The Role of Features and Context in Recognition of Novel Melodies

Daniel Müllensiefen
Goldsmiths, University of London, London, United Kingdom

Andrea R. Halpern
Bucknell University

We investigated how well structural features such as note density or the relative number of changes in the melodic contour could predict success in implicit and explicit memory for unfamiliar melodies. We also analyzed which features are more likely to elicit increasingly confident judgments of "old" in a recognition memory task. An automated analysis program computed structural aspects of melodies, both independent of any context, and also with reference to the other melodies in the testset and the parent corpus of pop music. A few features predicted success in both memory tasks, which points to a shared memory component. However, motivic complexity compared to a large corpus of pop music had different effects on explicit and implicit memory. We also found that just a few features are associated with different rates of "old" judgments, whether the items were old or new. Rarer motives relative to the testset predicted hits and rarer motives relative to the corpus predicted false alarms. This data-driven analysis provides further support for both shared and separable mechanisms in implicit and explicit memory retrieval, as well as the role of distinctiveness in true and false judgments of familiarity.

Received: February 2, 2013, accepted September 21, 2013.

Key words: implicit vs. explicit memory, computational modeling, automatic music analysis, true and false memories, distinctiveness

Remembering music is an important part of many people's lives, no matter what their musical background. In some ways, we have excellent memory for music. People maintain a large corpus of familiar tunes in their semantic memory. The representations are accurate in that someone can typically say if there is a wrong note in a familiar tune (Dowling, Bartlett, Halpern, & Andrews, 2008) and memory for tunes seems to last over one's lifetime (Bartlett & Snelus, 1981). On the other hand, encoding of new music is quite difficult (Halpern & Bartlett, 2010). Sometimes a tune sounds familiar but it turns out that it is only similar to one we knew in the past, creating false alarms. And people who have bought or downloaded some music only to discover the piece already in their collection have experienced the other kind of error, a miss.

Explaining success and failures in memory for music by applying well-understood memory principles has not always been successful, which raises the question of whether memory for music is special or different from memory for other kinds of information. For instance, type of encoding task seems not to affect overall recognition performance for unfamiliar tunes (Halpern & Müllensiefen, 2008; Peretz, Gaudreau, & Bonnel, 1998) and musical expertise does not always increase this sort of recognition memory (Demorest, Morrison, Beken, & Jungbluth, 2008; Halpern, Bartlett, & Dowling, 1995). However, in common with other materials, familiar tunes are generally recognized more accurately than unfamiliar tunes (Bartlett, Halpern, & Dowling, 1995). These predictors are largely concerned with the encoding situation, state of the rememberer, and some general aspects of the to-be-remembered items.

In contrast, our goal in this paper is to examine the extent to which two other factors can predict memorability of, in this case, real but unfamiliar pop tunes. One factor is the *features* of the tunes themselves. We take advantage of powerful statistical modeling techniques as well as automated feature extraction software to allow simultaneous evaluation of many features at the same time.

This discovery-driven approach assumes that stimuli in the world are composed of many kinds of features, and that people can employ statistical learning to encode those features. People certainly employ statistical learning in procedural tasks, like learning artificial grammars (Pelucchi, Hay, & Saffran, 2009) and motor sequences (Daselaar, Rombouts, Veltman, Raaijmakers, & Jonker, 2003), regardless of whether the features are processed consciously or not. The feature approach is well established in memory research. For example, Cortese, Khanna, and Hacker (2010) looked at recognition memory for over 2500 monosyllabic words, taking as

Music Perception, volume 31, issue 5, pp. 418–435, issn 0730-7829, electronic issn 1533-8312. DOI: 10.1525/mp.2014.31.5.418

predictors linguistic factors such as word frequency and imageability. They were able to predict 46% of the variance in hit rates, 15% of the variance in false alarm rates, and 30% of the variance in overall accuracy with eight feature predictors. Statistical learning has also been shown to operate in music; for instance, in learning a new musical scale (Loui & Wessell, 2006).

