On the other hand, the reduced model and the baseline model do not have statistically significant correlations ($p > .05$). The values of gMDL$_+$ also reflect this.

Measure	r	df	p	gMDL$_+$
Proposed	0.44174	33	0.0039	0.811
Reduced	-0.44841	33	1.0000	1.270
Baseline	0.00000	33	0.5000	1.270

Figure 3.9: Correlation with Shmulevich and Povel's (2000) data in Experiment 1

Discussion

My proposed model ($r = 0.44$) is much better than the Lempel-Ziv measure ($r = 0.15$) but not as good as the Povel-Shmulevich measure ($r = 0.75$). One way to look at it is that the Povel-Shmulevich measure deals only with beat music and does not generalise to polyphonic music; nor does it recognise any pitch-based features in the first place. Music like Gregorian chants are defined mainly by their pitches, so the Povel-Shmulevich

measure would not be applicable. Also, the reduced and baseline models performed very poorly in this experiment, suggesting that:

- The neural cancellation filter is important;

- The assumptions behind the baseline model (the applicability of Kolmogorov complexity theory) might be misguided.

With more data, I can expect to claim that my model is a reasonable trade-off between model specificity and sensitivity.

3.8.2 Experiment 2

My choice of the OPM format was based on Narmour's (1999) theory which was originally proposed for monophonic music. The validity of my extrapolation (to the polyphonic domain) is not yet demonstrated. It is possible that my model might not be able to handle polyphonic music well. To examine this question, I have replicated Conley's (1981) experiment below.

Method

Materials   The stimuli consisted of all sixteen Beethoven Eroica Variations from Conley's (1981) experiment (as shown in Figure 3.7). Conley (1981) used a Sviatoslav Richter recording (Angel S-40183) which is out-of-print and, even if available, would be extremely difficult to convert into proper MIDI files. To establish an approximate correspondence with her experiment, I used a publicly available MIDI performance by Bunji Hisamori (several versions are published on the Internet; I used his "Revision 2", dated July 1999).

Procedure   Stimuli were converted into the OPM format. The cognitive information $H(\cdot)$ of all 16 OPM files were then calculated. The correlation between the calculated values and the human judgements in Figure 3.7 were then reported (along with statistical significances and gMDL$_+$).

Results

Results are shown in Figure 3.10. For my proposed measure, the correlations are statistically significant for the Sophomore and Non-major groups ($p < .05$), but not for the

Graduate group ($p > .05$). All of the correlations with my reduced measure are statistically significant. None of the correlations with the baseline measure is statistically significant ($p > .05$).

Measure	Group	r	df	p	gMDL$_+$
Proposed	Graduate	0.415270	14	0.05500	1.830
	Sophomore	0.665640	14	0.00240	-0.670
	Non-major	0.597580	14	0.00730	0.298
Reduced	Graduate	0.489010	14	0.02700	1.360
	Sophomore	0.728240	14	0.00069	-1.860
	Non-major	0.656680	14	0.00290	-0.526
Baseline	Graduate	0.078651	14	0.39000	0.870
	Sophomore	0.231170	14	0.19000	0.870
	Non-major	0.318970	14	0.11000	2.180

Figure 3.10: Correlation with Conley's (1981) data in Experiment 2

Discussion

The results showed that my information measure is able to model the Sophomore and Non-major judgements reasonably well, but has marginally failed on the Graduate data. One possible explanation is that, as all the Beethoven Variations contained the same theme (Var. 1), there might be a priming effect where the main theme was memorised by the participants at least partially (more so by the Graduate group, assuming that they have a better memory for music than the other two groups).

3.8.3 Experiment 3

In this experiment, Conley's (1981) experiment is again replicated, but with conditional cognitive information (conditioned on the theme) in lieu of cognitive information. If there is a priming effect as suggested above, then the conditional cognitive information should correlate better with human ratings than unconditioned cognitive information (in Experiment 2).

Method

Materials   The stimuli were the same as in Experiment 2.

Procedure   The procedure was the same as that of Experiment 2, except that the conditional cognitive information is used (conditioned on the theme).

Results

Results are shown in Figure 3.11. The correlations are statistically significant for all three measures on all three groups ($p < .05$), and each of the correlations are higher than their counterparts obtained in Experiment 2.

Measure	Group	r	df	p	gMDL$_+$
Proposed	Graduate	0.48571	14	0.02800	1.3900
	Sophomore	0.68993	14	0.00150	-1.0900
	Non-major	0.62716	14	0.00470	-0.0882
Reduced	Graduate	0.54564	14	0.01400	0.8700
	Sophomore	0.73447	14	0.00060	-2.0000
	Non-major	0.66793	14	0.00230	-0.7080
Baseline	Graduate	0.41192	14	0.05600	1.8500
	Sophomore	0.40219	14	0.06100	1.8900
	Non-major	0.41614	14	0.05400	1.8200

Figure 3.11: Correlation with Conley's (1981) data in Experiment 3

Discussion

With the priming assumption, the results added support to my cognitive information model. By comparing the *r*-values in the two tables, we can see that the priming assumption produced the highest improvement for the Graduate group, while producing only negligible improvements for the Sophomore and Non-major groups. One possible interpretation is that the Graduate group had a better memory for music and were able to memorise the main theme while listening. Another possibility is that they might have learnt this theme before from Beethoven's Eroica Symphony.

Here again, the reduced model performed slightly better than the full model, and the baseline model has failed. The percentage of variances accounted for by my full model are 23%, 46% and 37%, respectively, which are not as good as Conley's best model (rate of rhythmic activity, accounting for 71%, 90% and 77% of the data variances). However, when I apply Conley's best model to Shmulevich and Povel's (2000) rhythmic patterns (Figure 3.6), I obtained the same constant for all 35 patterns because they all have the same number of sounding notes per minute. This certainly does not fit the data (in fact, the correlation coefficient is undefined in this case). Therefore, I argue that my model avoids overfitting and that it scores better in generalisability.

3.8.4 Experiment 4

In this experiment, I will replicate Heyduk's (1975) experiment to see if my complexity measures correlates well with Heyduk's human data. Note that the stimuli used here is again polyphonic.

Method

Materials The stimuli consisted of all four pieces from Heyduk's (1975) experiment.

Procedure The four pieces were entered into a computer and converted into the OPM format. The cognitive information was calculated for all four pieces. The correlation between computed information and human judgements (in Figure 3.8) is reported, along with statistical significance and gMDL$_+$.

Results

Results are shown in Figure 3.12. There are statistically significant correlations between my proposed model and human data as well as between the reduced model and human data ($p < .05$). The baseline model and the human data do not have statistically significant correlations ($p > .05$).

Measure	r	df	p	gMDL$_+$
Proposed	0.98768	2	0.0062	-3.640
Reduced	0.87465	2	0.0630	-0.371
Baseline	0.61580	2	0.1900	0.736

Figure 3.12: Correlation with Heyduk's (1975) data in Experiment 4

Discussion

The superiority of my proposed measure over the reduced and baseline models can be seen from the superior r-value and the gMDL$_+$ code length above.

3.9 General Discussion

While Experiment 3 provided strong evidence for a priming theory in perceived complexity, I leave further explanation of this conclusion to future work as it is really outside the scope of this thesis. At this juncture I will simply treat both as equally probable.

The sum of all the gMDL$_+$ values (see Subsection 3.3.1) so far is displayed in Figure 3.13.

Measure	$\sum_{1,2,4}$gMDL$_+$	$\sum_{1,3,4}$gMDL$_+$
Proposed	-1.40	-2.60
Reduced	-0.12	-0.94
Baseline	5.90	7.60

Figure 3.13: Meta-analysis of gMDL$_+$ code lengths

According to Figure 3.13, my proposed measure is the winning measure, because it has the lowest combined gMDL$_+$ (for both Experiment 1 + 2 + 4 and Experiment 1 + 3 + 4). Therefore, this measure was chosen for use in the next two chapters in the modelling of musical similarity and fitness.

This is the most important chapter of my thesis, in which I have proposed a detailed methodology for this thesis (meta-gMDL$_+$), motivated and presented a computational theory of musical memory, extended my theory to deal with two musical objects, proven that my extended theory obeyed the three laws of Shannon, and provided empirical evidence for my measures.

Pitt and Myung (2002) stated that if experimental data are noisy then standard goodness-of-fit measures (such as the square of the correlation coefficient) may not be the best way to compare models of cognition. While good models would require some goodness-of-fit, beyond a certain point the extra goodness-of-fit could mean overfitting, thus reducing the generalisability of the model (Pitt and Myung, 2002). This is the idea behind the MDL methodology. However, as I am only using one predictor, this methodology might be a bit overkill (all my models are parameterless so far). Still, it is a good idea to start with a flexible methodology, so the switch to more complex models would be easy, should the need arise.

3.10   Concluding Remarks and Future Work

My model takes ⟨onset, pitch, metrical level⟩ as its input. This means that duration, timbre, pedalling and other data are being thrown away. Future models may potentially benefit from incorporating these missing factors.

The predictive accuracy of my proposed model is not great. With variance accounted for as low as 16% in Experiment 1, there is certainly plenty of scope for improvement. For example, the way that ||STM|| is defined is not entirely justified. The square root bit is rather *ad hoc*. Future work should, *inter alia*, look into better ways to define ||STM|| as well as TDNN.

Finally, more work is needed on the psychological meaning of complexity. In Experiment 3, I have hinted at the possibility of perceived complexity as conditional information rather than plain information. Whether this should be the case remains unclear (more data is needed). Also, the link from cognitive information to reaction time can (and should) be investigated.

Chapter 4

Musical Similarity

4.1 Introduction

In the last chapter, I proposed an information-theoretic model of musical complexity that is constrained by theories of music cognition. I argued, and provided evidence with three experiments, that my model is better (according to gMDL$_+$) than a generic Kolmogorov complexity-based model (Cilibrasi et al., 2004), apparently because cognitively unconstrained models (such as Kolmogorov complexity) treat every detail in an object as equally important, thus ignoring the possibility that not all physical information is cognizable. In affirming the importance of cognitive constraints on memory and music perception (Atkinson and Shiffrin, 1968; Lerdahl and Jackendoff, 1983; Narmour, 1999; Temperley and Sleator, 1999), I have chosen a psychologically-motivated data representation for my cognitive information model, and have shown that this model provides the best overall fit to published human data on perceived musical complexity.

Next on my agenda is to derive models at the Psychology Layer (see Figure 1.1) that make use of my cognitive information model. I will first look at similarity, which is an important and well-published area of research (*cf.* Tversky, 1977; Medin et al., 1993). In memetics, the central concept of copying-fidelity (Dawkins, 1976) refers to the closeness between parent and child memes; the general concept of closeness is usually called *similarity* in music psychology (Cambouropoulos, 2001; Hofmann-Engl and Parncutt, 1998; Eerola et al., 2001). Given this connection, it would be interesting (and novel) to look at copying-fidelity from a psychological point of view. However, to date, there are no ready-made, psychological models of similarity that would allow me to plug cognitive information models into them directly. So, in this chapter, I first propose a unified theory of similarity by combining Tversky (1977) and measure theory (which would enable this plug-and-play functionality). I will then plug my cognitive information model into

this unified theory of similarity, and examine three parameterisations of it. To simplify terminology, I call each of these parameterisations a *model*. The best model amongst the three proposed is chosen using the same model-fitting strategy used in Chapter 3. A new experiment is described to further test the best-fitting similarity model. Whilst this best-fitting similarity model could also work as a model of copying-fidelity, the variance accounted for might be too low for an accurate reconstruction of (memetic) phylogeny (Graur and Li, 2000), so I will make no claims about phylogeny here.

The plan of this chapter is as follows:

1. To investigate well-established mathematical forms of set-theoretic and information-based similarity measures;

2. To review some music psychological experiments in similarity;

3. To propose a new similarity framework based on Tversky (1977) and measure theory; this unifying framework subsumes all the aforementioned similarity measures both set-theoretic and information-theoretic (including Tversky's);

4. To select the best parametrisation which fits published psychological data best;

5. To describe a new experiment to see if the best-fitting formula would also fit the experimental data well.

In following this plan, we need a further review of the state of the art in this area. I will first look at the mathematical background.

4.2 Review of Similarity Measures

4.2.1 Mathematical Forms

In order to understand similarity mathematically, we need a precise notion of how far apart two things are. Mathematically, a metric space (X, d) is the set X with a distance function $d : X \times X \mapsto \mathbb{R}$ such that the following axioms hold (Blumenthal, 1953):

1. $d(x, y) \geq 0$ with equality iff $x = y$ (positive definiteness),

2. $d(x, y) = d(y, x)$ (symmetry), and

3. $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality).

It was observed that not all similarity ratings obey the metric axioms (*cf.* Tversky, 1977), but hopefully there exist mathematical formulae to measure the degree of metric violations such that it is possible to test the metricity of human ratings. In the context of evaluating the tour quality of asymmetric travelling salesman problem (ATSP) benchmarks, Johnson et al. (2002) proposed three measures of tour quality, of which two are related to the metric properties. I will use these two measures to calculate metric violations in this chapter. The first relevant measure calculates the extent of symmetricity violation and the second one calculates the extent of triangle inequality violation (for a distance matrix $d_{N \times N}$). Johnson et al.'s first measure is

$$\frac{\sum_{j<i} |d_{ij} - d_{ji}|}{\sum_{j<i} |d_{ij} + d_{ji}|},$$

which can be interpreted as the Bray-Curtis distance (Faith et al., 1987) between the lower-triangular portion of a distance matrix and its upper-triangular counterpart. This measure ranges from 0 to 1 and equals 0 if and only if the matrix is completely symmetric. Their second measure is

$$\sum_{i \neq j} \frac{1 - d'_{ij}/d_{ij}}{N(N-1)}$$

where

$$d'_{ij} = \min \begin{cases} d_{ij}, \\ \min\{d_{ik} + d_{kj} : 1 \leq k \leq N\}. \end{cases}$$

Its value ranges again from 0 to 1 and reaches 0 if and only if the triangle inequality is obeyed.

In psychometry, Restle (1959) proposed a set-theoretic model of distance in which objects are represented by sets containing arbitrary elements. A measure function m (as in measure theory; see Section 4.3) is used to characterise the weight of each element:

$$D_{ij} = m[(S_i \cup S_j) \cap (\overline{S_i \cap S_j})].$$

Restle proved that D_{ij} obeys the three metric axioms shown above. Building on Restle's set-theoretic formulation, Tversky (1977) stated that a distance function need not satisfy the metric axioms; instead he proposed a feature-based model of similarity assuming the *matching function* axioms (while getting rid of the metric axioms and measure theory altogether):

1. $s(a,b) = F(A \cap B, A - B, B - A)$ (matching),

2. $s(a,b) \geq s(a,c)$ if $A \cap B \supset A \cap C$, $A - B \subset A - C$ and $B - A \subset C - A$ (monotonic-

ity).

These axioms say that similarity is expressible as a function of the objects' common and distinctive features, and that similarity increases if common features are added and decreases if distinctive features are added. This formulation can account for certain observed violations of metric axioms in human data (Tversky, 1977). Together with three auxilliary assumptions,[1] the matching function axioms lead to the celebrated *contrast model*,

$$S(a,b) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A)$$

where $\theta, \alpha, \beta \geq 0$, and its normalised counterpart, the *ratio model*,

$$S(a,b) = \frac{f(A \cap B)}{f(A \cap B) + \alpha f(A - B) + \beta f(B - A)}$$

where $\alpha, \beta \geq 0$. Here $f(\cdot)$ is restricted to a general linear transformation[2] reflecting the salience of various features. This transformation is sometimes called a *saliency function* (Cazzanti and Gupta, 2006). This model generalises some well-known set-theoretic similarity measures (Tversky, 1977): for example, if we denote $N_A = |A|$, $N_B = |B|$ and $C = |A \cap B|$, then the ratio model corresponds to the coefficients of Jaccard, Dice, Simpson and Braun-Blanquet, among others (*cf.* Cheetham and Hazel, 1969, pp. 1132–1133). See Figure 4.1 for a list of these coefficients along with the corresponding ratio model parameters.

Coefficient	Formula	Corresponding ratio model
Jaccard	$\frac{C}{N_A + N_B - C}$	$\alpha = \beta = 1$
Dice	$\frac{2C}{N_A + N_B}$	$\alpha = \beta = \frac{1}{2}$
Simpson	$\frac{C}{N_A}$, where $N_A \leq N_B$	$\alpha = 1, \beta = 0$
Braun-Blanquet	$\frac{C}{N_B}$, where $N_A \leq N_B$	$\alpha = 0, \beta = 1$

Figure 4.1: Similarity coefficients with corresponding ratio models

While the ratio model is very versatile, it is not without its problems. One problem is that the Tversky model does not work for feature sets that are fuzzy (Santini and Jain, 1999). In fuzzy logic (Zadeh, 1965), an object can have feature X and not have feature X at the same time. For instance, if X is "very tall", then the set membership may be ambiguous (with a graded degree of membership between 0 and 1 depending on height). This cannot be done with classical set theory, which says that the object is either "very

---

[1]These are the independence, solvability and invariance axioms; they are not relevant to the discussion here.

[2]Or "interval scale" as in Tversky's (1977, p. 332) original.

tall" or not "very tall". Even if we accept this limitation and continue to use classical
set theory, it would still be problematic if the set of features is not defined *a priori* but
determined adaptively (or even not defined at all, as in the case of my cognitive infor-
mation measure), since by definition, the Tversky model only applies to a predefined set
of features that describe the two objects in comparison. Of course, there have been work
on fuzzifying Tversky's measure (Bouchon-Meunier et al., 1996; Santini and Jain, 1999),
which extend the set membership function to a fuzzy truth value in $[0, 1]$ when applying
the contrast and ratio formulae. But we need something more—a general, non-negative
set function that is bounded by $[0, \infty]$—so that we can plug in arbitrary, non-negative
measures as a (quasi) set membership function. As it turns out, this is possible if we
recombine Tversky (1977) with measure theory (see Section 4.3).

