THE LEARNING TRAJECTORY OF MUSICAL MEMORY: FROM SCHEMATIC PROCESSING OF NOVEL MELODIES TO ROBUST MUSICAL MEMORY REPRESENTATIONS

A Dissertation

Presented to the Faculty

of the Graduate School of Cornell University

in Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

by

Kathleen Rose Agres

January 2013

THE LEARNING TRAJECTORY OF MUSICAL MEMORY: FROM SCHEMATIC PROCESSING OF NOVEL MELODIES TO ROBUST MUSICAL MEMORY REPRESENTATIONS

Kathleen Rose Agres, Ph.D.

Cornell University 2013

This dissertation utilizes a multi-method approach to investigate the processes underlying musical learning and memory. Particular emphasis is placed on schematic processing, musical structure, temporal aspects of learning, statistics-based predictive models, efficiency, and the role of musical expertise.

We employed a set of behavioral change detection studies with musician and non-musician participants to test what is encoded into gist memory upon hearing unfamiliar melodies varying in musical structure. These studies demonstrate that listeners abstract a schematic representation of the melody that includes tonally and metrically salient tones. In well-structured music, change detection performance improves when a musical event does not conform to the listener's schematic expectations. Musical expertise is also shown to benefit change detection, especially when the melodies conform to the conventions of Western tonal music.

In a study examining learning over a period of increasing musical exposure, we used an information theoretic approach to capture how the statistical properties of music influence listeners' musical memory. This work highlights how patterns and predictability can facilitate musical learning over time. In further investigation of what underlies this learning process, a series of neural network studies revealed that a compressed representation arose in the internal structure of a computational network as tonal and stylistic information were learned over time.

Population sparsity of the SRN’s hidden layer strongly predicted the sophistication of the network’s musical output as rated by human listeners.

Electroencephalography (EEG) methods were utilized to investigate the neural correlates of musical learning and memory, and to further explore the notion of increasing efficiency over the time-course of learning. These experiments suggest that the listener's implicit internal model of musical expectation is gradually developed and made increasingly accurate with repeated exposure to initially unfamiliar music.

Both the computational and EEG experiments illustrate how efficiency accompanies successful learning over time. These findings, as well as those from the change detection and information theory studies, provide evidence that schemata are formed as the probabilities of forthcoming music are gradually learned with increasing experience. Schematic expectations dynamically guide perception and influence memory, and generally allow for more efficient musical processing.

BIOGRAPHICAL SKETCH

Kat was born in San Diego, California in 1983. The daughter of two musicians, Nancy and Sam Agres, she took an early interest in music. She began playing the cello at age 8, and has held a strong passion for music ever since. After cultivating her love of science during high school, Kat attended Carnegie Mellon University (CMU) to pursue both psychology and music, and was awarded the degree of Bachelor of Humanities and Arts in Cognitive Psychology and Cello Performance in 2005. While at CMU, Kat also performed regularly with a rock band playing electric cello, worked in the laboratory of Professor Lori Holt for three years, briefly studied violin making with a luthier in Cremona, Italy, and became involved with music activism and music therapy. In her final year at CMU, Kat was awarded a Presidential Scholarship, as well as a yearlong fellowship from the National Institute of Mental Health. In 2006, Kat began her graduate studies of music cognition in the Psychology Department of Cornell University, where she was granted a Sage Fellowship. Her developing research interests included musical learning and memory, change detection, and patterns in neural activity during music listening. She was fortunate to be awarded the NIH-funded IMAGINE Training Grant during her fourth and fifth years at Cornell. During this fellowship, Kat used electroencephalography (EEG) to examine musical memory in the laboratory of Jason Zevin at Weill Cornell Medical College in New York City. Her other experimental methodologies included infant research (in the laboratory of Michael Goldstein), computational modeling (with Michael Spivey), and techniques based on Information Theory (with David Field, as well as Geraint Wiggins and Marcus Pearce). Kat is an avid music fan and chamber music performer. She also enjoys dancing, traveling, rock climbing, and the visual arts. Her postdoctoral work will be conducted at Queen Mary, University of London, in the laboratory of Marcus Pearce.

This dissertation is dedicated to my wonderful family and friends whose love, encouragement, and support made this PhD not only possible, but very enjoyable indeed.

ACKNOWLEDGEMENTS

I would like to give many thanks to my friends and colleagues in the Psychology Department, especially those in the laboratories of David Field and James Cutting. The insight, enthusiasm, and driving questions from these individuals led to a richer experience of graduate school and the scientific process. Thanks to Michael Hove, with whom many discussions of music cognition were shared, and to Hia Datta for many fascinating conversations about EEG and the brain, and for her help during late night (and daytime) trouble-shooting sessions.

In addition, a kind thank you to my collaborators: To Andrew Goldfine (WCMC) for his assistance with EEG coding and analysis techniques, to Samer Abdallah (University of London) for his brilliant computational contributions to our Information Theory study, and to Jordan DeLong, for his help with the second and third SRN studies (and answering many programming questions along the way). Also, I appreciate the help of Professor James Booth, for his statistical advice concerning my change detection studies, and Christine Lee, for her assistance in making stimuli and running participants in the second change detection study. To Geraint Wiggins and Marcus Pearce (University of London), my most sincere gratitude for extending the offer to collaborate (and host me in London), and for their kindness and invaluable guidance regarding our Information Theory project.

I am indebted to the Cognitive Science Program, and to Morten Christiansen and Julie Simons-Lynch, for their support and travel funding, making many conference presentations possible. I greatly appreciate having received the Cognitive Science Program's Summer Research Fellowship during 2011, which made simultaneous research at Cornell and WCMC feasible. Many thanks also to Barb Finlay, for her brilliant courses, and for her role in awarding me the NIH IMAGINE grant that opened many doors and enabled my training in neuroscientific methods at WCMC.

Finally, a very warm thanks to my Special Committee: Michael Spivey, for his exceptional kindness and support, and for introducing me to computational methods; Jason Zevin, for welcoming music cognition research into his lab, providing training on EEG techniques, and for sharing his knowledge of auditory perception; Michael Goldstein, for his enthusiasm and guidance as I investigated infant experimental psychology; David Pizarro for his optimism and support, and for spurring my interest in music and emotion; and David Field, my committee chair, for greatly insightful conversations, generous feedback on projects and papers, and for his incredible guidance and encouragement through the twists and turns of this PHD.

The EEG research within this dissertation was funded in part by a grant from the National Institute of Health, NIH T32HD05517.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Introduction to Musical Memory

Music provides a fascinating and rich domain for the investigation of human cognition – it is found in every culture, shares processing mechanisms with other domains, and has a formal structure that can be manipulated and tested. Further, the large range of musical ability found in most societies enables researchers to test the effects of expertise on music cognition and auditory perception more generally.

Although ample research within the field of music cognition has investigated memory for novel melodies and familiar tunes, there is a gap in knowledge regarding the process of *learning* music. This dissertation explores the process of learning novel melodies, with a focus on musical structure, schematic processing, and expectation in music (as assessed via information theory and measures of processing efficiency). The role of repeated exposure in encoding increasingly robust representations is explored, as well as the ways in which neural activity changes over the musical learning trajectory. In addition, learning and memory in music will be compared to other domains, especially vision, to highlight domain-general processing mechanisms, and to provide further insight into short-term musical memory.

Many mechanisms influence memory for music, with some weighted more heavily than others. Statistical learning, invariance, prediction, change detection, and the hierarchical processing of musical structure all impact memory representations. Of these, an emphasis will be placed on schematic processing, which is shown to unify various findings and theoretical explanations in music cognition. Arguably, schematic processing provides an essential framework for music perception and cognition, underlying auditory grouping, musical expectation, perception of melodic and rhythmic structure, and short-term memory.

The research within this dissertation utilizes several different methodologies for exploring musical learning and memory because every method has its limitations. Behavioral psychology, while insightful, can leave one without an understanding of underlying mechanisms. Computational approaches to cognition aim to reflect or predict neural processing and/or observable behavior. And neural activity that is recorded but not tied to perception or behavior is either meaningless or difficult to interpret. A multi-method approach allows for investigation at different levels of analysis (from the neural to computational to behavioral level). Compared to research relying heavily on one technique, this approach can provide a more global account by tying together converging findings and theoretical perspectives across areas. This dissertation utilizes behavioral and computational methods, information theory, and electroencephalography to explore the fundamental questions of *what* is learned and encoded in short-term musical memory and *how*.

1.1.1 Short-term Memory

Because this dissertation focuses on learning and memory in music, it is imperative to provide an overview of short-term and long-term memory in music. This also enables the reader to see how my studies fit into the literature. Although a comprehensive review of musical memory is beyond the scope of this chapter, a summary of important findings will provide a framework for understanding the concepts and mechanisms discussed later.

Many questions exist in the realm of music and memory: Do the musical features that are initially encoded in memory change over time? Does the memory representation that is formed after hearing a novel melody differ from that of a well-known melody? How do musical structure, repeated exposure, and musical training play a role? There is no over-arching theory of musical memory, but questions such as these have driven much interest in the area of memory for music.

In a typical short-term memory (STM) musical memory study, listeners will hear a pair of melodies and make a judgment about whether they are the same or different. When the melodies are brief, stylistic, not rhythmically complex, in a moderate tempo, and identical in terms of absolute pitch save one or two tones, the task is quite easy. When, however, the melodies lack conventional musical structure, are more than about two measures long, are presented at an extreme tempo, or are transposed (especially to distant musical keys), performance can decline dramatically (Halpern & Bartlett, 2010; Dowling, 2008).

Dowling and collaborators have a series of influential studies on short-term musical memory, and some of the early work on this topic proposed a two-component model of memory for melodies (Dowling, 1978; 1991). The first component of the theory lies within listeners' knowledge of the musical scale, which he refers to as a perceptual-motor schema. The second is the melodic contour of the melody (the pattern of ascending and descending pitch intervals), which, he claims, is stored separately and can function independently from memory of the exact interval sizes. Dowling maintains that these two components contribute to the reproduction and recognition of melodies (Dowling, 1978). According to this model, listeners extract the melodic contour upon hearing a melody, but the exact intervals within the melody may not initially be encoded (or immediately accessible).

To test his model, Dowling had musician and non-musician participants listen to 48 sets of melodies and respond whether they thought the two melodies of each set had an identical melodic contour (Dowling, 1978). The second (comparison) melody in each set was either a Target (transposition with exactly the same contour), “Tonal Answer” lure (a melody with the same contour, but different intervals), “Atonal Contour” lure (an atonal melody with the same contour, but different intervals), or a Random sequence (randomly selected tones from the diatonic scale, and a different contour). All of the participants successfully distinguished the Target from the Random sequence, and musicians (but not non-musicians) distinguished the Target from the Atonal Contour lure. Both groups, however, performed at chance with regard to the Tonal Answer melodies, showing that even musicians are fooled by tonal melodies featuring the same contour but different exact intervals. This provides support for the hypothesis that contour and interval size are stored separately in memory. Further evidence, he claims, stems from the fact that we can still recognize a tune such as *Twinkle Twinkle* when it is played in a minor key (we recognize the contour, despite the fact that some of the intervals are different) (Dowling, 1978).

In a later study, Dowling (1991) replicated and extended the above results by running an experiment that manipulated the delay period between the first and second melody of each test set. A continuous sequence of novel melodies was played, and each melody had either a strong sense of tonality, weak tonality, or no tonality (atonal). In the short-delay condition, participants were tested on whether the contours of two sequential melodies in the sequence (with a silence of 11 seconds in between) were the same or different. In the long-delay condition, the melodies were not adjacent in the sequence; the delay between comparison melodies (which was an average of 39 seconds) was filled with other melodies in the sequence. These intervening melodies had different tonalities than the comparison melodies being tested. Dowling found that for tonal melodies in the short delay condition, similar tonality (and contour) fooled participants, but that the false alarm rate for Same Contour lures diminished over the longer time period. Exact interval changes were more apparent after a longer delay, while contour changes were more detectable after the short delay. Therefore, it seems that contour (for tonal melodies) is represented initially in short-term memory, but this is gradually supplemented by the encoding of exact intervals in long-term memory. Surprisingly, this finding was true of melodies that were *not* repeated more than once. That is, regardless of exposure, the exact intervals encoded in memory representations for novel melodies became more stable over time (Dowling, 1991). Although these results provide a straightforward, parsimonious account of STM for melodies, there may be a few concerns with the experimental design. First, it is possible that the intervening melodies of different tonalities altered the memory of tonality for the first melody of the comparison set (thus resulting in less reliance on tonality for the comparison). Second, a confound of interference is present; the interleaved melodies may have prevented or disrupted encoding of the melody's contour.

Other work examining the recognition of transposed melodies has yielded a more complex picture than the one above. Lola Cuddy and colleagues have tested recognition of transposed melodies while manipulating different musical characteristics (e.g. Cuddy & Cohen, 1976; Cuddy, Cohen, & Mewhort, 1981). This has yielded the finding that many features can affect recognition and musical short-term memory, including "triadic structure, repetition of the tonic, leading tone to tonic ending, harmonic cadence (V-I), modulation within the sequence, and key-distance of transposition (tritone vs dominant)" (Krumhansl, 1991). Even the exact pattern of intervals in a *three*-note melody accounts for differences in listeners' recognition performance (Cuddy & Cohen, 1976). While this work may account for more variance in listeners' responses than the simpler model by Dowling (1978), it does not provide a holistic view, instead favoring the approach that nearly everything appears to influence short-term musical memory (given such a wide range of factors). This type of framework is more complex and difficult to test, and several of the findings, such as harmonic cadence and leading tone, could simply be considered aspects of the tonal component of Dowling's model. Also, it would be interesting to test whether the impact of specific interval patterns would decrease with longer musical sequences (at which point, more Gestalt-like schematic processing might prevail).

These musical features focus on tonal and harmonic influences on melody recognition, but rhythm also has an effect on short-term musical memory (Kidd, Boltz, & Jones, 1984). To explore the effect of rhythm on detection of pitch changes in melodies, pairs of 10-note melodies with either identical or dissimilar rhythms were played for participants. Listeners were told to focus on pitches and ignore the rhythm of the sequences, and respond whether the melodic content of the pair was the same or different. The authors speculate that rhythmic context can lead attention towards musically salient events within the melody. Rhythm, they argue, can prepare the listener for important patterns or relationships between musical events (Kidd, Boltz, & Jones, 1984). According to this "temporal expectancy hypothesis", rhythmic and temporal cues are taken to be a sort of priming tool during music perception (because expectations are formed or primed), which then affects measures of bias and discriminability, in the Signal Detection Theory sense (Kidd, Boltz, & Jones, 1984). This hypothesis was contrasted with the possibility that tempo (which was manipulated via slow, medium, and fast presentation rates of melodies) affects processing. If this were the case, task performance should improve at the slower tempo because this allows for more processing time. The results show that no significant differences in performance existed between presentation rates; tempo of the melodies did not affect accurate change detection. Evidence was found, however, to support the temporal expectancy hypothesis. When the rhythm of the initial melody was different than that of the comparison, listeners were significantly biased to report that the melodies were the same. Discriminability (d') was also affected in that performance was worse; listeners could not reliably tell change from same trials. The extent to which performance was degraded depended on the amount of temporal uncertainty resulting from the rhythmic context (Kidd, Boltz, & Jones, 1984). Therefore, the change in rhythm either influenced the encoding of the melodies into memory, or disrupted comparison of the melodies because the melodic representations were encoded in terms of both pitch and rhythm in memory.

The "surface characteristics" of music have been studied far less than pitch and rhythm, possibly because musical memory is largely invariant with respect to these properties, but there is some evidence that tempo and timbre affect melodic recognition. For example, after hearing a novel melody, if the tempo or timbre is altered in a subsequent exposure, measures of explicit memory for the melody are impaired (Halpern & Mullensiefen, 2008). Therefore, in short-term memory, tempo and timbre may be an important part of the memory representation for a novel melody. After more exposure to the melody, however, these characteristics may be less salient in long-term memory, as varying them can have little impact on melody recognition (Peretz & Zatorre, 2005).

1.1.2 Long-term Memory

Long-term memory (LTM) for music can be influenced by familiarity, nameability, and episodic and extra-musical associations, adding layers of complexity to the investigation of this type of musical memory (Halpern & Bartlett, 2010). In a typical LTM study, listeners will try to memorize a set of tunes, and 10-30 minutes later, the participants will be given a recognition test of old and new stimuli. When listeners study familiar tunes, they demonstrate better overall performance in the recognition task than a study in which all of the tunes are unfamiliar (Bartlett et al, 1995). But when familiar and unfamiliar tunes are both presented in the same listening set, performance declines, with false alarm rates increasing dramatically for familiar tunes (that is, there seems to be a strong tendency to judge familiar tunes as “old”). Halpern and Bartlett (2010) suggest this bias may be due to “subjective familiarity.” When familiar and unfamiliar tunes are tested separately, the listener can adopt a different familiarity criterion for well known versus unfamiliar tunes. When both types of tunes are tested together, however, it may be much more difficult to establish an accurate criterion.

Long-term memory for music is predominantly reliant on the melodic interval pattern of the music (which encompasses the key and contour of the melody), but, like musical STM, it is also influenced by rhythmic structure (Hebert & Peretz, 1997). In a study comparing the effects of contour and rhythm on recognition memory, familiar melodic excerpts (melodies that are stored in long-term memory) were played for listeners in two conditions: In the melodic condition, the interval pattern was kept the same, but all of the notes were isochronous (the same duration). In the rhythmic condition, the rhythmic content was preserved, but all of the notes were played on a single pitch (Hebert & Peretz, 1997). Better recognition of the familiar melodies was attained in the melodic condition, possibly because the melodic structure of the melodies was more salient than their rhythmic structure (Hebert & Peretz, 1997). All in all, however, the authors come to the intuitive conclusion that the most effective means of accessing long-term memory (accurately recognizing melodies) involves the correct combination of melodic (pitch) and temporal (rhythmic) information (Hebert & Peretz, 1997). Of course, surface features can also be stored in long-term memory, as evinced by our ability to recognize a specific rendition or performer (Peretz & Zatorre, 2005).

The fact that humans are able to identify a vast number of musical examples suggests that most individuals have a huge store of memory representations of familiar music. Peretz and colleagues refer to this musical memory corpus as the "musical lexicon" (Peretz, et al, 2009). The process of recognizing a tune, they argue, automatically engages a series of processing mechanisms that lead to access of the lexicon. First is the *access stage*, in which "the beginning of the music activates a series of potential tune candidates" that are based on the perceptual analysis of the pitch and temporal structure of the input. The *selection stage* reduces the number of candidates as more musical information unfolds, until one option emerges as the best fit. Lastly, there is also an *integration stage*, where the melody or phrase is placed within the broader musical context of the whole piece (Peretz, Gosselin, Belin, Zatorre, Plailly, & Tillmann, 2009). Because the concept of a musical lexicon has recently been introduced to the field, little is known about its capacity, but it is assumed to be quite vast. It would be extremely interesting to investigate the average storage capacity of the musical lexicon, for musicians and non-musicians, and compare this to the average capacity of the linguistic lexicon.

1.1.3 Interim Summary of Musical Memory

Empirical findings show that memory for music changes over time. Immediately upon hearing a new melody, the melodic contour and salient rhythmic events will be encoded. When a melody with the same contour is played for comparison, the melodies will often mistakenly be judged as the same, regardless of musical training. After some time (note that the exact length of time warrants investigation), the exact interval pattern is accessible in memory. In well-known melodies, such as those in most individuals' musical lexicon (e.g., *Twinkle Twinkle*), the memory is invariant with relation to key, tempo, and timbre. In addition, for specific musical pieces that are repeatedly heard, absolute features such as tempo and timbre are likely to be stored.

One of the primary goals of the research within this dissertation is to fill in the gap of knowledge between STM and LTM for music. Many studies have examined memory after initial exposure to novel melodies, and some have investigated the nature of memory for well-known music, but little work has explored the relationship between the two. Therefore, one of the main objectives of this dissertation is to address how memory representations change during the learning process as melodies in STM are more robustly encoded over time.

1.2 Domain General Processing

1.2.1 Insight from the Visual Modality

Studies of memory in non-musical domains can both provide a useful comparative context for understanding findings in music, but also inspire new directions for investigation. It can be especially interesting to examine cognition across modalities to test whether processing mechanisms are conserved and reused, or domain specific. Research has historically implied, for example, that we have a robust visual and auditory representation of the world. Visual memory can be surprisingly detailed and complete (Shepard, 1967), and recent findings continue to demonstrate that we have a large memory capacity for visual information (e.g., Brady, Konkle, Alvarez, & Oliva, 2008). Similar findings exist in the auditory modality. In the domain of music, as discussed earlier, listeners demonstrate a large musical lexicon (Peretz, et al., 2009), and even non-musicians are able to recall the absolute pitches of a piece with notable accuracy (Levitin, 1994). Despite these findings, there is also evidence to the contrary: Bartlett demonstrated decades ago that participants do not retain a verbatim account of speech or prose (Bartlett, 1932), and many studies in music perception have revealed listeners' poor ability to detect changes introduced to music (e.g., Halpern & Bartlett, 2010; Snyder, 2000; Dowling & Bartlett, 1981; Cuddy, Cohen, & Miller, 1979). Further, auditory recognition memory has been demonstrated to be inferior to visual recognition memory for a range of stimuli (Cohen, Horowitz, & Wolfe, 2009). Even in visual perception, viewers can be surprisingly poor at detecting changes to a visual scene, as shown through studies of change blindness (e.g., Simons & Rensink, 2005). Extensive work on this topic has helped to clarify this seeming paradox by demonstrating that viewers encode an incomplete representation of visual scenes (e.g., Rensink, O'Regan, & Clark, 1997; Simons & Levin, 1997; Simons & Ambinder, 2005), and instead tend to retain salient information and a general semantic understanding, or gist, of the scene (Oliva & Torralba, 2006; Oliva, 2005; Wolfe, 1998). Chapter 2 of this dissertation seeks to explore whether the same holds true in the auditory domain.

1.2.2 Gist Across Domains

Comparing memory representations in vision and audition may yield insight into whether processing mechanisms are shared across domains (For a review comparing visual and auditory change detection, see Snyder & Gregg, 2011). Because the early pathways of vision and audition utilize different physiological mechanisms, the most insightful comparison between the modalities will be with regard to high-level effects. The roles of saliency and gist memory in object and scene perception are of current interest in the vision literature (Oliva & Torralba, 2006; Oliva, 2005; Bar, 2004; Hollingworth, 2003; Rensink et al., 1997), but these topics have not yet received adequate attention in audition (but see Harding, Cooke, & Konig, 2007 for an account of auditory gist). Failure to detect change may be attributed to the changed item(s) conforming to the observer's schema or gist memory of the scene. Here, *musical gist* is defined as a memory representation for schematically consistent tones – a general abstraction that lacks full detail of the original stimulus.

Vision researchers have offered the possibility that viewers primarily encode the general schematic attributes of a visual scene upon brief initial viewing (Rensink et al., 1997; Oliva & Torralba, 2006). When a visual stimulus is categorized using a high-level semantic label (such as "beach at sunset"), the semantic context can both prime object recognition within the scene (Bar, 2004; Oliva & Torralba, 2007), and cause changes consistent with the schema to go undetected while schema-inconsistent changes are more likely to be remembered (Sakamoto & Love, 2004; Rensink, 2002). In situations in which schematic processing is compromised, such as the presentation of an unusual or scrambled context, change detection performance declines (Zimmermann, Schnier, & Lappe, 2010).

The high-level, semantic label can be thought of as a kind of conceptual gist, and interestingly, a distinction has been made in vision between *conceptual* and *perceptual* gists that may be a useful when applied to auditory perception. As defined by Aude Oliva, a perceptual gist is the "structural representation of a scene" that is constructed during perception, while a conceptual gist is broader and "includes the semantic information that is inferred while viewing a scene or shortly after the scene has disappeared from view" (Oliva, 2005). As one views examples of a scene, one gradually learns the statistical regularities of that scene. Within a beach scene, for example, one is likely to see sand and waves that are displayed with a particular spatial relationship. The greater number of beach scenes viewed, the richer the semantic, conceptual gist for "beach". The two types of gist are connected. A conceptual gist (or schema) may influence perception (and consequently the perceptual gist formed) by directing attention and creating expectations about objects that are likely to be present in the scene. And conversely, the conceptual gist is created and modified by perceptual experience. In music, higher level schematic knowledge is like a conceptual gist that guides perception and expectation of the forthcoming music. The abstracted memory formed upon hearing an excerpt is akin to a perceptual gist.

The distinction between perceptual and conceptual gist is also reminiscent of the distinction between familiarity and recollection, as outlined by Halpern and Bartlett (2010). That is, familiarity is described as a general sense that the music has been heard before, and is devoid of extra-musical contextual cues. Recollection is described as a conscious recall of particular features and contextual information about music. Recollection therefore contains a component of episodic memory, and can be related to Oliva’s perceptual gist. Familiarity, which is an abstraction from episodic context, is more akin to semantic memory, and the notion of conceptual gist.

1.2.3 The Role of Expertise

Successful change detection in the visual and auditory modalities can also depend on expertise within a particular domain. Evidence shows that experts can more quickly and efficiently encode chunks of information in their learned domain than their novice counterparts (Chase & Simon, 1973; Werner & Thies, 2000). Expertise seems to provide a high-level, robust processing framework that can facilitate pattern recognition and change detection. Therefore, trained musicians should demonstrate superior performance in change detection and musical memory tasks compared to novices. This question of expertise, as well as gist representations and schematic processing, will be explored in Chapter 2 using a change detection paradigm inspired by findings in visual perception.

1.3 Schematic Processing in Music

Dating back to Piaget (1926) and Bartlett (1932), investigations of schematic processing have contributed to our understanding of learning and memory. In the auditory modality, the research of Alan Bregman, W. Jay Dowling, Lola Cuddy, and others have explored the importance of "schema-based mechanisms" and the abstraction of tonal relationships (i.e., melodic contour, and the different function of pitches in the musical key) during music perception (Bregman, 1990; Cuddy, Cohen, & Mewhort, 1981; Dowling, 1978). Experience listening to commonly used musical devices or forms creates the mental framework we use for processing music (Gjerdingen, 1988; Lerdahl, 2001), and the underlying schemas are essentially a collection of rules that guide listeners' perception of music by continually creating expectations about the forthcoming music (e.g., Krumhansl 1990; Narmour, 1992; Lerdahl & Jackendoff, 1983; Huron, 2006). Although musicians may have more elaborated schemas, everyone exposed to Western classical music has implicitly learned these schemas. Again analogous with visual perception, some musical features can be encoded "veridically", like the particular quality (or timbre) of a singer's voice. Other musical features, such as pitch and rhythm, are not always remembered in detail. Rather than encode every feature of novel music, often only an abstraction (or gist), based on tonal and rhythmic schemata, is encoded which highlights certain salient features and general characteristics of the music (Snyder, 2000).