Music offers a rich array of features to analyze. Even in a simple melody line, features can be extracted relating to a variety of pitch and temporal aspects, on both an absolute and relational basis. For instance, just in the pitch domain, one could compute simple features like average pitch height, as well as relational features like the distribution of pitch intervals and note density. Pattern-level features like contour and motivic themes can also be computed from pitch information. Schulkind, Posner, and Rubin (2003) needed only four musical features to significantly predict at what note a well-known song was typically identified, accounting for about 30% of the variance. This approach is very powerful insofar that one need not make any assumptions about the differences among listeners in perceptual abilities, personality, encoding conditions or anything else, to explain the probability of remembering an item. Here, we started with many (nearly 60) features, and found clusters of related features that could best explain memory performance on a tune-by-tune basis.

A second factor is the *frame of reference* in which people are processing those features. We hypothesize that listeners either compare a tune in immediate memory to tunes they have heard in the immediate context and/or tunes that they are familiar with from their life experience. We furthermore propose that even the nonmusicians tested here can abstract statistical properties of both the items and contexts, and compare them. Thus, we analyzed how the frequency with which features occur, relative to defined sets of melodies, predicted various aspects of memorability. We investigated two contexts: the relative frequency of features with reference to the set of 80 unfamiliar melodies used in the study (local context), and relative frequency with reference to a large corpus of similar melodies from which we chose the testset melodies (global context) and that we presume models our listeners' experience with pop music in general.

People do use contextual information in memory for music. Using a gating paradigm, Bailes (2010) explained around 80% of the variance in her participants' data measuring the points at which more and less familiar songs were recognized. The most predictive features included information-theoretic features that make use of information derived from statistical distributions. Kopiez and Müllensiefen (2011) extracted melodic features from a set of songs by the Beatles and were able to relate musical structure to popularity as indexed by the chart success of cover versions of the songs by other artists. To our knowledge, however, no one has specifically included both the local and global context for a feature frequency analysis of memory performance.

In general, we hypothesized that relatively distinctive features would affect memory. The most obvious form of the hypothesis was that rarer features would attract attention, and increase processing of a melody. We know that distinctiveness (or its counterpart, typicality) of *items* can affect memory: Typical faces are rated as more familiar, but distinctive faces elicit lower false alarms on a recognition test than do typical faces (Bartlett, Hurry, & Thorley, 1984). But less is known about how *feature* distinctiveness may affect memory.

More specifically, we were interested in examining whether different aspects of memory would be more influenced by the local context (the melodies in the study) or global context (music in general). We know that people are sensitive to both types of context in other types of judgments, like expectancy ratings (Pearce & Wiggins, 2006). Use of local context requires abstraction of a set of tunes often presented just once in a learning set. But using local context is exactly what one should be doing in an episodic memory test. If local context is hard to process, this might account for some of the difficulty in learning new tunes.

Global context results from the large numbers of tunes in people's semantic memory, and even nonmusicians can abstract a fair amount about rules of Western tonality from repeated exposure, including probabilities of intervals (Pearce & Wiggins, 2006). Assuming they can apply these contexts during item recognition, this might be helpful in true recognition. On the flip side, global context might work against good discrimination of old from new items, if for instance, some features in a presented melody are also common in the corpus, and are thus mistakenly called "old."

Our first research question was whether the same factors would predict memory in explicit and implicit memory tasks. Explicit judgments refer to situations in which a person realizes he or she is retrieving a memory. For instance, asking someone to say whether tune had been presented earlier in the session requires explicit recognition. However, most memory retrievals in everyday life, including those for music, are done implicitly, or without awareness of retrieving, as when a skill improves with practice or when someone spontaneously sings along to a song on a radio. Here, we exploited the "mere exposure effect" (Zajonc, 1980), in which preference for

items often increases upon repeated exposure, whether or not people can recognize or even clearly perceive the information. When preference does increase for old compared to new items, and everything else about the materials is held constant or counterbalanced, this logically implies successful encoding. For instance, Halpern and O'Connor (2000) found that healthy older adults preferred twice-presented to never-presented tunes, despite chance explicit recognition of those tunes. Johnson, Kim, and Risse (1985) found a similar pattern in Korsakoff amnesia patients. These participants must have encoded the tunes to account for the increase in preference, but were impaired in consciously retrieving them.