In machine learning, there is another approach along Tversky's lines (Lin, 1998), but
developed independently of Tversky,[3] on creating a ratio-like model. Lin's (1998) mea-
sure, based on slightly different assumptions from Tversky's, is

$$\text{sim}(A, B) = \frac{\log P(\text{common}(A, B))}{\log P(\text{description}(A, B))},$$

where $P(\cdot)$ stands for the probability of truth of a proposition, $\text{common}(A, B)$ stands for
a proposition that states the commonality between A and B, and $\text{description}(A, B)$ stands
for a proposition that fully describes A and B. The advantage of this measure is that it
is information-theoretic (based on the negative logarithm of probabilities), so it is "uni-
versal" and can even "be used in domains where no similarity measure has previously
been proposed" (Lin, 1998, p. 296). The downside is that this measure is underspecified:
it could encapsulate just about anything, since "common" and "description" are left un-
defined as a price to pay for the "universality" of this measure. However, note that in the
actual examples that Lin (1998) gave, all four of them have the following form (a factor
of 2 is introduced because this measure needs to be normalised between 0 and 1):

$$\text{sim}(A, B) = \frac{2 \times \log P(f(A) \cap f(B))}{\log P(f(A)) + \log P(f(B))}.$$

Here the implicit definition of $P(\text{common}(A, B))$ is $P(f(A) \cap f(B))$ and that of
$P(\text{description}(A, B))$ is $\sqrt{P(f(A))P(f(B))}$. Written this way, Lin's measure can be seen
(at least in actual applications) as an information-theoretic extension of Dice's coefficient,

---

[3]Lin (1998) has briefly mentioned the contrast model by name (in the second sentence in the introduction),
together with a citation to Tversky (1977), but then in the main text, Lin (1998) reinvented the Tverskian-
axiomatic approach to similarity without crediting Tversky; furthermore, Lin's assumption 3 is very similar
to Tversky's matching axiom. So I believe that Lin has not read Tversky's paper.

a precursor of Tversky's ratio model. In fact, Lin's measure can always be written in the above form if we make the additional assumption that the features are probabilistically independent (Cazzanti and Gupta, 2006).

The most recent synthesis of set-theoretic and information-theoretic similarity is published by Cazzanti and Gupta (2006). They began with Tversky's contrast model with $\theta = 1, \alpha = \beta = \frac{1}{2}$ (same as Dice's coefficient and Lin's measure, except that the contrast model is used instead of the ratio one). Their novelty is in the use of Shannon mutual information (reviewed in Chapter 2) for the saliency function, and the incorporation of an extra object R (manifested as a set of features) which serves as the comparison context: $f(a \cap b) = I(R; a \cap b \subset R), f(a - b) = I(R; a - b \subset R)$, and $f(b - a) = I(R; b - a \subset R)$. With these parameters, the *residual entropy similarity* results:

$$s_{re}(a, b) = -H(R|a \cap b \subset R) + \frac{H(R|a - b \subset R)}{2} + \frac{H(R|b - a \subset R)}{2}$$

where $H(R|a \cap b \subset R) = -\sum_r P(R = r|a \cap b \subset R) \log P(R = r|a \cap b \subset R)$. Note that this measure, like Tversky's, need not obey the metric axioms.

Now I will review a number of information-theoretic distances (ones that obey the metric axioms) that have been proposed by different researchers over the years. These distances are well known in information theory but are not currently connected to Tversky's measure (in Section 4.3, I will propose a novel framework to link them together). These are the Rajski distance, the Horibe distance, the Kvålseth distance, the normalised sum distance, the normalised universal cognitive distance, and the normalised compression distance. Rajski (1961) defined a metric for discrete probability distributions, based on Shannon's (1948) information theory:

$$d(x, y) = \frac{H(x|y) + H(y|x)}{H(x, y)},$$

where x and y are probability distributions and $H(\cdot)$ denotes the Shannon entropy. Rajski's metric has a range of $[0, 1]$, with zero representing equality (up to isomorphism) and one representing statistical independence, and obeys the three metric axioms above. Rajski has also defined a coherence coefficient,

$$R(x, y) = \sqrt{1 - d^2(x, y)},$$

also in the range $[0, 1]$, with zero representing independence and one representing equality. A similar Horibe (1985) correlation coefficient and its related metric are defined as:

$$\rho(X, Y) = 1 - d(X, Y)$$

and

$$d(X, Y) = \begin{cases} H(X|Y)/H(X), & \text{if } H(X) \geq H(Y) \\ H(Y|X)/H(Y), & \text{if } H(X) \leq H(Y) \end{cases}$$

where $\rho(X, Y)$ is in the range $[0, 1]$ with zero indicating independence and one indicating isomorphism. Horibe proved that $d(X, Y)$ satisfies the three metric axioms, and gave an intuitive interpretation of $\rho(X, Y)$: assume without loss of generality $H(X) \geq H(Y)$, we have

$$[1 - \rho(X, Y)]H(X) = H(X|Y).$$

Thus, $\rho(X, Y)$ measures the relative reduction of uncertainty in X after knowing Y (Horibe, 1985). With regards Horibe's correlation measure, Kvålseth (1987) observed that

$$\rho(X, Y) = I(X; Y)/D,$$

where $D = \max\{H(X), H(Y)\}$. Kvålseth criticised Horibe's choice of D and demonstrated that a better D existed (at least in terms of statistical inferences):

$$D = \frac{H(X) + H(Y)}{2}.$$

The corresponding distance metric for Kvålseth's D is:

$$d(X, Y) = 1 - \rho(X, Y) = 1 - \frac{2I(X; Y)}{H(X) + H(Y)}.$$

Now I will review the remaining four measures on my list. Before I can do so, however, I need to introduce two more measures: the (unnormalised) sum distance and the (unnormalised) universal cognitive distance, on which their normalised versions are based. Bennett et al. (1998) defined these information distances based on Kolmogorov complexity.[4] They obey the metric axioms, but only approximately, for there is an "additive constant or logarithmic error term" involved (Bennett et al., 1998). The sum distance is equal to:

$$E_3(x, y) = K(x|y) + K(y|x) + O(\log(K(x|y) + K(y|x))).$$

---

[4]Refer to Section 2.3 for the definition of $K(x|y)$.

The universal cognitive distance is defined as:

$$E_1(x, y) = \max\{K(x|y), K(y|x)\}.$$

These two distances have been shown to be equal up to an additive logarithmic term (Bennett et al., 1998).

Bennett et al. (1998) proposed a property of "admissibility" on distance functions (defined as $\sum_{y:y \neq x} 2^{-D(x,y)} < 1$), and proved that for any admissible distance $D(x, y)$, we have $E_1(x, y) \leq D(x, y)$ up to an additive constant. In other words, E_1 is the optimal admissible distance (Bennett et al., 1998).

Li et al. (2003) extended the above information distances by normalising them to the range of $[0, 1]$. Their rationale is that two short strings with a distance n apart are probably not as similar as two long strings with the same distance n. They have normalised both the sum distance,

$$d_s(x, y) = \frac{K(x|y) + K(y|x)}{K(x, y)},$$

and the universal cognitive distance,

$$d(x, y) = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}},$$

and proved that the latter is more precise from a mathematical point of view.

Cilibrasi and Vitányi (2005) proposed a normalised compression distance which is an approximate version of the universal cognitive distance. By substituting the uncomputable Kolmogorov complexity K with a computable real-world compressor C, after some algebraic manipulation they arrived at

$$\text{NCD}(x, y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}.$$

Although independently developed, we can see that the normalised sum distance (Li et al., 2003) is effectively the Rajski (1961) metric, and that the normalised universal cognitive distance (Li et al., 2003) is effectively the Horibe (1985) metric, apart from a change in the underlying information theory (Shannon *versus* Kolmogorov). Therefore, we have a strong motivation for a unifying framework (proposed in Section 4.3).

4.2.2 Music Psychology

Since musical similarity is a psychological phenomena, a review of relevant literature is in order. Over the years, Deliège (1996) has proposed a psychological theory that examines

cue abstraction (feature salience), similarity, and category formation (*cf.* Cambouropoulos, 2001). Cambouropoulos extended Deliège's work into the computational domain and defined similarity as follows:

$$s_h(x,y) = \begin{cases} 1, & \text{iff } d(x,y) \leq h \text{ (similarity)} \\ 0, & \text{iff } d(x,y) > h \text{ (dissimilarity)} \end{cases}$$

where $d(x,y)$ denotes any distance function and h denotes a distance threshold.[5] While Cambouropoulos stated that "the distance between two objects can be calculated by various distance metrics", what he actually used in his paper is a version of the weighted Hamming distance where the weights are calculated adaptively by his *Unscramble* clustering algorithm (Cambouropoulos, 2001). However, as his distance function is adaptive, it is impossible to use a distance matrix as input (this is the biggest difference between *Unscramble* and many other distance-based clustering algorithms). Cambouropoulos' clustering algorithm was quite successful: in his Experiment 1 (Cambouropoulos, 2001), *Unscramble* was able to replicate the second task of Deliège's (1996) experiment without errors. In this task (Deliège, 1996), participants were first asked to memorise the two reference motifs A and B. They were then asked to classify 72 derivative motifs (of which 24 are distinct) into either family A or B. It was found that the musicians classified all of the motifs correctly, while the non-musicians classified 90% of the motifs correctly (Deliège, 1996).

Recall that my best-fitting information model deals with interval data by virtue of the neural cancellation filter (de Cheveigné, 1993). The first similarity model that deals with interval data is probably Hofmann-Engl and Parncutt's (1998). Hofmann-Engl and Parncutt (1998) investigated isochronous melodic similarity based on *normalised contour difference* (effectively the Hamming distance on contour, *e.g.*, up-up-down and up-down-up has a contour difference of two) and *normalised interval difference* (like the city block distance on intervals except that the absolute value is not taken, and furthermore the final value is normalised by the total number of intervals). They found that the correlation between normalised contour difference and participants' similarity ratings was low, but on the other hand, normalised interval difference accounted for 76% of the variance in the data. They concluded that interval differences are good predictors for melodic similarity. However, it should be noted that these experiments are somewhat synthetic (isochronous melodic fragments with 1–5 tones, with manipulations on tempo, transposition, inversion and order), thus its generalisability and ecological validity are somewhat

---

[5]For mathematicians, Cambouropoulos' formulation can be condensed for readability: define $s_h(x,y)$ as the characteristic function of x associated with the closed-h-ball centred at y.

suspect, and the high correlation may very well be a result of overfitting.

Eerola et al. (2001) motivated their work by the well-known fact that we are sensitive to the statistical properties of melodies. Eerola et al. (2001) began by questioning the sufficiency of statistical features in the predictions of melodic similarity. The statistical similarity is calculated as the city block distance between the distributions of tones, intervals, durations, two-tone transitions, interval transitions, and duration transitions of the two melodies. With a stimuli repertoire of fifteen folk songs, they performed similarity rating experiments on seventeen participants by asking them to rate on a 1–9 scale the similarity of every possible combination of pairs within the repertoire (with a randomised presentation order between pairs and within pairs). This procedure lasted about an hour. Their results showed that statistical musical properties (frequency of events using zero- and first-order statistics) could only account for 39% of the participants' similarity ratings, while descriptive variables (*e.g.*, tonal stability, mean pitch, number of tones) accounted for as much as 62% of the human ratings. Eerola et al. argued that this means statistical features were not effective predictors of music similarity (but nonetheless they admitted that further study is necessary). I disagree with this conclusion since they used only zero- and first-order statistics, which is not doing justice to the entire arsenal of statistical tools available to us. Therefore in this thesis I will use a more sophisticated model based on my new information theory.

Eerola and Bregman (2007) conducted more experiments in the same direction, this time with twenty phrases sampled from the Essen collection (Schaffrath and Dahlig, 2000; Schaffrath, 1997). Drawing on their experience with their previous experiment, they had made their stimuli much shorter, and allowed for up to 30% of pairings to be randomly omitted, such that their participants would not be over-fatigued. With twenty-two musically-trained participants, Eerola and Bregman correlated the human ratings with five similarity predictors: contour, pitch content, interval content, contour periodicity, and range. All five predictors were highly significant and the best single predictor is pitch content, accounting for 47% of the variance. In their discussion, Eerola and Bregman (2007) wrote:

> A re-analysis of a previous study [...] suggested that listeners use the most salient variation between stimuli as the deciding factor in similarity judgements [...] In an ideal similarity model, the features that contribute to similarity would be dynamically modified by the salient vairation within the context of comparison. (p. 227–228)

A connection between the quote above and my similarity framework is that salient features are dynamically discovered by a context-sensitive, cognitive information theory

that looks at global statistical regularities of both pieces at once. To test how well my measures approximate human cognition, I will replicate the second task of Deliège (1996), as well as both Eerola et al. (2001) and Eerola and Bregman (2007) computationally in this chapter.

4.3 A Novel Framework for Similarity

In this section, I will present a measure-theoretic analogue of Tversky's (1977) ratio model, motivated in part by Restle's (1959) use of measure theory and in part by measure-theoretic formulations of information theory (Hu, 1962). First we need a definition of a measure (Saks, 1937):

Definition 4.1 (Saks). *Let $\mathfrak{X}$ be any set. A non-negative function $\mu(X)$ defined on every subset of $\mathfrak{X}$ is a measure if*

$$\mu(\bigcup_n X_n) = \sum_n \mu(X_n)$$

for every pairwise disjoint sequence $\{X_n\}$ of subsets of $\mathfrak{X}$.

In simpler terms, a measure assigns arbitrary measurements (such as length, area, volume or counts) to all subsets of a set,[6] with the requirement that the measurements are non-negative and additive for disjoint sets. Intuitively, this means that area-proportionate Venn diagrams (area proportional to the measure of the subset) can be physically drawn on paper.

Now I will present my framework based on Tversky (1977). My framework is

$$S(a,b) = \frac{\mu(A \cap B)}{\mu(A \cap B) + \alpha\mu(A - B) + \beta\mu(B - A)}$$

where $\alpha, \beta \geq 0$ and $\mu(\cdot)$ is a measure. It has long been known that Shannon entropies have a measure-theoretic interpretation (Hu, 1962). Let a, b and c be dummy set variables associated with any random variables X and Y[7] via the following definitions:

$$\mu(\varnothing) = 0$$
$$\mu(\{a\}) = H(X|Y)$$

---

[6] Note that a measure can be defined on smaller collections of subsets, but these collections (called σ-algebras) are outside the scope of this thesis.

[7] Beyond the use of X and Y in the calculation of $\mu(\cdot)$, there are no further relationships between $\{a, b, c\}$ and $\{X, Y\}$.

$$\begin{aligned}
\mu(\{b\}) &= H(Y|X) \\
\mu(\{c\}) &= I(X;Y) \\
\mu(\{a,b\}) &= H(X|Y) + H(Y|X) \\
\mu(\{a,c\}) &= H(X) \\
\mu(\{b,c\}) &= H(Y) \\
\mu(\{a,b,c\}) &= H(X,Y).
\end{aligned}$$

The corresponding Venn diagram is shown in Figure 4.2.
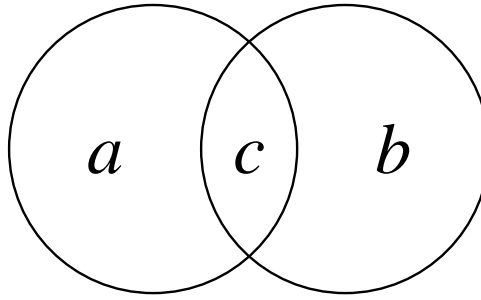
Figure 4.2: Venn diagram associated with X and Y

The Shannon entropies in the above formulation can be replaced by my cognitive information measures $H(\cdot)$, Kolmogorov complexities $K(\cdot)$, or Cilibrasi and Vitányi's (2005) compression-based complexities $C(\cdot)$ since they all obey (at least approximately) Shannon's three theorems (Hammer et al., 2000; Cilibrasi and Vitányi, 2005). I will assume that similarity measures are normalised to $[0,1]$, as per the normalised distance measures reviewed above. Additionally, I will define distance as the inverse of similarity, $d(X,Y) = 1 - S(X,Y)$. A corollary is that my framework encapsulates all the normalised information-theoretic distances reviewed above, as well as the set-theoretic ones originally covered by Tversky (see Figure 4.3; *cf.* Figure 4.1).

Distance	Corresponding ratio model
Rajski (1961)	$\alpha = \beta = 1$
Horibe (1985)	$\alpha = 0, \beta = 1$ where $H(X) \leq H(Y)$
Kvålseth (1987)	$\alpha = \beta = \frac{1}{2}$
Li et al. (2003), $d_s(x,y)$	$\alpha = \beta = 1$
Li et al. (2003), $d(x,y)$	$\alpha = 0, \beta = 1$ where $K(X) \leq K(Y)$
Cilibrasi and Vitányi (2005)	$\alpha = 0, \beta = 1$ where $C(X) \leq C(Y)$

Figure 4.3: Summary of information-theoretic distance measures

4.4 Three Competing Models of Musical Similarity

I will now define three competing musical distance measures based on Figure 4.3 and my newly proposed similarity framework, replacing the Shannon, Kolmogorov or Cilibrasi measures with my cognitive information measure. Three models follow:

$$\textbf{H1:} \quad d(x,y) \;=\; \frac{H(x|y) + H(y|x)}{H(x,y)}$$

$$\textbf{H2:} \quad d(x,y) \;=\; \frac{\max\{H(x|y), H(y|x)\}}{\max\{H(x), H(y)\}}$$

$$\textbf{H3:} \quad d(x,y) \;=\; \frac{H(x|y) + H(y|x)}{H(x) + H(y)}$$

These are the cognitive information analogues of the Rajski (H1), Horibe (H2), and Kvålseth (H3) measures, respectively. Here $H(\cdot)$ denotes my best-fitting cognitive information measure as defined in Chapter 3. It is clear that all three distances (H1–H3) are normalised to the range of $[0, 1]$, due to the fact that $H(X|Y) < H(X)$.

Recall that Li et al. (2003) called the Kolmogorov complexity version of H2 the "normalised universal cognitive distance". I have no problem with their term "universal", which comes from Kolmogorov complexity theory, but I strongly disagree with their calling it "cognitive". As I have said before, human minds obviously do not perform algorithmically optimal compression. This is evident in our inability to compress π (correct to say three billion digits) into our long-term memory—whereas the length of the shortest program that outputs three billion digits of π is much smaller than three billion. Hence I substitute the complexity measure with my proposed cognitive information measure, and hope to result in a better model of cognitive distance. The quality of these models can be empirically tested by the two experiments proposed in the next section.

Additionally, my musical similarity measure $S(x, y)$ equals $1 - d(x, y)$ (see Section 4.3), which can be interpreted as the opposite of musical distance. Musical similarity can be interpreted as a measure of cognitive independence between two pieces of music: identical pieces would yield a value of zero, while pieces that are cognitively independent will yield a value of one.