1.3.1 Overview of Musical Schemata

Although not always receiving due attention, schema theory has been mentioned in the field of music cognition for several decades. In *A Classic Turn of Phrase*, Gjerdingen (1988) gives a brief overview of schema theory and relates schematic processing to music. He cites examples of musical schemata that were explored by the pioneering musicologist, Leonard Meyer (although Meyer called them archetypes): the “gap-fill” schema and the “changing-note” schema (Gjerdingen, 1988). A gap-fill schema is essentially when a melodic leap is followed by an ascending or descending sequence of tones that fills the gap (created by the initial musical interval leap). A changing-note schema features two dyads (sets of notes), where the first dyad leads away from the tonic pitch, and the second dyad leads back to the tonic. Even musically untrained listeners are capable of distilling and identifying these schemata from listening to examples containing both types (Gjerdingen, 1988). He later argues that musical schemata are made of the specific set of features that create a “style structure”. Style structures, as elaborated by Eugene Narmour, are the arrangement of “style forms” (e.g., a melodic triad) into common contexts, as specified by their statistical frequency of occurrence. In sum, musical schemata are mental frameworks of musical knowledge that are developed from previous experiences with style forms in certain musical contexts, and that operate using both bottom-up and top-down integrative processes (Gjerdingen, 1988).

Previous experience, and the sets of musical expectations that are derived from it, are also central to Eugene Narmour’s (1992) work. Narmour describes these expectations as top-down, hierarchically organized *syntactic* schemata. The hierarchical structure is in direct contrast with how earlier researchers described the functional organization of schemata (as a web-like network of associations). Arguably, this may be attributed to the structure of music itself, which lends itself to hierarchical organization (e.g., Lerdahl and Jackendoff, 1996). In some sense, the entire Implication-Realization (I-R) model put forth by Narmour is built around schematic processing. The notions of Reversal and Continuation are expectations about how an implied melodic or harmonic trajectory will be realized. A Reversal is the expectation that after a large interval (equal to or larger than a perfect fifth) occurs, the next note is likely to frame a small interval in the opposite registral direction. Continuation occurs when a small interval (e.g., a major second) is followed by another small interval moving in the same registral direction (Narmour, 1992). When listening to specific musical examples, both of these features of the I-R model act as implicit schemata that guide the listener's syntactic expectations. The beauty of this model is that, because of the specific predictions therein, extensive research has been able to test its accuracy (several empirical tests will be reviewed later).

The way in which Fred Lerdahl, an influential music theorist and composer, describes schemata in *Tonal Pitch Space* is reminiscent of Narmour's views. Schemata, he says, are flexible constructs that depend on a convergence of multiple factors, some of which are central, and others that are peripheral (Lerdahl, 2001). When central factors converge, the example is prototypical of that schema. Lerdahl gives several examples of musical schemata, which seem to range from the sub-phrase level to over-arching structural forms, such as the schema for a sonata or rondo. To cite one of his examples, consider the form of a musical sentence: First there is a musical statement and response, and then there is a continuation that leads into a cadence (Lerdahl, 2001). For Lerdahl, then, a schema is a kind of well-used musical device or form, recognized by listeners because of their frequency of use, and often familiar to music theory.

Implicit in Lerdahl's perspective is that we can define schemata based on the statistical frequency of patterns in music. Along these lines, in *Music and Memory*, Bob Snyder (2000) describes musical schemata as networks of long-term memory associations that are essentially amalgamations of the statistical properties of music: semantic frameworks constructed from “the commonalities shared by different experiences” (Snyder, 2000). Over time, episodic memories gradually form a generalized schematic representation. In other words, specific instances become “fuzzy” memories as they blend into the existing schema; specific details are lost, but generalizability of the schemata is gained. Lastly, Snyder also suggests that schemata create expectations in the listener about how a musical sequence will continue.

1.3.2 Schematic Processing as a Guide For Musical Expectation

Schemata create a set of expectations in the listener, and research and theory surrounding tension/relaxation and melodic expectancy in music have formed a cornerstone of the field of music cognition. Melodic expectation, introduced by Leonard Meyer (1956) in his seminal work *Emotion and Meaning in Music*, is simply what a listener anticipates will come next in the melody. Meyer posited that an affective state is aroused in the listener when an expectation based on the musical context is not met. Researchers speak of tension and relaxation in music to describe musical motion – the ebb and flow of implication and realization as the music unfolds – and tension ratings are commonly used to assess melodic expectancy. When evaluating expectancy, tension ratings measure the degree of expectancy of the most expected event. Therefore, if a naive listener is experiencing a modern piece of music for the first time, they may not have a schema in place to make sense of what they hear, and they may have little idea of what might come next in the music. In this case, low tension ratings would reflect low musical expectancy.

In one approach to melodic expectation, the actual sequence of tension and relaxation in music is considered to be a schema (Bigand, 1993). Using Lerdahl and Jackendoff's (1983) theory as a starting point, the author speculates that metrical structure, grouping structure, and tonal hierarchies are combined to evaluate musical tension. The tonal hierarchy, or importance/stability among a key's scale degrees, is gradually learned from extensive exposure to Western music. Our implicit learning of this tonal structure develops from the statistical distribution of scale degrees and n-grams of scale degrees that exist in Western music (Krumhansl, 2000). Grouping structure and metrical structure lead to a time-span reduction of the music, which relays the structural importance of musical events (notes, chords, or rests) given a context. The time span reduction and tonal hierarchy together form a prolongational reduction analysis, which displays the perceived tension and relaxation of events in music as a hierarchy. To assess whether the prolongational reduction did in fact correlate with listeners' perception of musical tension, Bigand (1993) tested musicians and non-musicians in two experiments, described below.

Bigand (1993) was primarily concerned with the relative influence of tonal versus rhythmic structure (and their possible interaction) on the abstraction of tension/relaxation schemata. In the first study, melodies were segmented into melodic fragments (ending at different points in the melody) that varied in rhythmic and harmonic structure. The first set of melodies had the same melodic contour as the last set, but different implicit harmonies. Rhythm was the same across the sets of melodies, but was manipulated within the set. To measure the perceived stability (tension) while listening to the melodies, listeners rated the "completeness" of the melody fragments. Bigand argued that if the tonal hierarchy is driving schematic representations, ratings should differ between the first set of melodies and the second (which have different implicit tonal structures). Also, if metrical and durational structure impacts schematic processing, differences in tension should be found within the sets (which have differing rhythms and tone durations). Evidence was found to support both accounts: the harmonic structure *and* metrical structure guided listeners' abstraction of tension and relaxation in the melodies. Specifically, less tension was reported on notes that were tonally stable and had longer durations on metrically strong beats. In the second study, Bigand (1993) found that musically trained listeners perceived more nuances in melodic expectation than untrained listeners, but that both groups abstracted the patterns of tension and release in a similar manner.

In a different approach to testing melodic expectation, listeners heard part of a musical passage from Robert Schumann's "Du Ring an meinem Finger", and made predictions about what chord would likely follow (Schmuckler, 1989). This was done for 10 positions in the musical excerpt using the "probe position" technique. The expectation ratings were significantly correlated between listeners, and the "true continuation chords" (those actually next in the melodic sequence) were rated with significantly higher expectation than other continuation possibilities. These musically trained listeners were therefore fairly consistent in predicting the actual next chord in the sequence (Schmuckler, 1989). In the last study of this series, trained pianists were asked to perform how they thought unfinished melodic fragments (the same probe positions as before) would continue. There was a highly significant correlation between the first chord pianists played to continue the sequence (Schmuckler, 1989). This provides compelling evidence for the consistency of melodic expectancy across individuals. Schematic representations share specific features across listeners that govern similar processing and expectation in music. This is what enables a group of listeners to hear a new piece of music and have simultaneously elicited emotional reactions to the unfolding harmonies. It would be interesting to test whether a computational neural network would perform similarly to the pianists in terms of prediction and melodic continuation.

1.3.3 Invariance

Temporal and tonal schemata share an important property: invariance. If a listener can recognize a song from different performances, the schematic representation of the music in long-term memory must contain invariant properties (Peretz & Zatorre, 2005). A property displays relational invariance when it can be shifted or multiplied by a constant and still be recognized as an exemplar of the category. There are two main types of relational invariance in music – temporal (including tempo and rhythmic patterns) and tonal (both relative pitch translations and octave equivalence). Changing the tempo of a piece maintains the relationship of durations in the music. In other words, tempo invariance means that the same music can be recognized at different tempi because one is essentially multiplying all of the note durations by a constant. Rhythmic invariance could be considered a subset of temporal invariance, in that rhythms can often be recognized when the speed at which they are performed is changed. Tonal invariance can be found in the relationships between pitches, such that the notes in a melody, for example, can be shifted by an interval (e.g., a major third), and still be recognized as the same melody. Listeners demonstrate this sort of perceptual invariance very frequently; we hear the tune "Happy Birthday" sung in many different musical keys over the course of our lifetime, and yet we always correctly and effortlessly recognize the tune (and could do so even when lyrics are not present). A special kind of tonal invariance is octave equivalence (in this case, all of the notes in the tune would be shifted an octave). For example, an F# can be recognized at different pitch heights (F#s in different registers sound perceptually similar because they have the same *chroma*, or pitch class). Clearly, invariant properties are important to the listener for recognizing familiar music. Relational invariance is also important to musical performers, who often demonstrate the ability to play a piece in different tempos and keys. Incidentally, lack of temporal or tonal relational invariance means that the learning was not sufficiently general to successfully transfer knowledge of one tempo to another, or one key to another (Palmer, 1997). This can often be attributed to a lack of variance during exposure.

1.3.4 Schematic Processing as Perceptual Interference

In novel, uncommon situations, schematic expectations can provide an inappropriate framework that leaves the listener frustrated, bored, or confused. For example, when exposed to music from another culture that is built on a different scale system, listeners will try to cognitively organize the auditory information using their existing schemata. But because this framework is not suited to the different interval patterns or harmonic structure, the listener may not perceive the underlying structural regularities in the music, or may misremember features of the music (Dowling, 1978; also see Frances, 1958). According to Dowling (1978), "…if people in our Western European culture hear a melody from some non-Western culture using a non-Western scale, their reproductions of that melody will use their own Western scale…". An effect of the misuse of schemas on memory will be discussed in Chapter 2. This effect is also apparent in non-musical domains. In language, for example, Bartlett showed that participants who read a story containing elements they did not understand tended to recall the story's elements in more familiar terms that were in accordance with their schemata (Bartlett, 1932).

In regard to non-metrical and non-tonal sequences, studies show that information that does not fit into a schema is not comprehended as robustly as information that does.

Unstructured tonal patterns, which do not quite fit in with a listener's schema, will be recalled less accurately than structured sequences (Deutsch, 1980). There is also evidence that listeners are able to tap rhythms that fall into a metrical framework far more accurately than sequences that do not (Povel & Essens, 1985). In a study by Patel and colleagues (2005), listeners attempting to tap along to a "weakly metrical" sequence after hearing isochronous tones of the same meter were able to synchronize to the beat, but were not as accurate as in the strongly metrical condition. Listeners tried to apply their metrical schemata framework, but only with limited success.

Most often, schemata are indispensably useful for gaining a rich understanding and appreciation of music, but at times they can cause information to be misinterpreted or misremembered. This beckons an exploration of the relationship between schematic processing mechanisms and memory for music.

1.3.5 Summary of Schematic Processing in Music

Considering all of these different perspectives on schemata, we can extract (at least) five central points: First, schemata result from extensive musical experience, and involve an interaction of long-term memory associations and short-term/working memory. Second, frequently repeated patterns and musical features, from common cadences to tonality, are likely to be extracted. Third, musical schemata (comprising this knowledge of common patterns and rules) actively guide listeners' perception by creating a set of online musical expectations. Fourth, schematic processing influences what is encoded into a gist memory representation. And lastly, schemata allow for musical invariance, which makes this processing system extremely useful and robust.

1.4 Paradigms for Studying Schemata and Musical Memory

The previous sections have largely focused on behavioral approaches to music perception and memory, and understanding the function of schemas. Behavioral work is limited, however, in explaining how schemas are formed. Computational approaches offer a means of investigating the development of schemas, which in turn can lend insight into how novel musical information is perceived and encoded in memory. In addition, neuroscientific methods nicely complement behavioral and computational research by showing how differential neural processing can underlie different behavioral outcomes, and how the brain changes with experience.

The following section provides a general overview of the ways in which non-behavioral methods can inform our understanding of learning and memory in music. A more comprehensive background on computational models, information theory applications, and EEG findings will be provided at the beginning of those respective chapters of this dissertation.

1.4.1 Computational Modeling

It is very fitting to use computational approaches to help explain schematic processing and musical memory. Schematic processing is, in a sense, a distributed set of weights (or interconnected networks) pertaining to a particular domain or concept. After going through extensive training, a network's internal state reflects a statistical representation of the melodic and rhythmic patterns within the music. Once a model's internal representations are formed, it can be tested and compared to human listeners' performance on expectation and memory tasks. When the network is tested on novel input, it applies the statistical patterns it has extracted to the new repertoire – this is essentially schematic processing in action. One can also compare the learning trajectory of networks that are exposed to very different training corpora (reflecting, for example, music that varies in tonal structure), to model perception of different types of melodic sequences. Various approaches have tested the extent to which computational models are able to learn music and exercise schematic processing (through measures of expectation and prediction) in a manner similar to human listeners.

Researchers have implemented a range of symbolic and statistical approaches to model tonality and key-finding in music (e.g. Griffith, 1994; Mozer, 1994), temporal dynamics and rhythmic entrainment (Large & Palmer, 2004; Large & Kolen, 1999), and melodic expectancy (Paiement et al., 2009; Pearce & Wiggins, 2004). One of the most successful models in the field, called IDyOM for Information Dynamics of Music, is an n-gram model that features a weighted combination of higher- and lower-order models (Pearce & Wiggins, 2006). While Narmour (1992) claimed that the bottom-up processes of the Implication-Realization model were innate, Pearce and Wiggins (2006) argue that bottom-up features (namely, those giving rise to patterns of expectancy) can be accounted for from the statistical regularities that the network extracts from the input. Indeed, this unsupervised learning model has proven to be one of the most successful models of music perception that exists in the literature. For example, in one study, after training on a large corpus of folk songs, ballads, and Bach chorales, IDyOM was compared to behavioral findings on musical expectancy. Overall, the model performed similarly to human listeners, accounting for up to 83% of the variance in continuation ratings of melodies from empirical work (Pearce & Wiggins, 2006).

Modeling provides important converging evidence for how listeners perceive and process music. While empirical tests, such as those in Dowling (1978), can provide evidence for what is or is not likely to be stored in short-term memory, computational models can address *how* information can be encoded and extracted. More of this type research is needed to provide a greater understanding of how statistical regularities are learned and applied toward novel musical examples. It would be interesting and insightful, for example, to train a model on a corpus of Western tonal music and then test the network on atonal stimuli. This could shed light on how schematic processing influences the memory representation that is formed upon hearing an atonal or unstructured melody.

1.4.2 Information Theory

Information theory (IT) was developed by Claude Shannon in the 1940s, and has had far-reaching impact on everything from linguistics to genetics. IT addresses how much information is contained in a message being transmitted or stored. Entropy is a fundamental part of IT, and measures the average number of bits required to communicate or store one unit of a message. It is a quantification of the amount of *uncertainty* in predicting the value of the next unit of a sequence, and has recently been applied with notable success to the fields of music cognition and computational musicology.

With the previous discussion of melodic expectation in mind, it should be no surprise that information theoretic measures are well suited for characterizing listeners’ responses to music. Indeed, measures of surprisingness (entropy) have successfully modeled listeners’ expectations during music listening (Pearce & Wiggins, 2006). Predictive information, which measures the amount of information afforded by the current observation about future observations, has also played an important role in models of music information processing (Rohrmeier & Koelsch, 2012; Pearce & Wiggins, 2004; Abdallah & Plumbley, 2009). Information theory has contributed to some of the most successful computational models in the field, such as IDyOM (Pearce &

Wiggins, 2004), and has also been effective in predicting which tones will elicit certain event-related responses in an EEG signal, as described below.

1.4.3 Electroencephalography

Behavioral and computational methods lend insight into how the brain perceives and learns music, but neuroscientific methods are clearly needed for a more detailed and complete account. Although fMRI boasts fine spatial precision and has been used with increasingly frequency to examine brain activity during music listening, the temporal resolution (on the order of 1-2 seconds) is limiting for this type of temporal signal. Electroencephalography (EEG), however, offers fine temporal resolution (on the order of 1-2 ms), which makes this method very well suited for testing musical processing as every tone in a musical sequence is presented.

EEG has been used to study affective response in listeners, expectancy violations of tonality and harmony, differences in neural processing between musicians and non-musicians, and shared processing mechanisms between music and language, such as syntactic and semantic processing. Time-frequency approaches have been used, but many more studies have utilized methods examining event-related potentials (ERPs).

The majority of ERP studies in the area of music perception test some type of violation of musical expectation. Researchers have found that when the ending to a musical phrase is syntactically inappropriate, an early right anterior negativity (ERAN) occurs, which peaks 200ms on average after the unexpected chord (Koelsh & Friederici, 2003). This component is analogous to the ELAN in language (in response to syntactic violation of phrase structure), and the amplitude of the ERAN depends on the degree of surprise of the musical syntactic violation. The ERAN is often followed by a late frontal negative component, the N5, which has a maximum negativity 500-550ms post-onset. This component is believed to reflect the process of integrating the current tone or chord into ongoing the musical context (Steinbeis, Koelsch, & Sloboda, 2006; Koelsh & Friederici, 2003). The ERAN is more than a simple change detector, and can be distinguished from the MMN in the following manner: In studies that present an unexpected Neapolitan chord (a major chord built on the lowered 2nd scale degree) on either temporal position 3 or 5 of a sequence of five chords, the amplitude of the ERAN is modulated depending on the position of the chord (with a greater amplitude for position 5 of the sequence). This signifies sensitivity to the degree of structural violation of the unexpected element. The MMN does not display this sensitivity; its amplitude is unaffected by chord position (Koelsch, Gunter, Schroger, Tervaniemi, Sammler, & Friederici, 2001). The ERAN and N5 components can be thought of as neural responses to the violation of listeners' musical schemata. Interestingly, they have been displayed in both musicians and non-musicians, demonstrating that everyday exposure to music yields enough implicit knowledge of tonality and harmony to lead to these neural signatures of surprise without explicit training. Time-frequency approaches, such as those examining coherence and phase synchronization during music listening, can lend additional insight into music perception, but an overview of these methods and research findings will be reserved for Chapter 4.

When music EEG findings are taken together with behavioral and computational findings regarding melodic expectation and short-term memory, we can see that schematic processing creates a robust set of expectations, largely shared across listeners, that guides music perception in real time. In computational models, unexpected musical events are represented as being less probable. When human listeners' expectations are violated, memory for unexpected tones improves due to saliency, unless the music lacks predictable structure. These behavioral findings, in turn, are better understood in the context of EEG studies, which demonstrate how the brain displays heightened neural response to violations of schematic expectations.

1.5 Conclusions

In sum, although extensive research has examined short-term and long-term musical memory, little is known about the processing between these two states (that is, the trajectory of musical learning). This dissertation addresses what is likely to be stored in memory upon initial hearing of a novel melody, and how this representation becomes more richly detailed with exposure. Specifically, the following chapters address: 1) what is likely to be stored in an initial, melodic gist memory representation, 2) the effect of schematic processing on learning and memory, 3) the role of expertise in musical memory, 4) the trajectory of learning tonal patterns in music varying in structure, 5) the information theoretic properties of music that impact musical expectation and memory, and 6) changes in neural activity as music is learned over time.

CHAPTER 2

CHANGE DETECTION, SCHEMATIC PROCESSING, AND SHORT-TERM MEMORY IN MUSIC

2.1 Introduction

As discussed in the last chapter, recent research in vision has highlighted the importance of semantics, saliency, and gist representations in memory. Although the linguistic equivalent has been explored, few studies have investigated schematic short-term memory in the non-linguistic auditory domain. To this end, music has the advantage that there are relatively large differences in the expertise of individuals. We can therefore assess the contribution of expertise to the schematic processing of sounds, which is argued here to play a major role in change detection.

The present studies illuminate what is encoded into a gist memory representation by testing the degree to which listeners can detect single-tone changes to brief melodies.

In Experiment 1, musical memory was tested in both musicians and non-musicians using melodies varying in musical structure. Experiment 2 utilized a full-factorial design to examine several musical parameters, including tonality, pitch interval, metrical position, and note duration. The results suggest that listeners form a memory representation for schematically consistent tones, which may be referred to as the musical gist. In most cases, trained musicians form a more robust gist, and consequently, display greater change detection, than non-musicians. Surprisingly, however, schematically inconsistent tones in the initial melody can lead to worse change detection in musicians compared to their untrained counterparts.

2.1.1 Incomplete Memory Representations in Music

There exists clear evidence that memory for auditory details is fallible (e.g., Holst & Pezdek, 2006), as shown through studies of change deafness (Gregg & Samuel, 2008; Eramudugolla, Irvine, McAnally, Martin, & Mattingley, 2005; Vitevitch, 2003) and change detection paradigms in music (e.g., Jones, Boltz, & Kidd, 1982; Dowling & Bartlett, 1981; Cuddy, Cohen, & Miller, 1979). Work in speech perception and reading comprehension, dating back to Bartlett (1932) has provided evidence that listeners and readers do not form a completely robust memory representation. Rather, memory performance is fallible, often because recall conforms to general, gist-like properties of the stimulus type (for the same reason we are susceptible to false alarms in memory tests of word lists). In the non-speech auditory modality, change deafness research has demonstrated that change detection is facilitated when attention is directed towards the to-be-changed auditory object within the competing auditory scene (Eramudugolla et al., 2005). Further, Gregg and Samuel (2009) found that between-category changes are detected more frequently than within-category changes to sound objects, providing evidence that change detection often relies on high-level, schematic representations.

As discussed in the last chapter, schematic processing can offer a useful conceptual framework with which to understand the process of learning new music and lend insight into what is encoded in memory. Although the effect of schematic processing on memory has been demonstrated with speech and non-speech auditory objects, more research should examine non-vocal music, as this domain has the advantage of being a structured temporal sequence that is free of explicit semantic content. Whereas language uses an external reference frame to convey meaning, music does not, and therefore has the potential to uncover basic processing mechanisms of the auditory system. Furthermore, music provides a means of quantifying levels of expertise: individuals with varying years of musical training can be tested to elucidate the impact of training on musical memory and change detection. This paper aims to explore whether the high-level mechanisms underlying schematic processing and change detection in vision and speech extend more broadly to musical memory.

Several different methods may be employed to assess what is likely encoded in short-term musical memory. In a study by Welker (1982), participants listened to variations on a musical theme (not hearing the prototypical theme itself), and were instructed to abstract the "central tendency of all the melodic patterns" they heard. In a subsequent false recognition test, participants displayed the most false positives for the prototypical theme, which infers that they successfully abstracted the theme's gist from its variations (Welker, 1982). Gist memory may be more directly assessed in trained musicians using methods of recall or reproduction. In a study examining the ability to recognize invariant musical structure, professional pianists were asked to improvise on a set of melodies (Large, Palmer, & Pollack, 1995). Across pianists, these improvisations were largely based on the structurally important melodic and rhythmic events from the initial melodies (structurally important as outlined by the hierarchical models of music theory, e.g., Lerdahl and Jackendoff, 1983). This provides further evidence that people are able to extract, and in this case, reproduce through performance, a musical gist. As an account for why listeners abstract this generalized backbone of the music, Large et al (1995) suggest that memory often serves the same purpose in music as language, where the global meaning (or gist) of the conversation is more important than the underlying details. Across domains, our perceptual and memory systems largely function to remember the general semantic information, or gist, along with salient details of the percept.

Although the concept of an auditory gist has been offered in the literature (e.g., Harding, Cooke, & Konig, 2007), little is known about its content and level of detail. Therefore, because the term *gist* has a vague and fuzzy connotation, another goal of this paper is to explicate features that contribute to a musical gist, and explore whether variations in the completeness of our auditory representations are *systematic*. Two central hypotheses of the paper are that, firstly, schematic processing often dictates what will be encoded in a gist-like memory for novel music, and will therefore also lend insight to change detection performance; and secondly, trained musicians will be more likely to detect changes in brief, unfamiliar melodies than non-musicians. Previous research also suggests that atypical musical contexts, such as melodies lacking in typical tonal structure, will hinder change detection, and changes that violate tonality will facilitate change detection (Bartlett & Dowling, 1988).

In sum, the present studies explore whether the high-level mechanisms responsible for short-term memory and change detection in vision, especially schematic processing, saliency, and expertise, underlie those employed for musical change detection. The following studies concentrate on the contributions of these factors on musical change detection by arguing that listeners form a schematic, gist-like memory of music that may enhance processing efficiency, but at the expense of detail. Experiment 1 tests the impact of musical structure and musical expertise on musical change detection.

2.2 Experiment 1: Musical Structure

Just as the statistical properties of images can dictate how difficult it is to detect a change within the image, musical structure can determine music's memorability. The following study addresses the role of musical structure in change detection, while also examining whether musical training creates more robust schemata for processing musical structure, yielding better change detection.

Empirical research motivated by music theory has helped elucidate how listeners perceive musical structure. For example, tonality, the relationship between pitches in the musical key, provides listeners with a framework for perceptually organizing tones in time. This perceptual framework allows for more efficient processing and encoding of Western tonal information. Musical events at the top of the hierarchy (that is, structurally stable notes) are more likely to be encoded in memory than events low in the hierarchy. After rating the endings of musical phrases on a scale of low to high expectancy, listeners were found to subsequently remember high-expectancy melodies better than mid- and low-expectancy melodies (Schmuckler, 1997). In other words, just as in the visual modality, schematically central events (those which are highly predicted) are more strongly encoded than "peripheral" events in the domain of music (Schmuckler, 1997).