A large body of research has shown that implicit and explicit judgments are affected by different manipulations, suggesting a functional dissociation. For instance, levels of processing manipulations usually are ineffective in implicit tasks, and perceptual matching of items at study and test often benefits implicit more than explicit tasks (Schacter, Chiu, & Ochsner, 1993). This distinction has neurological correlates, in that people with temporal lobe amnesia have been shown to have normal implicit retrieval in priming tasks at the same time they have greatly impaired explicit retrieval (Graf & Schacter, 1985). Likewise, different patterns of activation emerge in neuroimaging studies on explicit and implicit retrieval (Schacter & Buckner, 1998). Although not everyone agrees that these outcomes necessitate two entirely different memory systems (Reder, Park, & Kieffaber, 2009), it seems clear that implicit and explicit tasks call upon different processing considerations.

Our approach allows us to test whether the two tasks dissociate in the sense that they will be sensitive to different sets of features. If different feature sets predict memorability in the implicit compared to the explicit task, this would imply that the retrieval processes are dissociable in the two tasks (as the tunes for the two tasks were only encoded once). For instance, if context is more important in predicting one type of retrieval, this necessarily implies a more global computational process.

Our second research question pertains only to explicit recognition. In recognition memory studies, "yes" responses imply that the item has induced a feeling of familiarity (and, sometimes, recollection (Yonelinas, 2002)) whereas "no" responses imply the opposite, a feeling of novelty. We asked whether particular musical features or relative frequencies might affect attributions of familiarity or novelty to items at test, whether that attribution is correct or not. In other words, can some features and contexts of melodies drive recognition responses over and above whether the item had actually been presented or not? We anticipated that melodies with features that are common with respect to the testset of melodies, or the large parent corpus from which we drew the melodies, would evoke judgments of "old," regardless of whether the melody was actually old, and vice-versa for judgments of "new."

We can furthermore ask whether different features and contexts engender invalid judgments of oldness that lead to false alarms and invalid judgments of novelty that lead to misses. The FA rate overall and in different types of new items can illuminate the type of information people encode, under the presumption that similarities in representation between new items and old ones can lead to these errors. For instance, false alarms to never-presented prototypes (e.g., Welker, 1982 for music; Cabeza & Kato, 2000, for faces) suggest that prototypes are formed during category learning.

Less attention has been paid to the other kind of recognition memory error: misses (an illusion of novelty). However, it is not a foregone conclusion that valid vs. invalid judgments of familiarity and novelty are influenced by the same factors, especially in nonverbal and artistic domains such as music. A dissociation between recognition responses for old and new items has been found at the neural level. For example, Henson and colleagues (Henson, Hornberger, & Rugg, 2005) have studied the old/new effect using fMRI. Hits compared to correct rejections were associated with large areas of activation including parietal lobes, posterior cingulate, and some frontal regions, whereas correct rejections activated a medial temporal area (perirhinal cortex) more so than hits. This last finding suggests that new items engender an encoding response more than do previously experienced items, which itself might be a basis for discrimination of old from new items.

Thus our second analysis was designed to, in fact, detect whether correct vs. incorrect performance with old and new items is driven by different factors. Different features might predict errors with old and new items, but it is also possible that the local and global contexts might contribute differentially. For instance, because new items are, by definition, not experienced in the testset until the actual test, the influence of the pop music corpus might be more important in predicting misses than false alarms.

To summarize, we presented nonmusicians with melodies from unfamiliar pop tunes, followed by an explicit (recognition on a 6-point confidence scale) and implicit (ratings of pleasantness) memory test. We used pop songs because they represent a kind of music quite familiar to people lacking music training and have melodic structures that are tractable to analyze. We

examined overall explicit and implicit memory aggregated over items to help answer the first research question. We then looked at old and new items in the explicit task to help answer the second research question. From this entirely discovery-driven approach, we hoped to shed light on both the general ability of untrained listeners to process statistical properties of musical items, as well as the particular ways these properties can predict item-by-item memory performance.