Finally, the metric properties of my proposed measures are not known. Here I simply report the degree of metric violations using Johnson et al.'s measures reviewed above. If the violation is nearly zero then it provides indirect evidence that perceived similarity may be defined on a metric space.

In the next subsection, I present the results of three experiments to select the most

promising cognitive model.

4.4.1 Model Selection with Three Experiments

In the experiments below, I will simply correlate the human data in Deliège (1996), Eerola et al. (2001) and Eerola and Bregman (2007) with my similarity measures. I will also report their metric properties where possible. The best-fitting similarity measure is then selected using meta-gMDL$_+$ (proposed in Chapter 3).

Correlation with Deliège's (1996) Experiment

Only her second task is replicated.[8] My model predictions are $\Delta = d(\text{music, A's prototype}) - d(\text{music, B's prototype})$ for each of the 26 motifs and for all three distance measures. The idea is that if music is closer to A's prototype then $\Delta \leq 0$, else $\Delta > 0$. The point-biserial correlations between my model predictions and the true classifications[9] are then calculated. Results are shown in Figure 4.4. The correlations are statistically significant for H2 and H3 ($p < .05$).

Measure	r	df	p	gMDL$_+$
H1	0.29870	24	0.069	2.49
H2	0.38291	24	0.027	1.93
H3	0.42199	24	0.016	1.55

Figure 4.4: Point-biserial correlation with Deliège's (1996) data

Correlation with Eerola et al.'s (2001) Experiment

Recall that Eerola et al. (2001) performed similarity rating experiments by asking participants to rate the similarity of all combination of pairs given a repertoire of fifteen songs. The correlations of Eerola et al.'s mean human ratings with my hypotheses are shown in Figure 4.5. All correlations are statistically significant ($p < .05$). Metric violations are shown in Figure 4.6.

Correlation with Eerola and Bregman's (2007) Experiment

Recall that Eerola and Bregman (2007) conducted another experiment in the same direction, but now with Essen folksong phrases instead of complete songs and that all partici-

---

[8]Deliège did four tasks in total (the first of which is irrelevant to similarity), so two more tasks remain to be replicated in future work.

[9]Recall that the musicians classified all of the motifs correctly in the original experiment (see Section 4.2.2).

Measure	r	df	p	gMDL$_+$
H1	0.42047	103	4.0×10^{-6}	-4.01
H2	0.45405	103	5.7×10^{-7}	-5.82
H3	0.41780	103	4.6×10^{-6}	-3.87

Figure 4.5: Correlation with Eerola et al.'s (2001) data

Data	Asymmetry	Triangle inequality violation
Human	0.0000	0.0034
H1	0.0025	0.0000
H2	0.0000	0.0000
H3	0.0059	0.0000

Figure 4.6: Metric violations pertaining to Eerola et al.'s (2001) data

pants were musicians. The correlations of their mean human ratings with my hypotheses are shown in Figure 4.7. All correlations are statistically significant ($p < .05$). Metric violations are shown in Figure 4.8.

Measure	r	df	p	gMDL$_+$
H1	0.51983	188	7.6×10^{-15}	-22.6
H2	0.45283	188	2.7×10^{-11}	-14.6
H3	0.49928	188	1.1×10^{-13}	-19.9

Figure 4.7: Correlation with Eerola and Bregman's (2007) data

Data	Asymmetry	Triangle inequality violation
Human	0.000000	0.0035
H1	0.006200	0.0000
H2	0.000029	0.0000
H3	0.016000	0.0000

Figure 4.8: Metric violations pertaining to Eerola and Bregman's (2007) data

Discussion

The sum of gMDL$_+$ code lengths are shown in Figure 4.9. The three measures seem to be very close to each other, but for the purpose of this chapter, the winner is the one with the smallest total description length, which is H1. All reported metric violations are minor, suggesting that the metric axioms hold. In the next subsection, I will present an experiment to test H1.

Measure	$\sum$gMDL$_+$
H1	-24
H2	-19
H3	-22

Figure 4.9: Meta-analysis of gMDL$_+$ code lengths for similarity

4.4.2 Validation of H1

The aim of this experiment is to test whether H1 can be psychologically validated for pairs of polyphonic music presented to the participants. My hypothesis is that there is a correlation between H1 and human judgements of similarity. I assumed that the participants are representative with respect to similarity judgements, and controlled for age, sex and musical training.

Method

Participants   Participants were recruited from the University of Sheffield through a university-wide volunteers' e-mail list. The eligibility criteria are that they be aged 18–64, have at least one healthy ear, and not have musicogenic epilepsy. They were paid eight pounds sterling per hour for their participation. Ethical approval was granted by the Department of Psychology Ethics Sub-Committee at the University of Sheffield. Thirty participants signed up for this experiment with informed consent (10 males and 20 females). The mean age was 36 years (SD = 11) and the mean years of musical training was 4.9 years (SD = 8.6).

Materials   The stimuli consisted of fifteen short MIDI files downloaded from the Internet prior to the experiment. The main critera of inclusion are that their lengths must be within the range of fifteen to twenty seconds each, and they must have the same timbre, tempo and key, and preferably be all by the same composer, to reduce confounding effects. I settled on some historical piano rolls performed by piano masters of the past century, meticulously scanned in and MIDI-fied by Terry Smythe, at his website `http://members.shaw.ca/smythe/rebirth.htm`. The fragments were shown in Figure 4.10. These fragments satisfied the following constraints:

- Composed by a single composer: Frédéric Chopin (1810–1849);

- In the key of C sharp minor and/or D flat major;

- Each fragment lasted 15–20s each;

- All phrases are complete phrases.

	Title	Pianist	Roll No.	MIDI Ticks
1.	Etude, Op. 10, No. 4			
	(a) Presto (bb. 1–12)	T. Lerner	Ampico 6854	100496–107928
2.	Etude, Op. 25, No. 8			
	(a) Vivace (bb. 1–8)	A. Cortot	Duo-Art 6740	44–6536
3.	Nocturne, Op. 27, No. 2			
	(a) Lento sostenuto (bb. 2–5)	J. Hofmann	Welte 668	2209–7870
4.	Scherzo, Op. 39			
	(a) A tempo risoluto (bb. 25–56)	L. Godowsky	Ampico 5111	10835–19162
	(b) Meno mosso (bb. 156–171)	L. Godowsky	Ampico 5111	46499–53950
	(c) Tempo I (bb. 573–604)	L. Godowsky	Ampico 5111	201082–210432
5.	Waltz, Op. 64, No. 1			
	(a) Molto vivace (bb. 1–36)	E. d'Albert	Ampico 5060	1–8130
	(b) Sostenuto (bb. 38–65)	E. d'Albert	Ampico 5060	11791–19353
6.	Waltz, Op. 64, No. 2			
	(a) Tempo giusto (bb. 1–16)	L. Godowsky	Ampico 5495	1007–8781
	(b) Piu mosso (bb. 33–48)	L. Godowsky	Ampico 5495	17151–21165
	(c) Piu lento (bb. 65–80)	L. Godowsky	Ampico 5495	25904–33618
7.	Impromptu, Op. 66			
	(a) Allegro agitato (bb. 5–16)	H. Bauer	Duo-Art 6058	2871–9458
	(b) Moderato cantabile (bb. 43–48)	H. Bauer	Duo-Art 6058	27053–34199

Figure 4.10: Chopin excerpts used in the experiment

Fragments were extracted using the `midicopy` utility (part of the `abcmidi` package), which allows selective copying of a small part of a MIDI file. In addition, I increased the MIDI velocities of the Duo-Art and Welte rolls by 32, due to the fact that mechanical pianos have a slightly different response curve from software synthesisers.

For the playback of MIDI files, I chose a realistic piano SoundFont file called `Akai-SteinwayIII.sf2`, downloaded from `http://www.sf2midi.com/`.

Procedure An information sheet and consent form were given to the participants (see Appendix A). Verbal instructions were also given. Once the consent form was signed, the instructions in Figure 4.11 were shown to the participants on a computer screen on which they would begin the experiment by pressing a button.

Once they have clicked on the button to begin, the main protocol started. The protocol is mainly taken from Eerola et al. (2001). There are $C(15, 2) = 105$ distinct pairs given a corpus of fifteen pieces, and all such pairs were tested. To reduce order effects, both the within-pair and between-pair orders were randomised as per Eerola et al. (2001).
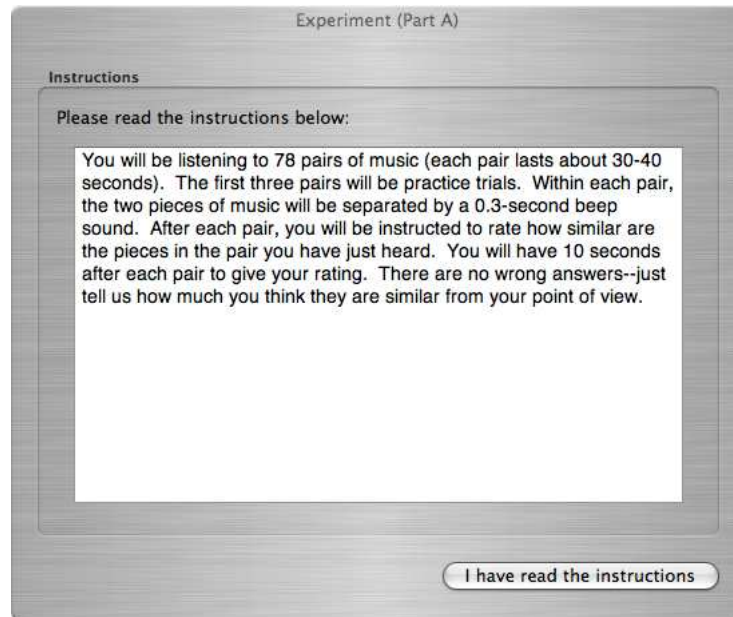
Figure 4.11: On-screen instructions

For each pair of music (say X and Y), the participant was first presented with X (Figure 4.12), followed by 0.6s of silence, followed by a 0.3s beep at 2093Hz, followed by another 0.6s of silence, followed by Y. The participant was then asked to rate the similarity between X and Y on a 1–9 scale (see Figure 4.13). The rating buttons are greyed-out until the rating phase, in order to prevent participants from pressing the buttons prematurely. The human ratings were recorded alongside the computational predictions of musical similarity, $1 - d(X, Y)$. All of the above were presented using an iTunes-like computer GUI. The first three trials were practice trials where the results are not analysed. This procedure was repeated until all 105 combinations of pairs have been exhausted. At the end, the human ratings were correlated with model predictions and the statistical significance are reported below.

As a last-minute addition, an "Obscure Names" toggle has been added to the Experiment menu (Figure 4.14), after hearing the feedback from a participant who felt that showing the name of the piece was giving too much information away. With the "Obscure Names" toggle set to on, the name display will be replaced by "Excerpt No. X". Half of the participants have this toggle set to on, in order to control for possible effects of knowing the name of the piece. The whole set of data can still be analysed at once, but if the resulting correlations are statistically insignificant, then the two datasets can be reanalysed separately.

Figure 4.12: Listening to the first piece

Figure 4.13: Rating the similarity

Data Analysis Here the unit of analysis is the mean of all participants' rating on each piece. This decision is based on theoretical grounds due to the population nature of memes—the same meme exists not only within a brain but potentially across many dif-

Figure 4.14: Listening to the first piece (with the names obscured)

ferent brains as well.

As the practice trials do not count, they are removed and treated as missing data. Since the order of presentation is by design completely random (not biased towards particular pieces), the missing values of the first three trials can also be regarded as missing completely at random. So the strategy I adopt below is to replace each missing value by the mean of other participants' ratings on the same piece (this is also the strategy used by Eerola and Bregman, 2007). Finally, for intra-rater reliability, Cronbach's (1951) α is reported. Cronbach's α is a type of intraclass correlation (ICC) which can be used to measure the realiability of the means of human ratings under the assumption of a two-way Judge $\times$ Target ANOVA model without interactions (Shrout and Fleiss, 1979; McGraw and Wong, 1996), and the assumption that absolute agreement of the ratings is not required ("9" from one participant could be consistent with "8" from another after a shift of anchor). If Cronbach's $\alpha < 0.8$, partition the data according to age, sex, musical training and name obscuration, and reanalyse the data separately.

Results

One participant "[a]nswered the trial questions and I think the first 2 or 3 of the experiment incorrectly—got the scale mixed up" (anon.). Assuming that the ratings of this participant is consistent with the rest of the population, the corrected scale should yield

the highest intra-rater reliability. Based on this assumption, I corrected this participant's ratings by inverting the scales for the first n responses only, for each $1 \leq n \leq 10$ (with one inversion at $n = 1$ and successively more and more inversions towards $n = 10$), incrementally. The inversions are tabulated in Figure 4.15.

n	Before	After
1	2	8
2	1	9
3	8	2
4	8	2
5	7	3
6	2	8
7	5	5
8	4	6
9	3	7
10	9	1

Figure 4.15: Scale inversions for one anonymous participant

I then look for the value of n that maximises Cronbach's α (see Figure 4.16). Here the best fit is $n = 2$ ($\alpha = 0.94767$), so I take this as the final correction. After this correction, I correlate the mean human ratings with my H1 model (see Figure 4.17). There is a statistically significant correlation, $r(76) = 0.57106$ ($p < .05$), meaning that my H1 model accounts for 33% of the variance in the average human ratings.

n	α	F	df1	df2	p
0	0.94685	18.8	77	2233	$< 10^{-37}$
1	0.94696	18.9	77	2233	$< 10^{-37}$
2	0.94767	19.1	77	2233	$< 10^{-37}$
3	0.94757	19.1	77	2233	$< 10^{-37}$
4	0.94740	19.0	77	2233	$< 10^{-37}$
5	0.94722	18.9	77	2233	$< 10^{-37}$
6	0.94701	18.9	77	2233	$< 10^{-37}$
7	0.94701	18.9	77	2233	$< 10^{-37}$
8	0.94711	18.9	77	2233	$< 10^{-37}$
9	0.94698	18.9	77	2233	$< 10^{-37}$
10	0.94653	18.7	77	2233	$< 10^{-37}$

Figure 4.16: Cronbach's α after inversion of the first n trials for one anonymous participant

It was found that the human data obey the metric axioms, and H1 approximately obeys the metric axioms. See Figure 4.18 for violations of metric properties.
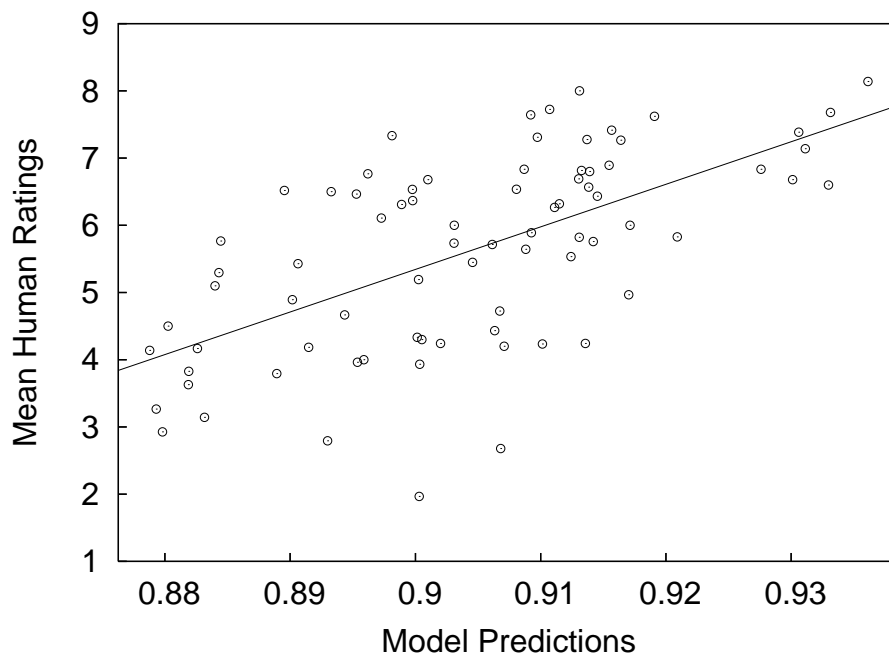
Figure 4.17: Scatterplot of results for all 105 pairs of pieces

Data	Asymmetry	Triangle inequality violation
Human	0.0000	0.0000
H1	0.0015	0.0000

Figure 4.18: Metric violations with human data

Discussion

A reasonable objection to my methodology is the use of averaged distance without accounting for the order of presentation within each pair. This always results in the asymmetry of zero as shown in Figure 4.18. To show that my human data really are metric, I reanalyse the data by averaging the values separately for each order of presentation so that we can really test the human data for asymmetry. First I redo the corrections as before (see Figure 4.19).

As expected, this again suggests that $n = 2$ is the best fit. I then look at the metric violations with this correction applied. Results show that there are very minor violations (see Figure 4.20; *cf.* Figure 4.18). My conclusion is that human data really are metric; the asymmetry observed by Tversky (1977) is probably due to confounding factors such as the built-in directionality of the question "is X similar to Y".

With regards inter-rater reliability, Shrout and Fleiss (1979) interpreted Cronbach's α

n	α	F	df1	df2	p
0	0.98893	90.3	168	4872	$< 10^{-37}$
1	0.98895	90.5	168	4872	$< 10^{-37}$
2	0.98907	91.5	168	4872	$< 10^{-37}$
3	0.98906	91.4	168	4872	$< 10^{-37}$
4	0.98901	91.0	168	4872	$< 10^{-37}$
5	0.98899	90.8	168	4872	$< 10^{-37}$
6	0.98898	90.7	168	4872	$< 10^{-37}$
7	0.98898	90.7	168	4872	$< 10^{-37}$
8	0.98899	90.8	168	4872	$< 10^{-37}$
9	0.98897	90.7	168	4872	$< 10^{-37}$
10	0.98890	90.1	168	4872	$< 10^{-37}$

Figure 4.19: Corrections with asymmetric data

Data	Asymmetry	Triangle inequality violation
Human	0.079	0.00018

Figure 4.20: Metric violations with asymmetric human data

as a de facto correlation coefficient but which is not generalisable to another group of 30 participants (participants as fixed effects). On the other hand, McGraw and Wong (1996) interpreted this value not as a real correlation coefficient in the sense of variance accounted for, but as a general value of consistency which could be generalised to any other population of 30 participants (participants as random effects). As the ability to interpret Cronbach's α as a percentage of variance is not essential here, I choose the interpretation of McGraw and Wong (1996) so I can generalise the inter-rater reliability to another group of 30 participants. The high value of α provides further evidence for the validity of the similarity construct in the context of 30 participants listening to 10–20s fragments of Chopin pieces.