In addition to the general schematic framework that contributes to global gist memory representations, unexpected events can activate domain-general novelty detection mechanisms and be highly memorable. While listening to chord sequences, greater activation is found in a distributed network of cortical areas for deviant (out-of-key) chords, showing that unexpected events elicit larger brain responses (Koelsch, et al., 2002). We therefore hypothesize that when an unexpected, salient event draws the listener's attention, change detection should be worse than when the sound is in the perceptual background.

The role of tonal structure has been of interest within the field of music cognition for many years. In a study by Cuddy, Cohen, and Miller (1979), for example, listeners were asked to compare transpositions of melodies consisting of three tones, where the melodies were embedded within diatonic (in the musical key) or non-diatonic (outside the key) contexts, or given no additional tonal context. The sequences were either exact transpositions or differed by one tone. Change detection suffered when a non-diatonic context was present, and incorrect transpositions were not detected as reliably when the altered tone was within the key. This provides evidence that tonal structure enables effective schematic processing, and that tone changes are more detectable when they violate the surrounding tonal structure. The research did not, however, test sequences completely lacking in tonal structure, or examine the effect of changing a non-diatonic tone in the initial melody to a diatonic tone in the comparison melody.

Dowling (1978; 1991) has suggested that in addition to the key and listeners' tonal schemas, short-term musical memory is predominantly influenced by melodic contour (the pattern of rising and falling pitches in the melody). Whereas tonal schemas reflect high-level (implicit) knowledge about pitch relationships, such as how a major key sounds different than a minor key, the melodic contour contains gross information about the shape of the musical phrase. This may help create the perceptual framework upon which musical anchors can be placed in memory. The present studies therefore always maintain the same contour between melodies. In addition, whereas the research of Cuddy, Dowling, Bartlett, and colleagues most often involve the comparison of transposed melodies, absolute pitches (not relative pitches) will be used in the present experiments for a more direct comparison to studies in visual perception.

Many music perception studies examine the effect of musical expertise; that is, musicians versus non-musicians. Through training and years of performing, musicians acquire very robust schemas. Therefore, it is useful to give musicians difficult tasks that strain their schematic processing, to determine when the functionality of the schemas breaks down. Testing musicians can also be useful because of their ability to perform, recall, or transcribe what they have heard.

When musicians are asked to notate musical sequences in perceptual and memory tasks, their performance is compromised when less structure is present (Deutsch, 1980; Roberts, 1986), with both tonal and temporal structure aiding accurate transcription.

The present work further investigates the role of musical structure in perception and memory of musicians and non-musicians. In the following study, listeners heard a brief, two-measure melody followed by a comparison melody that may or may not have contained a changed tone. Unlike the studies by Cuddy and colleagues, melodies were not transposed, both to more accurately parallel studies of change blindness, and to see how listeners performed when all but one small change remained constant. On trials containing a changed tone, the smallest interval of change was one semitone, which is at least seven to ten times larger than the threshold for hearing differences in pitch (Wier, Jesteadt, & Green, 1977; Tervaniemi, Just, Koelsch, Widmann, & Schröger, 2004). Thus, the tones in isolation would never be confused with one another. This study was directed at understanding the properties of melodic contexts that prevent listeners from hearing these relatively large pitch changes.

The effects of two factors were explored on the ability to detect changes of a tone within a melody. One factor was musical structure, in which some melodies were stylistic and conformed to musical conventions, some melodies were non-stylistic, and others were generated randomly. When the tonal structure is ambiguous, schematic processing is more difficult, and less detail should be encoded in memory. Thus, when less musical structure that is present, change detection performance should decrease. The second factor was musical expertise; the performance of non-musicians was compared to professional musicians, with the expectation that musicians show superior change detection performance compared to non-musicians.

2.2.1 Method

2.2.1.1 Participants

Two groups of listeners participated in the experiment: Non-musicians and professional musicians. The non-musicians were 15 Cornell undergraduates who volunteered to participate in the experiment in exchange for extra credit in a psychology course. They had little musical training (average number of years studying an instrument = 2.9 yrs, std = 3.1 yrs). Of those who had once studied music, none of the non-musicians were currently performing.

The 11 professional musicians were members of the Indianapolis Symphony Orchestra, received $20 for their participation, and had extensive musical training and performance experience (average = 44.9 yrs, std = 8.6 yrs).

2.2.1.2 Stimuli

Seventy-two melodies were used in the experiment. They were composed specifically for the study and were unfamiliar to the listeners. The melodies were two measures long (in 4/4 time). The rhythms varied between melodies, but were subject to the constraint that each measure includes two quarter notes (long tones) and four eighth notes (short tones). The melodies were in the musical key of C, G, D, or F Major. The experiment contained 36 *stylistic* melodies as well as 36 melodies not conforming to traditional rules of Western tonal music, including 18 *non-stylistic* melodies and 18 *random* melodies. Stylistic melodies conformed to normal conventions of Western classical music. Non-stylistic melodies sounded awkward, containing strange melodic jumps or unusual tonal progressions. Random melodies were created using a random number generator to choose tones in the two octave range beginning with the tonic of the key. For example, in the key of C Major, a random number of ‘1’ would correspond to C (262 Hz), ‘2’ to D (294 Hz), ‘3’ to E (330 Hz), and so on. Examples of the three kinds of melodies are shown below in Figure 2.1.

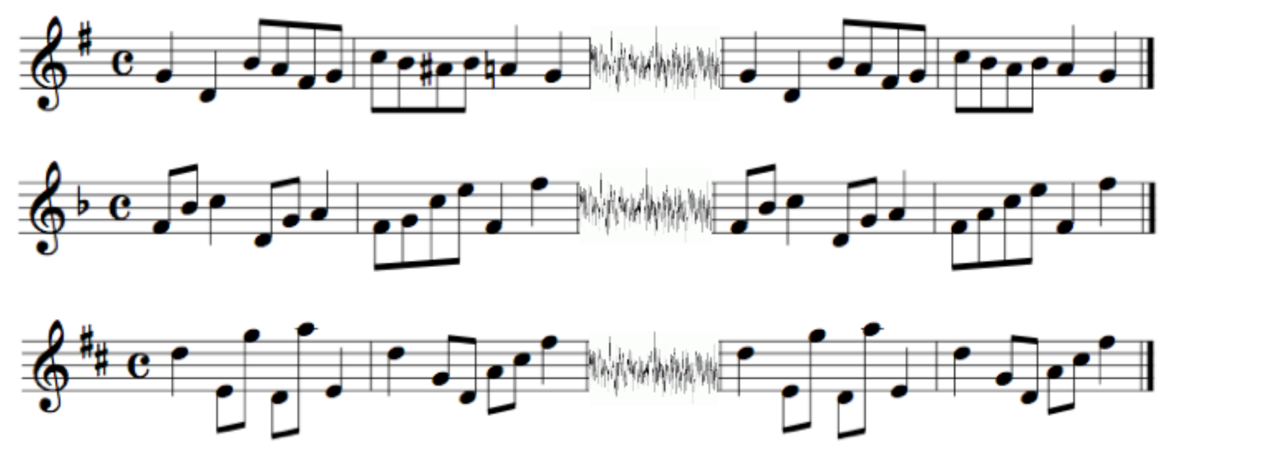


Figure 2.1. Example stimuli used in Experiment 1. Each stimulus contains a two-measure melody, 500ms of white noise, and a comparison two-measure melody. The melodies are, from top to bottom: Stylistic (change condition), non-stylistic (change condition), and random (same condition).

2.2.1.3 Apparatus

The melodies were created in Digital Performer 4.5, saved as MIDI (Musical Instrument Digital Interface) files, and then converted into .wav files using a MIDI to .wav converter. All of the melodies were of a solo piano timbre. Five hundred ms of white noise was inserted between the first and second melodies using Cool Edit 2000. The melodies were presented on a Del Inspiron laptop using E-Prime software, which randomized the 72 trials and also collected the responses. They were delivered to participants through Bose noise-canceling headphones at a comfortable listening volume.

2.2.1.4 Procedure

Participants read the instructions, and then a brief practice session began to familiarize participants with the general procedure. The three melodies in the practice session were different from those used in the experiment. The participants then heard the 72 trials of the experiment, and their task was to judge whether the two melodies in each trial were the same or different. Each trial consisted of a four-second-long melody, 500 ms of white noise, and then another four-second-long melody. The second melody was either an exact repetition of the first melody (in the *same* condition) or an altered version with one tone changed (in the *change* condition). There were six same trials per melodic condition, yielding a total of 18 same trials. In the remaining 54 change trials, the interval between the to-be-changed tone in the first melody and the corresponding changed tone in the second melody ranged from one semitone (a minor 2nd) to seven semitones (a perfect 5th). Within a trial, the rhythm of the two melodies was the same. The changed tone never occurred on the very first or very last tone, but could occur anywhere else within the two measures. Also, the change always preserved the contour (the pattern of rising and falling in pitch) of the original melody. The trials from each melodic condition, including same and change trials, were randomized and presented in one listening session. Participants were given unlimited time to respond, and upon the participant's response the next trial began. No feedback about performance accuracy was given. Responses were made on a computer interface with 6 buttons displayed with '1' labeled *absolutely sure same* and '6' labeled *absolutely sure different*; numbers in between were used to express degrees of certainty. Participants were encouraged to use the full range of the scale.

2.2.2 Results

A mixed design 3 X 2 X 2 (Melody Type X Change X Musical Expertise) ANOVA was performed on the data, with musical expertise as the between-subjects variable, and melody type and change as the within-subjects variables. Mean ratings were analyzed to determine the participants' performance; better performance consisted of ratings closer to '6' on change trials and ratings closer to '1' on same trials.

2.2.2.1 Effects of Melody Type and Musical Expertise

The results for change trials are shown in Figure 2.2. There was a significant effect of melody type, F(2,48) = 17.78, p < .001, for both professional musicians and non-musicians. There was also an effect of musical expertise, with professional musicians outperforming non-musicians, F(1,24) = 4.85, p < .05. Though professional musicians were more adept than non-musicians, both groups were able to perform the task, showing a significantly different response for change than same stimuli, F(1,24) = 76.30, p < .001. The statistical interaction between melody type and musical expertise was significant, F(2,48) = 8.44, *p* = .001. Non-musicians performed poorly on both non-stylistic change trials and random change trials, and a linear contrast showed that there was no statistical difference in performance between these conditions for non-musicians, F(1,48) = .54, *p* = .46. Professional musicians, however, performed significantly better on random change trials than non-stylistic change trials, F(1,48) = 11.80, *p* < .01. Performance on same trials (Figure 2.2B) clarifies this somewhat counter-intuitive finding.

For same trials, an interaction was present between musical expertise and melody type such that professional musicians were more likely than non-musicians to judge same random stimuli as different, F(2,48) = 6.34, *p* < .01. The professional musicians' poor performance on random same trials stemmed from a strong bias to report that these trials were changed. Consequently, the results were examined in terms of signal detection theory.

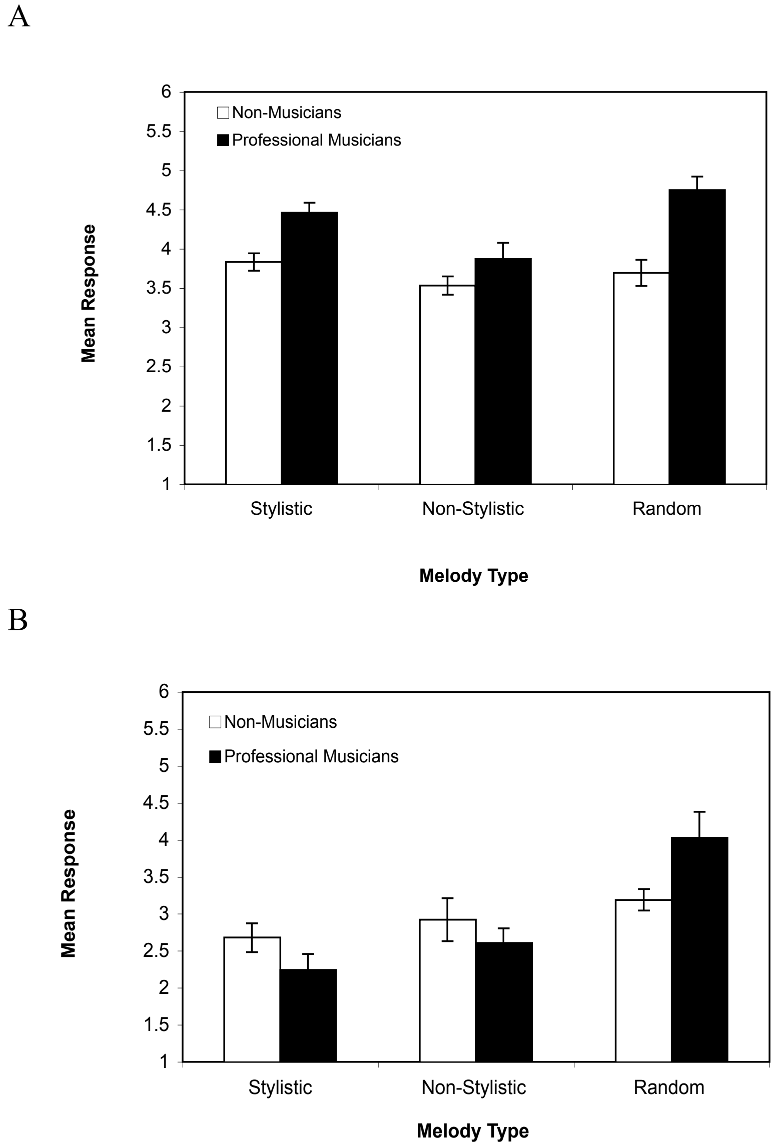


Figure 2.2. Experiment 1 mean responses for non-musicians and professional musicians across conditions. (A) Mean responses for change trials by melody type, where responses closer to '6' indicate better performance. (B) Mean responses for same trials by melody type, where responses closer to '1' indicate better performance. Error bars represent standard error of the mean.

In addition, because the response scale included a measure of certainty of response, the data were dichotomized into 1-3 responses (responded “same”) and 4-6 responses (responded “change”) simply to confirm that participants’ “sureness” did not confound the mean ratings of change detection. The dichotomized data do show the same pattern of results as the mean response data reported above. This verifies that the above findings were not, for example, due to professional musicians displaying more confidence in their responses than non-musicians. As a result, the following analyses will continue to use mean response data.

2.2.2.2 Signal Detection Analysis

As hypothesized, due to their extensive training and performance of classical music, professional musicians consistently outperformed non-musicians, $F(1, 24) = 4.85$, $p < .05$, with the exception of the random same trials, for which they were outperformed by non-musicians. Figure 2.3 shows the results of the signal detection analysis. The criterion values are plotted in the graph above and the discriminability, d', values are plotted in the graph below. Figure 2.3A shows a strong criterion shift for random melodies for the professional musicians, demonstrating that they had a large bias to judge random same melodies as different.[1] In addition to the main effect of melody type, $F(2,48) = 11.17$, $p < .001$, there was a significant interaction between melody type and musical expertise, $F(2,48) = 5.29$, $p < .01$, reflecting a criterion shift. Once this criterion shift is taken into account, one can see that professional musicians were more

[1] Although there were more Change trials than Same trials, there was no significant difference in Criterion (response bias) between Musicians and Non-musicians *within* Stylistic and Non-stylistic conditions, and no significant difference in Criterion *between* Stylistic and Non-stylistic conditions. The significantly different response bias demonstrated by Musicians for Random melodies appears to be due to a psychological bias to report “different” for these trials (note that Non-musicians did not share this bias for Random melodies).

successful overall in discriminating between same and change trials than non-musicians, F(1,24) = 6.99, p = .014. This is apparent in Figure 2.3B, which shows d', the ability to distinguish between same and change trials. There was a main effect of melody type, F(2,48) = 4.64, p = .014, with discriminability diminishing with decreased musical structure. The interaction between melody type and musical expertise was not significant, F(2,48) = 0.71, p = .499.

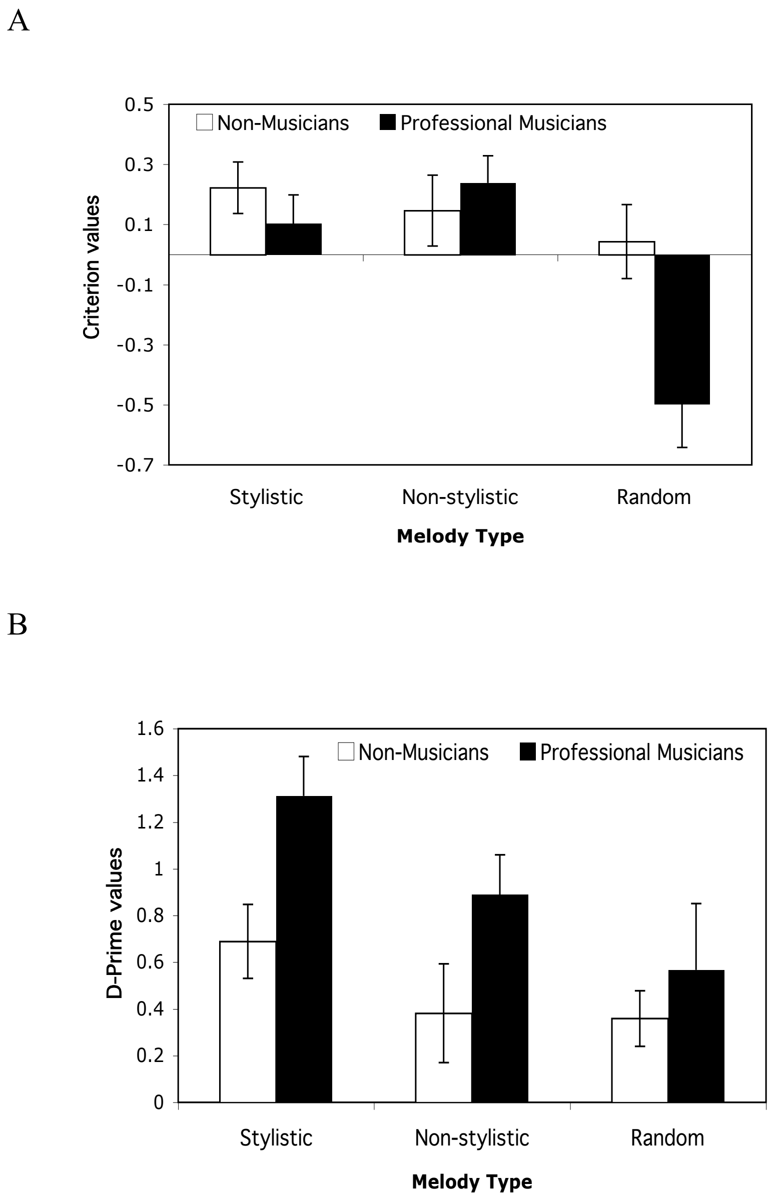


Figure 2.3. Results of signal detection theory analyses for Experiment 1. (A) Criterion values for professional and non-musicians across melody type. (B) Discriminability (d') values for professional and non-musicians across melody type. Error bars represent standard error of the mean.

2.2.2.3 Summary

Results confirmed that tonal structure has a considerable effect on listeners' ability to detect relatively large changes in melodies. All participants found the task more difficult when the musical structure was less conventional. This verifies the hypothesis that tonality is a strong factor in the global processing of melodies; when the tonality was ambiguous in the less well-structured melodies, performance on the task was impaired. Thus, it appears as though knowledge about musical style facilitates memory when the melody is conventional and hinders memory when it is unconventional. Musical expertise amplifies this effect, as musicians demonstrated better overall performance with stylistic melodies compared to non-stylistic and random melodies.[2] Musicians' poor performance with Random Same melodies indicates the powerful effect of schematic expectation during music perception: Musicians' schemata, which are more highly developed than those of non-musicians, will be inappropriate when applied to Random melodies.

Schemata help the listener organize and remember a musical percept, and performance deteriorates when the listener's schemas are inappropriate or non-applicable. Although the global characteristics of music provide insight into musical memory, examining the specific types of changes that go unnoticed can provide a richer account of what is stored in short-term musical memory.

[2] It should be noted that, in regard to the age difference between the professional musicians and non-musician undergraduates, research has shown that age and musical experience do not appear to interact with respect to musical change detection performance (experience strongly outweighs the effect of age) (Dowling, Barlett, Halpern, & Andrews, 2008).

2.3 Experiment 2: Specific Musical Factors

The importance of musical structure to memory and change detection is clear, but most often listeners will encounter music with normal (stylistic) structure. To gain a fuller understanding of which melodic properties have the greatest impact on musical memory under normal listening conditions, different musical characteristics must be thoroughly tested: Music research has shown various parameters, such as contour, tonality, and metrical emphasis, to play a role in melodic change detection. Some work has examined the effect of temporal parameters on the perception of pitch and tonality (Jones, Boltz, & Kidd, 1982), but few studies have systematically tested these parameters within one paradigm, which is necessary to assess which play the largest role in change detection. If salient musical elements are more likely to be encoded in the musical gist, then they should be remembered more accurately, and changes to these elements should be detected more reliably. Experiment 2 sought to explore which musical characteristics were most salient and reliably encoded.

This experiment used a complete factorial design with four factors: *Tonality*, the *interval* of pitch change, the *position* of pitch change, and *rhythm*. Also, because musical expertise plays a role in how efficiently and effectively music is encoded in memory, two groups were tested: Professional musicians and non-musicians. Timbre (the type of instrument playing) and dynamics (the loudness of the music) may also affect musical memory, but these factors were not manipulated in this experiment.

To assess the role of tonality in change detection, trials with scale and non-scale tones were used. On change trials, a scale tone could be changed to either a different scale tone or a non-scale tone, and a non-scale tone could be changed to a scale tone. The prediction is that a non-scale tone in the first melody will not likely be encoded in the musical gist and the change to a scale tone will be difficult to detect. In contrast, a change from a scale tone to a non-scale tone should be easy to detect, given the violation of the overall tonality in the second melody. Finally, changes from one scale tone to another may be very difficult to detect if the gist strongly encodes scale membership and less strongly encodes the particular tones in the melodies.

The interval of pitch change was systematically varied in this experiment and ranged from one to four semitones. If the gist of a melody encodes tones, not only in terms of tonality, but also in terms of the category of interval change, then the results for minor and major seconds (m2 and M2, or one and two semitones) should be similar to one another, and those of minor and major thirds (m3 and M3, or three and four semitones) should be similar to one another.

Rhythm and metrical position were manipulated to test whether metrical emphasis and note duration play an important role in detecting changes. This experiment used two different rhythms, as follows, rhythm 1: ♩ ♫ ♩ ♫ | ♩ ♫ ♩ ♫, and rhythm 2: ♫ ♩ ♫ ♩ | ♫ ♩ ♫ ♩. Either the fourth, fifth, or sixth tone, called position 1, 2, and 3, respectively, could be changed within these two rhythms. Examples of the stimuli can be found in Figure 2.4.

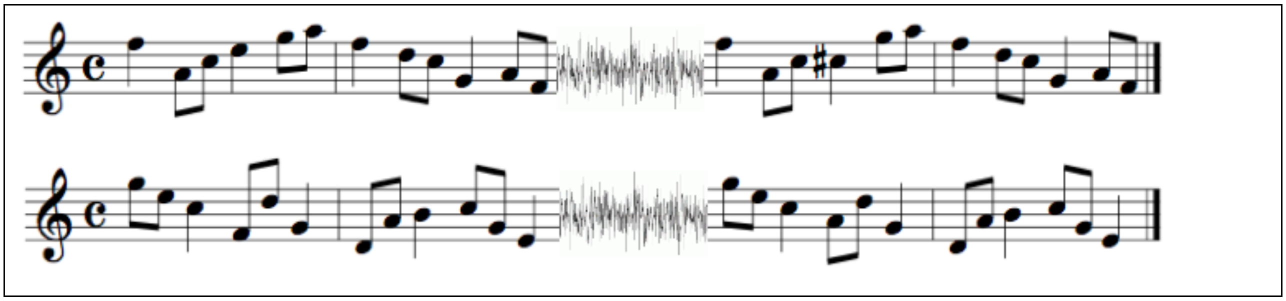


Figure 2.4. Examples of change stimuli for Experiment 2. A scale – non-scale trial containing a minor third non-scale tone change on the first metrical position of rhythm 1 is shown above. A scale – scale trial containing a major third tone change on the first metrical position of rhythm 2 is depicted beneath.

Considering the indication of the relative stress of the different metrical positions (e.g. Lerdahl & Jackendoff, 1983), metrically stressed tones were predicted to draw more attention and become encoded in the listener's gist; therefore, changes to the stressed positions of the measure should be easier to detect. In particular, position 1 of rhythm 1 (the third beat of the first measure) should be particularly strongly encoded, as this position has both the metrical emphasis of being on a "strong beat", and has a long note-duration (a quarter note).

2.3.1 Method

2.3.1.1 Participants

Two groups of listeners participated in the experiment: Non-musicians and professional musicians. The non-musicians were 20 Cornell undergraduates who volunteered to participate in the experiment in exchange for extra credit in a psychology course. They had little musical training (average years playing an instrument = 1.6 yrs, std = 1.9 yrs), and none were currently playing an instrument. The 16 professional musicians were members of the Indianapolis Symphony Orchestra, received $20 for their participation, and had many years of musical training and performance experience (average = 43.9 yrs, std = 7.4 yrs).

2.3.1.2 Stimuli

Two melodies were composed for each combination of four within-subjects variables (rhythm, interval, position, and tonality). All melodies were stylistic and in the musical key of C major. As in Experiment 1, each trial contained 2 two-measure long melodies separated by white noise, with the same timing. All together, there were 192 Change trials and 96 Same trials.

Two rhythms were used as shown: ♩ ♫ ♩ ♫ (quarter-eighth-eighth), and ♫ ♩ ♫ ♩ (eighth-eighth-quarter). Position refers to the serial position within the melody of the tone that was altered on change trials; it was one of the last three positions in the first measure of the melody (position 4, 5, or 6). Interval refers to the interval between the to-be-changed tone in the first melody and the changed tone in the second melody, which could be either 1, 2, 3, or 4 semitones (minor 2nd, major 2nd, minor 3rd, or major 3rd, respectively). The change was such that the two melodies had the same contour.

Tonality refers to whether or not the tones were in the key of C major. There were three tonality conditions. In scale – scale trials, a scale tone in the first melody was changed to another scale tone in the second melody. In scale – non-scale trials, a scale tone in the first melody was changed to a non-scale tone in the second melody. And lastly, in non-scale – scale trials, a non-scale tone in the first melody was changed to a scale tone in the second melody. The last two conditions were formed by reversing the order of the two melodies within each trial.