Method

PARTICIPANTS

Participants were 34 undergraduate students from Bucknell University with a mean age of 19.12 year ($SD = 1.07$) with either no or only moderate amounts of music training. All were enculturated to Western music. Music training was indexed by years of private instrumental lessons and ranged from 0 to 6, with a mean of 1.13 years ($Mdn = 1$ year, $SD = 1.18$ year).[1]

MATERIALS

Eighty short single-line melodies were selected from a large database of 14,063 commercial pop songs covering many different popular genres and periods from the 1950's to 2006. All vocal melodies of the 14,063 songs were first segmented into 473,034 melody phrases using the David Temperley's Grouper algorithm (Temperley, 2001) which has been shown to be a relatively accurate segmentation tool for melodies (Pearce, Müllensiefen, & Wiggins, 2010). From this large pool of melody phrases, 80 were automatically selected to meet as many as possible of nine criteria to ensure a selection of standard, and "normal-sounding" melodic vocal phrases from commercial pop music.[2]

The 80 selected items were divided at random into two subsets of 40 melodies each so as to allow counterbalancing of old and new items. Tunes were presented in a synthesized piano timbre and recorded onto CD for later presentation.

PROCEDURE

Participants were tested in small groups. They first filled out a questionnaire on their musical background. Then, the 40 to-be-remembered melody items were presented on a high quality CD player over speakers, with 2s between each item. As this was an incidental paradigm, participants were asked to rate each melody on a 3-point scale of familiarity, from *unfamiliar* to *familiar*. These ratings were merely a cover task; familiarity ratings seemed reasonable to participants and served to insure attention. No mention was made of a subsequent memory test.[3]

After all melodies had been rated, the test phase commenced immediately, in which the 40 old melodies were mixed randomly with 40 new melodies. After each melody was played, two judgments were required. The recognition judgment asked participants to indicate whether they had heard the melody in the first part of this experiment on a 6-point confidence scale ranging from 1 (*definitely did not hear the tune before*) to 6 (*definitely heard this tune before*). Intermediate ratings corresponded to intermediate levels of confidence. The second judgment concerned the pleasantness of the melody item and asked for a rating from 1 (*very unpleasant*) to 7 (*very pleasant*). The difference in pleasantness judgments for old compared to new items served as an indicator for implicit memory. Both judgments were made within a 4-s response window.

Half the participants received one subset of tunes as old items, and rated each melody first for pleasantness and then for recognition. The other half of the participants received the other subset of tunes as old items, and made the judgments in reverse order.

[1] This research was carried out before the Goldsmiths Musical Sophistication Index (Müllensiefen, Gingras, Musil, & Stewart, 2014) was available as a standardized self-report instrument for the extent of individual assessing musical training.

[2] Selected phrases should not straddle a song section (e.g., verse-chorus transition), they should stay within length boundaries of a minimum of 6 and a maximum of 13 notes as well duration boundaries 4 s (minimum) to 8 s (maximum); phrases should also not straddle a tempo nor a time signature change in the song, they should remain within a pitch range spanning a maximum of 17 semitones (an octave and a fourth) where all notes should come from singable absolute pitch ranges for male and female voices between E2 (MIDI pitch 40) and C6 (MIDI pitch 84). In addition, the last note of a phrase should be followed by a rest and it should not contain any overly long notes covering more than a third of the entire duration of the entire phrase. No more than one phrase from the same song was used and overly popular tunes were avoided in the selection to ensure that participants were unfamiliar with the tunes. The selected phrases naturally varied in tempo and tonality as well as in rhythm and pitch patterns. Phrases could be selected from any part of a song but care was taken to avoid familiar motives e.g., from a chorus that would make their origin easily recognizable. The database of MIDI transcriptions is hosted at Goldsmiths, University of London. A description of the contents of the database can be found in Müllensiefen, Wiggins, and Lewis (2008). It was acquired for research purposes from Geerdes MIDI Music (http://www.midimusic.de/), a commercial supplier of MIDI and karaoke files.

[3] The unfamiliarity of the songs was validated, first, by the low familiarity scores from the experiment (grand mean = .5 on scale from 0 to 2) and also by a new group of 7 college-aged listeners who were asked to rate familiarity on a 1 - 7 scale and try to name all the songs. The overall familiarity rating was very low (grand mean = 1.19/7; the highest familiarity rating for a given song was 2.14) and only one person named two songs out of the 80 songs correctly.