4.5 General Discussion

I have proposed a novel framework of similarity combining Tversky's ratio model and measure theory. I demonstrated that the new framework subsumed a large range of previously proposed distance measures both set-theoretic and information-theoretic. I then proposed three parametrisations of this framework and have shown that the first one (H1) is marginally better, at least for the data used here. This better measure was shown to fit human data quite well. I showed that similarity ratings are metric (*contra* Tversky), at least for the human data in Eerola et al. (2001), Eerola and Bregman (2007),

and the validation experiment described here.

Medin et al. (1993) stated that similarity comparison is a dynamic, context-specific global search process in which features are not independent. Eerola and Bregman (2007) said that "in an ideal similarity model, the features that contribute to similarity would be dynamically modified by the salient variation within the context of comparison" (p. 228). My similarity measure fits this description (at least in part), because:

1. My feature space is adaptively generated by T-decomposition (Titchener, 2000);

2. It is a global search in the sense that T-composition can "see" both objects in comparison simultaneously (in a concatenated form).

4.6   Concluding Remarks and Future Work

I defined distance as the one minus similarity, because Rajski and Horibe did so. But this definition is controversial because there exists human data where the inverse relation between similarity and difference does not hold (*cf.* Tversky, 1977). Secondly, the correlation with human data is not that high—I suspect that this is due to an inaccurate Information Layer underneath. Both issues await future research.

As always, more experiments with human participants are required. Also, the Application Layer should be explored. For example, Baronchelli et al. (2005) listed a few things that one could do with information distances measures (with text documents):

1. Language recognition. This can be done by minimising the cross-entropy between a corpus and an unknown document. This cross-entropy is calculated by using the first document to create the compression dictionary, then use this dictionary (instead of the one induced from the second document) to compress the second document. While the authors have "English, French, Italian..." in mind, we could in theory substitute it with "English folk songs, French folk songs, Italian folk songs..."

2. Authorship attribution. Again we can minimise the cross-entropy to get the closest author to a document.

3. Author trees. Using a phylogeny inference package (see Graur and Li, 2000; Felsenstein, 1993), one could recreate an authorship tree from a distance matrix.

4. Language trees. Same analysis, but for a collection of the same text(s) written in different languages.

Due to possible horizontal pathways of cultural transmission, the use of information distance measures could lead to inaccurate authorship and language trees. To get around this, Baronchelli et al. (2005) proposed that one could use "Swadesh list techniques" where one would eliminate the borrowed words from a language corpus before the doing the analysis. This might be an interesting research topic for its musical analogue.

Chapter 5

Musical Fitness

5.1 Introduction

After musical similarity in the last chapter, I will now investigate another model at the Psychology Layer (see Figure 1.1), this time *musical fitness*.

In memetics, cultural fitness refers to the relative success of cultural replication and corresponds to a meme's "intrinsic appeal to a brain" (Jan, 2000a). It is often said that musical beauty is in the subjective ear of the listener, yet it is also true that there are cognitive constraints on musical listening (Lerdahl and Jackendoff, 1983; Lerdahl, 1988; Narmour, 1999). One could even argue that such constraints are universal regardless of cultural background. Despite Berlyne's (1974) early efforts, there have been few empirical studies to date on mathematical models of musical beauty. Is it possible to have a mathematical model of musical fitness that is scientifically valid? I answer in the affirmative by introducing subjectivity into a model of musical fitness proposed in this chapter.

Musical fitness is an important concept in evolutionary musical models, for both cognitive modelling and in practical applications such as evolutionary musical composition. Following the same plan as the previous chapter, this chapter is organised as follows:

1. To investigate relevant research in theoretical and empirical aesthetics;

2. To review other mathematical forms that will be used in this chapter;

3. To propose three competing models of musical fitness (and a training corpus representing the listener's musical knowledge);

4. To select the best fitting model;

5. To design a falsifying experiment to see if the best model holds.

I will first review theoretical and empirical aesthetics.

5.2 Review of Theoretical and Empirical Aesthetics

5.2.1 The Birkhoff Formulation and Beyond

The mathematical quantification of beauty dates back to Birkhoff (1933) who proposed an objective measure of beauty,

$$M = \frac{O}{C},$$

where O denotes order and C denotes complexity. Eysenck (1942) observed that Birkhoff's formulation did not agree with empirical findings (where C was shown to be positively correlated with M instead of being negatively correlated). Therefore, he proposed an improved formula

$$M = O \times C,$$

where O denotes unity and C denotes diversity. Eysenck's formula was dismissed by Katz (1994, p. 201) as "inadequate" because unity is inherently subject-dependent; however, I have addressed Katz's criticism in my proposed version of unity, in the next section, by accounting for subjectivity explicitly (where both O and C are subject-dependent).

Katz (1994) himself proposed a more complex model of musical affect based on a "connectionist operationalization" of the idea of unity in diversity. His model consists of a bank of bandpass filters connected to a multilayer neural network. Katz argued that the degree of unity in diversity can be measured by the activation levels in the neural network (high activation means high affect). With this model, one of his experiments showed that affect is an inverted U-shaped curve as a function of melodic complexity (determined by note range). However, Katz's model is currently limited to monophonic melodies within the range of an octave, and it is not clear whether his model could be adapted for polyphonic pieces with a larger range. Furthermore, his model has not been validated with human data. Nevertheless, Katz's was able to demonstrate that "the model's response to degraded versions of [good] melodies decreases with the degree of degradation" (Katz, 1994, p. 219), which seems reasonable at first glance. I will return to his degradation experiment in Subsection 5.3.4.

Of course, Katz is by no means the first to address subjectivity. Moles (1968, p. 162) for example defined "artistic value" as

$$V = f(|H - H'|)$$

where H denotes the message's originality and H' denotes the receptor's capacity (here $f(\cdot)$ is an unspecified decreasing function).

The idea of beauty as "unity in variety", "order in complexity" or "unity in diversity", although usually attributed to Birkhoff (1933) and Eysenck (1942), can be traced back for several centuries to philosophers like Francis Hutcheson (Berlyne, 1974), and this will be the lineage of research to which I will be contributing in this chapter.

5.2.2 Algorithmic Aesthetics

Stiny and Gips (1978) coined the term "algorithmic aesthetics". Among other things, they defined an entity called an "evaluation algorithm" as part of a larger aesthetic system. Given an interpretation X of an object, an "evaluation algorithm" E outputs the object's aesthetic value, *e.g.*, $E(\text{"Danse Macabre"}) = 666$. By this definition, the value of $E(X)$ depends only on how the object is interpreted, and therefore may vary from interpreter to interpreter. As an extended example, Stiny and Gips (1978) defined the evaluation algorithm

$$E_Z(\langle \alpha, \beta \rangle) = \frac{L(\beta)}{L(\alpha)}$$

where $\langle \alpha, \beta \rangle$ denotes an interpretation, α denotes the input component of the interpretation, β denotes the output component of the interpretation, and $L(\cdot)$ denotes the length of its argument. Then they linked E_Z to the aesthetic notion of unity in variety (where α is the information for construction and β is the description), and furthermore to Kolmogorov complexity (where α is the shortest program reconstructing β; see Section 2.3 for a review). Stiny and Gips (1978) did not attempt to validate their work empirically, but to be fair, psychology is orthogonal to their aim—they were effectively taking an artificial life stance.

Koshelev (1998) noted that while Birkhoff's original formula of beauty (see above) can be shown to work for specific classes of objects (such as simple melodies), a more general formalisation that is applicable to any arbitrary object is missing. Motivated by the general nature of Kolmogorov complexity theory, and independently from Stiny and Gips (1978), Koshelev (1998) formalised Birkhoff's idea to the general case by proposing that:

$$
\begin{aligned}
O &= 2^{-l(p)}, \\
C &= t(p),
\end{aligned}
$$

where p is chosen from the space of all possible programs that generates the object, $l(p)$ is the length of p, and $t(p)$ is the running time of the program such that $M = O/C$ takes the

maximum value. Koshelev (1998) noted that this value is precisely the reciprocal of the object's Levin complexity (Section 2.3). However, as Levin complexity is not computable in practice, they also proposed a practical alternative:

$$O = \text{length of the wavelet compressed object,}$$
$$C = \text{length of the zip compressed object.}$$

As this alternative measure was proposed without experimental validation, there is no way to tell, as it stands, whether this is a good model or not. It is also not clear whether this computable alternative has any mathematical relation at all to the theoretical one.

5.2.3 The Wundt Curve and Its Contenders

D. E. Berlyne is the father of modern experimental aesthetics (the subject of the next subsection). Also of note is his theoretical reinterpretation of the *Wundt curve*, which represents the inverted-*U* relation between "hedonic value" and "arousal potential" (see Figure 5.1b), hypothesised to be the summation of two opposing activities (reward and aversion) in the brain (Berlyne, 1974). Berlyne (1974) also related this hypothesis to the theories of Birkhoff and Eysenck (see above). The Wundt curve lies at the heart of both theoretical and experimental aesthetics and describes the following phenomenon: if the listener does not understand the music at all, then the music cannot be appreciated; on the other hand, if the listener can understand the music absolutely fully, then the music would sound boring.

However, the Wundt curve is not without competition. Two major contenders are identified by Walker (1973). The first one is a double-inverted-*U* function (see Figure 5.1c), while the second one is a monotonic increasing function (see Figure 5.1a). Walker stated that "with psychological complexity and preference theory, adaptation results in a gross temporary reduction in the complexity of an event and a correlated reduction in preference" and this is characterised by the double inverted-*U* shape. The adaptation-level theory is first proposed by Helson (1947) and the double-inverted-*U* shape is first proposed by McClelland et al. (1953), where it assumes that small discrepancies from the optimal adaption-level equal pleasingness, whereas big discrepancies from it equal negative affect (resulting in a butterfly-shaped curve). Haber conducted experiments with cold and hot water to support this theory (Haber, 1958).

Secondly, the monotonic increasing function can occur in two cases: one is Walker's example of piles of money, which is the more the merrier. Another one is the mere exposure hypothesis (Zajonc, 1968), which states that we prefer stimuli that we are more

(a) Monotonically increasing (Zajonc, 1968)   (b) Inverted U-shaped (Berlyne, 1974)

(c) Double inverted U-shaped (Haber, 1958)
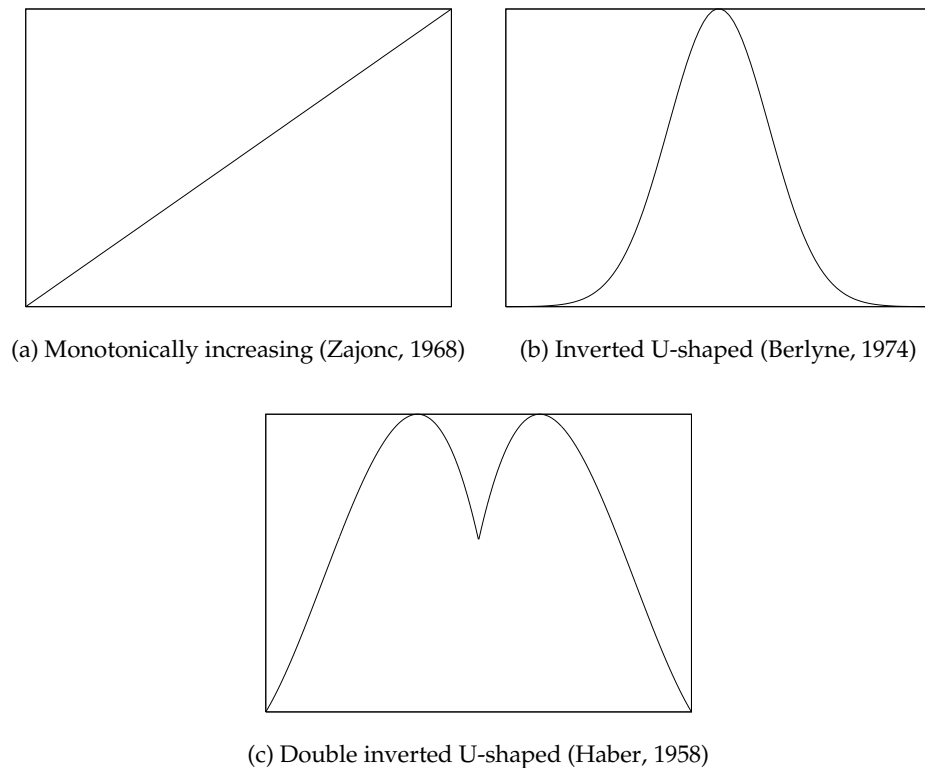
Figure 5.1: Three types of preference functions (Walker, 1973)

familiar with. Both of these contenders will be further explored in this chapter.

5.2.4 Experimental Aesthetics

Neuroscience

In neuroscience, Birbaumer et al. (1996) performed an EEG experiment consisting of three blocks (melody, rhythm, melody & rhythm) each with three types of stimuli (periodic, low-dimensional chaos, high-dimensional chaos). They found that both periodic and high-dimensional chaotic music elicited higher EEG dimensions, which "reflects the [higher] number of independently active neuronal cell assemblies" (Birbaumer et al., 1996, p. 275). Furthermore, in response to low-dimensional chaotic music in the rhythmic blocks, participants who preferred classical music responded with higher EEG dimensions whereas participants who preferred popular music responded with lower EEG dimensions. In Birbaumer et al.'s (1996) own words, "complex music produces complex brain activity in complex people, simple music excites simple brain activity in simple

people [*sic*]" (p. 268).[1]

Jeong et al. (1998) performed a similar EEG experiment with a slightly different set of stimuli. Their experiment is divided into three blocks (as shown in Figure 5.2) with four types of stimuli ($1/f$, white, brown and constant). Their stimuli were prepared as follows (Jeong et al., 1998, p. 218):

1. "**White music**. We need only one imaginary die with 120 sides to produce white music. We successively throw the die. The sequence is made from the selected number on the die. Each value has the same probability of 1/120 of being [chosen], and one quantity is not affected by any of its preceding quantities."

2. "**Brown music**. The first note [...] is determined by a random number generator. The next note of the pitch, or the duration, for brown music is determined by throwing a die with three sides $(+1, 0, -1)$. For $+1$, the fluctuating quantity (pitch or duration) increases by one step. For 0, it stays the same, and for -1, it decreases one step."

3. "**$1/f$ music**. We use twenty dice, each with six sides, to produce $1/f$ music. First, we throw all twenty dice and calculate their sum. For the next trial, we randomly choose seven dice and throw only those chosen dice again. We recalculate the sum of all twenty dice; then, we repeat the procedures [*sic*] as many times as we like."

In layman's terms, white music is the most complex (totally unpredictable), $1/f$ is in the middle, while brown music is the simplest (highly predictable).

Jeong et al.'s (1998) stimuli had a frequency range of 100–800Hz (120 notes) and a durational range of 0.1–2s (Jeong et al., 1998). Each trial lasted 30s. Their results revealed significant differences in EEG dimensions during $1/f$ music perception. Augmented by their participants' self-reports, they have established a perfect negative correlation between brain activity and aesthetic pleasingness (Jeong et al., 1998). Jeong et al. (1998) also related their results to Birkhoff's theory of aesthetics:

> The reason interesting music has $1/f$ spectra for its pitch and its duration is partially answered by the 'theory of aesthetic value' propounded by the American mathematician Birkhoff. Birkhoff's theory states that for a work of art to be pleasing and interesting, it should be neither too regular and predictable nor too irregular and unpredictable. (p. 224)

---

[1]For political correctness, "simple people" could be rewritten as "musically less sophisticated participants".

	Melody	Rhythm
Block 1	$1/f$	$1/f$
	White	White
	Brown	Brown
Block 2	$1/f$	Constant
	White	Constant
	Brown	Constant
Block 3	Constant	$1/f$
	Constant	White
	Constant	Brown

Figure 5.2: Jeong et al.'s (1998) experiment

Music Psychology

In music psychology, Vitz (1964) was the first to investigate the relationship between the information rate of tone sequences (in bits per second) and human ratings of pleasantness. In this experiment, Vitz (1964) expected an inverted-U relationship but instead found that the mean human rating is a monotonically increasing function of information rate (*cf.* Figure 5.1). Vitz (1966) then tried a more nuanced criterion for stimulus variation, where stimuli were generated with six predefined levels of randomness involved (called the "magnitude of stimulus variation"). Vitz then obtained a Wundt curve by plotting subjective ratings against magnitude of stimulus variation. Using ecologically more valid stimuli (four original piano compositions in increasing complexity), Heyduk (1975) also obtained this inverted-U curve by plotting mean liking ratings against compositional complexity.

North and Hargreaves (1995) shifted the focus of research from complexity to familiarity. Using 60 excerpts of popular music, they not only obtained an inverted-U curve with subjective complexity, but also a monotonically increasing function with subjective familiarity.

Tan et al. (2006) studied the effect of repeated hearings on liking. They used two types of stimuli: "intact" compositions consisted of complete piano pieces, whereas "patchwork" compositions consisted of excerpts from three different piano works patched together into one. Their results are shown in Figure 5.3. Although they did not mention the double inverted-U curve and instead chose to explain their findings in Berlyne's terms, it can be seen that the "patchwork" compositions induced a monotonically increasing curve whereas the "intact" compositions induced a curvilinear trend suggestive of a double inverted-U relationship (see Figure 5.1). Unfortunately, there are not enough points on the hearing axis, so the curvilinear relationship has only weak support.
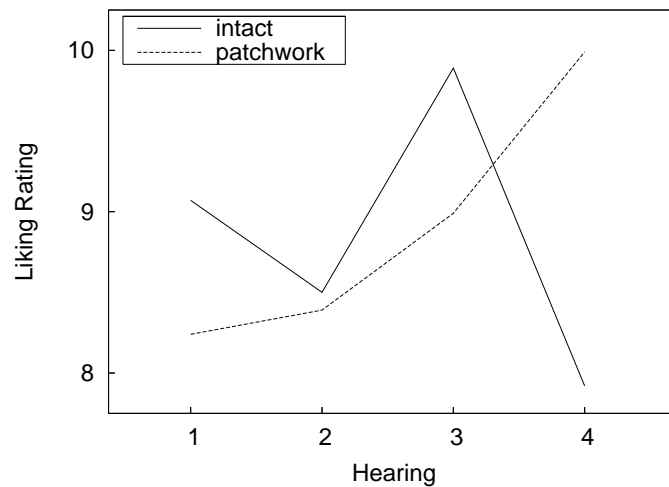
Figure 5.3: Liking versus exposure (after Tan et al., 2006)

Historiometry

Although not strictly experimental, *historiometry* is a science where "hypotheses about human behaviour are tested by applying quantitative analyses to data concerning historical individuals" (Simonton, 1997, p. 107). Simonton advocated the application of historiometry to the analysis of musical creativity (which includes the product, person, and period aspects). Here I will only focus on analyses of the product aspect, which includes composer identification, quantification of melodic originality, and quantification of aesthetic success (Simonton, 1997); of particular relevance to this thesis is melodic originality and aesthetic success.