2.3.1.3 Procedure and Apparatus

The procedure and apparatus were the same as Experiment 1, but because this experiment contained many more trials, the experiment was presented in three 15-17 minute blocks that were counterbalanced across participants.

2.3.2 Results

The mixed design of Experiment 2 consisted of four within-subject variables, tonality, rhythm, position, and interval, as well as one between-subject factor, musical expertise. Because there are only two trials in each variable combination, it is not possible to examine dichotomized data in a meaningful way. Also, the signal detection analysis used in Experiment 1 is not tailored for this kind of multi-factorial design, especially because interactions between factors are expected.[3] Further, an omnibus ANOVA can yield spurious high-order interactions when many variables are present; therefore, a regression analysis was performed to determine which factors were most important in contributing to the overall pattern of results. The regression analysis showed highly significant effects for musical expertise F(1,2582) = 415.41, *p* < .0001, and tonality, F(2,2582) = 307.69, *p* < .0001. In addition, there was a significant effect of interval, F(3,2582) = 3.73, *p* = .01. There was no effect of rhythm or position, so the subsequent analyses collapsed over these variables. Using the factors shown to be significant through the regression analysis, a mixed design 3-way ANOVA (2 Tonality X 4 Interval X 2 Musical Expertise) was performed. Unless otherwise noted, the statistics reported below are for change trials only, and because of the quantity of data, which can inflate small effects, a p-value of .01 was chosen as the significance threshold with which to present the following results.

2.3.2.1 The Effect of Tonality and Musical Expertise

Tonality had a very large effect on the ability to detect changes, F(2, 68) = 152.34, p < .001, with changes from a scale to a non-scale tone easiest to detect. The role of musical training on the perception of tonality (shown in Figure 2.5) was also of interest. Musical expertise was

[3] That said, a SDT analysis in which all within-subjects factors were collapsed over tonality was run for comparison to Experiment 1, and to confirm the strong effect of tonality. This analysis yielded an average D-Prime of 1.47 (musicians) and 0.40 (non-musicians) for trials containing a non-scale tone, and 0.56 (musicians) and 0.26 (non-musicians) for trials containing only tones in the scale. This included significant main effects of tonality, F(1,34) = 47.08, *p* < .0001, and musical expertise, F(1,34) = 33.02, *p* < .0001, as well as a significant interaction between these two factors, F(1,34) = 25.25, *p* < .0001. In addition, tonality yielded a significant main effect on Criterion, F(1,34) = 234.40, *p* < .0001. Also, musicians displayed a significant negative bias for trials containing a non-scale tone, making the interaction between Musical Expertise and Tonality significant for Criterion as well, F(1,34) = 45.55, *p* < .0001.

found to be highly significant, F(1, 34) = 25.13, p < .001, with professional musicians outperforming non-musicians. Changes from a scale tone to a different scale tone, and changes from a scale tone to a non-scale tone were equally difficult for non-musicians to detect. Professional musicians, however, were better at detecting changes from a scale tone to a non-scale tone rather than to another scale tone. This is reflected in the significant interaction between tonality and musical expertise, F(2, 68) = 22.99, p< .001, which was expected due to the extensive musical training and experience with tonality that professional musicians acquire (see Figure 2.5).

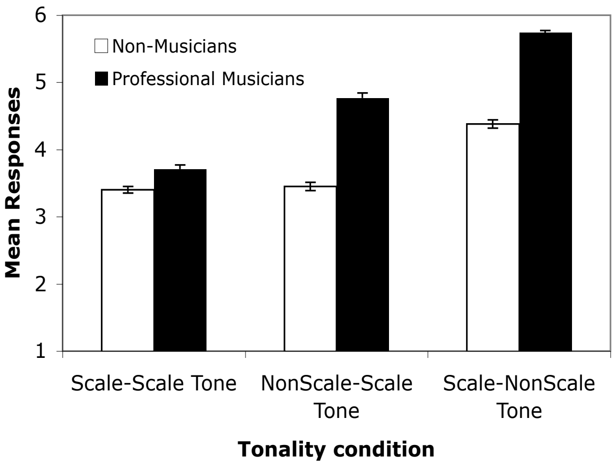


Figure 2.5. The effect of tonality on change detection. Mean responses (where '6' represents "sure of change") for professional and non-musicians across levels of tonality. Error bars represent standard error of the mean.

The interaction between tonality and musical expertise stems from the professional musicians' relatively low mean for the condition in which a scale tone is changed to another scale tone. In fact, they performed just slightly above 3.5 (the mid-point of the response range). This is a striking finding, as some of these changes are as large as four semitones (a major third), which provides additional evidence that listeners employ schematic processing which causes within-scale changes to become less noticeable. Apparently, even highly trained musicians encode melodies largely in terms of whether or not the tones are in the key – a tonal gist. Consequently, when a scale tone is substituted for another, the change goes undetected (assuming the contour of the melody is preserved).

2.3.2.2 The Effect of Interval

The interval of pitch change also played a significant role in change detection, F(3,102) = 6.64, p < .001. The interaction between interval and musical expertise was not significant, F(3,102) = 1.11, *p* = .35. As predicted, performance for both groups depended on the interval of change. Larger intervals (minor and major thirds) were detected more frequently than smaller intervals of change (minor and major seconds). Interestingly, though, changes of a major second, not a minor second, were the least detectable, perhaps due to its frequency in Western classical music (as there are five major seconds in a musical scale, and only two minor seconds, for example). A linear contrast comparing the mean of major seconds to that of minor seconds, minor thirds, and major thirds was highly significant, F(1, 102) = 15.22, p < .001.

2.3.2.3 The Interaction of Rhythm and Position

Although the regression analysis yielded no main effect of rhythm or position, initial inspection of the data did suggest an interaction between the factors in terms of metrical and durational emphasis. Therefore, a separate ANOVA was performed for rhythm and position. The ANOVA collapsed over tonality and interval, yielding a mixed 3-way ANOVA (2 Rhythm X 3 Position X 2 Musical Expertise). As expected, there was no main effect of rhythm on change detection, F(1, 34) = .015, p = .90. There was, however, a main effect of position, F(2, 68) =

10.43, p < .001, driven by the interaction between rhythm and position (see Figure 2.6). The highly significant interaction between rhythm and position, F(2, 68) = 11.15, p < .001, was due to the excellent performance for detecting change on position 1 of rhythm 1. In rhythm 1, the first position is a long tone (quarter note), while positions 2 and 3 are short tones (eighth notes). In rhythm 2, positions 1 and 2 are short tones, and although position 3 is a long tone, it does not occur on a strong beat. A linear contrast yielded significantly better performance (a higher mean response) for position 1 of rhythm 1 as compared to all other positions, F(1,68) = 26.36, p < .001. Thus, the combination of metrical and durational emphasis appears to facilitate detection of a changed tone.

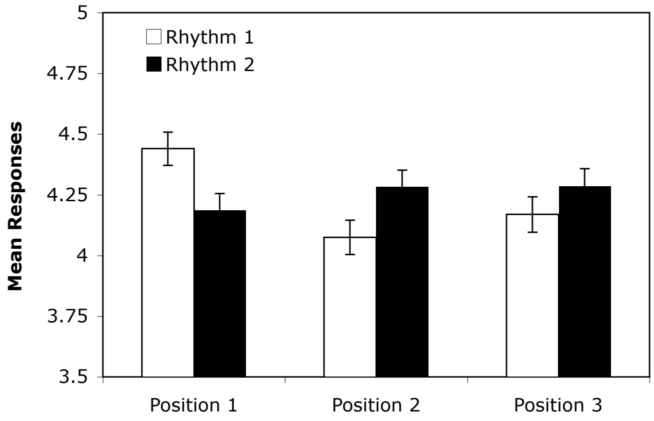


Figure 2.6. Effect of rhythm and position on change detection. Mean responses (where '6' means "sure of change") for the variables rhythm and position. Error bars represent standard error of the mean.

2.3.2.4 Same Trials

Also of interest in this study are the patterns of performance for same trials. A mixed 3-way ANOVA (3 Tonality X 2 Rhythm X 2 Musical Expertise) was performed for this purpose. Position and interval were omitted from the analysis because they had no meaning for same trials (there was no interval of change, or a position at which a tone was changed). No effect of rhythm or any interactions with rhythm were found to be statistically significant. There was, however, a significant effect of tonality, F(1,34) = 14.91, p < .001, with listeners performing worse on melodies containing a non-scale tone (they responded with a false positive). Recall that for change trials, listeners failed to detect a change in the non-scale – scale tonality condition. Because the non-scale tone in the first melody was not included in the listener's gist of the melody, changing this tone to a scale tone in the second melody often went undetected. The same effect describes this interesting finding: Listeners failed to encode the non-scale tone of the first melody, so that upon listening to the comparison melody, the non-scale tone sounded out of place (even though the comparison melody was identical to the first melody).

2.4 Discussion

The studies discussed above lend insight into the musical equivalent of findings within visual perception, and explain the failure to detect change in terms of schematic processing and gist memory. A number of interacting parameters were shown to underlie musical change detection. Listeners do not encode detailed information about all of the characteristics of music; rather, they abstract a gist based on musical schemata. Tonal and metrical structure seem to give listeners a template on which to build their gist, and when a lack of musical structure or style is present, listeners are worse at encoding features of the music. When the melodies presented in Experiment 1 were random or non-stylistic (lacking in tonal structure), both professional musicians and non-musicians could not reliably encode features of the music necessary for change detection. Further, professional musicians, who are more heavily reliant on using their internalized musical schemas, have more difficulty than non-musicians when no tonal structure is present (e.g., the random melodies in Experiment 1). This surprising finding suggests that a very elaborated schematic processing framework, when employed inappropriately, can lead to a compromised gist of the melody.

Experiment 2 investigates what is likely incorporated in memory when normal musical structure is present. Predicting the exact content of memory representations for novel music is challenging, but this study offers preliminary evidence about what is likely to become encoded in a musical gist. When hearing an unfamiliar melody, the listener encodes a general representation of the tonality (e.g., the musical key and salient tonal anchors in the melody) and melodic contour. Musicians are able to encode a greater level of detail about tonality than non-musicians due to their robust schematic processing. Musicians are more likely, for example, to encode a non-scale tone in the initial melody and detect whether a non-scale tone was present in the comparison melody. Still, pitch information is not encoded for every tone, and changes within a comparison melody will often go undetected if they uphold the gist of the tonality established by the initial melody (that is, the changed tone is within the key). These findings both replicate and extend those of Dowling (1978) and Bartlett and Dowling (1988) by showing that non-scale tone changes can create an interesting perceptual dissociation in listeners: A scale tone in the initial melody changing to a non-scale tone in the comparison melody is usually very obvious to listeners. But when a non-scale tone in the initial melody is changed to scale tone in the comparison melody, often no change is detected, especially in non-musician listeners. This is striking, because the very same melodies presented in the opposite order (in which the non-scale tone is presented in the comparison melody) almost always leads to change detection.

In addition to stressing the importance of tonal information to gist memory, Experiment 2 also demonstrates how temporal properties of music guide listeners' perception and memory.

Rhythm and metrical structure can emphasize a group of notes or a passage of music, making these sections more probable candidates for inclusion in the gist. As shown in Experiment 2, long tone durations that occur on metrically stressed beats (i.e., the downbeat) are more likely to be encoded in memory.

The present studies offer new insight into what is encoded in listeners' memory upon hearing unfamiliar music. The gist representation that is formed may provide an account of why listeners will fail at detecting changes to music (a kind of 'musical change deafness'). This research also suggests a number of promising lines for further investigation: How do these findings extend to learning longer musical sequences? How does attention mediate these effects? What are the neural correlates of musical change detection, and are there, for example, different neural signatures for tonal change detection vs. metrical change detection? Also, computational modeling has confirmed that patterns and statistical properties of music can be implicitly learned through experience (Tillman, Bharucha, & Bigand, 2000), but this method has not yet been applied to modeling gist memory for novel music. Computational approaches may yield valuable insight by essentially tracing the trajectories of gist representations through state space, and these models can be trained to reflect varying amounts of expertise.

In conclusion, the failure to detect change may be driven by similar perceptual processes across modalities. Expertise in a domain can lead to richer memory representations, but even then perceivers tend to "offload" processing demands on the stable and often highly predictable environment/context by treating the world as an external memory source (O'Regan, 1992; Spivey, 2007). This reduces the amount of information being processed and updated, and creates a more efficient means of perceiving the world. The lack of continual updating, however, will occasionally give rise to phenomena such as change blindness and change deafness. Schematic processing and gist-like memory allow for efficient and flexible processing, within the realm of music cognition and across domains more generally.

CHAPTER 3

POPULATION SPARSITY OVER THE MUSICAL LEARNING TRAJECTORY OF A SIMPLE RECURRENT NETWORK

3.1 Introduction

As discussed in the last two chapters, musical structure and schematic processing play fundamental roles in short-term memory for melodies. But how are schemata, structure, and musical rules learned over time? The following two chapters look at the trajectory of learning with increasing exposure to music. The current chapter focuses on computational approaches to this topic, examining a simple recurrent network as it gradually learns the statistical properties and tonal structure of music over time.

Computational models of music have been a useful in describing, clarifying, and predicting various aspects of music perception. Many different architectures and learning algorithms have been investigated, including supervised and unsupervised learning approaches. Supervised models utilize some type of "teacher" in which inputs are paired with desired outputs. Although this technique has been successful in modeling behavioral findings, supervised learning approaches are often considered to be less ecologically valid models of human cognition than unsupervised approaches. In unsupervised learning, the network acquires knowledge solely from the input training corpus, learning from differences between its own predictions about the input and the actual properties of the corpus. This chapter will primarily focus on Elman's Simple Recurrent Network (SRN) model (Elman, 1990), which was originally developed to process and predict the appearance of sequentially ordered stimuli, making the SRN a particularly useful model for learning the structure of music.

3.1.1 Overview of computational models of music perception

Computational methods have clarified various aspects of human music perception, such as the way in which pitch information is processed and perceived by the brain. Self-Organizing Map models (SOMs) have been used to distill the statistical regularities of music to find tonal centers and map out harmonic relationships (Tillman, Bharucha, & Bigand, 2000; Leman, 1995; Page, 1993). Unlike other artificial neural networks, these models use a neighborhood function to preserve topological features of the state space of the input (and large models using thousands of nodes can display emergent properties of the corpus). In one SOM approach, Marc Leman uses a conceptual framework of short-term and long-term memory to assess how tonal centers are found in music (Leman, 1995). Given the network's performance, Leman is able to conclude that schemata are formed from long-term exposure to music and guide perception of music in the short term (real-time listening). The way in which a model's structure mirrors the input's structure is of particular interest to the forthcoming discussion on efficient coding and sparse representation.

In another SOM model examining tonal relationships, pitch regularities were learned from mere exposure to musical sequences, and a hierarchical organization relating tones, chords, and keys was formed (Tillman, Bharucha, & Bigand, 2000). When tested and compared to empirical findings in music perception, such as probe tone ratings, the psychological distance between keys, and Dowling's (1978) study (discussed in Chapter 1), the model successfully mirrored human performance (Tillman, Bharucha, & Bigand, 2000). This model was able to not only extract statistical regularities from tonal input (to learn what Dowling would refer to as scale and contour schemata), but also successfully apply the general schematic structure it had gleaned from musical exposure to various novel test items. While performing these tests, information about the time course of schematic processing was discovered. First, bottom-up activations about tones were sent through the network. This was followed by feedback from higher-level representations propagating to lower levels and modulating their activity (the key layer influenced the chord layer, which in turn influenced the tone layer) (Tillman, Bharucha, & Bigand, 2000). The model therefore provides insight into the mechanisms responsible for schematic processing in humans; that is, we can think of tonal and harmonic processing as a cascade of information processing from high-level schemata down to low-level expectations.

Again, the notion of multiple layers of representation emerging within the network's internal structure is very interesting, and highlights these models' ability to reflect structure in the signal. Another model utilizing multiple levels of processing is called IDyOM (Information Dynamics of Music; Pearce, 2005). IDyOM is a variable-order Markov model that learns information about the sequential structure of music through unsupervised learning. After completing training, the model is given a melody, one musical element at a time, and predicts the probability of each successive element (e.g., pitch, onset time etc). Like previous work by Conklin and Witten (1995), IDyOM utilizes a multiple viewpoint framework that combines a long-term model with a short-term model. The long-term component is trained on a large music corpus to model the extensive exposure akin to an adult listener with years of experience listening to music. The short-term model is acquired during the current listening session, and reflects musical expectation based on local structure in the tune.

While the above studies compare the model's performance to previous empirical findings on tonal processing and melodic expectancy, the following study is one of very few to compare participants' short-term, gist-like memory representations with a model's extraction of gist. Large Palmer, and Pollack (1995) developed a neural network model with a complex connectionist architecture called RAAM (Recursive Auto-Associative Memory). The model was programmed to take musical input and extract its gist by applying rules derived from Lerdahl and Jackendoff's (1996) prolongational reduction analysis. The network learns to parse elements in a musical sequence and create a distributed representation of their time-span segmentation (the music's hierarchical structure). After this nested structure is encoded, the network goes through "reconstruction" algorithms (data compression) to recreate the elements of structural importance of the original sequence (the gist). In an effort to compare the network's performance on this task with human performance, skilled pianists were recruited to perform variations of the corpus of melodies through improvisation. The network's structural reduction of the melodies was in accordance with the musicians' performances, in that the features of greater importance were successfully extracted from the melodies (Large, et al., 1995). In sum, the RAAM network and behavioral findings show that a computational network can mimic musicians' extraction of a theme (the gist) from a set of variations.

The last topic of computational study worth mentioning here is composition. Composition is a useful means of testing how much musical structure a network has distilled after exposure to a training corpus. To this end, Mozer (1994) developed the CONCERT model, which was a modified Elman (1990) network. The model was trained using melodic sequences to extract the notes in the scale that would be musically appropriate and stylistic for use during composition. While ratings of this network were better than compositions chosen from a transition table, they still were still notably aesthetically lacking (Mozer, 1994). The model employed for the experiments within this chapter is also an Elman network, although one with a simpler architecture, which enables more transparent analysis of the internal state of the network.

While most studies have concentrated on the success of a network's end state or ultimate compositional ability, the following studies focus on the internal state of the network as it learns. Subjects' ratings of the network's compositions were collected and examined as a means of validating the network's learning, and diagnostic properties of the network are examined as well, such as mean squared error (MSE) and measures of efficient coding.

3.1.2 Compression and sparse coding

Compression is a coding strategy in which a minimal number of units represent a stimulus. Because these units may be consistently active over time, this representation is not necessarily *sparse*. Sparse distributed coding is a strategy in which a population of nodes completely encode a stimulus using the minimum number of active units. Taken to an extreme, this strategy is similar to the concept of a 'Grandmother Cell' that responds robustly to only one stimulus, and thus has a very low average firing rate. This is directly in contrast to a fully distributed system where every neuron takes part in encoding every stimulus and fires an average of half of the time.

Sparse coding allows a distributed system to efficiently learn and encode structure in the world. The benefits of efficient coding strategies have been reviewed in depth (Field, 1994; Olshausen and Field, 2004), but this work will concentrate on two of them. First, as shown in studies of neural systems, encoding stimuli using relatively few neurons allows for a complete representation without the biological demands of having every neuron fire (Levy & Baxter, 1996). Therefore, compressed coding will be investigated. Second, this compressed code develops in order to efficiently mirror the structure of the signal or stimuli. While there are various approaches to measuring sparse code (for example, see Willmore, Mazer, & Gallant, 2011), the present studies will focus on population sparsity.

By examining the architecture of a neural network over its learning trajectory, we can investigate how the network's coding efficiency changes with experience. Given the conventions of Western tonality (e.g. common chord progressions), as outlined by music theory, the progression of tones obeys rules and patterns. Standard transitions impose order; notes do not skip randomly around the musical state space. When a SRN receives this structured musical input, it learns how best to efficiently code the information therein.

The developing internal structure of the network is of prime concern, but of equal importance is how the network's output reflects its internally changing structure. For external validation of the network's ability to produce increasingly stylistic output with training, listeners were recruited to rate the sophistication of the network's novel compositions. This external evaluation confirmed the network's internal measures of population sparsity and learning.

3.2 Experiment 1

In this study, we tested how a Simple Recurrent Network learns tonal structure over time, with a focus on what internal changes occur in order to produce increasingly sophisticated compositions. This experiment explores the relationship between the efficiency of the SRN's hidden layer activations and the ability of the SRN to learn and predict the next note in a musical sequence. To elucidate the relationship between sparse population code and the sophistication (complexity and style) of the network's compositions, participants rated the novel compositions from several points along the learning trajectory. We hypothesized that the population sparsity of the network would increase over training, and that subject ratings would similarly increase.

3.2.1 Method

3.2.1.1 Network Architecture

Matlab software was used to program and run the SRN. The network was given one note at a time during training; it learned musical structure by predicting the next note in the sequence, and then compared its prediction with the actual next note in the training melody. The error signal (difference between predicted and actual) was then backpropogated through the network.

The network was trained on five simple, 8-measure long melodies composed specifically for this study (see Figure 3.1). They were monophonic, of a piano timbre, and contained no rhythmic variation (all of the tones were quarter notes). Notes were held at equal duration in order investigate the probabilistic distribution of tonal relationships during training.

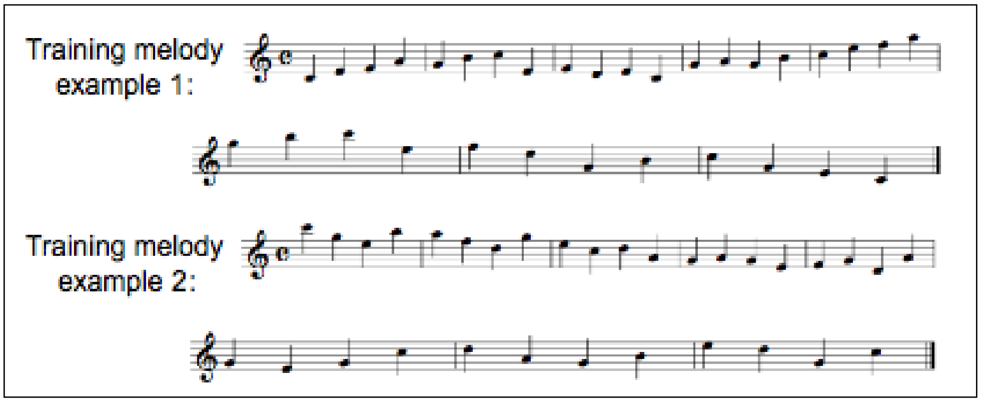


Figure 3.1. Examples of training melodies used as input.

The input and output layers of the network consisted of 15 nodes each, while the context and hidden layers contained 30 nodes (see Figure 3.2). The format of the input was such that one note (which was represented by turning on a corresponding node of the 15 present in the input layer) would be presented per timestep. For every timestep, the network predicted the next note in the training series, and each epoch of learning was comprised of 32 timesteps. The network randomly selected one of the five training melodies for every epoch. Hidden and output layer activations were transformed using a logistic function, 1/(1+e^(-x)), and varied between 0 and 1. Because the last note of one training melody is not musically related to the first note of the next training melody, the context layer activations were reset after each epoch of training. The learning rate of the network was 0.15 and momentum was 0.9 (this term is multiplied by the previous weights to compute the current weights).

Sparsity was measured in the hidden layer of each network by looking at the proportion of hidden layer nodes with an activation value greater than .3. These values were averaged over six iterations of the network, and were measured at 5, 25, 75, 150, 300 and 450 epochs.

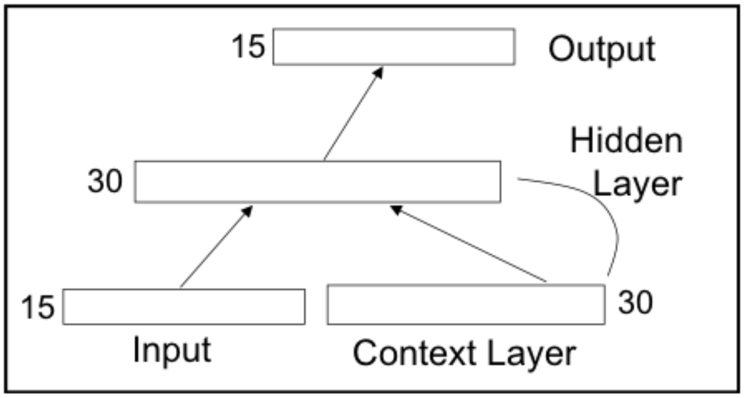


Figure 3.2. SRN architecture used in Experiment 1.

3.2.1.2 Behavioral Study 1

External validation is required to draw any conclusions regarding the relationship between increasing sparsity over training and improvement in the quality of the network's compositions. Therefore, listeners rated ten sample compositions from epochs 5, 25, 75, 150, 300, and 450. These compositions were created by inputting the note 'Middle C' at each of these benchmark epochs. The network then predicted the next note, which was in turn fed back into the network as input. This method of sequence prediction is a strength of the SRN architecture, and has been used primarily to study grammatical aspects of language (Elman, 1991).

3.2.1.3 Participants

Twenty Cornell undergraduates volunteered to participate in the experiment for extra credit in a psychology class. All participants had normal hearing, and had an average of 6.2 ± 3.7 years of musical training.

3.2.1.4 Materials

After completing a particular number of epochs of training, sixteen notes of the network's compositional output were recorded. Ten examples were recorded from each level of training (5, 25, 75, 150, 300, or 450 epochs). Each compositional sample was manually transferred from Matlab to Finale, a music software program, and converted into .wav sound files. All compositions were set to a piano timbre, and rhythm was kept constant (each tone was one quarter note in duration). Each trial consisted of a 16-note composition (four-measures in 4/4 time), and was 8 seconds in duration.

3.2.1.5 Procedure

After reading the instructions, a brief practice session consisting of four trials preceded the experiment. No feedback was given during the practice or experimental trials; the practice session simply functioned to familiarize participants to the types of melodies they would be rating. The practice trials were drawn from different points along the learning trajectory, including 5, 25, 75, 150, 300, and 450 epochs, and were different from those included in the experiment. The sixty experimental trials were completed without interruption and presented in random order using E-Prime software. After listening to each trial, the listener rated the composition on a ‘goodness’ scale from 1 to 7, where ‘1’ represented a “poor example of classical music” and ‘7’ represented an “excellent example of classical music”. Participants were urged to use the whole scale as they found appropriate. The experiment was administered on a Dell Inspiron laptop running E-Prime software, and participants wore Bose Noise Canceling headphones set to a comfortable listening volume.