Analytical features of musical structure. The analysis of the experimental results focuses on musical features of the test items that are used as predictor variables in the subsequent statistical models. A few relevant features are described very briefly below both in terms of how they are computed as well as how they relate to cognitive concepts of musical information processing.

The features used in this exploratory study and implemented in the software toolbox FANTASTIC[4] are implementations of formal analytic procedures from a wide range of cognitive or computational music studies and the music theory literature, such as the feature approaches that have been applied to explain similarity ratings (Eerola, Järvinen, & Toiviainen, 2001) and complexity ratings (Eerola, Himberg, Roivianen, & Louhivuori, 2006) for Western and non-Western melodies or features to describe tonality perception in Western music (Temperley, 2007). Other features were adapted from or inspired by similar approaches in computational linguistics (Baayen, 2001), text retrieval (Baeza-Yates & Ribeiro-Neto, 2011), natural language processing (Landauer, McNamara, Dennis, & Kintsch, 2007), and by the work of Huron (2006), who explained several types of cognitive processing of melodies via melodic features and the frequencies with which these features occur in the musical corpora that constitute a listener's musical environment. The conceptual references for all features are given in the software documentation (Müllensiefen, 2009).

The features implemented in FANTASTIC can be divided into first-order features, reflecting only characteristics of the melody item analyzed, and second-order features, reflecting characteristics of the melody item analyzed in the context of a specific corpus of melodies. Most first-order features have a second-order counterpart that captures the same music-structural dimension but also incorporates information about the relative frequency of the particular feature value. We further distinguish between *summary features* (computing a structural aspect of a melodic phrase without taking information about the note order in that phrase into account) and *m-type features* (reflecting the usage and repetition of melodic motives in a melodic phrase by observing note order). We describe one feature per subcategory in the following section to provide a feel for the type of information those categories can capture. The formal definitions of all features mentioned here and in the results section are given in the Appendix. Melodies were represented by the onset time and pitch value of each note, the latter using equal temperament. All other musical attributes necessary for the analyses are derived from this basic pitch and time representation.

First-order features: Summary features. Summary features reflect the hypothesis that a melody is analyzed and abstracted at encoding and that cognitively relevant aspects are each reduced to and stored as one feature value or a few feature values only. This allows for a compact representation of a melody item in memory and facilitates retrieval as well as similarity judgments and rapid categorization of new musical input (for similar approaches in music see Deliège, 1996, and Eerola et al., 2001; for feature-based approach describing visual shapes, see Chong & Treisman, 2005). In our implementation, one structural aspect of a melody is usually represented by one value on a feature scale or one class label in case of a categorical feature. The different summary features implemented in FANTASTIC capture most aspects that are believed to be important in processing of melodies: absolute pitch, pitch intervals, durations (rhythm), melodic contour, tonality, and global characteristics.

As an example, note density (`note.dens`) is one of the simplest summary features and its calculation is simply the number of notes in the melody divided by the duration of the melody in seconds (i.e., the difference between the onset of the last note and the onset of the first note). Despite its simplicity, note density was shown to be an effective predictor of psychological constructs such as similarity perception in earlier cognitive studies (Eerola et al., 2006).

First-order features: M-type features. M-types (or *melody-types*) are meant to capture the hypothesis that in addition or in contrast to reductive features that summarize aspects of melodic structure by single numbers, listeners make use of representations of short melodic subsegments (the *m-types*) when cognitive processing involves a memory component. We assume the frequency with which melodic subsegments appear in a melody and in a musical corpus to be very important for cognitive processing and treat them in various ways similar to word types in linguistic processing. The concept of m-types is related to the *n-grams* concept in natural speech processing, automatic text processing, and text compression. N-gram models have been demonstrated to be very successful in explaining data from implicit learning experiments (Saffran, Johnson, Aslin, & Newport, 1999) and also in modeling several aspects of music processing (Pearce & Wiggins, 2006; Wiggins, Pearce, & Müllensiefen, 2009). M-types operate on pitch intervals and duration ratios of consecutive notes and thus represent pitch and rhythmic information at

[4] Source code and full documentation available at http://www.doc.gold.ac.uk/isms/mmm/?page=Software%20and%20Documentation

the same time. M-types can be thought of as short melodic motives from which a full melody is built (see the Appendix for a more comprehensive explanation of how m-types are derived from a melody). The m-types of a melody are tabulated in a frequency table and the resulting empirical distribution is usually similar in shape to a negative exponential function and very similar to word frequency distributions arising from the frequency analysis of natural language text. Thus, in order to describe a melody in terms of its usage of the m-type vocabulary, we make use of descriptive statistical features developed in computational linguistics and natural language processing. We consider m-types of different lengths (presently length 1, i.e., sequences of 2 notes, to length 5, i.e., sequences of 6 notes) and subsequently take the mean of the feature values for the different lengths considered.