Simonton (1997) quantified melodic originality by first tabulating the frequencies of two-note transitions in 15,618 classical themes (truncated to the first six notes of each theme). He then calculated the improbability of each theme from the probabilities of its constituent two-note transitions. Simonton (1997) called this improbability the *repertoire melodic originality*. He also quantified aesthetic success of a composition by its number of appearances in "catalogues of recorded performances, music appreciation textbooks, student scores, concert and record-buying guides, thematic dictionaries, anthologies of great music, and music histories, dictionaries, and encyclopedias" (Simonton, 1997). Simonton found a link between melodic originality and aesthetic success: "the popularity of a composition is an inverted backwards-*J* function of originality" and he noted that this inverted-*J* shape is remarkably similar to the famous Wundt curve (Simonton, 1997).

5.2.5 Other Mathematical Forms

While not previously connected to aesthetics, the following complexity measures will be refered to in this chapter and therefore are summarised below.

For a system with N states associated with the probabilities $\{p_1, p_2, \ldots, p_N\}$, López-Ruiz et al. (1995) defined the disequilibrium in the system as

$$D \;=\; \sum_{i=1}^{N}(p_i - 1/N)^2.$$

The López-Ruiz complexity of this system is then defined as

$$C \;=\; HD,$$

where H denotes the Shannon entropy. A normalised version is also given by López-Ruiz et al.:

$$\overline{C} \;=\; \overline{H}D,$$

where $\overline{H}$ is the normalised Shannon entropy $\overline{H} = \sum_{i=1}^{N} p_i \log p_i / \log N$. For $N = 2$ with $p_1 = x$ and $p_2 = 1 - x$, López-Ruiz et al. showed that $\overline{C}(\overline{H})$ is an inverted-U function by plotting $\overline{C}$ against $\overline{H}$.

Shiner et al. (1999) proposed a family of complexity measures that takes both order and disorder into account,

$$\Gamma_{\alpha\beta} \;=\; \Delta^\alpha \Omega^\beta = \Delta^\alpha (1 - \Delta)^\beta = \Omega^\alpha (1 - \Omega)^\beta,$$

where the disorder is defined by $\Delta = S/S_{\max}$ and the order is defined by $\Omega = 1 - \Delta$. Here S denotes the Shannon entropy. With different combinations of α and β, the complexity curve can be a monotonically increasing, an inverted-U, or a monotonically decreasing function of disorder (Shiner et al., 1999). Unnormalised versions are also given by Shiner et al. (1999):

$$G_{\alpha\beta} \;=\; \begin{cases} S^\alpha \Omega^\beta = S^\alpha (1 - S/S_{\max})^\beta \\ \Delta^\alpha (S_{\max} - S)^\beta = (S/S_{\max})^\alpha (S_{\max} - S)^\beta \\ S^\alpha (S_{\max} - S)^\beta. \end{cases}$$

This measure is said to be a generalisation of the López-Ruiz complexity (Shiner et al., 1999).

5.3 Three Competing Models of Musical Fitness

In this section, I present my new models of musical fitness. Following Eysenck (1942), I define *musical fitness* as "unity in diversity between music and listener", with three competing mathematical realisations. Conceptually, unity represents understandability or predictability, while diversity represents novelty or unpredictability. Unity in diversity implies striking a balance between the two.

But first, note that Eysenck's objective definition of complexity does not necessarily hold in music. Indeed, twelve-tone music is less complex for trained ears than for the untrained. A single formula for all participants would be fairly hard to justify. Therefore, I instead assume that both unity and diversity are subject-dependent and that subject dependency is mostly a function of memory (ignoring the emotional aspects). This is motivated by the everyday experience that complex pieces, once learned, becomes less complex from a subjective point of view. This highlights the importance of familiarity in my model, which I call the "to each their own" principle, where both unity and diversity are subjective functions of the listener's brain. This principle is supported by neuroscientific evidence, where "complex music produces complex brain activity in complex people, simple music excites simple brain activity in simple people [*sic*]" (see Subsection 5.2.4).

5.3.1 Unity in Diversity Between Music and Listener

I define unity as $I(x;y)$, diversity as $H(x|y)$, and familiarity as normalised unity, $F(x,y) = \frac{I(x;y)}{H(x)} = 1 - \frac{H(x|y)}{H(x)}$, where x is the music (information source), y is the listener (destination), and $I(x;y)$ and $H(x|y)$ are my cognitive information measures (defined in Chapter 3). The listener is represented by a corpus of music (see below). It is clear from the definition of $I(x;y)$ that $F(x,y) \in [0,1]$. $F(x,y)$ can be seen as the proportion of information in x accounted for by y. In addition, I define a forced-choice binary variable "I know this piece" taking the probabilities $P(\text{"I know this piece"}=\text{True}) = F(x,y)$ and $P(\text{"I know this piece"}=\text{False}) = 1 - F(x,y)$. In other words, I model the probability P of a forced choice self-reported "I know this piece" as the familiarity of the piece to the listener—the more familiar the piece is, the higher the probability. Now, I propose three alternative hypotheses for the relationship between musical fitness and my cognitive information measures:

H1: Fitness equals unity times diversity (after Eysenck, 1942):

$$
\begin{aligned}
\text{unity} \times \text{diversity} &= I(x;y) \times H(x|y) \\
&= H(x)F(x,y) \times H(x)[1 - F(x,y)]
\end{aligned}
$$

$$= H(x)^2 F(x,y)[1 - F(x,y)]$$

See Figure 5.4. With x fixed H1 is the Wundt curve and with y fixed it is a monotonically increasing quadratic curve. This supports both the inverted-U and monotonically increasing interpretations (*cf.* Figures 5.1b and 5.1a). This measure is mathematically equivalent to Shiner et al.'s third G_{11} measure (see Section 5.2.5) with $S = I(x;y)$ and $S_{\max} = H(x)$.
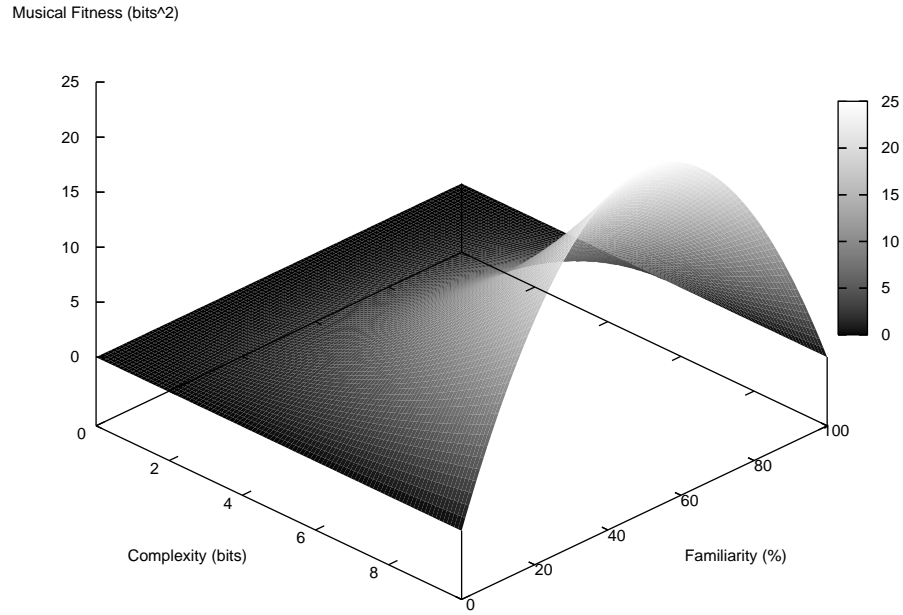
Figure 5.4: Unity times diversity versus familiarity (H1)

H2: Fitness equals unity divided by diversity (after Birkhoff, 1933):

$$\frac{\text{unity}}{\text{diversity}} = \frac{I(x;y)}{H(x|y)}$$
$$= \frac{F(x,y)}{1 - F(x,y)}$$

See Figure 5.5. This supports the monotonically increasing interpretation (*cf.* Figure 5.1b). An alternative (but mathematically equivalent) formulation is that musical fitness is defined as the odds[2] on the truth of "I know this piece". This measure is mathematically equivalent to Shiner et al.'s $\Gamma_{1,-1}$ measure with $\Omega = F(x,y)$.

---

[2]The statistical odds on a proposition with probability p is $\frac{p}{1-p}$.
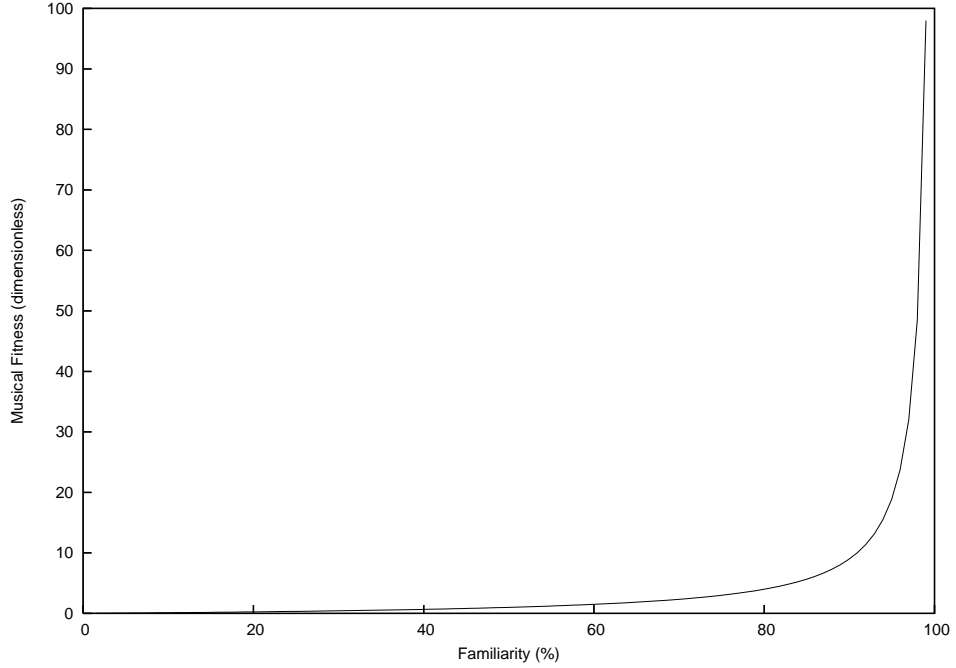
Figure 5.5: Unity divided by diversity versus familiarity (H2)

H3: Fitness is the normalised López-Ruiz complexity (see Section 5.2.5) of the "I know this piece" variable:

$$
\begin{aligned}
C(P) &= H(P)D(P) \\
&= -\{F(x,y)\log_2 F(x,y) + [1 - F(x,y)]\log_2[1 - F(x,y)]\} \times 2\left[F(x,y) - \frac{1}{2}\right]^2 \\
&= -\{F(x,y)\log_2 \frac{F(x,y)}{1 - F(x,y)} + \log_2[1 - F(x,y)]\} \times 2\left[F(x,y) - \frac{1}{2}\right]^2
\end{aligned}
$$

See Figure 5.6. This supports the double inverted-U interpretation (*cf.* Figure 5.1c). Furthermore, $C(P)$ almost fits Moles's (1968) model of artistic value (reviewed in Section 5.2.1), if we make the additional assumption that musical originality is $1 - F(x,y)$ and the listener's capacity is $\frac{1}{2}$:

$$
\begin{aligned}
C(P) &= -\{F(x,y)\log_2 F(x,y) + [1 - F(x,y)]\log_2[1 - F(x,y)]\} \times 2\left[F(x,y) - \frac{1}{2}\right]^2 \\
&= -\left\{\left[\frac{1}{2} + \left|1 - F(x,y) - \frac{1}{2}\right|\right]\log_2\left[\frac{1}{2} + \left|1 - F(x,y) - \frac{1}{2}\right|\right] \right. \\
&\quad \left. + \left[\frac{1}{2} - \left|1 - F(x,y) - \frac{1}{2}\right|\right]\log_2\left[\frac{1}{2} - \left|1 - F(x,y) - \frac{1}{2}\right|\right]\right\} \times 2\left[1 - F(x,y) - \frac{1}{2}\right]^2
\end{aligned}
$$

$$= \quad f\left(\left|1 - F(x,y) - \frac{1}{2}\right|\right).$$

The reader can verify that the expression on the second line is equivalent to the one on the first line, by considering seperately the cases $F(x,y) > \frac{1}{2}$ and $F(x,y) \leq \frac{1}{2}$. I write "almost fit" here because $f(\cdot)$ is not a monotonically decreasing function as Moles had originally specified.
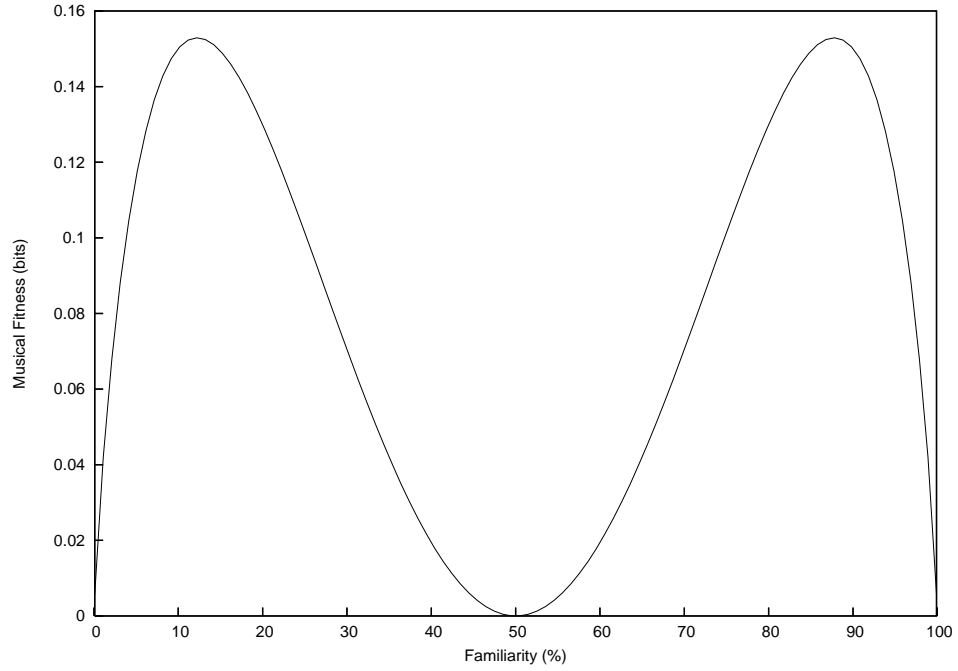
Figure 5.6: López-Ruiz complexity versus familiarity (H3)

5.3.2 A Simple Model of Human Listeners

My subjective measures require a model of the listener. For simplicity, in this chapter I will use eighty-eight songs (all songs beginning with K0xxx and K1xxx) from the Essen Kinder collection (Schaffrath and Dahlig, 2000; Schaffrath, 1997), and concatenate them together to form a single listener set.

5.3.3 Model Selection with Three Experiments

In the experiments below, I will correlate the human data in Vitz (1966), Heyduk (1975) and Jeong et al. (1998) with my model predictions. The best-fitting fitness measure is then selected using meta-gMDL$_+$ (see Chapter 3).

Correlation with Vitz's (1966) Experiment

A program was written to generate 64 pieces each of six levels of randomness (a total of 384 pieces). The average fitness for each level is calculated and correlated with Vitz's human data. Results are shown in Figure 5.7. The correlations are statistically significant for all three measures ($p < .05$).

Measure	r	df	p	gMDL$_+$
H1	0.84667	4	0.0170	-0.724
H2	0.94361	4	0.0023	-2.990
H3	0.86956	4	0.0120	-1.070

Figure 5.7: Correlation with Vitz's (1966) data

Correlation with Heyduk's (1975) Experiment

The fitness for each of Heyduk's four pieces is calculated and correlated with Heyduk's human data. Results are shown in Figure 5.8. The correlations are statistically significant for H3 only ($p < .05$). Note that this is the only experiment with stimuli that qualify as western tonal music.

Measure	r	df	p	gMDL$_+$
H1	0.74572	2	0.130	0.424
H2	0.74938	2	0.130	0.410
H3	0.93187	2	0.034	-1.180

Figure 5.8: Correlation with Heyduk's (1975) data

Correlation with Jeong et al.'s (1998) Experiment

Here I will demonstrate that there is a correlation between my model predictions and the negation[3] of Jeong's EEG results. I perform 128×3 (white, brown, $1/f$) runs of random music generation (monophonic, lasting thirty-two seconds at sixty beats per minute). For each type of music, I calculate the average musical fitness of all 128 pieces. In order to avoid the potential confounding factor of biased random numbers, a sophisticated scheme known as "Dynamic Creation of Mersenne Twisters" (Matsumoto and Nishimura, 2000) is used. Then I correlate the mean musical fitness with the negation of Jeong's EEG dimensions. Results are shown in Figure 5.9. The correlations are statistically significant for H2 only ($p < .05$).

---

[3]Negation because aesthetic superiority is measured by the lowering of brain activity (Jeong et al., 1998).