3.2.2 Results and Discussion

3.2.2.1 Network Internal Structure

By examining the activations of the hidden layer at different stages along the learning trajectory, we see that efficiency increases over time. As the network completes more epochs of training, the population sparsity increases (that is, the number of active nodes in the hidden layer decreases). This trend of increasing compression is shown below in Figure 3.3.

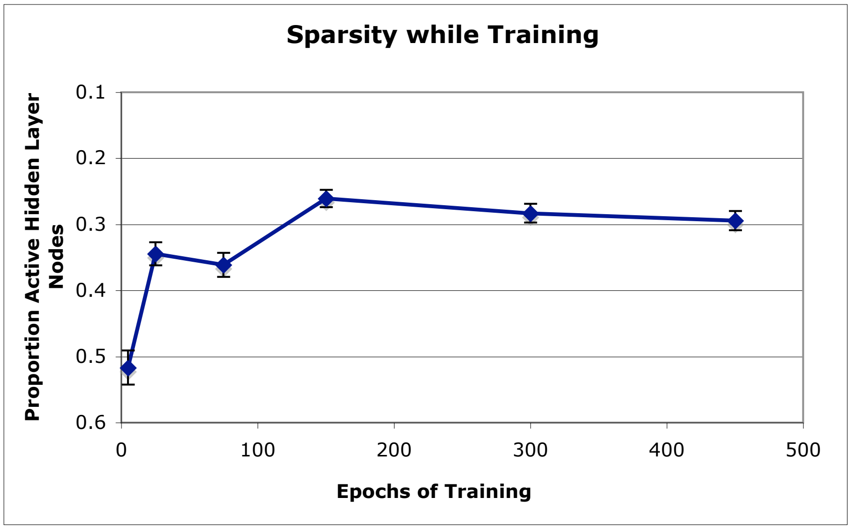


Figure 3.3. The proportion of active hidden layer nodes (population sparsity) over the learning trajectory.

As shown above, during the early stages of the network's development, there is a dramatic increase in efficiency of the hidden layer representations, as indicated by a reduction in the proportion of hidden nodes with activations greater than .3 (note inverted Y axis). Again, these values are derived by taking the average over six networks of the proportion of hidden activations above .3 (for each training epoch in question). After rapidly distilling structure from the training melodies, this decreasing trend begins to plateau around 150 epochs of training.

3.2.2.2 Behavioral Results

To assess how well the internal measure of compression corresponds to the sophistication of the network's compositions, we tested whether population sparsity was an informative predictor of listeners' goodness ratings. Indeed, listeners displayed a general preference for melodies produced after more epochs of training (see Figure 3.4).

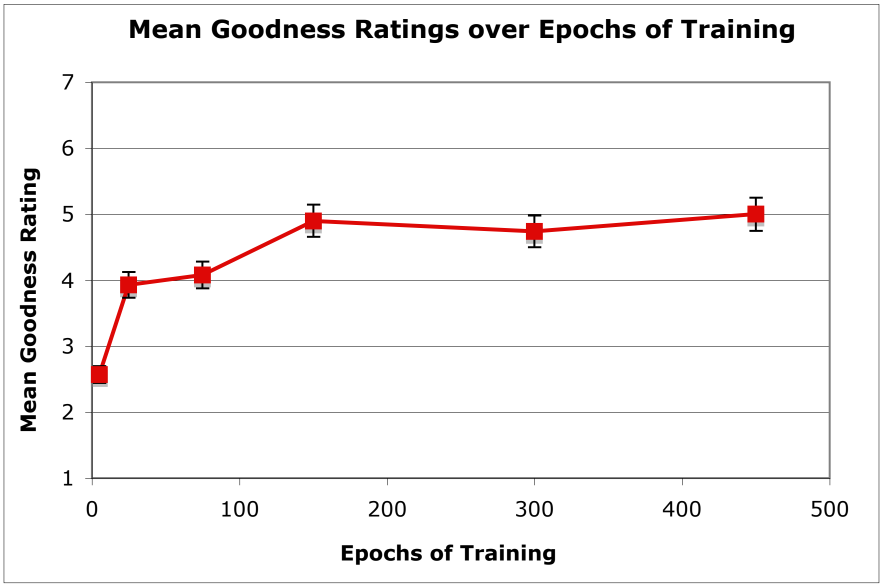


Figure 3.4. Average of listeners' goodness ratings over epochs of training.

Because the population sparsity measurements and goodness ratings followed roughly the same trend over time, population sparsity did prove to be an excellent predictor of how sophisticated the melodies sounded to listeners, $R^2 = .95$, F = 84, $p < .001$.

3.3 Experiment 2

The second experiment examines the same network structure as the first, but utilizes more complex input stimuli, many more training epochs, and employs a new sparsity metric. Three movements from J.S. Bach's Suite No.1 in G Major for Unaccompanied Violoncello were selected for the network's training input because they are musically complex and sophisticated, yet monophonic (there is a single, unaccompanied voice). The Prelude, Allemande, and Courante were chosen because they can all be performed at a similar tempo. These pieces are more complex than those used in the first experiment because each features different note durations and musical themes.

In addition to musical changes, a new sparsity metric was adopted from single-cell recording (Rolls & Tovee, 1995), in which the square of the mean activation for each node is divided by the mean of the squares (Figure 3.5). While the metric used in Experiment 1 is mostly equivalent, the Rolls sparsity metric is used pervasively in the literature. Both the previous population sparsity .3 criterion and the Rolls sparsity metric will be used to assess the efficiency of the hidden layer activations in this experiment.

$$S = \frac{\left(\frac{1}{n}\sum_{i} r_i\right)^2}{\frac{1}{n}\sum_{i} r_i^{2}}$$

Figure 3.5. Equation for Rolls sparsity metric, where n is defined here to be the number of hidden layer nodes, and r is the rate activation for each node.

3.3.1 Method

3.3.1.1 Network Architecture

The same basic SRN architecture from Experiment 1 was used in this study. Because of the increased complexity of the musical input, MIDI numbers and note durations were combined into the input for each timestep. This was encoded in the input and output by turning on one pitch node and one duration node per note. Duration values were represented by sixteen nodes, with each node being representative of a note duration ranging from a 16th note to a whole note. Due to this increase in complexity of the input (a larger pitch range and rhythmic information), the number of nodes in each layer was increased. The input and output layers now consist of 144 nodes (128 MIDI notes and 16 durations), and the hidden and context layers contain 64 nodes.

This same network architecture was used for two different training techniques. The Normal network was fed a 32-note sequence, randomly selected from one of the movements of Bach, for each epoch of training. A second network, the Bigram network, was also trained on 32 notes per epoch, but the sequence of notes lacked musical structure: After an initial note was randomly chosen from one of the movements of Bach, the network's predictions of the next note in the sequence were compared with the actual next note. Then, however, the Bigram network skipped to another random note within the musical corpus (thus, the network was only able to learn musical structure via a series of bigrams). This effectively limits the Bigram network's predictive capability to the note played immediately prior, thereby reducing the amount of structure the network is able to learn. Context layer activations were reset in both the Normal and Bigram networks after each training epoch.

The SRN was run ten times for both network types, and for all 20 of these networks, the hidden layer was captured at training epochs 5, 50, 500, 5 thousand, 50 thousand, 500 thousand, and 5 million, and the activations were used to measure the population sparsity of the network's internal structure. Also, a 32-note-long composition was created at each of these benchmark training epochs.

3.3.1.2 Behavioral study 2

3.3.1.3 Participants

Ten Cornell undergraduates volunteered to participate in the experiment for extra credit in a psychology class. All participants had normal hearing, and had an average of 2.4 ± 2.7 years of musical training.

3.3.1.4 Materials

For each level of training tested (5, 50, 500, 5 thousand, 50 thousand, 500 thousand, and 5 million epochs), ten 32-note compositions were recorded for both the Normal and Bigram networks. Each compositional sample was manually transferred from Matlab to Finale and converted into a wav sound file. The compositions were all of a piano timbre, and the compositions' rhythmic variation was included. Because of the increased complexity of the musical material, each trial consisted of a 32 tones. Due to some variation in note duration, the trials were of slightly different lengths (average length = 12 sec).

3.3.1.5 Procedure

The same procedural protocol was used as in the first study: After reading the instructions, a brief, four-trial practice session preceded the experiment. These practice trials included an example from 50, 5k, 500k, and 5m epochs, and were different from any test trials in the experiment. A total of 140 test trials were presented, with the 70 trials from the Normal network and 70 trials from the Bigram network combined into one large block of trials and presented in random order. Listeners rated each composition on a goodness scale from '1' to '7' as outlined for the first experiment. The experiment was administered on a Dell Inspiron laptop running E-Prime software, and participants wore Bose Noise Canceling headphones set to a comfortable listening volume.

3.3.2 Results and Discussion

3.3.2.1 Network Internal Structure

As predicted, the internal representations of both networks do become more efficient as the network learns structural relationships inherent in the music (see Figure 3.6). This pattern continues until roughly 1 million training epochs, even while adopting the alternative Rolls (1995) metric of sparsity.

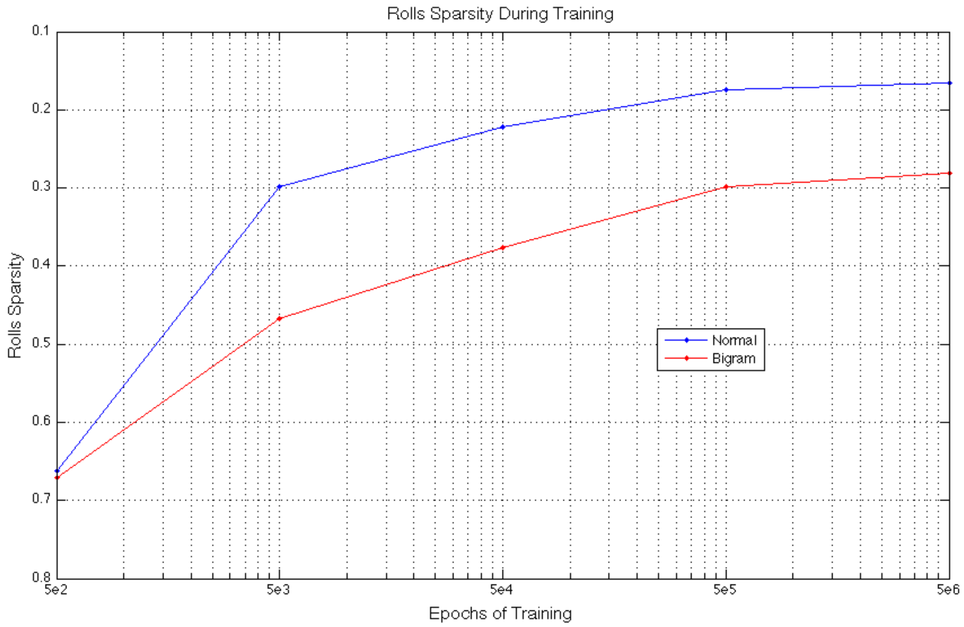


Figure 3.6. Rolls sparsity metric over epochs of training for the Normal (blue) and Bigram (red) networks.

The hidden layer of the Normal network displays greater population sparsity than that of the Bigram network. In order to shed light on the nature of the hidden layer activations of the network *while composing*, population sparsity was also examined while the network produced output. Both networks display an increase in efficiency at 5,000 epochs, but return to a less compressed state by 5 million epochs. Though both networks display similar degrees of population sparsity, the Bigram network exhibited more compression during composition at 50,000 and 500,000 epochs (see Figure 3.7). The Bigram network also created simpler melodies than those of the Normal network. This is mainly due to the fact that while the Normal network is more efficient at encoding the stylistic structure from which it is trained, it has more difficulty encoding its own output during composition. The Bigram network does not have this limitation, as the structure it learns during training is similar to what it is capable of composing. In addition, the Mean Squared Error (MSE) of both networks decayed quickly and reached a plateau with little variation by 30,000 epochs of training. The Bigram network's MSE was slightly lower than that of the Normal network.

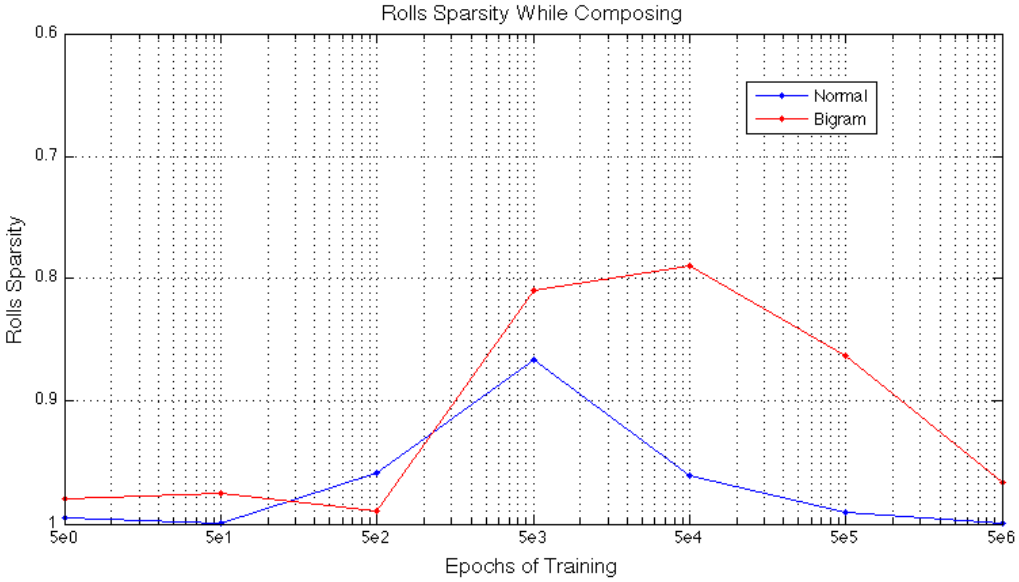


Figure 3.7. Rolls sparsity metric while composing after increasing exposure to the training corpus.

3.3.2.2 Behavioral Results

Interestingly, the compositions of the Bigram network are better rated by participants than those of the Normal network, $R^2 = .95$, $F = 19.30$, $p < .01$, as shown below in Figure 3.8.

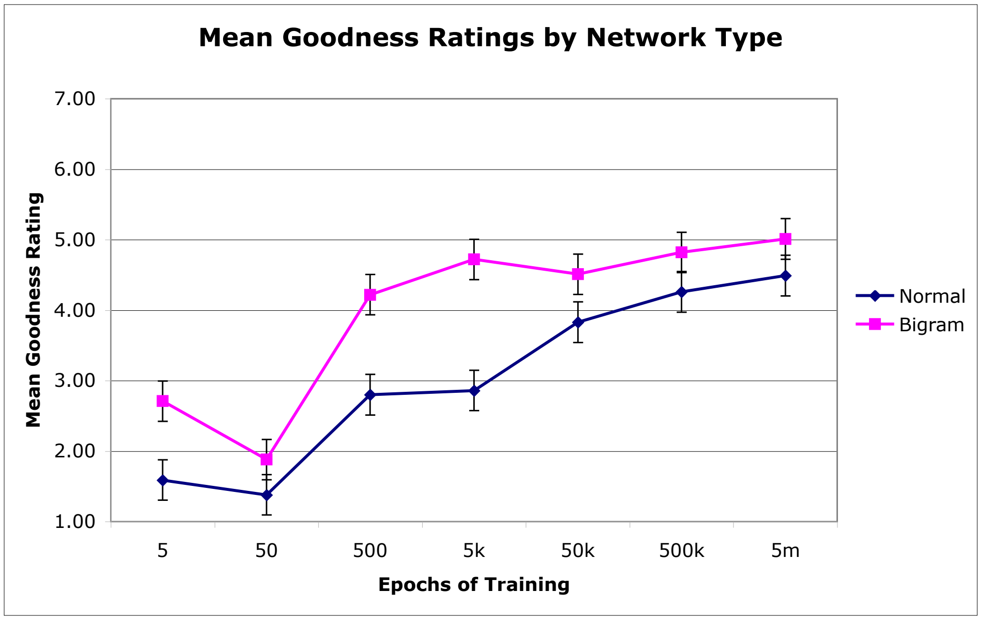


Figure 3.8. Participant mean response over epochs of training for the Normal and Bigram networks.

A comparison was made between the .3 criterion population sparsity measure and the Rolls sparsity metric (from training) in predicting the behavioral data. The .3 criterion was not a significant predictor of goodness ratings for the Normal network, R^2 = .57, F = 3.93, p = .14, but was significant for the Bigram network, R^2 = .81, F = 12.65, p < .05. The Rolls sparsity metric performed similarly: It was not a significant predictor of ratings for the Normal network, R^2 = .62, F = 4.87, p = .11, but was significant for the Bigram network, R^2 = .77, F = 9.99, p = .05.

3.4 Experiment 3

The Bigram network in Experiment 2 seemed to learn enough about tonal relationships through simple bigram information to perform somewhat similarly compared to the Normal network in terms of sparsity and Goodness ratings. This led us to question how much information could be gathered from the simple statistical distribution of tones in a corpus. That is, when given only unigram information, how well can a network perform (in terms of compositions) and how will the hidden layer change over time? To this end, the following study examines the performance of a truly random network compared to a normal network.

Although the Bach corpus of the previous experiment was more ecological and musically complex than the simple melodies of Experiment 1, the large octave range and note durations (yielding 144-node input/output layers), along with compressive hidden/context layers, yielded compositions that were less stylistic than the compositions of Experiment 2. Therefore, the following experiment used the same corpus as Experiment 1. Because the first study only examined the network to 450 epochs of training, this network was tested up to one million epochs.

When comparing Goodness ratings to both graphs of population sparsity, one can easily see that that efficiency during training is generally a better predictor of goodness ratings than efficiency during composition (the upward trend shown in goodness ratings is also displayed in sparsity during training but not composition). Therefore, only population sparsity during training was assessed in this final study.

3.4.1 Method

3.4.1.1 Network Architecture

The SRN architecture used in Experiment 1 was also used in this study. That is, the SRN had 15 input/output nodes, with a localist representation for each of the 15 notes in the two-octave range from C4 to C6 (using a C major scale), and 30 hidden/context layer nodes. The input melodies were a set of five simple melodies in the key of C, each eight measures (32 quarter notes) long. This study therefore also used isochronous tone sequences for training and compositional output (only pitch information was encoded, rhythm and note duration were held constant). Additionally, the learning rate was .15 and the momentum was .9, just as in Experiment 1.

The SRN was trained in two different ways, resulting in Normal and Random networks. Input for the Normal network consisted of the five simple melodies. The Random network was trained on a random permutation of melodies from the same corpus (to maintain the same distribution of tones in the corpus). During training, a melody was randomly chosen for every epoch, where one epoch equals 31 timesteps. In the Random network, a random permutation of the chosen melody was created before each training epoch, thereby preserving the distribution of pitches in the melody but eliminating the normal transitional probabilities between notes in the music.

Sparsity of the hidden layer activations were observed at epochs 10, 100, 1 thousand, 10 thousand, 100 thousand, and 1 million epochs of training. To gain the measurement of population sparsity using the Rolls sparsity measure as before, the network was run ten separate times to each benchmark epoch. At each of these six benchmark epochs, the network's hidden layer activations and weights were recorded, and a 16-note composition was produced.

3.4.1.2 Behavioral study 3

3.4.1.3 Participants

Ten Cornell undergraduates volunteered to participate in the experiment for extra credit in a psychology class. All participants had normal hearing.

3.4.1.4 Materials

For each level of training tested (10, 100, 1 thousand, 10 thousand, 100 thousand, and 1 million epochs), ten 16-note compositions were recorded for both the Normal and Random networks. Like the previous studies, each composition was manually transferred from Matlab to Finale and converted into a wav sound file. The compositions were all isochronous and of a piano timbre. The experiment was administered on a Dell Inspiron laptop running E-Prime software, and participants wore Bose Noise Canceling headphones set to a comfortable listening volume.

3.4.1.5 Procedure

The same procedural protocol was used as in the previous studies: After reading the instructions, a brief, four-trial practice session preceded the experiment. Then a total of 120 test trials were presented in the study, with 60 trials from the Normal network and 60 trials from the Random network combined into one large block of trials and presented in random order. Like the previous studies, listeners rated each composition on a goodness scale from '1' to '7'. The experiment was administered on a Dell Inspiron laptop running E-Prime software, and participants wore Bose Noise Canceling headphones set to a comfortable listening volume.

3.4.2 Results and Discussion

3.4.2.1 Network Internal Structure

To measure compressed coding of the hidden layer, Rolls sparsity was measured at each of the benchmark epochs of training. As one would predict, the Normal network's hidden layer activations became increasingly sparse over time. It is interesting to note that the Random network, like the Bigram network in Experiment 2, also became significantly more compressed over time, although not as much as the Normal network (see Figure 3.9 below). When more musical structure is present, the network's internal structure can be represented more efficiently.

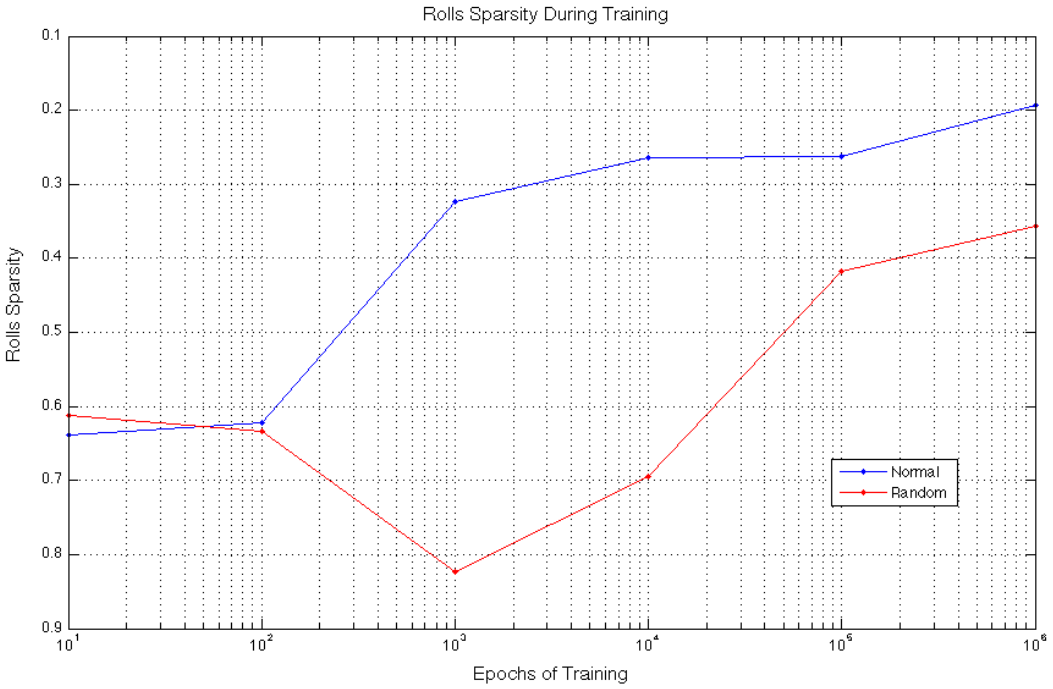


Figure 3.9. Rolls sparsity for Normal and Random networks over increasing epochs of training.

Compressed coding in the Random network can be attributed to the SRN learning the statistical probability distribution of tones to which it is exposed. Upon investigation of the network's predictions for the next tone in the sequence during training, the Random network appears to nearly always choose the "tonic" tone (a "C" in the key of C Major), which is the most statistically frequent tone in the corpus. This is the most consistent and successful means of reducing the network's MSE. Using the hidden layer activations from the six benchmark epochs from each network, were also able to make a preliminary examination of lifetime sparsity. If a network demonstrates lifetime sparsity, the activations of hidden layer nodes display not only population sparsity, but also selectivity in their response. We found preliminary evidence that the Normal network activations were selective to input; the nodes showed evidence of a distributed representation over time. This trend was not found for the Random network – a small number of nodes were active, but this compressed representation was not distributed over time (the activation of most of the nodes stayed close to 0, while a very small number of nodes remained consistently active). We cannot definitively say here that the Normal network displays lifetime sparsity, as the activations from many more epochs would be required for conclusive results, but the trend is interesting and a full analysis of lifetime sparsity will be conducted in future work.

3.4.2.2 Behavioral Results

Compositions from the Normal network are rated far better than those of the Random network. Although the internal structure of the Random network becomes more compressed over time, the Goodness ratings are poor regardless of the number of epochs of training (see Figure 3.10 below).

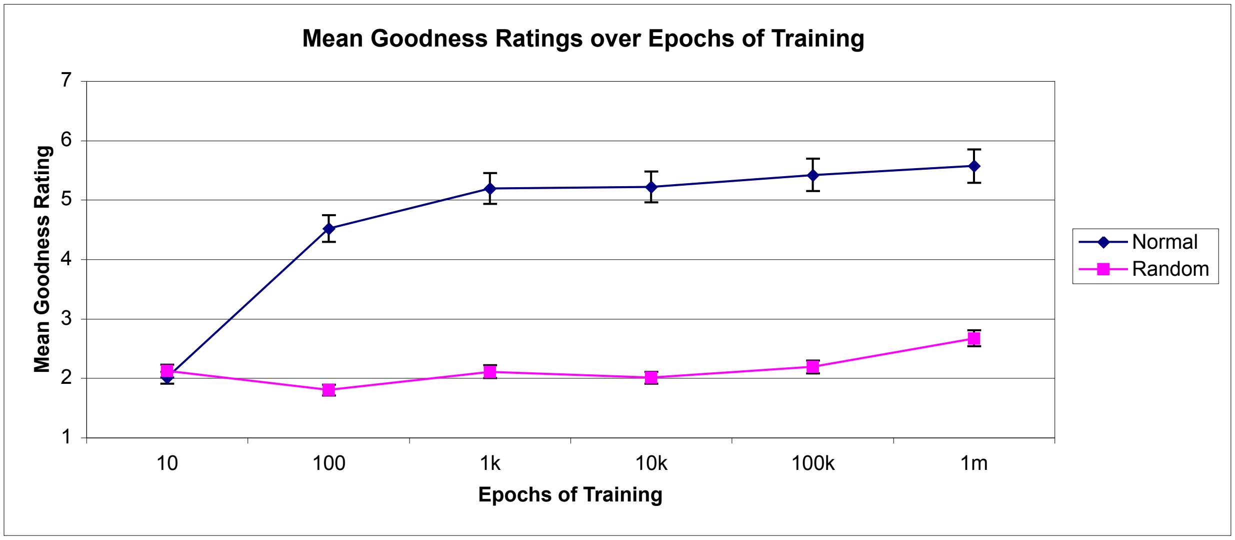


Figure 3.10. Participant mean responses over epochs of training for the Normal and Random networks.