A straightforward feature is m-type productivity (`mean.productivity`), which has been adapted from morphological productivity, a feature widely used in quantitative linguistics (Baayen, 1992). It measures the number of short motives (m-types) only used once in a melody divided by the number of all m-types contained in that melody. Productivity is often understood to be an indicator of sophistication or complexity.

Second-order features. In many studies of verbal memory, frequency information (e.g., word frequency) has been found to be an important predictor for memory accuracy as well as speed in lexical decision tasks. Similarly, our second-order features look at frequencies of melodic features in the melody under study that can be derived from a corpus of melodies, which is why they are also referred to as *corpus-based features* (Müllensiefen, 2009). Within this approach, a corpus of melodies can be taken to approximately represent the music-structural knowledge of a listener who has been exposed to the music of the corpus. Thus, the corpus-based approach allows one to model listening context by proxy, given a sufficiently large collection of music (see Müllensiefen, Wiggins, & Lewis, 2008, for conceptual and technical details of the corpus-based approach in music cognition). This is conceptually similar to vector space approaches in psycholinguistics such as the Latent Semantic Analysis framework (Landauer et al., 2007), where cognitive verbal tasks involving semantic processing are simulated and predicted using statistical information from large text corpora.

Many of our second-order features reflect the commonness or prevalence of a first-order summary feature value for a given melody in the context of a specific corpus. Technically, first-order values of summary features are replaced by their corresponding relative frequencies or probability densities in the corpus. The estimation of probability density values for continuous and quasi-continuous features is realized by using a density estimation function with a kernel smoothing method.

As a simple example we consider the first-order summary feature `glob.duration`, measuring the length of a melodic phrase in seconds. Most melodic phrases in our large collection of pop melodies are between 1 s and 4 s long. Thus, for example, a melody having a common duration (e.g., 1.5 s) for the first-order feature `glob.duration` will be assigned a relatively high probability density value for the corresponding second-order feature dens.glob.duration (the prefix dens indicating a second-order feature based on probability densities), and consequently high values of glob.duration (of 5 s and more) will be assigned low probability density values for the second-order feature `dens.glob.duration`.

For features based on m-types, second-order features generally indicate how frequently the m-types of a given melody are used in a corpus ("does the melody consist of very common or very unusual m-types?") or how the usage (i.e., the frequency distribution) of m-types in a given melody differs from their usage in a corpus ("do m-types that are very common in a corpus, also occur rather frequently in the given melody and is the reverse true as well?"). All second-order m-type features are prefixed by mtcf, abbreviating "m-type corpus frequency." The Spearman correlation of 'term' and 'document' frequencies[5] feature (`mtcf.TFDF.spearman`) uses the well-known rank-based correlation coefficient to indicate whether m-types in a given melody and within a corpus have a similar rank number when ordered by the respective frequencies. Values for this feature can vary between -1 and 1.

One interesting opportunity that the computation of second-order features within this corpus-based approach offers is the possibility of modeling different cognitive listening contexts by employing different corpora. In this study we used second-order features with two different contexts: both the full corpus of 14,063 pop songs representing the history of Western commercial pop music and the 80 stimulus items used as experimental materials. This allowed us to assess the context

[5] We adopt the relevant terminology from computational linguistics and information retrieval here (see, e.g., Baeza-Yates & Ribeiro-Neto, 2011) to make the provenance of these computational principles clear. With *term frequency* we denote the frequency with which an m-type occurs in a melody item and by *document frequency* we mean the number of melodies in a corpus that contain the m-type at least once.