Measure	r	df	p	gMDL$_+$
H1	0.845900	1	0.180	0.0207
H2	0.996570	1	0.026	-3.5400
H3	-0.99235	1	0.960	-0.0589

Figure 5.9: Correlation with Jeong et al.'s (1998) data

Discussion

The sum of gMDL$_+$ code lengths are shown in Figure 5.10. H2 has the minimal total description length and is thus selected. However, note that if the stochastic experiments are removed, then H3 will be the winner instead (see Figure 5.8 only). Because of this, I have decided to check for both hypotheses in the next section.

Measure	$\sum$gMDL$_+$
H1	-0.28
H2	-6.10
H3	-2.30

Figure 5.10: Meta-analysis of gMDL$_+$ code lengths for fitness

5.3.4 Validation of H2 and H3

Using my fitness model I now extend Katz's degradation experiment on a larger scale. The original Katz hypothesis states that the number of mutations is roughly inversely proportional to fitness (in other words, H2), but I will also test for H3 below.

Method

Participants Participants were recruited from the University of Sheffield through a university-wide volunteers' e-mail list. The eligibility criteria are that they be aged 18–64, have at least one healthy ear, and not have musicogenic epilepsy. They were paid eight pounds sterling per hour for their participation. Ethical approval was granted by the Department of Psychology Ethics Sub-Committee at the University of Sheffield. Thirty participants signed up for this experiment with informed consent (10 males and 20 females). The mean age was 36 years (SD = 10) and the mean years of musical training was 4.0 years (SD = 6.0).

Materials The stimuli consisted of K2445 and K3027 from the Essen Kinder collection, mutated into 178 mini-pieces. For each song, degraded melodies are created "by re-

placing a set number of randomly chosen notes with a random note in the pitch range between the highest and lowest note in the song, conforming to the key of the melody" (Katz, 1994, p. 220). The requirement that the mutations should conform to the key will be reversed to *enhance* Katz's position and to *diminish* mine—at least intuitively, atonal mutations should be worse than tonal mutations. Twenty mutants will be created for each set number of mutations (so as to improve the quality of the statistical sampling). The set number of mutations is set to 5, 10, 15 and 20 mutations.

Procedure An information sheet and consent form were given to the participants (see Appendix A). Verbal instructions were also given. Once the consent form was signed, the instructions in Figure 5.11 were shown to the participants on a computer screen on which they would begin the experiment by pressing a button.
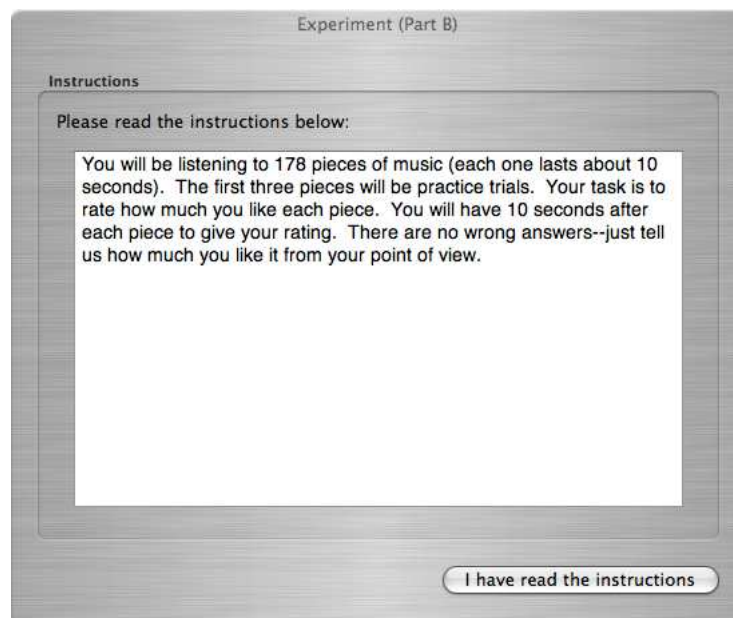
Figure 5.11: On-screen instructions

The order of presentation was randomised to reduce priming effects. The participant rated how much they like each song in a 1–9 scale (see Figure 5.12), with a ten-second inter-stimulus silence interval. The rating buttons were greyed-out until the rating phase, in order to prevent participants from pressing the buttons prematurely. The procedure was repeated until all songs (mutants as well as originals) were presented. The first three pieces were practice trials and the results were not analysed. At the end of this experiment, the results were correlated to model predictions. For the playback of MIDI files, I had chosen the same sound font used in the previous chapter.

Figure 5.12: Listening to the first piece

Data Analysis The unit of analysis is the mean of all participants' rating on each piece, and the practice trials were treated as missing data which were replaced by the mean of other participants' ratings on the same piece (as in Chapter 4). Finally, for inter-rater reliability, Cronbach's (1951) α was reported.

Results

The Cronbach's α is 0.857, meaning that the mean of the human ratings are consistent. Results are shown in Figure 5.13. The correlations are statistically significant for both H2 and H3 ($p < .05$).

Measure	r	df	p
H2	0.66223	176	3.9×10^{-24}
H3	0.66943	176	8.5×10^{-25}

Figure 5.13: Musical fitness results

Discussion

Both H2 and H3 have statistically significant correlations with human data ($p < .05$). H2 accounted for 44% of the variance in the data, whereas H3 accounted for 45% of the

variance. In order to find out which model is better, quadratic fits of the human data and the model predictions are plotted in Figure 5.14. From the figures, it can be seen that H3 fits the shape of the human data better. In other words, human ratings follow the middle portion of the double inverted-*U* curve (*cf.* Figure 5.1c), at least in this experiment.

(a) Human (K2445)

(b) Human (K3027)

(c) H2 (K2445)
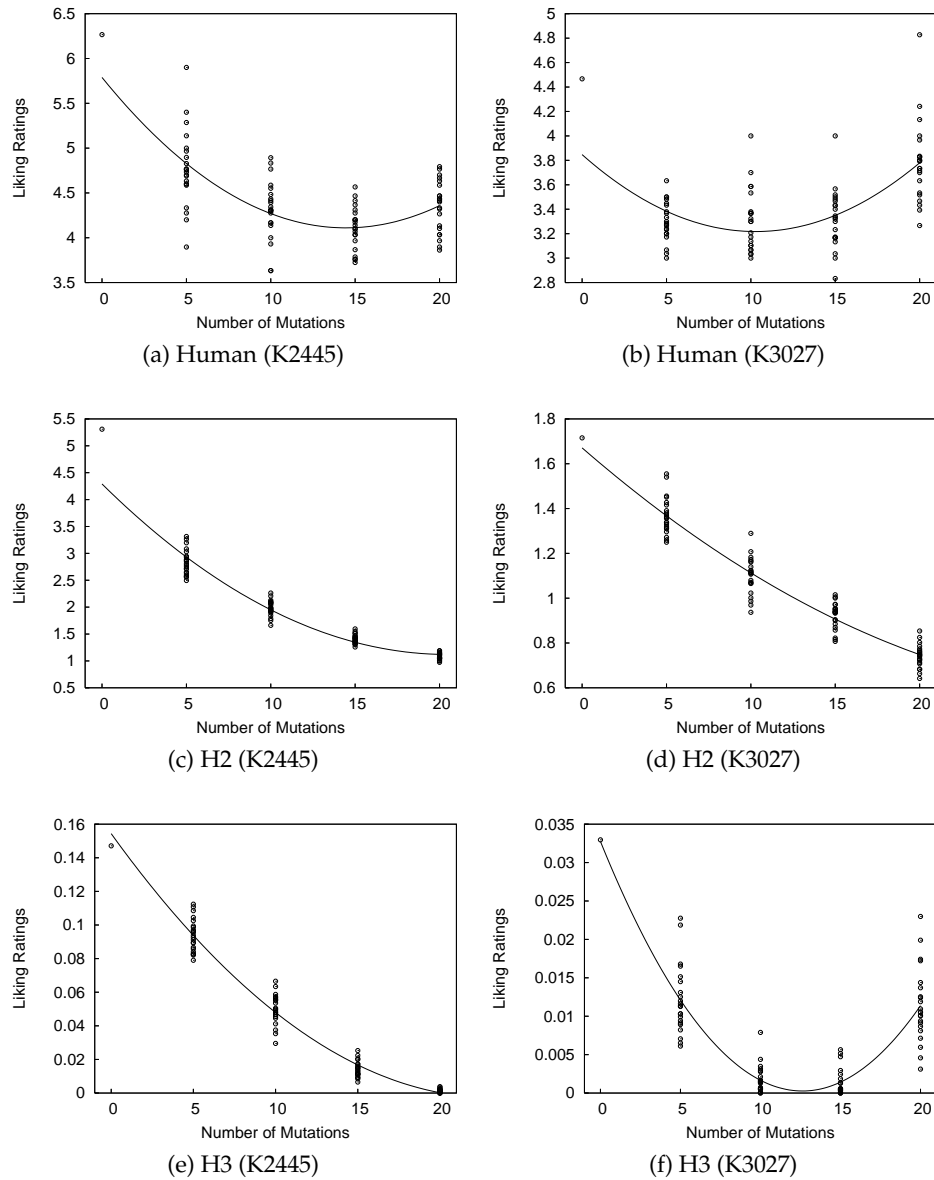
(d) H2 (K3027)

(e) H3 (K2445)

(f) H3 (K3027)

Figure 5.14: Quadratic curve fits (by song)

5.4 General Discussion

I have proposed three computational models of musical fitness and have shown that H2 is better in general but H3 is the best in a western tonal music context. A possible explanation for this discrepancy can be found in the data of Tan et al. (2006), where "intact" compositions generated a double inverted-*U* curve, whereas "patchwork" compositions generated a monotonic increasing curve (see Section 5.2.3). The results of this chapter challenged the perceived superiority of the Wundt curve (Berlyne, 1974; Walker, 1973), a model which performed the worst in the model selection experiments.

5.5 Concluding Remarks and Future Work

More work is needed to create new formulae of fitness and to experimentally validate them. In fact, I will go further and say that we need a lot more data before we can properly evaluate the results in this chapter.

With regards to computational creativity (Application Layer), the fitness measure could be applied to creative systems that compose music, and the value of the creative product could be evaluated by a panel of expert human judges, perhaps using the consensual assessment technique (Amabile, 1982). Also, given an arbitrary creative system with mathematical definitions of complexity and familiarity in non-information-theoretic terms, there is a possibility to reverse engineer the Information Layer (loosely speaking). Where this is possible, a hard test of the generalisability and independence of the layers would be to reverse engineer the Information Layer and use it to predict higher level properties at the Psychology Layer, such as fitness.

Chapter 6

Discussion

6.1 Introduction

In this thesis, I have offered a synthetic, multi-layer account of music cognition based on both information theory and Dawkins' memetic theory of culture (specifically, that musical culture is made up of atomic units of evolutionary information transmission analogous to the gene). My starting point is Jan's (2000b) premise that musical culture is memetic (which enables me to interpret copying-fidelity and aesthetic fitness of music in Darwinian terms).

Of course, I acknowledge that memetics is not a silver bullet for answering all scientific questions about music. Indeed, even the validity of the gene-meme analogy is in dispute (see next section). Nonetheless, it is my hope that at least some anti-memeticists would reconsider memetics as a useful discipline after seeing the preliminary evidence provided in this thesis.

While previous research in memetics has suffered from vagueness, unfalsifiability, and incompleteness, this thesis is the first to start with a falsifiable definition of a meme and try to disprove it with psychological experiments.[1] In so doing, the vagueness and unfalsifiability in traditional computational memetics is circumvented and replaced by the vigour of psychology. So far, my research has shown that the memetic strategy is sound, and the modelling results are encouraging.

---

[1]Pocklington and Best (1997) was probably the only paper to start with a falsifiable definition of a meme; unfortunately, the authors do not follow it up with psychological experiments—they just say, effectively, that such and such are the memes extracted from a text corpus using their (falsifiable) meme extraction algorithm, and that those memes exhibited certain statistical properties which suggests that they are likely to be units of (Darwinian) selection. But they stopped short of psychological validation.

6.2   A Short Reply to Bruce Edmonds

Bruce Edmonds, publisher (and one of the founding editors) of the *Journal of Memetics—Evolutionary Models of Information Transmission*, famously pronounced the death of memetics, before he teminated the publication of the journal (Edmonds, 2005). Because of his authority in memetics, and because my thesis does not address his challenges, a short reply is in order. Let me give a bit of history first.

Edmonds (2002) posed three challenges to the memetics community. He claimed that if these challenges are not met, memetics will not survive. Challenge 1 is "a conclusive case study" where the meme "needs to be something physical and not in the mind". Challenge 2 is "a theory for when memetic models are appropriate" where the theory must not be based on "unobtainable information" such as "the composition of mental states". Challenge 3 is "a simulation of the emergence of a memetic process" without a "built-in" imitation process. Three years later, Edmonds (2005) could not find any substansive answers to his challenges, so he went on to make his famous claims based on the absence of evidence. I will refute his claims, point by point, below (Edmonds, 2005):

1. "I do claim that the failure to answer those challenges was indicative of the poverty of the memetics project resulting in a lack of demonstrable progress which, in turn, has meant that it has failed to interest other academics": on the contrary, I believe that the failure to answer was indicative of the poverty of those challenges. Edmonds' first two challenges assumed that memes are physical and not in the mind, and that composition of mental states are unobtainable information and should not be relied upon in a memetic theory. But psychologists and anthropologists disagree with this stance. Plotkin (2000) and Henrich and Boyd (2002) advocated putting cognition into a theory of cultural evolution, and they all said that memes are in the brain. And so do I in this thesis. As for Edmonds' third challenge, I believe that it was misplaced—the lack of evolvability of memetic evolvability is not a necessary condition for the demise of memetics; it could very well be that our capacity for memes is an evolutionary accident. Remember that life on earth is highly improbable. Why must the ability to imitate be treated differently? If memetic capacity is easily evolvable, why is *Homo sapiens sapiens* the lone subspecies on earth that has produced the likes of Bach and Beethoven?

2. "Here I distinguish between [...] the 'broad' and the 'narrow' approaches to memetics [...] The later, narrow, sense involves a closer analogy between genes and memes—not necessarily 100% direct, but sufficiently direct so as to justify the epithet of 'memetics'. What has failed is the narrow approach": here the boundary

between "broad" and "narrow" is too vague and subjective. What is sufficiently direct to me might not be so to Edmonds. Since I took the stance that memes are in our minds, memes would be very different from genes in terms of longevity, fecundity and copying fidelity. Would a 50% copying fidelity count as not direct enough? How about 25%? Is there a peer-reviewed, internationally accepted criterion of how close is close enough? I am not aware of one. Also note that Henrich and Boyd (2002) have provided simulation evidence that even individuals with low copying fidelity could still give rise to a culture with high copying fidelity (at the population level).

3. "I claim that the underlying reason memetics has failed is that it has not provided any extra explanatory or predictive power beyond that available without the gene-meme analogy [...] The ability to think of some phenomena in a particular way (or describe it using a certain framework), does not mean that the phenomena has those properties in any significant sense": the same can be said about most terms in quantum physics or even in genetics. Just what exactly is the definition of a gene? Answers differ, even within the field of genetics (Hull, 2000)! Yet this does not stop the term "gene" from being useful. I can imagine that if Edmonds reads this thesis, he might claim that all that he saw is an information theory of musical memory (no memes here). But if Edmonds wants to disprove the usefulness of the "meme" framework at this early stage of memetics, he needs to work a lot harder to prove that the memetic framework is redundant and has already been subsumed completely by other disciplines. Then and only then, could he invoke Occam's razor.

4. "[T]here is a successful community of social simulators who study, among other things, evolutionary models of information transmission. Similarly there is work in computer science, applying evolutionary ideas to computational processes and work in theoretical biology studying non-genetic evolutionary processes. Thus this wider work will continue as subsets of other projects, but not under the discredited label of memetics": those "wider works" are clearly successful examples of memetics research. Edmonds is apparently discrediting memetics by confining memetics to an unnecessarily small niche, thus excluding all the successful examples, then making his claims by appealing to the absence of evidence. There is no evidence.

5. "[T]he meme-gene analogy [...] has been a short-lived fad whose effect has been to obscure more than it has been to enlighten. I am afraid that memetics, as an identifiable discipline, will not be widely missed": Edmonds' claims here are both

unsubstantiated and premature. The scientific study of memes is still in its infancy, and memetics is still awaiting its Watson and Crick. The science of memetics may very well turn out to be a fad in the future, but currently we have no concrete evidence of that.

With the "not in the mind" ban lifted, I believe that Edmonds' first two challenges still need to be met. Specifically, I believe that Edmonds' desire for a theory of applicability (Challenge 2) is well-intentioned and indeed necessary, but I think it is premature to expect such a theory to surface anytime soon, as we are still determining the boundaries of applicability of the memetic theory and it will be a long journey.

6.3   Implications and Prospects of Research

This thesis has provided a research programme of computational memetics (not necessarily limited to music) that is based on a multi-layer, information-theoretic and cognitive approach to memetic modelling.

This thesis has contributed significantly to the computational aspects of music memetics. Recall that Best's (2001) original taxonomy of computational memetics consists of simulation, computational theory, and population memetics. Best (2001) noted, in relation to computational theory, that "[w]hile we do not, to my knowledge, have as strong a result specific to memetics, this line of computational memetic inquiry is quite valuable and should pay off handsomely in tomorrow's research outputs." This thesis can be seen as a reply to this call, unifying computational theories of memetics, similarity, and aesthetics under my cognitive information framework (Chapter 3), supplemented by solid experimental work (Chapters 4 and 5).

This research has indirectly contributed to other areas of research, which may be potentially significant (detailed research on any of these would however remain future work):

1. In artificial intelligence, evolutionary methods for algorithmic composition have a long history, but their cognitive plausibility has been questioned (Papadopoulos and Wiggins, 1999). By asking direct scientific questions pertaining to the evolutionary mechanisms of music transmission, this research will bring us one step closer to a gold standard for knowledge representation and fitness criteria in evolutionary music composition. Indeed, aesthetic selection and artificial creativity are two of the "open problems in evolutionary music and art" (McCormack, 2005), deemed to be grand challenges worthy of sustained future research.

2. In biomusicology, this research contributes a novel approach to cognitive musicology and experimental aesthetics. To my best knowledge, this is the first thesis to combine evolutionary, cognitive, affective and computational modelling in musicology under one unifying framework.

3. In music education, my complexity measure might be useful for the construction of an aural training curriculum that gets harder progessively, to the ultimate goal where students could transcribe free jazz by ear with accuracy. The increment of difficulty can now be done algorithmically with my complexity measure, rather than by guesswork.