When the population sparsity for both Normal and Random networks is correlated with goodness ratings, Rolls sparsity is shown to be a highly significant predictor of subjects' responses, $R^2 = .57$, $F = 13.00$, $p < .01$. The correlations become weaker when breaking down the correlation by network type: Normal sparsity barely reaches significance as a predictor for goodness ratings for those melodies, with $R^2 = .65$, $F = 7.43$, $p = .05$. Random sparsity, however, is not a significant predictor of Ratings, $R^2 = .29$, $F = 1.62.00$, $p = .27$. Therefore, it seems compressed coding may be a more useful heuristic of learning and compositional sophistication for structured, stylistic music. Also, the correlations between population sparsity and goodness ratings may be lower in the present study because, especially in the Normal network, the goodness ratings largely plateau after several thousand epochs of training.

3.5 General Discussion

Examining how neural networks learn musical structure can point to ways in which humans learn music. These studies provide evidence that a compressed coding strategy is the optimal way for neural network models to encode musical information.

The Normal, Bigram, and Random networks all display increasingly efficient internal representations over their developmental trajectory. Listeners' ratings follow a general increase that corresponds with the amount of training that a network has received as well as the population sparsity of the network's hidden layer while learning. While we expected that subject ratings would increase with training, the increasingly compressed representations during training shows that the learning algorithm of the networks also picked up the sparse structure of the input. While many approaches attempt to build compression and sparsity into the model, it is interesting that compressed coding simply arises in the present networks as they learn.

The structure of music may lend itself to efficient coding. Of the vast number of notes that may be selected for use in the composition, only a subset of them are appropriate given the tonal and harmonic structure. Tonality is hierarchically organized, and its foundation is centered around a particular group of tones (i.e., the tonic triad). This inherent organization can be optimally encoded with sufficient training, and it is this musical structure that is largely responsible for the strong correlation between the population sparsity that develops in the hidden layer and listeners' goodness ratings. This type of computational model can be used to capture not only the structure of music, but also the common musical patterns and rules that listeners internalize over time. Future work with this SRN will model listeners' schemata: after different types of training (modeling differing amounts and types of musical experience), the network will be prompted to produce the trajectories of schemata through state space.

Considering again the prior experiments, the Normal and Bigram networks from Experiment 2 show the difference in hidden layer efficiency that results from differing amounts of structure in the network's input. The Bigram network did exhibit less efficiency while training, a hallmark of less structure being present in the signal (because transitional relationships *between* bigrams were random). While the Normal network is more efficient during training, the Bigram network interestingly shows more compression during some stages of composition, and receives better ratings overall. This may be because while the Normal network has a more compressed representation during training, it is more likely than the Bigram network to enter into a repetitive series of notes while composing (such as the tonic triad) because it was trained on melodies with a longer musical context (utilizing information from more previous time-steps during training). Similarly, the Random network of Experiment 3 converges on a solution of selecting the most probable tone – the tonic – for prediction and composition, which impacts compression and explains the Random network's poor goodness ratings.

3.6 Future Directions

Although the previous studies confirm that the current architecture of this network is effective and informative, future work will explore how to optimize parameters of the SRN for better performance. For example, follow-up experiments can implement more recent advances in recurrent neural network architectures that encode time information in different ways. Some of the newer models used to generate and predict musical output are "Long Short Term Memory" networks (Eck & Schmidhuber, 2002) and Echo State Networks (Jaeger, 2001). Additionally, the model could use an interval-based representation rather than a pitch-based representation to examine whether differences in learning and composition would arise.

Continuing to explore the different internal characteristics of a network during composition versus training may also yield interesting results. The counterintuitive fact that the Bigram network in the second study exhibited greater compression during composition and higher goodness ratings shows that the process of composition in a SRN may be more multifaceted than previously appreciated. When a network feeds itself its own output during composition, the inherent complexity of the recurrent loop generates highly variable output that warrants further investigation. Future work will also measure the lifetime sparsity of Normal and Random networks by examining the hidden layer activations from many epochs of training. We would expect both Normal and Random networks to display population sparsity (compression), but only Normal networks to demonstrate lifetime sparsity.

3.6.1 Learning Algorithms

A network's learning algorithms are essential to how well the network can perform. A learning rate that is too low makes the network inefficient (many epochs of training are required to learn the structure of the input). Conversely, setting the learning rate very high leads to rapid learning of the patterns in the current epoch, but can create instability in the network. Learning in the current epoch can completely override traces of prior learning, called catastrophic interference. This produces a network that cannot generalize beyond one training example or learn general patterns in the larger training corpus. Therefore, finding the optimal learning rate, which likely depends on the structure and quantity of the input, is essential.

Motion through the SRN uses a logistic function, which is a fairly simple approach. A more complex but possibly more accurate means of modeling musical structure would be to use nonlinear functions instead of sigmoid transformed linear maps. The gradient descent error function can, especially when used in a high-dimensional state space, cause problems in terms of getting "caught" in local minima and converging upon non-optimal solutions. To move beyond this problem, non-gradient methods can be used, such as Expectation Maximization (Mark Andrews, personal correspondence).

3.6.2 Relative Size of Network Layers

Of considerable importance is the relative size of the input/output layers and hidden/context layers. The above model uses an expansive hidden layer, but one that compresses the input might lead to different discoveries about the representation of tonal structure. Tones that function similarly in the key may be clustered together in state space, for example.

In Elman (1990), the input and output layers were each comprised of 31 nodes, while the hidden and context layers each had 150 nodes. He also specifies that the XOR network used 6 input/output nodes and 20 hidden/context nodes. Therefore, it might prove beneficial in future implementations of this network to use hidden/context layers that are three to five times as large as the input/output layers.

3.6.3 Combining Short-term and Long-term Models

Lastly, another direction that could be extremely effective would be to employ an architecture that weights and combines multiple networks (that are the product of different types of training) in order to more accurately reflect human listeners' performance. This approach is used in one of the most successful models of musical expectation, IDyOM (Pearce, 2005). As mentioned previously, IDyOM is variable-order Markov model that employs a multiple viewpoint framework combining short- and long-term models.

The "short-term memory" network would be similar to the current SRN, but possibly with a higher learning rate, and would reflect learning for one melody. The "long-term memory" model would feature much lower momentum and learning rates, and would learn tonal structure across many different training examples. This network would represent the implicit learning that listeners demonstrate after years of exposure to Western tonal music (that is, a statistical distribution of the likelihood of tones in a key; Krumhansl, 1990). Combining these two networks will allow the model to use both well-learned statistical regularities in music, as well as local statistics from the current melody, to produce even more sophisticated musical compositions.

The benefit to this approach is that modeling tonal learning with a multi-layer SRN may yield more human-like performance than a model using a Markov-based approach. IDyOM's short-term memory layer includes a memory trace of all the previous inputs (tones or chords) from training. Arguably, this is not the most accurate model of how the brain processes and remembers tonal information. Because recurrent networks feature a parameter of decay that is reflected in the context layer, the SRN may provide a more accurate model of human memory than a Markov approach.

CHAPTER 4

INFORMATION THEORY AND ELECTROENCEPHALOGRAPHY: AN INVESTIGATION OF FACTORS CONTRIBUTING TO AND REFLECTING SUCCESSFUL MUSIC RETENTION IN ADULT LISTENERS

4.1 Introduction

We experience the world in time, dynamically finding structure in sequences of sensory events. Music is a fruitful domain for exploring the mechanisms responsible for learning structured sequences, a task that subserves a wide range of human behaviors. Research by Krumhansl (1990), Pearce & Wiggins (2006), Huron (2006), and others shows that listeners implicitly acquire knowledge about the rules and structure of music. As shown in Chapter 3, computational modeling lends insight into this process of learning over time. When a simple recurrent network is exposed to different statistical properties of music (i.e. as in the Normal, Bigram, and Random networks), tonal structure is clearly shown to play a large role in the network's compositional success, as well as its internal structure. To take this research a step further, the musical structure of the input can be manipulated according to specific parameters more nuanced than "Normal" and "Random" (which can only lead to gross differences in processing). Using computational methods, tonal and statistical structure can be manipulated more systematically to help reveal the ways in which these properties may interact and influence human learning and memory.

The following chapter examines the process of learning novel music over time, with a focus on expectation and musical structure, using two very different methodologies. The first is a behavioral study using carefully constructed tone sequences that vary across information theoretic measures (such as entropy and predictive information). Expectation ratings are collected during listening sessions, and a memory test is given after each listening session. This approach enables us to examine how the statistical structure of music, as measured by information theory, affects expectation ratings of tones, as well as memory for specific exemplars, over a period of increasing exposure. The second method uses electroencephalography (EEG) to measure the brain's electrical activity as tunes varying in musical structure are played in several listening sessions. Both event-related potential (ERP) techniques and time-frequency analyses are used to examine neural changes over time. The combination of information theory and neuroscience methods explicates the mechanisms responsible for music perception and memory.

4.2 Information Theory Behavioral Experiment

Information theory (IT) has been instrumental in explaining phenomena across a wide range of domains, such as engineering, linguistics, neurobiology, and music. Information-theoretic measures such as *entropy*, a measure of uncertainty, have successfully described and predicted how the human brain anticipates forthcoming sensory input (e.g., Manning & Schutze, 1999; Abdallah & Plumbley, 2009). Within music, an emphasis on anticipation and prediction has existed since the 1950s, and statistics-based approaches to learning have been influential for decades (consider Krumhansl & Kessler, 1982; and Saffran, Johnson, Aslin, & Newport, 1999). The probabilistic output of IDyOM (one of the computational models described in the Chapter 3), for example, has been used to derive information theoretic properties such as information content and entropy, which have been shown to accurately reflect and predict listeners' expectations (Pearce & Wiggins, 2006; Pearce et al., 2010).

While statistical and computational approaches have modeled human performance on a variety of music perception tasks, these approaches have not yet been extended to modeling the learning *trajectory* of listeners: we do not yet know how information-theoretic measures capture musical learning over increasing exposure to musical exemplars, and how much exposure is necessary to learn the statistical regularities of novel music. The following study addresses these questions.

In the present study, computational techniques were used to create a set of tone sequences varying systematically across information theoretic measures. Varying the sequences' statistical structure allows us to assess which factors have the greatest impact on music perception and memory. We focused on testing how well three information theoretic factors, *surprise* (entropy), *coding gain*, and *predictive information* (see Abdallah & Plumbley, 2009), captured listeners' expectancy of tones and memory for tone sequences. Surprise, as outlined in (Abdallah & Plumbley, 2009), is the entropy of the predictive distribution, measured as the negative log probability of *x* given the context *z*: $-\log p_{X|Z}(x|z)$. Coding gain quantifies how much the last observation (x_{t-1} not including prior history) helps the listener predict the current (known) observation x_t. Predictive information quantifies how much the current observation helps the listener predict the future (x_{t+1}) given all past observations. The average of each of these three measures was computed for every tone sequence in the present study (henceforth referred to as whole-sequence statistics).

To investigate the processes underlying musical learning, listeners were exposed to tone sequences and tested on recognition memory over several listening sessions. In each listening session, participants heard tone sequences and rated the expectedness of a tone (termed the "Probe tone") within each sequence. Probe tones varied in terms of suprisingness (-log probability) across sequences. A recognition memory test followed each listening session. This format enabled us to compute information theoretic measures for every tone sequence, and compare the effect of these measures on Probe tone ratings. We also examined how IT measures reflected recognition performance in the test sessions. We hypothesized that sequences featuring generally high-entropy would be difficult to remember, and Probe tones would be rated with lower expectancy. Because each tone sequence was presented in every listening session, we also aimed to clarify the learning *trajectory* of tone sequences; that is, how music represented in short-term memory gradually becomes more richly encoded in long-term memory, and how IT measures influence this process over time.

4.2.1 Method

4.2.1.1 Participants

Twenty-three students participated in this study for extra credit in a psychology course.

4.2.1.2 Materials and Procedure

After receiving written and verbal instructions, participants listened to three approximately 15-minute long listening sessions, each followed by a brief test session. In each of the three listening sessions, participants heard 24 tone sequences and were asked to rate the melodic expectancy of a particular tone (the Probe tone) within each sequence. This tone was identified visually on the computer screen via a clock counting down on the subsequent tones of the sequence. When the clock reached midnight, participants rated the expectancy of the concurrent tone on a scale from 1 to 5, where '1' represented highly unexpected and '5' represented very expected.

Each listening session was followed by a test session. Sixteen test stimuli were presented in each of the three test sessions, where 8 sequences were familiar (had been presented previously) and 8 were unfamiliar. After each test sequence, participants responded "Yes" or "No" to whether they had heard the sequence before. Upon responding, the listener made a confidence rating on a scale from 1 to 5 where '1' represented not confident and '5' represented very confident.

The 24 tone sequences of the listening sessions were comprised of 24 isonchronous tones, played in a piano timbre. Each tone was 500ms in duration, yielding sequences that were 12-seconds-long each. The sequences were generated with an alphabet of 7 pitches (representing one octave of the diatonic scale). A first-order Markov transition matrix was derived (Pearce, 2005) from the scale degrees of Canadian folk songs/ballads, Chorale melodies, and German folk songs in a major key (the same corpus as described in Table 2 of Pearce and Wiggins, 2006). Sequences used in the study were generated using random sampling from this transition matrix, and subsets were selected from different quadrants of the subjective 3-dimensional information space formed by the information theoretic measures described above: average Surprise (entropy), average Coding Gain, and average Predictive Information (averages were computed for each tone sequence). We use the term *subjective* because the statistical measures depend on the model's training – if the model were trained on a corpus of non-Western music, it would produce very different tone sequences that reflect the statistical properties of that genre. In addition, the IT measures are the product of an *adaptive* model; the model's predictions are updated with every musical event in order to better reflect listeners' perception and expectations.

As a perceptual "reset", a distinct 500ms white noise clip was played after every tone sequence in the listening and test sessions. The study was administered on a MacBook Pro laptop, and stimuli were presented and responses collected using Psychtoolbox (Psychophysics Toolbox Version 3) within the programing environment of MATLAB 2010a (MathWorks, Inc). Participants listened to stimuli over headphones set to a comfortable listening volume.

4.2.2 Results and Discussion

4.2.2.1 Listening Sessions

For the listening sessions, an ANOVA was performed with Probe tone, Surprise, Coding Gain, Predictive Information, and Listening Session as independent measures, and Expectation Ratings as the dependent measure. Listeners were included as a random effect in the analysis. There was a highly significant main effect of Probe Tone, F = 181.74, p < .0001, with higher-entropy tones rated as less expected. As for the whole-sequence IT measures, there were also main effects of Surprise, F = 3.92, p < .05, and Predictive Information, F = 9.67, p < .01. In addition to these main effects, there were also a significant interaction between Probe Tone and all three of the IT measures of sequences statistics: Surprise X Probe Tone, F = 22.34, p < .0001, and Coding Gain X Probe Tone, F = 35.72, p < .0001, and Predictive Information X Probe Tone, F = 91.65, p < .0001, were all highly significant. Listening Session did not contribute significantly to the results.

In the graphs provided below, the average Expectation Rating for each melody was calculated by collapsing over participants to more clearly display main effects (on a continuous rather than discrete scale). Probe Tone had the largest effect in the study and had a highly significant linear relationship with Average Expectancy Rating, R^2 = .69, F = 154.20, p < .0001. As shown in Figure 4.1 below, high expectancy tones do receive reliably higher expectation ratings than low expectancy tones (again, Probe Tone is a measure of the –log probability of the rated tone).

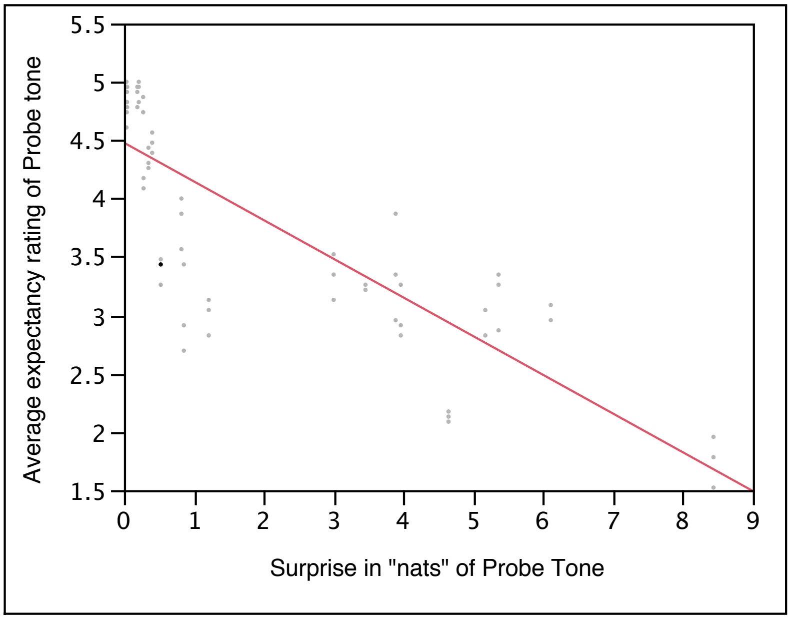


Figure 4.1. Surprise (in nats) of probe tone as a predictor of average expectancy ratings of probe tones

In terms of whole-sequence statistics, both Surprise and Predictive Information were also significant predictors of Average Expectancy Ratings. As shown in the top graph of Figure 4.2 below, Surprise (-log probability calculated for the entire tone sequence) was correlated with Average Expectancy Ratings such that more predictable sequences (lower Surprise values) yielded higher Expectancy ratings of probe tones, R^2 = .29, F = 28.87, p < .0001. The second graph of Figure 4.2 displays the correlation between Predictive Information and Average Expectancy Ratings, R^2 = .34, F = 36.23, p < .0001. The third graph shows Coding Gain and Expectancy Ratings, again significant in this analysis, R^2 = .37, F = 41.54, p < .0001.

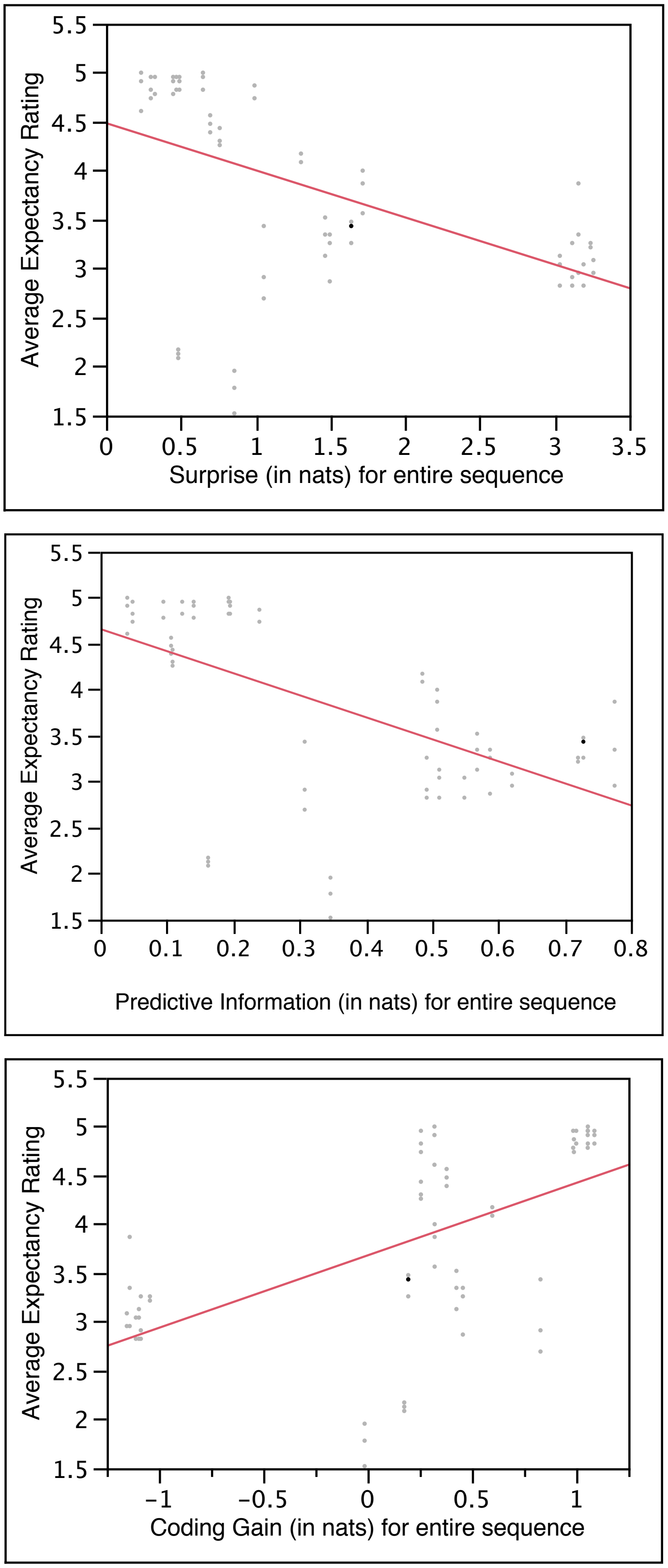


Figure 4.2. The main effects of Surprise, Predictive information, and Coding Gain on Average Expectancy Ratings during the listening sessions.

Sequences with high average Surprise values contain uncertainty; the tones comprising these sequences have high information content on average. Therefore, it is logical that sequences containing many “surprising”, unpredictable tones would yield lower expectancy ratings as shown above.

These sequences only contain 24 tones, relatively few for approximating the transition matrix created by the IDyOM model and establishing measures of predictability. Imagine sequences constructed to display high-predictive information: to be high average predictive information, each successive tone in the sequence must have high information content. Therefore, analysis of these sequences shows that they are almost indistinguishable from completely random sequences. We believe this is why participants have trouble rating tones as highly expected in high-predictive information sequences.

Coding gain is a measure of how much additional information was gained from the last observation in helping to predict the current (known) observation. Positive values for coding gain infer a *reduction* in surprisingness given the current observation. Therefore, the greater overall coding gain of the sequence, the more predictable the sequence, which suggests an increased ability to rate tones as having high expectancy.

4.2.2.2 Test Sessions

Data from the test sessions are reported in Table 4.1 below as percent correct response. Signal detection analysis was performed but yielded no significant results for D-prime or Criterion; therefore, subsequent analysis will use Percent Correct Response as the dependent measure. Chance performance would be .5, and the similarity of performance for Familiar and Unfamiliar items indicates little bias towards either response.

Listening Session	Familiar Correct	Familiar Incorrect	Unfamiliar Correct	Unfamiliar Incorrect
Session 1	0.67	0.33	0.64	0.36
Session 2	0.63	0.37	0.65	0.35
Session 3	0.70	0.30	0.65	0.35

Table 4.1. Test performance (percent correct) for familiar and unfamiliar sequences across listening sessions.

Because Yes/No responses were collected in the test sessions, a logistic regression was performed with Surprise, Coding Gain, Predictive information, Familiarity (new or old stimulus), and Listening Session as factors, and Correct Response as the dependent variable. All three whole-sequence statistics showed significant main effects: Surprise was a significant main effect, $\chi^2 = 17.10$, $p < .0001$, as well as Predictive information, $\chi^2 = 13.37$, $p < .0001$, and Coding Gain, $\chi^2 = 4.26$, $p < .05$. The only significant interaction including Familiarity was with Predictive information, $\chi^2 = 14.18$, $p < .001$. Listening Session interacted with each of the whole-sequence IT measures: Surprise X Listening Session, $\chi^2 = 7.24$, $p < .05$, Predictive Information X Listening Session, $\chi^2 = 9.42$, $p < .01$, and Coding Gain X Listening Session, $\chi^2 = 7.43$, $p < .05$, were all significant interactions.

Confidence ratings are not reported here, as they were very similar across listening sessions and stimulus types: The average confidence ratings across subjects varied between 2.80 and 3.14 for all conditions.

The logistic regression highlights the significant roles that measures of entropy and predictability have on musical learning and memory. The three information theoretic measures examined here, Surprise, Predictive information, and Coding Gain, were all significant predictors of learning over time (as evinced by their significant interactions with Listening Session). In the first Listening Session, Surprise has little effect on the correctness of participants' responses. In the subsequent listening sessions, a trend was displayed between increasing Surprise and number of correct responses (p < .01). Similarly, Coding Gain did not have a significant effect on response in the first listening session, but was *negatively* correlated (p < .01) with Correct response in the second and third listening sessions. Predictive information showed a positive correlation with Correct response that was not significant until the third listening session, in which greater predictive information led to more correct responses (p < .05). Follow-up studies need to be conducted to explore these complex information dynamics, but it is clear that the information theoretic measures investigated in this study interact dynamically with both expectancy and learning over a period of increasing exposure to novel tone sequences.

4.2.3 Future Directions of IT Research

Because it is impossible to perform an exhaustive behavioral investigation of which exemplars and rules listeners learn, computer models must be further developed to simulate and predict the process of musical learning. To this end, both IDyOM and the Simple Recurrent Network model (Agres, DeLong, & Spivey, 2009) discussed in the previous chapter will be optimized and then tested on IT measures and compared to human listeners. Future work will also test memory differences between ecologically valid melodies and experimentally controlled tone sequences with an expectation that stylistic, ecological exemplars will be more easily remembered than high entropy sequences.

4.3 Bridging Information Theory and EEG

Listeners are adept at learning the statistical rules underlying musical sequences. The present study demonstrates the difficulty in committing many novel tone sequences to memory (as shown by the relatively poor memory task performance). This may be due to the type of stimuli used; clearly listeners are able to learn a vast number of songs and themes, therefore more ecological stimuli may lead to better learning and memory performance. Also, language research (a domain in which listeners have been shown to be proficient in statistical learning of phonological sequences) has revealed that people tend to remember the *semantics* of what is said, not a verbatim account. Therefore, it may prove more insightful to test listeners' learning of semantics (musical rules and underlying statistics) across exemplars rather than the individual exemplars themselves.