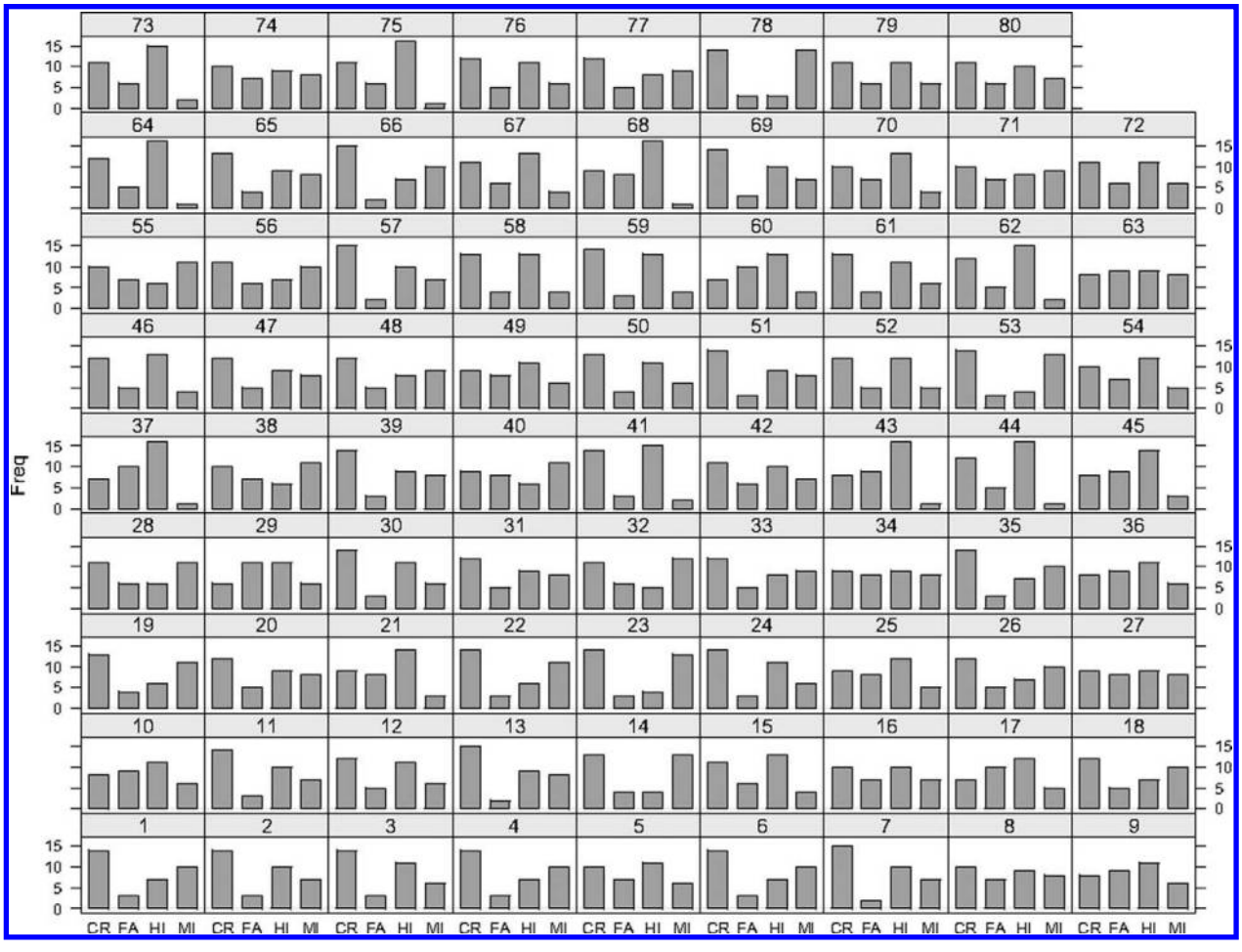

FIGURE 1. Distribution of correct rejections (CR), false alarms (FA), hits (HI), and misses (MI) across the 80 melody items. Some items (e.g., 78, 14, 53) generate a bias towards 'new' responses with a high number of correct rejections and misses. Other items (e.g., 60, 37, 29, 17) almost always trigger 'old' responses, regardless of whether they are correct or not. In addition note that some items appear to be easy, generating a large number of correct responses (e.g., 3, 41, 44,) while other items can be considered difficult given the almost equal proportions in all four categories, which indicates guessing (e.g., 27, 34, 71).

that participants potentially used when performing the experimental tasks.