4. In music information retrieval, my similarity measure might have immediate applications to musical databases and archives, for example in searching and querying.

5. My similarity measure might also be used to create a phylogenetic tree, with implications for music history and musicology, for example author identification (but look at the end of Chapter 4 about horizontal meme transfer).

6. In cultural ecology, one could look at the geographical distribution of songs as in Aarden and Huron (2001). My thesis adds a similarity measure and a fitness measure to Aarden's toolbox, so that we can now ask for the geographical distribution of songs that matches fuzzily with a polyphonic query (utilising my similarity measure), or the distribution of songs that would sound best (fitness exceeds a certain threshold) to those of us encultured in a certain musical tradition.

7. Also in cultural ecology, one could look at the biodiversity of memes, for example using Faith's phylogenetic diversity (PD) measure (Faith, 1992). PD can be calculated simply by constructing a phylogenetic tree of the pieces using my similarity measure, then add up all the branch lengths (Faith, 1992). We can apply this number in different ways, such as:

   • Cultural environmental protection: the negation of PD could be used as a measure of cultural hegemony. If it is higher than a certain threshold, then it might be a good time for intervention.

   • Concert programming: consider that a soloist has decided to do a recital on a single composer, with the additional assumption that this composer only composes works of the highest quality. This gives us unity, but what about diversity? By using PD we will be able to, within the confines of the soloist's active repertoire, select out a subset of that repertoire that maximises the cultural diversity.

6.4 Concluding Remarks

This research has provided evidence for the soundness of the computational memetic approach. It is my hope that the reader would, by now, be convinced that this approach is fruitful and that it opens up a great many possibilities for future work.

Chapter 7

Conclusion

7.1 Introduction

In this chapter I summarise my contributions to knowledge.

7.2 Summary of Methodologies

The aim of this thesis is to propose and examine a novel, multi-level cognitive information theory. New methodologies are proposed to fulfill this aim. The following is a summary.

7.2.1 Multiple-Layer Approach

My research strategy is to divide the modelling effort into four layers, proposed in Chapter 1 (see p. 13). The *Data Layer* corresponds to perceptual inputs, the *Information Layer* corresponds to my information measures, the *Psychology Layer* includes similarity and aesthetic fitness, and the *Application Layer* includes creative systems and cultural ecology (not explored in the previous chapters). With the exception of the Data Layer (which has no dependencies), all higher-level layers depend on the lower-level layers.

With this scheme, information becomes the ultimate foundation underlying all explanations of musical behaviours above the Data Layer level. The advantages of multi-layer modelling is that we can potentially reuse the same *Application* and *Psychology Layers* to describe phenomena in a different domain while changing only the *Information* and *Data Layers*.

The problem with the multi-layer approach is that it is not clear whether different areas of the brain share the same cognitive processing mechanisms. It might be premature

to allow for a theory with a plug-and-play information-theoretic substrate (music, chess, *etc.*). It is outside the scope of this thesis to talk about the memetics of chess (de Sousa, 2002), but the theory presented here will be heavily bolstered if the very same formulae for similarity and fitness (brilliancy) would also work for chess moves without modifications.

In the Data layer, I simply assumed a postprocessed MIDI file containing onset, pitch, and metrical position data. This is operationally useful, but it lacks important features such as loudness, so this representation would be inadequate for tasks requiring loudness discrimination (or other information that it does not have).

The main focus of this thesis is on the Information and Psychology layers. For these two layers, I proposed competing models and selected the best fitting models based on published experimental data. In addition, for the Psychology layer, I designed, conducted and analysed new experiments in order to validate my models.

First, for the Information layer, I started from the well-established modal memory model and propose, in precise computational terms, a putative information theory of memory. Three competing models each outputted a single number denoting the cognitive complexity of a piece of music, which could be interpreted as the theoretical number of bits required to store the piece into memory. The model best fitting human data was selected, and higher-level psychological theories were built on top of this model. In this thesis, I investigated two such theories at the Psychology layer, musical similarity and musical value.

In the similarity chapter, I proposed three competing models of musical similarity based on the information measure described above, and selected the best one using a model-fitting strategy. In the same chapter, I presented a new experiment to check if the selected model would break. The experiment used polyphonic piano music as stimuli, with 30 participants each rating the similarity of 75 pairs of musical fragments. I also proposed an overarching similarity framework that subsumes a large number of existing similarity measures both set-theoretic and information-theoretic.

In the fitness chapter, I again proposed three competing models and have the best fitting model selected. Then a new experiment was designed to check the predictive power of this model, with 30 participants each rating how much they like each of the 175 musical fragments based on two mutated Essen folksongs.

7.2.2 Minimum Description Length Principle

The minimum description length (MDL) principle (see Chapter 3, p. 36) is used throughout this thesis, both in model selection and in the models themselves. The MDL principle

states that a model that minimises the description length of a dataset is the best fitting model (Rissanen, 1978; Barron and Rissanen, 1998). This is also known as the Occam's razor or the principle of parsimony (Barron and Rissanen, 1998). MDL is better than traditional goodness-of-fit measures because MDL penalises overfitting (Pitt and Myung, 2002; Grünwald, 2005). MDL might suffer from overgeneralisation (imprecise models) but that could be a good thing if cross-domain modelling of universals is the ultimate goal—we do not want domain-specific "noise" to get in our way. Another way to put this is that it is a trade-off between domain-specific and universal modelling and we prefer a universal theory.

The meta-gMDL$_+$ selection criterion (proposed in Chapter 3) is used throughout this thesis. In meta-gMDL$_+$, the best fitting model is taken to be the one with the lowest value of gMDL$_+$.

Not only is the model fitting related to MDL, but so are the models themselves. My proposed information measures are measures of compressibility so they are related to the MDL principle (in the sense of Ernst Mach who "proposed that the goal of perception [...] is to provide the most economical explanation of sensory data", Chater, 2005), or in other words, perception equals compression.

I believe that this thesis is the first piece of work that adheres to the MDL principle at two different levels simultaneously.

7.3 Summary of Results

For both the Information and Psychology layers (p. 13), competing models were proposed and the best fitting models were selected based on published experimental data. In addition, for the Psychology layer, two experiments were designed, conducted and analysed in order to validate the proposed models. Here is a summary of the results.

7.3.1 Cognitive Information

In Chapter 3, I proposed three competing models of musical information, and selected the best fitting model based on Conley's (1981), Shmulevich and Povel's (2000) and Heyduk's (1975) human data. The best fitting model ($\sum_{1,2,4}$gMDL$_+$ = -1.4, $\sum_{1,3,4}$gMDL$_+$ = -2.6) is the one based on the T-complexity measure (Titchener, 2000) after neural cancellation filtering (de Cheveigné, 1993). I have also proposed an alternative model of human listening (using conditional information rather than just information, on the grounds that listeners would unconsciously remember for example the theme when listening to a set of variations on the same theme). Using this listener model, the data fit is somehow im-

proved, but not by much. Note that Conley's stimuli (Beethoven's Eroica variations) are ecologically valid. While this model is rough (only accounting for 17% of the variance in Conley's Graduate data), it is a good enough starting point for my similarity and fitness measures.

7.3.2 Musical Similarity

In Chapter 4, I proposed three competing models of musical similarity based on various information distances adapted to use my information measures, and selected the best fitting model using Cambouropoulos's (2001), Eerola et al.'s (2001) and Eerola and Bregman's (2007) human data. The best fitting model H1 ($\sum$gMDL$_+$ = -24) is the one based on Kvålseth's (1987) information measure. A new experiment was carried out to validate the selected model. The experiment used polyphonic piano-roll music as stimuli, with 30 participants each rating the similarity of 75 pairs of musical fragments. The model fit is statistically significant, $r(76) = 0.57106$ ($p < .05$), accounting for 33% of the variance in the mean human ratings. Results also showed, *contra* Tversky (1977), that human data obey the metric axioms.

7.3.3 Musical Fitness

In Chapter 5, I again proposed three competing models and used model-fitting to select the best. The best fitting model H2 ($\sum$gMDL$_+$ = -6.1) was the one based on Birkhoff (1933), but if stochastic music was disallowed, then H3 ($\sum$gMDL$_+$ = -2.3), the model based on López-Ruiz et al. (1995), would be the best fit instead. A further experiment was carried out to validate these two models, with 30 participants each rating how much they like each of the 175 musical fragments based on two mutated Essen folksongs. Both models fit the data well: H2 accounted for 44% of the variance in the data, $r(176) = 0.66223$ ($p < .05$), whereas H3 accounted for 45% of the variance, $r(176) = 0.66943$ ($p < .05$). Quadratic curve fitting revealed that H3 fits the shape of experimental data better. The results of this chapter challenged the received view that the Wundt curve is the best model for experimental aesthetics.

7.4 Summary of Other Contributions

Firstly, this thesis contributed a method for the meta-analysis of correlational studies, based on a non-standard correlation model with non-negativity constraints. I expect this method to be applicable outside this thesis as well.

Secondly, this thesis proposed a working Data Layer model, namely the OPM (Onset, Pitch, Metrical Level) model. The bundling of Onset and Metrical Level is first done by Temperley and Sleator (1999) in their meter-finding algorithm. Therefore, this thesis may be seen as further corroborating evidence for Temperley and Sleator's representation.

Thirdly, it was known that the assumptions made by many information-theoretic analyses of music were untenable or at least hard to justify (Cohen, 1962; Sharpe, 1971). My model addresses one of these limitations: most languages are not describable by finite-state Markov sources such as *n*-gram models (Chomsky, 1956); while this is not sufficient evidence to falsify the information-theoretic approach to music *per se*, it shows that this approach (when applied only to smaller basic units such as single notes) is inadequate and insufficient (Sharpe, 1971). By using a compression-based model of information that looks at all possible pairs of notes, this limitation is circumvented.

Fourthly, the cognitive information results had shown that a cognitively constrained information theory is better than a generic compression algorithm like Cilibrasi et al. (2004). This confirms that one cannot ignore representation in music cognition research.

Fifthly, the overarching framework proposed in the similarity chapter (subsuming most set-theoretic, Shannon-based and compression-based similarity measures) might be useful in other contexts such as music information retrieval.

Sixthly, the bridge between information and musical fitness, while not new (Meyer, 1957; Moles, 1968), has been questioned (Cohen, 1962). Cohen maintained that information alone cannot account for the musical experience; the musical experience bit belongs to aesthetics. I believe that this assertion is too strong; it could very well be that an universal information measure would simultaneously be a good model of aesthetic fitness. We need more research of course, but at least the results of this thesis fail to refute my view.

Finally, the musical fitness results had shown that, despite its popularity, the inverted-*U* formulation of fitness did worst in the model selection phase. In other words, it does not seem to model human data well. Of course, given the poverty of empirical evidence, more evidence is needed.

Appendix A

Information Sheet and Consent Form

A.1 Information Sheet

- The expected duration of this experiment is two hours, split into two one-hour slots with a fifteen-minute break in between.

- You have the right to withdraw from this experiment at any time without incurring any penalty.

- The purpose of this experiment is to learn more about human judgements of similarity and liking in music, and to see whether a computational measure of musical complexity could account for the variances in the human judgements of similarity and liking.

- There are no known discomforts or risks involved in this experiment, except for people who suffer from musicogenic epilepsy or related neurological disorders. People with musicogenic epilepsy are likely to have epiletic fits when they listen to music. If you have this medical condition, you will not be allowed to participate in this experiment.

- Publications arising from this experiment will *not* contain personally identifiable information. Such information will be kept confidential and will be destroyed within six months of this experiment.

A.2 Consent Form

I have been informed about the nature and potential risks of this research. I declare that I do not suffer from musicogenic epilepsy or other neurological disorders related to music. I agree to participate voluntarily.

Signature ....

Date ....

Appendix B

Computability and Kolmogorov Complexity

B.1   Computability

The Church-Turing thesis (Church, 1936; Turing, 1936) states that a function is *computable* if there exists a non-halting program that computes it. An example of an uncomputable function is the *halting function*—given an arbitrary program, decide if it halts. Turing (1936) proved that the halting function is uncomputable.[1]

B.2   Kolmogorov Complexity

To calculate the Kolmogorov complexity $K(x)$ (see Section 2.3), one searches for the shortest program that generates x. This is computationally equivalent to running all possible programs and searching for the smallest one outputting x. Since there exists programs that will halt, by the Church-Turing thesis, $K(x)$ is uncomputable.

B.3   Super-Turing Computation

Copeland and Sylvan (1999) argued that the Church-Turing thesis (as stated above) is false. For them, the thesis only delineates what is computable by "orthodox computing devices" (p. 48) such as Turing machines. Copeland and Sylvan (1999) introduced the notion of computability by "heterodox computing devices" (p. 48), which could poten-

---

[1]Note that uncomputability is different from intractability; intractability means that the function cannot be computed in polynomial-time in the size of the input (Garey and Johnson, 1979).

tially compute the halting function (pp. 60–62). The first heterodox computing device is introduced by Weyl (1949):

> [...] there is no reason why a machine should not be capable of completing an infinite sequence of distinct acts of decision within a finite amount of time; say, by supplying the first result after 1/2 minute, the second after another 1/4 minute, the third 1/8 minute later than the second, etc. In this way it would be possible, provided the receptive power of the brain would function similarly, to achieve a traversal of all natural numbers and thereby a sure yes-or-no decision regarding any existential question about natural numbers! (p. 42)

There exists at least two more heterodox devices that are capable of computing the halting function (Copeland and Sylvan, 1999, pp. 60–62). Whilst these devices are outside the scope of my thesis[2], I will simply note that their existence would render Kolmogorov complexity computable. Nevertheless, given that the aforementioned devices are unrealisable with present-day technologies, I will disregard super-Turing computation and maintain that Kolmogorov complexity is uncomputable for the purpose of this thesis.

---

[2]Interested readers are referred to Copeland and Sylvan (1999) and the references therein for more details.

Glossary

aesthetics The study of beauty, both philosophical and empirical.

bit strings Vectors in $\{0,1\}^n$.

city block distance The city block distance between two vectors **x** and **y** is $\sum_i |x_i - y_i|$.

clustering Methods of grouping things together based on their closeness.

conditional probability The probability of one event occurring given that the other event also occurs.

connectionism Cognitive modelling using *neural networks*.

correlation matrix A matrix of correlations between all pairs of data.

diagonal matrix A matrix A in which $a_{ij} = 0$ whenever $i \neq j$.

EEG See *electroencephalograph*.

eigenvalues The eigenvalues of a complex matrix A are the solutions of $|A - \lambda I| = 0$.

electroencephalograph A device for measuring the electrical activities on the scalp.

excitatory neurons Neurons whose firing would cause other neurons to fire.

Hamming distance The Hamming distance between two *bit strings* is equivalent to their *city block distance*.

hippocampus Part of the brain primarily responsible for memory.

in vitro In glass. See also *in vivo*.

in vivo In the living body. See also *in vitro*.

inhibitory neurons Neurons whose firing would stop other neurons from firing.

joint probability The probability of two events occurring together.

linear independence A set of vectors is linearly independent if none of them can be written as a linear combination of the others.

magnetoencephalograph A device for measuring the magnetic field over the head.

MEG See *magnetoencephalograph*.

neural networks A network of simple processing units with a learning algorithm.

neuronal Of or related to neurons.

neuroscience The scientific study of *neuronal* systems in humans and other animals.

principal component analysis A method to reduce a large number of variables into a few statistically uncorrelated components. The components are sorted such that the first principal component has the largest variance. The principal component analysis can be computed using the *singular value decomposition*.

rank The rank is defined as the maximal number of *linearly independent* rows or columns in a matrix.

similarity measures A numerical measure telling you how close together two items are (roughly the opposite of distance measures).

singular values The singular values of a matrix A are the square roots of the *eigenvalues* of AA^*.

singular value decomposition A complex matrix A can be decomposed into $A = U\Sigma V^*$ where U and V are *unitary* and Σ is a *diagonal matrix* containing the *singular values* of A.

statistical independence Two events X and Y are statistically independent if their *joint probability* $p(X, Y)$ can be factorised into $p(X)p(Y)$.

time series A vector of numerical samples taken from discrete time points.

unitary A matrix A is unitary if $A^*A = AA^* = I$.

weighted Hamming distance The weighted Hamming distance between two *bit strings* $\mathbf{x}$ and $\mathbf{y}$ is $\sum_i w_i|x_i - y_i|$ where $\mathbf{w}$ represents the weights.

Bibliography

Aarden, B. and Huron, D. (2001). Mapping European folksong: geographical localization of musical features. *Computing in Musicology*, 12:169–183.

Amabile, T. M. (1982). Social psychology of creativity: a consensual assessment technique. *Journal of Personality and Social Psychology*, 43(5):997–1013.

Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: a proposed system and its control processes. In Spence, K. W. and Spence, J. T., editors, *The Psychology of Learning and Motivation, Vol. 2*, pages 89–195. Academic Press, New York.

Baronchelli, A., Caglioti, E., and Loreto, V. (2005). Artificial sequences and complexity measures. *Journal of Statistical Mechanics*, P04002.

Barron, A. and Rissanen, J. (1998). The minimal description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44:2743–2760.

Bateson, G. (1973). *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution and Epistemology*. Paladin, London.

Bennett, C. H., Gacs, P., Li, M., Vitányi, P. M. B., and Zurek, W. (1998). Information distance. *IEEE Transactions on Information Theory*, 44(4):1407–1423.

Berlyne, D. E. (1974). The new experimental aesthetics. In Berlyne, D. E., editor, *Studies in the New Experimental Aesthetics: Steps Toward an Objective Psychology of Aesthetic Appreciation*, pages 1–25. Hemisphere, Washington, DC.

Best, M. L. (2001). Towards computational memetics. *Journal of Memetics–Evolutionary Models of Information Transmission*, 4. [http://jom-emit.cfpm.org/2001/vol4/editorial.html].

Birbaumer, N., Lutzenberger, W., Rau, H., Braun, C., and Kress, G. M. (1996). Perception of music and dimensional complexity of brain activity. *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering*, 6(2):267–278.

Birkhoff, G. D. (1933). *Aesthetic Measure*. Harvard University Press, Cambridge, MA.

Blumenthal, L. M. (1953). *Theory and Applications of Distance Geometry*. Oxford University Press, London.

Boon, J. P. and Decroly, O. (1995). Dynamical systems theory for music dynamics. *Chaos*, 5(3):501–508.