An area of music cognition that warrants much more investigation is the relationship between musical rule-learning (statistical learning of musical structure and schemata) and learning of particular musical exemplars. We see from this IT study that learning individual sequences is possible, but challenging. Again, however, this may be due to the nature of the stimuli (tone sequences are difficult to commit to memory). In the following EEG studies, we use more ecological stimuli – folk and jazz tunes – with the prediction that this more typical musical structure will assist memory. In the previous IT study, it is likely that participants were learning the *rules* describing the underlying transition matrices rather than the particular exemplars themselves. Therefore, to further examine participants' ability to learn specific exemplars, and to test the use of ecological melodies, the following EEG studies probe listeners' memory for novel folk tunes (and their randomized counterparts) over time. The use of stylistic and randomized tunes allows us to test the role of structure over the course of learning. The following EEG studies explore whether repeated same-day listening sessions result in successful encoding of unfamiliar melodies, and lend insight into the neural dynamics of the learning process for these more ecological stimuli.

4.4 Electroencephalography (EEG) Experiments

To gain a more accurate account of human music learning, it is crucial to corroborate behavioral and computational approaches with studies of brain activity. Research in speech and music shows that listeners model their auditory environment to form expectations about future input. Although imaging techniques such as fMRI can show fine-grained spatial activation of brain areas underlying perceptual and cognitive tasks, the temporal resolution is insufficient to study in detail responses to individual notes in music played at natural tempi. EEG, however, permits us to study the electrical activity of the brain with excellent temporal resolution, making this method especially appropriate to examining neural responses during music listening (e.g., Williamson & Egner, 2004; Tervaniemi *et al*, 2001).

Our EEG research examines how listeners' expectations of forthcoming tones predict memory performance and influence early portions of the auditory evoked response (AER) (Naatanen, 1992). Relatively few EEG studies of music perception have investigated neural responses underlying learning as music becomes familiar with repeated exposure, and how brain responses change accordingly with the increasingly detailed musical representation that is formed in memory. Recent work has shown that low probability tones in unfamiliar, well-structured melodies elicit a negative Event-Related Potential (ERP) between 400-450ms post-tone onset, as well as increased beta band activity (Pearce et al, 2010). The following studies examine how ERP response and oscillatory activity change as melodies are repeatedly presented.

4.5 EEG Experiment 1: ERP Response

As made very clear through the previous section on information theory, prediction and expectation play a fundamental role in auditory perception. In language, expectancy effects are evident in a number of phenomena observed in the EEG literature (e.g., the MMN, N400, and P600). Researchers in music have made the connection between prediction and music perception, cognition, and emotion for decades (e.g., Meyer, 1953; Narmour, 1990; Huron, 1990). The brain does not passively process incoming information, rather, listeners continually form a predictive model of their auditory environment.

Only recently has on-line learning in the non-speech auditory domain been addressed using electroencephalography (EEG). In a statistical learning study by Abla, et al. (2008), listeners heard 'tone words' over 3 learning sessions. The transitional probabilities (TP) between the three tones of a word are greater than the TPs between words. Once these probabilities are learned, the second tone of a word should be highly predictable upon hearing the first tone. The authors hypothesized that these differences in TPs would be reflected in neural response. Indeed, they found that, after hearing the first tone of a word, listeners' N1 and N4 ERP amplitude increased while the transitional probabilities within words *were being learned*. Once these probabilities were known, the ERP amplitudes decreased (Abla, et al, 2008). This interesting finding suggests that during the process of learning, more effortful neural processing is present, but once the statistical properties of the stimulus are learned, fewer resources need to be expended.

The following study explores the impact of learning and familiarity on the amplitude of obligatory components of the auditory evoked response (AER), specifically the N1 component. We hypothesized the following: First, a difference in the AER should be observed between normally structured melodies and scrambled melodies because structured music is more predictable. As the listener learns a melody, she will form increasingly more specific predictions about the melody (e.g., what tones to expect in a subsequent musical phrase). Therefore, the amplitude of the N1 component should increase during musical learning, and become attenuated once the music is learned (see Abla et al, 2008; Loui et al, 2009; Kim et al., 2011). For melodies lacking musical structure, we should not observe a decline in the N1 amplitude, because it should be nearly impossible to form a stable predictive model. These predictions are expressed graphically in Figure 4.3 below.

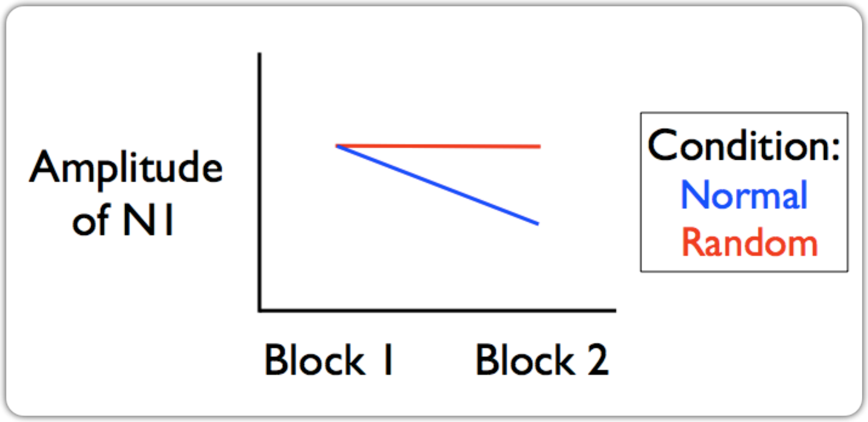


Figure 4.3. We hypothesize that the amplitude of the N1 component will gradually decline because fewer neural resources will need to be recruited as familiarity increases.

4.5.1 Method

4.5.1.1 Participants

Ten adult volunteers (6 female and 4 male) with normal hearing and minimal musical training participated in the study.

4.5.1.2 Materials and Procedure

Participants listened to monophonic Irish folk tunes, played in a plucked guitar timbre, during two listening sessions (Block 1 and Block 2). Both sessions featured the same four Normal tunes and four Randomized versions of those tunes. Normal tunes were alternated with

Random tunes, and presented in a different order for Block 1 and Block 2. There was no overt task or presentation of visual stimuli; listeners were told to focus on committing the tunes to memory. The folks songs were drawn from the Nottingham Folk Music Database.

Finale PrintMusic 2010 software was used to create WAV files of the songs. Praat software was then used to isolate the tones of the melodies and create a WAV file for each tone. The tones were either presented sequentially (in the Normal condition) or in a pre-specified random order (in the Random condition) using E-Prime Software. The tunes were played at a tempo of 90 bpm to allow enough time after tone onsets to collect ERP responses. Each listening session was approximately 12 minutes long, and listeners were able to take short breaks between listening sessions if they desired.

4.5.1.3 Data Acquisition and Preprocessing

EEG was recorded using a 128-channel EGI Hydrocel geodesic sensor net with a Cz reference. Data were sampled at 500 Hz/channel and impedances were kept below 60 kΩ by applying saline solution to dry electrode sponges when necessary. Using BESA (Brain Electrical Source Analysis, MEGIS Software, Gräfelfing, Germany) software, eye blinks were corrected (multiple source eye correction method), and channels with pervasive artifacts were spline interpolated (this was kept below 10% of channels per subject). Segments of data still containing large artifacts were rejected by hand. The data were then bandpass filtered (0.3–30 Hz) and segmented (−100 to 800 ms) to obtain ERPs before averaging. Lastly, the data were re-referenced to an average reference using Brain Vision Analyzer software (Brain Products, Munich, Germany) and filtered at 1 Hz.

4.5.1.4 Data Analysis

Brain Vision Analyzer software was used to create grand averages for both conditions. Based upon the N1 topography, peak amplitude for this obligatory component was found for every subject and condition at the fronto-central electrodes for which N1 amplitude was the greatest (electrodes numbers 6 and 11).

4.5.2 Results and Discussion

An ANOVA was conducted including the factors Listening Block (1 and 2) and Song Type (Normal and Random), with Subject as a random variable, and peak N1 amplitude as the dependent variable. For electrode 11, localized centrally and towards the front of the scalp, there was a significant main effect of Song Type, F(1,27) = 4.29, $p < .05$, and a significant interaction between Listening Block and Melody Type, F(1,27) = 21.66, $p < .0001$. For electrode 6, which is central but more posterior to electrode 11, the analysis yielded a significant interaction between Block and Melody Type, F(1,27) = 4.98, $p < .05$. As expected, the N1 amplitude decreases as participants learn the Normal melodies, but increases as listeners struggle to form an accurate predictive model for the Random melodies. This effect was apparent for frontal central channels, and the grand average ERPs for both conditions (as well as the N1 peak topography) are shown below for a channel in this region (see Figure 4.4).

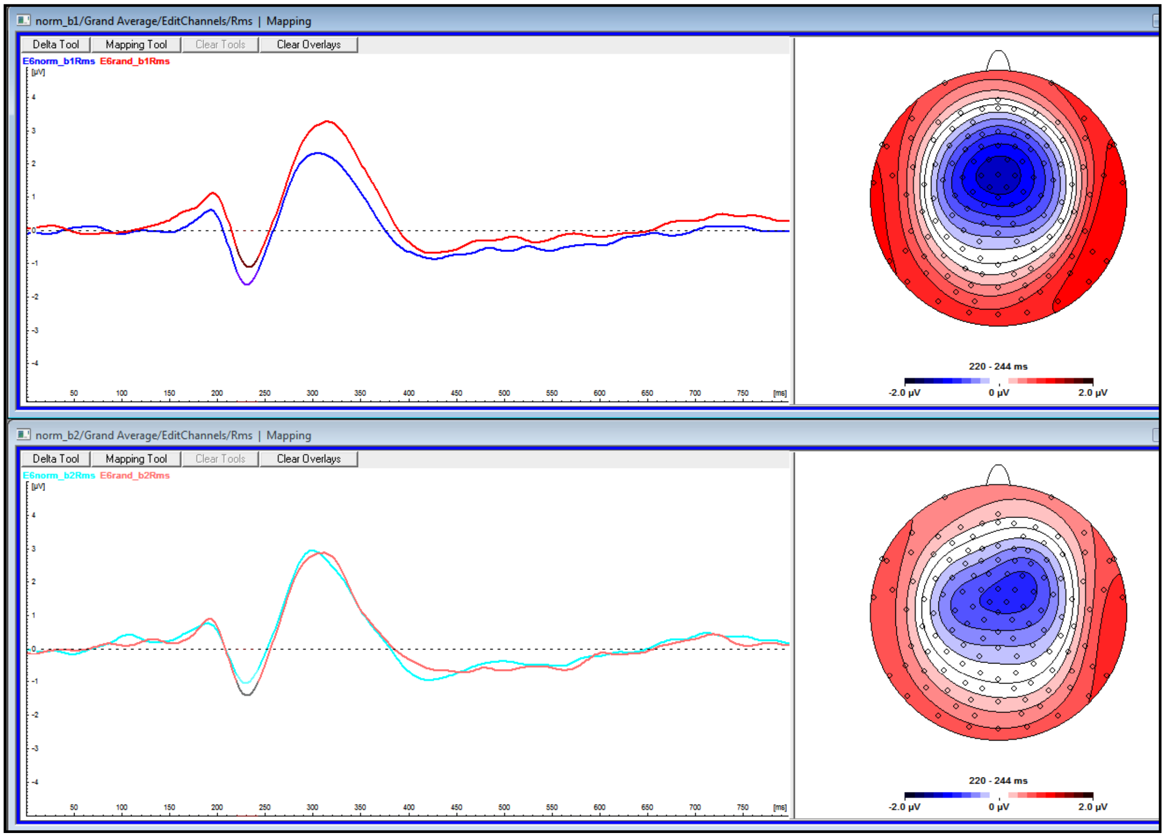


Figure 4.4. Grand average ERP responses (for electrode 6) and N1 topography per condition, across blocks. The top graph displays Normal (blue) versus Random (red) response for Block 1, and the graph below displays Normal (light blue) versus Random (pink) response for Block 2. Negative is plotted downwards.

Our results lend support to the claim that as learning occurs, and a stronger internal predictive model is formed, the brain gradually becomes more efficient (see Figure 4.5). Because Normal folk tunes are highly predictable in nature, a larger N1 amplitude was initially seen (as compared to the Random "melodies"). This amplitude decreased once the listeners became familiar with the melodies by Block 2. Conversely, the N1 amplitude for the Random melodies increased over time, presumably because it is very difficult to predict forthcoming tones in this unstructured music.

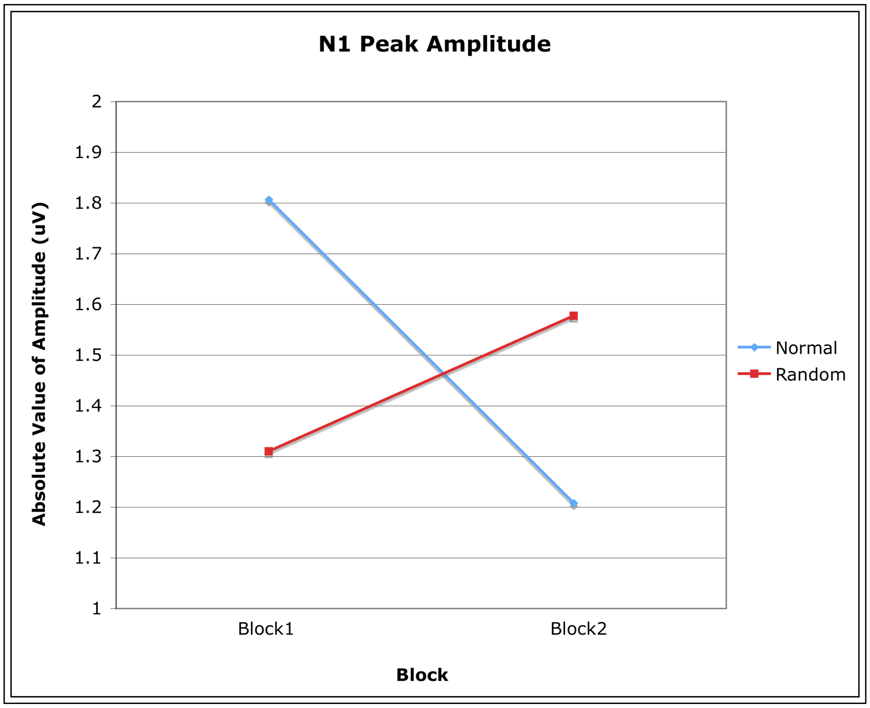


Figure 4.5. Peak amplitude of the N1 component for Normal and Random sequences.

In sum, this study provides evidence that increasing familiarity with Normal melodies results in decreased N1 amplitude. As predictability increases, the brain's response becomes more efficient. No such effect is observed for repeated Random "melodies", presumably because it is unlikely that a predictive model is formed.

4.6 EEG Experiment 2: Time-Frequency Analysis

Investigating event-related responses is only one approach to understanding how neural activity may reflect the processing demands of incoming stimuli. The oscillatory neural dynamics of music listening have been studied far less than event-related responses. Therefore, rather than focus solely on ERP techniques, we also examined changes in the power spectra of the alpha (8-12 Hz) and beta (15-30 Hz) frequency bands as Irish folk tunes became familiar over time. Contrary to previous research, which alludes to alpha as an idling of the brain, some studies have shown that an increase in alpha activity has been correlated with the number of items held in working memory (Jensen, et al, 2002). If processing Random sequences places greater demand on working memory due to the lack of musical structure, we expect to see more alpha activity for these sequences compared to Normal sequences. In the domain of music, more beta band activity has been demonstrated for improbable (high information content) compared to probable (low information content) tones within melodies (Pearce, et al, 2010). Consequently, we also expect to see greater beta band activity for randomized sequences.

In the following study, participants heard Normal and Random folk melodies over the course of three listening sessions. Participants also completed a recognition memory test after the final listening session. To summarize our hypotheses, we predict: 1) more alpha and beta activity during Random rather than Normal sequences, because it is more effortful to try to learn Random sequences; 2) better performance on the memory test for Normal compared to Random sequences; and 3) a correlation between alpha/beta band activity and performance on the memory test.

4.6.1 Method

4.6.1.1 Participants

Nineteen non-musician adult volunteers with normal hearing participated in this experiment.

4.6.1.2 Materials and Procedure

Participants listened to monophonic Irish folk tunes, played in a piano timbre, during three 16-min listening sessions (Blocks 1, 2, and 3). All three blocks featured the same six

Normal tunes and six Randomized versions of those tunes (played in a different order in each block). Like the previous experiment, there was no overt task or presentation of visual stimuli; listeners were told to stay alert and focus on memorizing the tunes. The folks songs were drawn from the Nottingham Folk Music Database.

Finale PrintMusic 2010 software was used to create WAV files of the songs. Praat software was then used to isolate the tones of the melodies and create a WAV file for each tone. The tones were either presented sequentially (in the Normal condition) or in a pre-specified random order (in the Random condition) using E-Prime Software. The tunes were played at a tempo of 120 bpm, faster than the previous study because ERP responses were not analyzed. Listeners were encouraged to take a short break between listening sessions. After the third listening session, participants ran in a brief memory test in which four-measure-long excerpts of Normal and Random songs were presented. Listeners heard 24 Normal and 24 Random excerpts, of which half were Familiar (heard during the Listening Blocks) and half were Unfamiliar. After hearing each excerpt, participants simply responded ‘Yes’ or ‘No’ as to whether they had heard the phrase before.

4.6.1.3 Data Acquisition and Preprocessing

EEG was recorded using a 128-channel EGI Hydrocel geodesic sensor net with a Cz reference. Data were sampled at 500 Hz/channel and impedances were kept below 60 kΩ by applying saline solution to dry electrode sponges when necessary. Eye blinks were corrected (multiple source eye correction method), and channels with pervasive artifacts were spline interpolated (this was kept below 10% of channels per subject) using BESA software rather than

Independent Component Analysis (ICA)[4]. Segments of data still containing large artifacts were rejected by hand. The data were then bandpass filtered (0.3 - 50 Hz), re-referenced as a 27-channel Laplacian montage, and exported for time-frequency analysis using Matlab, EEGLAB software (Arnaud Delorme and Scott Makeig, version 10), and the Chronux toolbox (developed in the laboratory of Partha Mitra). A Laplacian montage was used because, unlike an average reference, this approach preserves information about surrounding channels. Average referencing can introduce artifacts to all the channels and generally decrease the localization of sources across the scalp (compare Figure 2 and Figure S1 of Cimenser, et al, 2011). Because a Laplacian reference only utilizes activity from surrounding electrodes, it provides a more accurate account of localized neural activity.

4.6.1.4 Data Analysis

Based on the Chronux toolbox, and with the help of Dr. Andrew Goldfine (Weill Cornell Medical College), scripts were developed to run the Multi-Taper Method (Thomson, 1982) on

[4] *Use and limitations of ICA:* ICA has been widely implemented to remove blinks and other artifacts from EEG data, and is also often used to find neural signatures of perceptual and cognitive processing. This analysis is well suited for ERP studies that have brief trials, and works best with relatively clean data. In my current EEG studies, I perform time-frequency analyses on relatively long swaths of data (several-minute long datasets), in which inconsistencies are present (different types of artifacts occur). Because these data are quite noisy in some cases, ICA was found to not be an adequate processing tool. For example, the first component in an ICA analysis of EEG data should almost always isolate the blink artifact. But due to noise in the data (bad channels and muscular artifacts), the blink artifact was distributed over many components, making it nearly impossible to remove blinks without potentially disrupting neural data.

In addition to the difficulty of isolating blink artifacts, ICA was found to distort the oscillatory activity of interest in my experiments. To reduce the amount of noise and inconsistencies in the data (to improve the efficacy of ICA), ICA was run on data from one listening session/condition at a time. Although less data produces cleaner components, a major confound is introduced via this approach: By running ICA separately for each condition/listening session, and because the artifactual components have traces of neural activity, I would not be able to claim that differences between conditions were due to different processing mechanisms in the brain. In other words, by removing (different) noisy blink components for each dataset, more alpha activity (for example) may have been inadvertently removed in one dataset compared to another. Because running ICA on an entire subject's data (all listening sessions and both conditions) is not an option due to all the different types of artifacts present in that quantity of data, ICA was determined to me an in-effective tool for this type of study/analysis. Therefore, preprocessing of data was only possible using BESA's blink and artifact correction tools.

these data in order to obtain accurate power spectra (PS) of oscillatory activity during the listening sessions. A frequency resolution of 2 Hz was used. Also, a Two-Group Test was performed on the data to isolate significant frequency differences in the PS between conditions and blocks.

4.6.2 Results and Discussion

A regression analysis of Melody Type (with Subject as a random factor) and Memory Performance yielded a significant difference in Hit Rate (correctly identifying a sequence as "familiar") between Normal and Random sequences, F(1,18) = 24.61, p < .0001, with listeners performing better on Normal trials. The average Hit rate for Normal sequences was 82% (+/-3.0% std error), whereas the average Hit rate for Random sequences was at chance, 53.5% (+/-4.4% std error). There was no significant difference in D-Prime, however there was a main effect of Criterion, F(1,18) = 51.42, p < .0001, with an average Criterion value of -.79 Normal sequences and .25 for Random sequences.

Preliminary time-frequency results have not yielded a significant correlation between memory performance measures and average alpha or beta band activity, possibly because only one recognition memory test was given at the end of the entire experiment (as opposed to testing memory after each listening session). The average peak oscillatory activity for alpha and beta across subjects is shown below in Figure 4.6.

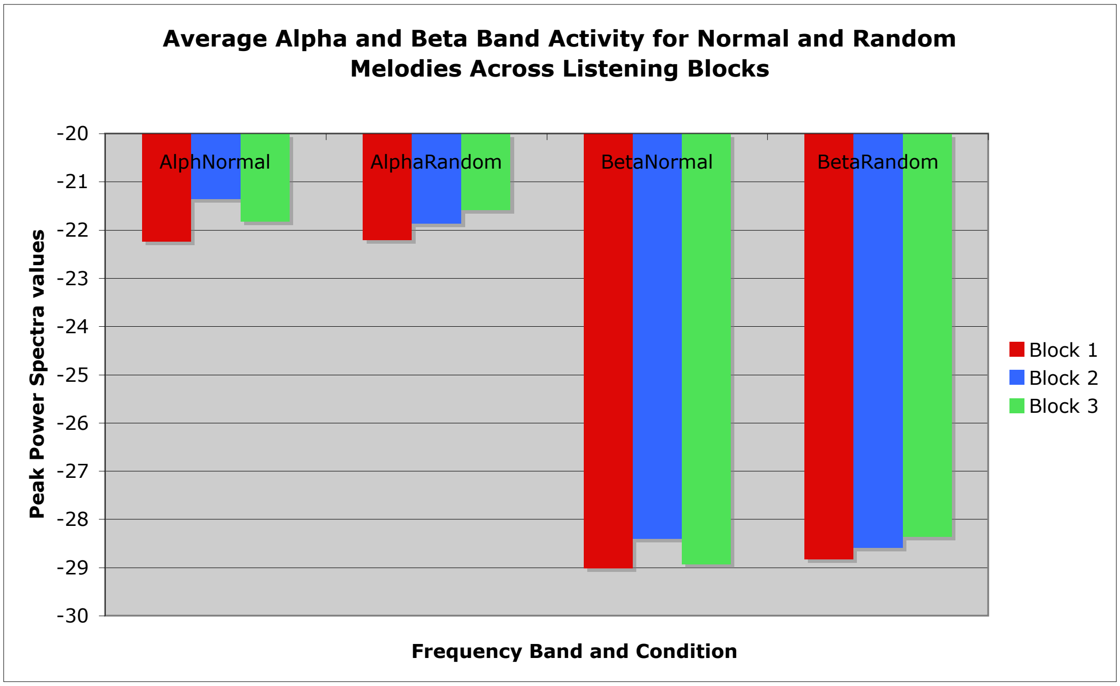
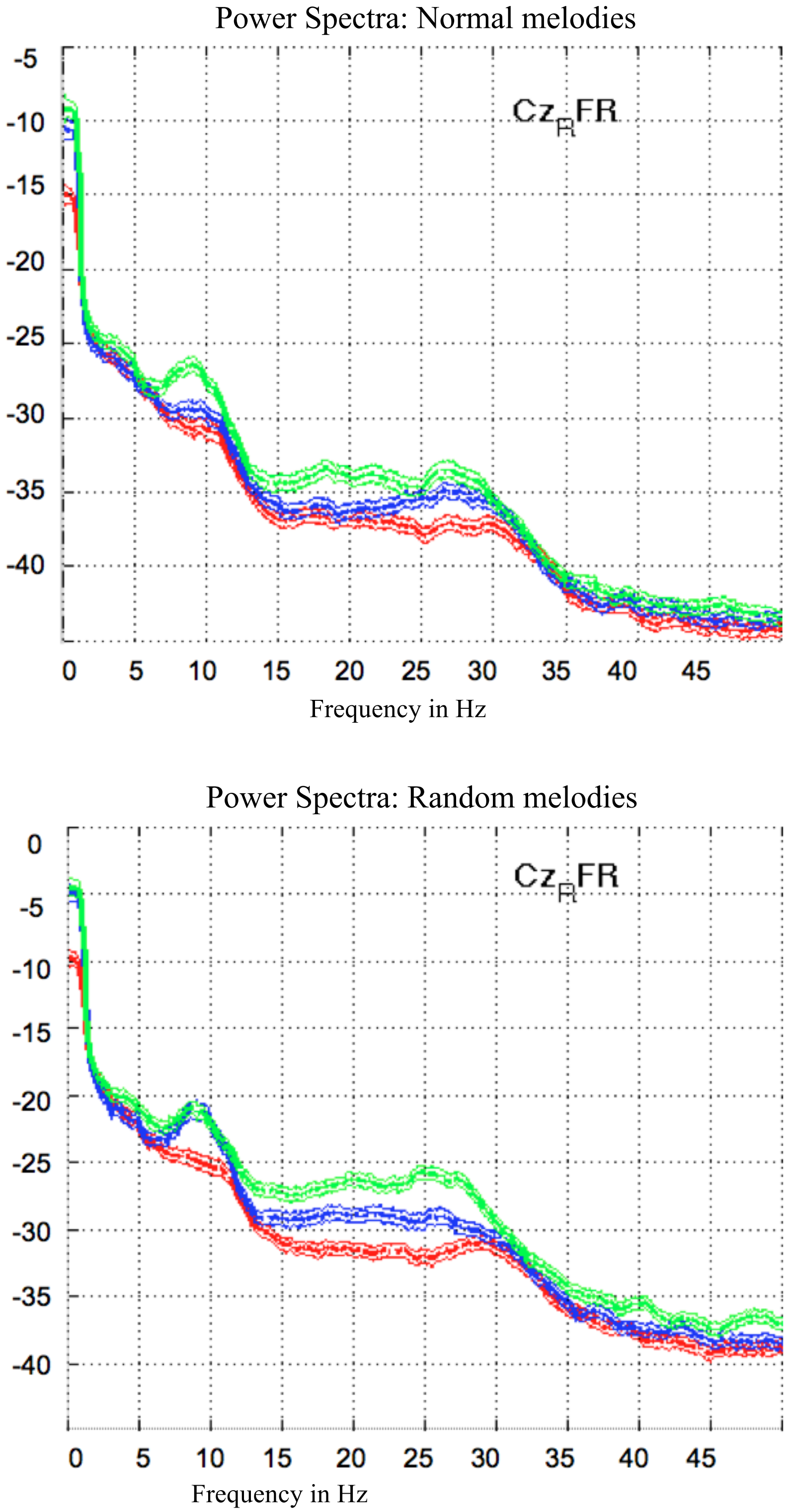


Figure 4.6. Peak values for average oscillatory activity for Normal and Random sequences in each of the three listening blocks.

Despite these negative findings, the results are still promising, as it appears that significant differences in alpha and beta band activity between conditions may have been washed out due to individual differences. One listener's power spectra results are shown below in Figure 4.7, in which significant differences between blocks are clearly shown in alpha activity (between 6-12Hz) and beta activity (15-30Hz).


Power Spectra: Normal melodies

Power Spectra: Random melodies

Figure 4.7. Example power spectra from one subject at central electrode Cz for listening sessions 1 (red), 2 (blue), and 3 (green). Top graph: Normal melodies. Bottom graph: Random melodies.

The participant above displayed increasing alpha and beta activity over the course of the three listening blocks, with greater overall power in the alpha and beta bands for Random sequences. Also, increased alpha activity was elicited during blocks 2 and 3 compared to block 1 for Random sequences. This increased activity is not shown until block 3 for Normal sequences. In addition, there were greater increases in beta activity over the course of the experiment for Random compared to Normal sequences. These findings need clarification through follow-up research, but give preliminary evidence supporting increased alpha and beta activity as sequences are learned over time. The greater increase in oscillatory activity for Random compared to Normal sequences shown above may be due to the higher processing demands of learning sequences lacking in musical structure.

In sum, we expected listeners to be unsuccessful at learning Random sequences, and support for this hypothesis was found in the memory test, where listeners performed poorly on Random test trials. In addition, we expected listeners to have difficulty creating predictive models for Random sequences, and hypothesized that this lack of accurate expectation would be reflected in neural activity (a less accurate predictive model should lead to greater alpha and beta activity). Although we cannot yet make claims from the group time-frequency results, several participants did demonstrate increasingly greater alpha and beta band activity for Random sequences over the course of the listening sessions. This may possibly result from interesting individual differences, such as variable amounts of musical training and experience. Like the Alba, et al (2008) study demonstrates, significant differences in task performance can be reflected in participants' neural activity. Therefore, future studies in this line of investigation will "bin" data according to 'years of musical training' and memory task performance. In addition, future studies will incorporate a memory test after every listening session to examine the time-course of musical learning at a finer level of detail.

4.7 Conclusions and Future Directions

Information Theoretic approaches have elucidated various aspects of music perception, such as the melodic expectancy of forthcoming music (e.g., Pearce, et al., 2010). In the IT study described above, three subjective information theoretic factors, Surprise, Predictive information, and Coding Gain, all significantly influenced expectation ratings of Probe tones during the listening sessions. Generally, sequences that were more difficult to predict (higher surprise/entropy, etc) gave rise to worse memory performance in participants. There was also an increasing impact of these factors on memory for exemplars throughout the study. The effect of entropy became more pronounced as listeners repeatedly heard melodies (sequences with low overall Surprise were more likely to be remembered by the third listening session, for example).

The effect of unpredictability was also apparent in my EEG studies. When listeners are unable to form an accurate predictive model of forthcoming input, neural processing is more effortful: the amplitude of N1 response increases for Random compared to Normal sequences over time. In addition, preliminary evidence shows increases in alpha and beta band activity as novel sequences are learned over a period of increasing exposure.

Although the above EEG studies only manipulated global musical structure by using Normal and Randomized melodies, future work will use tone sequences varying on the information theoretic measures described in the previous section. Over the course of learning, we predict that sequences should elicit more efficient neural oscillatory response over time. We expect, for example, that melodies with low predictive information and high entropy should impose greater processing demands than those with low entropy and high redundancy. Lastly, it would be interesting to test learning and memory for the above ecological melodies compared to those tailored to reflect particular information theoretic properties.

CHAPTER 5

GENERAL DISCUSSION: PREDICTION AS AN EFFICIENT AND INDISPENSIBLE COMPONENT OF MUSICAL LEARNING

5.1 General Discussion

This dissertation centers around the dynamic process of learning music over time – what is likely to be encoded upon first hearing a novel melody (in terms of musical characteristics and statistics), and how mental representations change with increasing exposure to music. Schematic processing, predictive mechanisms, and increased efficiency of representation were all highlighted as important aspects of musical learning. Evidence for these findings was drawn from behavioral research, computational and information theoretic methods, and electroencephalography.

Like memory for language or visual experience, musical memory is multifaceted and complex. Sometimes sections of music are remembered "verbatim", while other passages are encoded in a more general, schematic framework. Expertise can play a large role in musical learning, as differences between musicians and non-musicians (such as the accuracy of a set of musical predictions) affect the level of detail retained in memory. The robustness of memory also depends on the amount of time lapsed since exposure to novel music, with the emphasis on musical contour in short-term memory giving way to a more precise interval pattern encoding in long-term memory. Music recognition is also influenced by familiarity with the stimulus and episodic associations. In addition, musical memory is of course subject to general memory constraints, such as the limits of working memory span. All of these facets of musical memory may be understood in terms of a few key concepts described below.

5.1.1 Experience Enables Efficiency

One general "goal" of evolution is to be as structurally and metabolically as efficient as possible. When considering cognitive function, this is manifested in terms of predictive processing - finding meaningful patterns in the environment allows for expedited processing of future input. As Moshe Bar attests, the brain is proactive – it needs not form new predictions "from scratch", but almost always relies on existing "scripts" in memory, which are the product of previous experience and associative processing (Bar, 2011). Although some resources are required, especially initially, to create a mental predictive model, an accurate set of expectations decreases demands on resources later. Schemata are long-term predictive models – they are based on extensive experience and allow the perception of stylistic (well-structured) music to be more efficient. While most often an indispensible part of musical perception and learning, schematic processing is occasionally detrimental to memory by causing a poverty of detail. Similarly, an inaccurate or inappropriately applied schematic framework can lead to errors of encoding or recall. When a predictive model is consistently erroneous, it is altered to reflect the statistical properties of the signal. A tradeoff exists between the level of *informativeness* that prediction affords, where precise expectations are more useful in guiding perception, and *accuracy* of prediction, which often stems from more general and abstract expectations.

5.1.2 Schematic Processing as a Predictive Model

Arguably, one of the most important aspects of our perception and memory is our ability to use schematic processing to understand, efficiently encode, and predict what we experience. Schematic processing is the top-down mediation of input that guides the listener's expectations during music perception (thereby also influencing memory). More robust schematic processing allows for a more structured and detailed gist to be encoded in memory. Given the flood of information constantly bombarding our senses, using schemata to process and selectively store incoming information is more efficient and flexible (for comparisons and abstractions) than encoding all of the information verbatim.

If an adult with Western musical experience listens to a theme and variations for the first time, his musical schemata will guide his perception of the music through a series of implicit expectations about the rising tension and cadential release in the music (Gjerdingen, 1988). In this way, schemata scaffold the listener's perception of music by providing a framework of knowledge and expectations. Upon hearing the theme, which is often a salient and repeated event in the music, the theme's contour and foundational metrical features will be stored as a gist in memory. While hearing the variations, in which the thematic material is repeated in different ways (such as transpositions and alternate rhythms), the listener forms an increasingly more detailed representation of the principal theme. The melodic contour will be supplemented by specific interval patterns in long-term memory, and a schematic representation of the theme's prolongational reduction (pattern of tension and relaxation based on the harmonic and metrical structure) will be encoded (Lerdahl & Jackendoff, 1983). This creates a sort of melodic invariance that allows the theme to be recognized across different instantiations. It is the type of schematic-predictive model that is general and flexible enough to be widely applicable and accurate for general expectations.

Research in music cognition has outlined several types of melodic expectation (common chord progressions, melodic motion, etc) and predictive processing, but many of the current approaches to modeling this tension/relaxation in music are essentially based on a snapshot in time. Time-span reduction analysis, for example, creates predictions about rising and waning tension without regard to how the listener's expectations change over time (Lerdahl & Jackendoff, 1983). The information-theoretic approach of Chapter 4 offers a useful remedy by modeling adaptive statistical predictions in the listener that change dynamically with every subsequent event in the music. In addition, computational models and EEG can capture the mechanisms underlying the evolving experience of learning music over time.

In sum, schematic processing and gist-like representation are crucial for the comprehension and appreciation of music. Future work should capitalize on the recent advances in knowledge about statistical learning and computational modeling, as well as neuroscientific approaches, to further address how schemata are developed and memory representations dynamically change over time.

5.2 Summary of Findings and Conclusions

The studies within this dissertation explored the process of forming musical expectations, and the role of schematic expectations in guiding perception and memory. In order to explore the memory representation formed upon hearing a melody for the first time, I conducted a set of behavioral change detection studies (Chapter 2). Inspired by visual change blindness research, these experiments investigated the relationship between musical schemata, tonal and rhythmic structure, and musical expertise. Both musicians and non-musicians were tested to assess whether training facilitates the use of schematic processing and thus the formation of a more detailed memory representation.

The results of the first behavioral study showed that professional musicians were significantly more successful at detecting changes than non-musicians for melodies containing at least a minimal amount of structure. Generally, less tonal structure (eg. non-stylistic and randomized melodies) resulted in compromised change detection for the two groups. When sequences did not conform to listeners' schemata, they were either left with a poverty of information encoded in memory (i.e., for random sequences), or an inaccurate representation (i.e., for diatonic changes). The second experiment revealed tonality to have a particularly large effect on memory performance, with note duration, interval of change, and musical expertise also contributing significantly.

Clearly, the memory representation for a novel melody is quite different than that of a well-known melody (or one that has been presented repeatedly). Memory for novel music will generally be less detailed and more schematic than long-term musical memory. To gauge how memory representations change over the course of the learning trajectory, I conducted a series of SRN studies.

The computational studies discussed in Chapter 3 examined the learning trajectory of tonal representations in music while observing changes in the network's internal structure over time. Our simple recurrent network demonstrated that sparse population coding is an efficient and effective way to distill information about musical structure. Three experiments examined the learning trajectory of a simple recurrent network upon exposure to musical corpora differing in statistical structure (Normal, Bigram, and Random networks). By having listeners rate the network's own novel musical output from different points along the learning trajectory, the experiments compared the networks' internal representations to behavioral data. We found that the hidden layer representations of tonal structure become more efficiently represented (in terms of population sparsity) as the network learns, and that this sparsity is strongly correlated with listeners' judgments of the networks' compositions. We argue that sparsity underlies the network's success: It is the mechanism through which musical characteristics are learned and distilled, and facilitates the network’s ability to produce more complex and stylistic compositions over time. Future work will clarify which type of sparsity arises within the network’s internal representations, population or lifetime sparsity (or both). Population sparsity describes a network in which only a few nodes are active all the time. Lifetime sparsity, which may be a more interesting measure of sparsity, occurs when a small set of active nodes are distributed in representing information over time. We may calculate this measure in future work by tracking the activations of hidden layer nodes throughout the course of training. Our hypothesis is that both population and lifetime sparsity are present in networks exposed to normal, stylistic music, but only population sparsity arises in networks exposed to random sequences of tones.

Considering further the internal structure of these SRNs, and comparing these computational models with the previous change detection findings, we hypothesize that the SRN essentially learns musical schemata over the course of training. Future work will prompt the SRN to produce computational schemata (trajectories through tonal state space) after exposure to a corpus of stylistic, non-stylistic, or random music. The networks can also be used to produce their own gist memory after exposure to brief melodies. Comparing this output to listeners’ behavioral findings will be some of the first computational work to model gist memory.

The SRN’s process of acquiring statistical regularities from exposure may also be compared to the listeners’ process of learning the statistical regularities of tone sequences in Chapter 4. In the information theoretic study presented therein, sequences of varying predictability were presented over the course of three listening sessions. The measures of Surprise, Predictive information, and Coding Gain were all found to influence both melodic expectation during listening and memory at test. Generally, sequences that were highly unpredictable led to lower expectation ratings of probe tones and worse memory performance.

Interestingly, memory performance for sequences that were high-entropy or low average Coding Gain significantly improved from the first to third test session. This may be because listeners these sequences were the most challenging to commit to memory, and therefore initially very poor performance improved with exposure. Like the SRN of Chapter 3, listeners slowly learned the statistical tonal regularities of tone sequences over time. To test whether the listeners' memory representations also become more efficient, a series of EEG studies was conducted.

The two EEG experiments outlined in Chapter 4 examined the N1 obligatory component and alpha and beta band activity as musical sequences varying in structure were learned over time. These studies provide preliminary evidence that while novel tone sequences are being learned, processing demands are high. Once an accurate predictive model is formed (i.e., for structured sequences), fewer resources need to be recruited for perception (the N1 amplitude decreases for Normal tunes). However, when little structure is present in the signal, as is the case with randomized tone sequences, processing demands increase as the brain attempts to form a useful model. Arguably, the differential N1 response between Normal and Random melodies and the preliminary oscillatory findings in Experiment 2 suggest the absence of schematic processing and expectation in melodies lacking musical structure. In other words, normal structure enables the formation and utilization of predictive models that ultimately result in increased efficiency of musical processing mechanisms. This finding is reminiscent of a TMS cortical mapping study in which the volume of neural substrate dedicated to finger movement increased as novice participants learned a pattern on the piano. After weeks of training, the volume of the dedicated motor representation reduced in size, showing increased efficiency once the pattern was learned (Pascual-Leone, 2006). It should be noted that in this study, as well as imaging studies of repetition suppression, for example, it is difficult to determine whether the decrease in response is due to neural *fatigue*, *sharpening*, or *facilitation* (for an overview of these models see Grill-Spector, Henson, & Martin, 2006). Similarly, in my EEG studies, further research would be needed to determine whether the observed decrease in neural activation is due to an overall reduction in responsiveness (all the neurons are still firing, but a decreased rate), to fewer neurons responding (at the same rate), or to increasing sparsity (the neurons have a distributed and selective response).

In sum, the studies described above elucidate the process of musical learning over time, with particular emphasis on schematic processing, musical structure, statistics-based predictive models, increased efficiency, and the role of musical expertise. The SRN and EEG experiments illustrate how increased efficiency underlie successful learning over time. These findings, as well as results from the IT study, provide evidence that schemata are formed as the probabilities of forthcoming music are gradually learned with increasing experience. The set of expectations that schemata provide dynamically guide perception and influence memory, sometimes at the expense of accuracy and detail, but most often to support flexible and efficient cognitive processing.

REFERENCES

Abdallah, S., & Plumbley, M. (2009). Information dynamics: patterns of expectation and surprise in the perception of music. *Connection Science, 21,* 89-117.

Agres, K., & Krumhansl, C. (2008). Musical Change Deafness: The Inability to Detect Change in a Non-speech Auditory Domain. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 969-974), eds. B. C. Love, K. McRae, & V. M. Sloutsky. Austin, TX: Cognitive Science Society.

Anderson, R. (1977). The Notion of Schemata and the Educational Enterprise: General Discussion of the Conference. In *Schooling and the Acquisition of Knowledge,* eds. Richard Anderson, Rand Spiro, and William Montague. Hillsdale, NJ: Erlbaum.

Anderson, R., Reynolds, R., Schallert, D., and Goetz, E. (1977). Frameworks for Comprehending Discourse. *American Educational Research Journal*, *14(4)*, 367-381.

Atallah, H., Frank, M., & O'Reilly, R. (2004). Hippocampus, cortex, and basal ganglia: Insights from computational models of complementary learning systems. *Neurobiology of Learning and Memory, 82*, 253-267.

Bar, M. (2011). The proactive brain. *Predictions in the brain*. New York, NY: Oxford University Press.

Bartlett, F. (1932). *Remembering: A Study in Experimental and Social Psychology.* Cambridge: Cambridge University Press.

Bigand, E. (1990). Abstraction of two forms of underlying structure in a tonal melody. *Psychology of Music, 18(1)*, 45-59.

Bigand, E. (1993). The influence of implicit harmony, rhythm and musical training on the abstraction of "tension-relaxation schemas" in tonal musical phrases. *Contemporary Music Review, 9(1)*, 123-137.

Brewer, W., & Nakamura, G. (1984). The Nature and Function of Shemas. In *The Handbook of Social Cognition*, eds. R. Wyer and T Srull. Hillsdale, NJ: Erlbaum.

Cimenser, A., Purdon, P., Pierce, E., Walsha, J., Salazar-Gomez, A., Harrell, P., Tavares-Stoeckela, C., Habeeba, K., & Brown, E. (2011). Tracking brain states under general anesthesia by using global coherence analysis. *PNAS*, *108*, 8832-8837.

Cole, M., Cole, S., & Lightfoot, C. (2005). *The Development of Children.* Fifth Edition. New York: Worth Publishers.

Cuddy, L., Cohen, A., & Mewhort, D. (1981). Perception of structure in short melodic sequences. *Journal of Experimental Psychology: Human Perception & Performance, 7(4),* 869-883.

Deutsch, D. (1980). The processing of structured and unstructured tonal sequences. *Perception & Psychophysics, 28*, 381-389.

Dowling, J. (1978). Scale and Contour: Two Components of a Theory of memory for Melodies. *Psychological Review, 85(4),* 341-354.

Dowling, J. (1991). Tonal strength and melody recognition after long and short delays. *Perception and Psychophysics, 50(4),* 305-313.

Eck, D., & Schmidhuber, J. (2002). A First Look at Music Composition using LSTM Recurrent Neural Networks. *Technical Report IDSIA-07-02*, Instituto Dalle Molle di studi sull intelligenza artificiale, Manno, Switzerland.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14(2)*, 179-211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7*, 195-224.

Essens, P., & Povel, D.J. (1985). Metrical and nonmetrical representations of temporal patterns. *Perception & Psychophysics, 37(1),* 1-7.

Field, D. (1994). What is the Goal of Sensory Coding? *Neural Computation, 6,* 559-601.

Graesser, A., & Nakamura, G. (1982). The impact of a schema on comprehension and memory. In *The psychology of learning and motivation: advances in research and theory*, ed. G. Bower, 16, 59-109.

Greenberg, M., Westcott, D., & Bailey, S. (2004). When Believing is Seeing: The Effect of Scripts on Eyewitness Memory. *Law and Human Behavior*, 22(6), 685-694.

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *TRENDS in Cognitive Sciences, 10(1)*, 14-23.

Halpern, A., & Mullensiefen, D. (2008). Effects of timbre and tempo change on memory for music. *The Quarterly Journal of Experimental Psychology, 61(9),* 1371-1384.

Hebert, S., & Peretz, I. (1997). Recognition of music in long-term memory: Are melodic and temporal patterns equal partners? *Memory and Cognition, 25(4)*, 518-533.

Jaeger, H. (2001). The "echo state" approach to analysing and training recurrent neural networks. In *GMD Report 148, German National Research Center for Information Technology.*

Kaiser J., & Lutzenberger W. (2005). Human gamma-band activity: A window to cognitive processing. *NeuroReport, 28,* 207-211.

Kidd, G., Boltz, M., & Jones, M. (1984). Some effects of rhythmic context on melody recognition. *American Journal of Pyschology, 97(2)*, 153-173.

Koelsch, S., & Friederici, A. (2003). Toward the neural basis of processing structure in music: Comparitive results of different neurophysiological investigation methods. *Annals of the New York Academy of Sciences, 999,* 15-28.

Koelsch, S., Gunter, T., Schroger, E., Tervaniemi, M., Sammler, D., & Friederici, A. (2001). Differentiating ERAN and MMN: An ERP study. *NeuroReport: Neurophysiology, Basic and Clinical, 12,* 1385-1389.

Krumhansl, C. (1999). Music Psychology: Tonal Structures in Perception and Memory. *Annual Reviews of Psychology, 42*, 277-303.

Krumhansl, C. (2000). Rhythm and Pitch in Music Cognition. *Psychological Bulletin, 126(1)*, 159-179.

Krumhansl, C., & Kessler, E. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review, 89*, 334-368.

Large, E., Palmer, C., & Pollack, J. (1995). Reduced Memory Representations for Music. *Cognitive Science, 19,* 53-96.

Leman, M. (1995). *Music and Schema Theory: Cognitive Foundations of Systematic Musicology*. New York: Springer.

Lerdahl, F. (2001). *Tonal Pitch Space.* New York: Oxford University Press.

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

Levy W.B., Baxter R.A. (1996). Energy efficient neural codes. *Neural Computation, 8,* 531-543.

Manning, C. & Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Mavromatis, P. (2005). A hidden Markov model of melody production in Greek church chant. *Computing in Musicology, 14*, 93–112.

Meyer, L. (1956). *Emotion and Meaning in Music.* Chicago: Chicago University Press.

Margulis, E. (2005). A Model of Melodic Expectation. *Music Perception, 22(4)*, 663-714.

Minsky, M. (1975). A Framework for Representing Knowledge. In *The Psychology of Computer Vision,* ed. Patrick Winston. New York: McGraw-Hill, 211-277.

Mozer, M. (1994). Neural Network Music Composition by Prediction: Exploring the Benefits of Psychoacoustic Constraints and Multi-scale Processing. *Connection Science, 6,* 247-280.

Narmour, E. (1992). *The analysis and cognition of melodic complexity: The implication-realization model*. Chicago: University of Chicago Press.

Noland, K., & Sandler, M. (2006). Key Estimation Using a Hidden Markov Model. In *Proceedings of the International Conference on Music Information Retrieval,* Victoria, Canada.

Olshausen B., & Field D. (2004). Sparse Coding of Sensory Inputs. *Current Opinion in Neurobiology*, *14*, 481-487.

Page, M. (1993). Modeling Aspects of Music Perception Using Self-organizing Neural Networks. Unpublished doctoral dissertation. University of Wales.

Palmer, C. (1997). Music Performance. *Annual Reviews of Psychology, 48*, 115-38.

Pascual-Leone, A. (2006). The brain that plays music and is changed by it. *Annals of the New York Academy of Sciences, 930,* 315-329.

Patel, A. (2008). *Music, Language, and the Brain*. New York: Oxford University Press.

Patel, A., Iverson, J., Chen, Y., & Repp, B. (2005). The influence of metricality and modality on synchronization with a beat. *Experimental Brain Research, 163(2),* 226-238.

Pearce, M., Wiggins, G. (2006). Expectation in Melody: The Influence of Context and Learning. *Music Perception, 23(5),* 377–405.

Peretz, I., & Zatorre, R. (2005). Brain Organization for Music Processing. *Annual Review of Psychology, 56,* 89-114.

Peretz, I., Gosselin, N., Belin, P., Zatorre, R., Plailly, J., & Tillmann, B. (2009). Musical Lexical Networks: The Cortical Organization of Music Recognition. *The Neurosciences and Music III—Disorders and Plasticity: Annals of the New York Academy of Sciences, 1169*, 256–265.

Povel, D.J., & Essens, P. (1985). Perception of temporal patterns. *Music Perception, 2(4),* 411-440.

Raphael, C. (1999). Automatic segmentation of acoustic musical signals using hidden Markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21*, 360–370.

Rolls, E.T. and Tovee, M.J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *Journal of Neurophysiology*, *73(2)*, 713-726.

Rumelhart, D. (1980). Schemata: The Building Blocks of Cognition. In *Theoretical Issues in Reading Comprehension,* eds. Rand Spiro, Bertram Bruce, and William Brewer. Hillsdale, NJ: Erlbaum.

Rumelhart, D., & Ortony, A. (1977). The Representation of Knowledge in Memory. In *Schooling and the Acquisition of Knowledge,* eds. Richard Anderson, Rand Spiro, and William Montague. Hillsdale, NJ: Erlbaum.

Saffran, J., Johnson, E., Aslin, R., & Newport, E. (1999). Statistical Learning of Tone Sequences by Human Infants and Adults. *Cognition, 70*, 27–52.

Schank, R., & Abelson, R. (1977). *Scripts, plans, goals, and understanding: an inquiry into human knowledge structures*. New York: Halsted Press.

Schellenberg, G., Iverson, P., & McKinnon, M. (1999). Name that tune: identifying popular recordings from brief excerpts. *Psychonomic Bulletin and Review, 6,* 641–646.

Schmuckler, M. (1989). Expectation in Music: Investigation of Melodic and Harmonic Processes. *Music Perception, 7(2)*, 109-150.

Schmuckler, M., & Boltz, M. (1994). Harmonic and rhythmic influences on musical expectancy. *Perception and Psychophysics, 56(3)*, 313-325.

Snyder, B. (2000). *Music and Memory: An Introduction*. Cambridge: MIT Press.

Steinbeis, N., Koelsch, S., & Sloboda, J.A. (2006). The role of harmonic expectancy violations in musical emotions: Evidence from subjective, physiological, and neural responses. *Journal of Cogntive Neurosci*ence, *18,* 1380–1393.

Todd, P. (1989). A connectionist approach to algorithmic composition. *Computer Music Journal*, *13(4)*, 27-43

Todd, P. (1999). Evolving musical diversity. In *Proceedings of the AISB'99 Symposium on Creative Evolutionary Systems*, 40-48. Sussex, UK: Society for the Study of Artificial Intelligence and Simulation of Behavior.

Tuckey, M., & Brewer, N. (2003). How Schemas Affect Eyewitness Memory over Repeated Retrieval Attempts. *Applied Cognitive Psychology, 17*, 785-800.

Welker, R. (1982). Abstraction of themes from melodic variations. *Journal of Experimental Psychology: Human Perception and Performance, 8(3)*, 435-447.

Willmore, B., Mazer, J., & Gallant, J. (2011). Sparse coding in striate and extrastriate visual cortex. *Journal of Neurophysiology, 105*, 2907-2919.