Bouchon-Meunier, B., Rifqi, M., and Bothorel, S. (1996). Towards general measures of comparison of objects. *Fuzzy Sets and Systems*, 84:143–153.

Brown, S., Merker, B., and Wallin, N. L. (2000). An introduction to evolutionary musicology. In Wallin, N. L., Merker, B., and Brown, S., editors, *The Origins of Music*, pages 3–24. MIT Press, Cambridge, MA.

Cambouropoulos, E. (2001). Melodic cue abstraction, similarity and category formation: a computational approach. *Music Perception*, 18(3):347–370.

Castelfranchi, C. (2001). Towards a cognitive memetics: socio-cognitive mechanisms for memes selection and spreading. *Journal of Memetics–Evolutionary Models of Information Transmission*, 5. [http://jom-emit.cfpm.org/2001/vol5/castelfranchi_c.html].

Cazzanti, L. and Gupta, M. R. (2006). Information-theoretic and set-theoretic similarity. In Barg, A. and Yeung, R. W., editors, *Proceedings of the 2006 International Symposium on Information Theory*, pages 1836–1840. IEEE, Piscataway, NJ.

Chan, T.-S. T. and Wiggins, G. A. (2002). Computational memetics of music: memetic network of musical agents. In Britta, M. and Mélen, M., editors, *Proceedings of the ESCOM 10th Anniversary Conference on Musical Creativity [CD-ROM]*. Université de Liège, Liège, Belgium.

Chater, N. (2005). A minimum description length principle for perception. In Grünwald, P. D., Myung, I. J., and Pitt, M. A., editors, *Advances in Minimum Description Length: Theory and Applications*, pages 385–409. MIT Press, Cambridge, MA.

Cheetham, A. H. and Hazel, J. E. (1969). Binary (presence-absence) similarity coefficients. *Journal of Paleontology*, 43(5):1130–1136.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, 2:113–124.

Church, A. (1936). An unsolvable problem of elementary number theory. *American Journal of Mathematics*, 58(2):345–363.

Cilibrasi, R., Vitányi, P., and de Wolf, R. (2004). Algorithmic clustering of music based on string compression. *Computer Music Journal*, 28(4):49–67.

Cilibrasi, R. and Vitányi, P. M. B. (2005). Clustering by compression. *IEEE Transactions on Information Theory*, 51(4):1523–1545.

Cohen, J. E. (1962). Information theory and music. *Behavioral Science*, 7(2):137–163.

Conklin, D. and Witten, I. (1995). Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24(1):51–73.

Conley, J. K. (1981). Physical correlates of the judged complexity of music by subjects differing in musical background. *British Journal of Psychology*, 72:451–464.

Copeland, B. J. and Sylvan, R. (1999). Beyond the universal Turing machine. *Australasian Journal of Philosophy*, 77(1):46–66.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16:297–334.

Dawkins, R. (1976). *The Selfish Gene*. Oxford University Press, Oxford.

de Cheveigné, A. (1993). Separation of concurrent harmonic sounds: fundamental frequency estimation and a time-domain cancellation model of auditory processing. *Journal of the Acoustic Society of America*, 93(6):3271–3290.

de Sousa, J. D. (2002). Chess moves and their memomics: a framework for the evolutionary processes of chess openings. *Journal of Memetics–Evolutionary Models of Information Transmission*, 6. [http://jom-emit.cfpm.org/2002/vol6/de_sousa_jd.html].

Deliège, I. (1996). Cue abstraction as a component of categorisation processes in music listening. *Psychology of Music*, 24:131–156.

Dirst, M. and Weigend, A. S. (1994). Baroque forecasting: on completing J. S. Bach's last fugue. In Weigend, A. S. and Gershenfeld, N., editors, *Time Series Prediction: Forecasting the Future and Understanding the Past*, pages 151–172. Addison-Wesley, Reading, MA.

Edmonds, B. (2002). Three challenges for the survival of memetics. *Journal of Memetics–Evolutionary Models of Information Transmission*, 6. [http://jom-emit.cfpm.org/2002/vol6/edmonds_b_letter.html].

Edmonds, B. (2005). The revealed poverty of the gene-meme analogy: why memetics per se has failed to produce substantive results. *Journal of Memetics–Evolutionary Models of Information Transmission*, 9. [http://jom-emit.cfpm.org/2005/vol9/edmonds_b.html].

Eerola, T. and Bregman, M. (2007). Melodic and contextual similarity of folk song phrases. *Musicæ Scientiæ*, Discussion Forum 4A:211–233.

Eerola, T., Järvinen, T., Louhivuori, J., and Toiviainen, P. (2001). Statistical features and perceived similarity of folk melodies. *Music Perception*, 18(3):275–296.

Eysenck, H. (1942). The experimental study of the "good Gestalt": a new approach. *Psychological Review*, 49:344–364.

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, 61:1–10.

Faith, D. P., Minchin, P. R., and Belbin, L. (1987). Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio*, 69:57–68.

Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle.

Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, New York.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5(10):3–8.

Glass, G. V. (1977). Integrating findings: the meta-analysis of research. *Review of Research in Education*, 5:351–379.

Goodenough, W. H. (1957). Cultural anthropology and linguistics. In Garvin, P. L., editor, *Report of the Seventh Annual Round Table Meeting on Linguistics and Language Study*, pages 167–173. Georgetown University Press, Washington, DC.

Graur, D. and Li, W.-H. (2000). *Fundamentals of Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Grünwald, P. (2005). Introducing the minimum description length principle. In Grünwald, P. D., Myung, I. J., and Pitt, M. A., editors, *Advances in Minimum Description Length: Theory and Applications*, pages 3–21. MIT Press, Cambridge, MA.

Haber, R. N. (1958). Discrepancy from adaptation level as a source of affect. *Journal of Experimental Psychology*, 56(4):370–375.

Hammer, D., Romashchenko, A., Shen, A., and Vereshchagin, N. (2000). Inequalities for Shannon entropy and Kolmogorov complexity. *Journal of Computer and System Sciences*, 60(2):442–464.

Hansen, M. H. and Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454):746–774.

Helson, H. (1947). Adaptation-level as frame of reference for prediction of psychophysical data. *American Journal of Psychology*, 60(1):1–29.

Henrich, J. and Boyd, R. (2002). On modeling cognition and culture. *Journal of Cognition and Culture*, 2:87–112.

Heyduk, R. G. (1975). Rated preference for musical compositions as it relates to complexity and exposure frequency. *Perception and Psychophysics*, 17(1):84–91.

Hofmann-Engl, L. and Parncutt, R. (1998). Computational modeling of melodic similarity judgments: two experiments on isochronous melodic fragments. [http://freespace.virgin.net/ludger.hofmann-engl/similarity.html].

Hofstadter, D. R. (1985). *Metamagical Themas: Questing for The Essence of Mind and Pattern*. Basic Books, New York.

Hopcroft, J. E., Motwani, R., and Ullman, J. D. (2000). *Introduction to Automata Theory, Languages, and Computation* (2nd ed.). Addison-Wesley, Boston.

Horibe, Y. (1985). Entropy and correlation. *IEEE Transactions on Systems, Man and Cybernetics*, 15:641–642.

Hu, K. T. (1962). On the amount of information. *Theory of Probability and its Applications*, 7:439–447.

Hull, D. L. (2000). Taking memetics seriously: memetics will be what we make it. In Aunger, R., editor, *Darwinizing Culture: The Status of Memetics as a Science*, pages 43–67. Oxford University Press, Oxford.

Jan, S. (2000a). The memetics of music and its implications for psychology. In Woods, C., Luck, G., Brochard, R., Seddon, F., and Sloboda, J., editors, *Proceedings of the 6th International Conference on Music Perception and Cognition, Staffordshire, England*. [CD-ROM].

Jan, S. (2000b). Replicating sonorities: towards a memetics of music. *Journal of Memetics–Evolutionary Models of Information Transmission*, 4. [http://jom-emit.cfpm.org/2000/vol4/jan_s.html].

Jeong, J., Joung, M. K., and Kim, S. Y. (1998). Quantification of emotion by nonlinear analysis of the chaotic dynamics of electroencephalograms during perception of 1/f music. *Biological Cybernetics*, 78(3):217–225.

Johnson, D. S., Gutin, G., McGeoch, L. A., Zhang, W., and Zverovitch, A. (2002). Experimental analysis of heuristics for the ATSP. In Gutin, G. and Punnen, A. P., editors, *The Traveling Salesman Problem and Its Variations*, pages 445–487. Kluwer Academic Publishers, Dordrecht.

Katz, B. F. (1994). An ear for melody. *Connection Science*, 6(2–3):299–324.

Kitcher, P. (1987). Confessions of a curmudgeon. *Behavioral and Brain Sciences*, 10(1):89–99.

Kolmogorov, A. N. (1965). Three approaches to the quantitive definition of information. *Problems of Information Transmission*, 1:3–11.

Kolmogorov, A. N. (1968). Logical basis for information theory and probability theory. *IEEE Transactions on Information Theory*, 14(5):662–664.

Koshelev, M. (1998). Towards the use of aesthetics in decision making: Kolmogorov complexity formalizes Birkhoff's idea. *Bulletin of the European Association for Theoretical Computer Science*, 66:166–170.

Kraehenbuehl, D. and Coons, E. (1959). Information as a measure of the experience of music. *Journal of Aesthetics and Art Criticism*, 17(4):510–522.

Kvålseth, T. O. (1987). Entropy and correlation: some comments. *IEEE Transactions on Systems, Man and Cybernetics*, 17:517–519.

Large, E. W., Palmer, C., and Pollack, J. B. (1995). Reduced memory representations for music. *Cognitive Science*, 19(1):53–96.

Lerdahl, F. (1988). Cognitive constraints on compositional systems. In Sloboda, J. A., editor, *Generative Processes in Music: The Psychology of Performance, Improvisation, and Composition*, pages 231–259. Oxford University Press, Oxford.

Lerdahl, F. and Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. MIT Press, Cambridge, MA.

Levin, L. A. (1973). Universal sequential search problems. *Problems of Information Transmission*, 9:265–266.

Levitin, D. J. (1994). Absolute memory for musical pitch: evidence from the production of learned melodies. *Perception and Psychophysics*, 56:414–423.

Levitin, D. J. and Cook, P. R. (1996). Memory for musical tempo: additional evidence that auditory memory is absolute. *Perception and Psychophysics*, 58:927–935.

Li, M., Chen, X., Li, X., Ma, B., and Vitányi, P. M. B. (2003). The similarity metric. In Farach-Colton, M., editor, *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 863–872. SIAM, Philadelphia, PA.

Licklider, J. C. R. (1951). A duplex theory of pitch perception. *Experientia*, 7:128–134.

Lin, D. (1998). An information-theoretic definition of similarity. In Shavlik, J. W., editor, *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA.

Lomax, A. (1980). Factors of musical style. In Diamond, S., editor, *Theory and Practice: Essays Presented to Gene Weltfish*, pages 29–58. Mouton, The Hague.

López-Ruiz, R., Mancini, H. L., and Calbet, X. (1995). A statistical measure of complexity. *Physics Letters A*, 209:321–326.

Lőrincz, A., Szatmáry, B., and Szirtes, G. (2002). The mystery of structure and function of sensory processing areas of the neocortex: a resolution. *Journal of Computational Neuroscience*, 13:187–205.

Lynch, A. and Baker, A. J. (1994). A population memetics approach to cultural evolution in chaffinch song: differentiation among populations. *Evolution*, 48:351–359.

Matsumoto, M. and Nishimura, T. (2000). Dynamic creation of pseudorandom number generators. In Niederreiter, H. and Spanier, J., editors, *Monte Carlo and Quasi-Monte Carlo Methods 1998*, pages 56–69. Springer, Berlin.

McClelland, D. C., Atkinson, J. W., Clark, R. A., and Lowell, E. L. (1953). *The Achievement Motive*. Appleton-Century-Crofts, New York.

McCormack, J. (2005). Open problems in evolutionary art and music. In et al., F. R., editor, *Applications on Evolutionary Computing*, pages 428–436. Springer, Berlin.

McGraw, K. O. and Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30–46.

Medin, D. L., Goldstone, R. L., and Gentner, D. (1993). Respects for similarity. *Psychological Review*, 100:254–278.

Meyer, L. B. (1957). Meaning in music and information theory. *Journal of Aesthetics and Art Criticism*, 15:412–424.

Moles, A. (1968). *Information Theory and Esthetic Perception (J. E. Cohen, Trans.)*. University of Illinois Press, Urbana.

Narmour, E. (1999). Hierarchical expectation and musical style. In Deutsch, D., editor, *The Psychology of Music*, pages 441–472. Academic Press, San Diego.

Neisser, U. (1967). *Cognitive Pychology*. Appleton-Century-Crofts, New York.

North, A. C. and Hargreaves, D. J. (1995). Subjective complexity, familiarity, and liking for popular music. *Psychomusicology*, 14:77–93.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15:267–273.

Papadopoulos, G. and Wiggins, G. (1999). AI methods for algorithmic composition: a survey, a critical view and future prospects. In Patrizio, A., Wiggins, G. A., and Pain, H., editors, *Proceedings of the AISB'99 Symposium on Musical Creativity*, pages 110–117. Society for the Study of Artificial Intelligence and Simulation of Behaviour, Brighton.

Patel, A. D. and Balaban, E. (2000). Temporal patterns of human cortical activity reflect tone sequence structure. *Nature*, 404:80–84.

Phillips, W. A. (1997). Theories of cortical computation. In Rugg, M. D., editor, *Cognitive Neuroscience*, pages 11–46. MIT Press, Cambridge, MA.

Pinkerton, R. C. (1956). Information theory and melody. *Scientific American*, 194(2):77–86.

Pitt, M. A. and Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, 6(10):421–425.

Plotkin, H. (2000). Culture and psychological mechanisms. In Aunger, R., editor, *Darwinizing Culture: The Status of Memetics as a Science*, pages 69–82. Oxford University Press, Oxford.

Pocklington, R. and Best, M. L. (1997). Cultural evolution and units of selection in replicating text. *Journal of Theoretical Biology*, 188:79–87.

Poppel, E. (1989). The measurement of music and the cerebral clock: a new theory. *Leonardo*, 22(1):83–89.

Rajski, C. (1961). Entropy and metric spaces. In Cherry, C., editor, *Information Theory*, pages 41–45. Butterworths, London.

Restle, F. (1959). A metric and an ordering on sets. *Psychometrika*, 24:207–220.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14:465–471.

Saks, S. (1937). *Theory of the Integral*. Hafner, New York.

Sanger, T. D. (1989). An optimality principle for unsupervised learning. In Touretzky, D. S., editor, *Advances in Neural Information Processing Systems 1*, pages 11–19. Morgan Kaufmann, San Mateo, CA.

Santini, S. and Jain, R. (1999). Similarity measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):871–883.

Schaffrath, H. (1997). The *Essen Associative Code*: a code for folksong analysis. In Selfridge-Field, E., editor, *Beyond MIDI: The Handbook of Musical Codes*, pages 343–361. MIT Press, Cambridge, MA.

Schaffrath, H. and Dahlig, E. (2000). EsAC folksong database. [http://www.esac-data.org/].

Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology*, 188:415–431.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656.

Sharpe, R. A. (1971). Music: the information-theoretic approach. *British Journal of Aesthetics*, 11(4):385–401.

Shiner, J. S., Davison, M., and Landsberg, P. T. (1999). Simple measure for complexity. *Physical Review E*, 59:1459–1464.

Shmulevich, I. and Povel, D.-J. (2000). Measures of temporal pattern complexity. *Journal of New Music Research*, 29(1):61–69.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

Simonton, D. K. (1997). Products, persons, and periods: historiometric analyses of compositional creativity. In Hargreaves, D. J. and North, A. C., editors, *The Social Psychology of Music*, pages 108–122. Oxford University Press, New York.

Snyder, B. (2000). *Music and Memory: An Introduction*. MIT Press, Cambridge, MA.

Stanton, P. K. and Sejnowski, T. J. (1989). Storing covariance by the associative long-term potentiation and depression of synaptic strengths in the hippocampus. In Touretzky, D. S., editor, *Advances in Neural Information Processing Systems 1*, pages 394–401. Morgan Kaufmann, San Mateo, CA.

Stiny, G. and Gips, J. (1978). *Algorithmic Aesthetics: Computer Models for Criticism and Design in the Arts*. University of California Press, Berkeley, CA.

Tan, S.-L., Spackman, M. P., and Peaslee, C. L. (2006). The effects of repeated exposure on liking and judgments of musical unity of intact and patchwork compositions. *Music Perception*, 23(5):407–421.

Temperley, D. and Sleator, D. (1999). Modeling meter and harmony: a preference rule approach. *Computer Music Journal*, 23(1):10–27.

Titchener, M. R. (2000). A measure of information. In Storer, J. A. and Cohn, M., editors, *Proceedings of the Data Compression Conference (DCC'00)*, pages 353–362. IEEE Computer Society, Los Alamitos, CA.

Toop, R. (1993). On complexity. *Perspectives of New Music*, 31(1):42–57.

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42:230–265.

Tversky, A. (1977). Features of similiarity. *Psychological Review*, 84:327–352.

Vitz, P. C. (1964). Preferences for rates of information presented by sequences of tones. *Journal of Experimental Psychology*, 68(2):176–183.

Vitz, P. C. (1966). Affect as a function of stimulus variation. *Journal of Experimental Psychology*, 71(1):74–79.

Walker, E. L. (1973). Psychological complexity and preference: a hedgehog theory of behavior. In Berlyne, D. E. and Madsen, K. B., editors, *Pleasure, Reward, Preference: Their Nature, Determinants, and Role in Behavior*, pages 65–97. Academic Press, New York.

Wallin, N. L. (1991). *Biomusicology: Neurophysiological, Neuropsychological, and Evolutionary Perspectives on the Origins and Purposes of Music*. Pendragon, Stuyvesant, NY.

Weyl, H. (1949). *Philosophy of Mathematics and Natural Science*. Princeton University Press, Princeton.

Yang, J. and Speidel, U. (2005). A T-decomposition algorithm with $O(n \log n)$ time and space complexity. In Hanly, S. and Schlegel, C., editors, *Proceedings of the 2005 International Symposium on Information Theory*, pages 23–27. IEEE, Piscataway, NJ.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8:338–353.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, 9(2, Pt. 2):1–27.