Results

To obtain an initial overview of the data we categorized the explicit recognition ratings as hits (ratings 4-6 for old items), correct rejections (ratings 1-3 for new items), misses (ratings 1-3 for old items), and false alarms (ratings 4-6 for new items) and aggregated them for all 80 melody items. Figure 1 gives the distribution of the four response categories for each of the 80 items and confirms our starting assumption that items vary to a considerable degree in eliciting "old" (hits, false alarms) vs. "new" responses (correct rejections, misses) and correct (hits, correct rejections) vs. incorrect (false alarms, misses) judgments.

Explaining this variability on the basis of structural features of the melodies is the overall goal of the following analyses. This overall goal is implemented as two research questions with two separate analyses using different analytical techniques. For the first research question and in order to identify clusters of features for predicting explicit and implicit memory accuracy, we used partial least squares regression (PLSR; Wold, 1975) which, in effect, combines advantages of principal component analysis and linear regression. Using PLSR (see Garthwaite, 1994, and Haenlein & Kaplan, 2004, for introductions and background to the PLSR method) we first combined correlated features to form components

(similar to factor scores in factor analysis) and then used these components to predict memory accuracy. We had two main goals that we pursued within the first research question: We wanted to assess the amount of variability in the explicit and implicit memory scores that could be explained by features of the melodic structure alone, i.e., how well could we predict the degree to which individual melodies were remembered just from their structural features? The other goal was to determine the degree to which the same components predicted implicit vs. explicit memory performance. This analysis was carried out on the data aggregated over the 80 items.

To answer the second research question about whether old vs. new judgments are predicted by the same features or differ by context, we analyzed the explicit rating data from each participant for each item without aggregation. Here the focus was not on the amount of variance that could be explained but rather to identify individual features that would elicit subjective feelings of oldness or newness in the participants. We used two linear mixed models for the analysis of the responses on the 6-point rating scale for old and new items separately. A series of variable selection steps was employed to identify individual features in the three contexts that best explained the pattern of old and new judgments.

RESEARCH QUESTION 1: ACCURACY IN IMPLICIT AND EXPLICIT MEMORY TASKS

Participant-wise results. We first aggregated participants' responses over those melody items that were played in the exposure phase ("old items") and those that were not played in the exposure phase ("new items"). We derived as the dependent variable measuring explicit memory the Area Under the Memory Operating Characteristic (AUC) from their ratings on the "heard this tune before" scale. This measure is an unbiased estimate of discrimination weighted by confidence, and varies from 1.0 (perfect) to .50 (chance; Swets, 1973). The dependent measure for implicit memory was the mean difference between old and new items of their ratings on the pleasantness scale. For explicit memory we obtained a mean AUC of .68 ($SD = .10$, 95% CI [.64 .71]) and for implicit memory a mean old-new difference of .29 ($SD = .24$, 95% CI [.20 .37)]. Thus, participants performed well above chance (= .50 and 0 for AUC and old-new, respectively) in both tasks. The amount of the participants' music training as gathered from the background questionnaire was not significantly correlated to either explicit AUC scores ($r = .17$; $t(32) = 0.99$, $p = .33$) nor to implicit old-new differences ($r = .18$; $t(32) = 1.05$, $p = .301$).

Item-wise results. The raw memory scores (correct vs. incorrect judgment) and the pleasantness ratings differed slightly between the two sets of melodies that we used for counterbalancing. The differences were significant but this was due to the large sample size of 1,360 judgments points for each set of melodies. Because the effect sizes for both comparisons were very small ($r_{memory} = .06$; $r_{pleasantness} = .05$) we collapsed the results from both melody sets.[6] The mean of the item-wise AUC scores (measuring explicit memory performance) was .68 ($SD = .11$) and ranged from .42 to .92. The pleasantness-difference score (measuring implicit memory performance) had a mean 0.29 ($SD = 0.45$) and a range from -0.77 to 1.24. The huge ranges for both dependent variables confirms that melodies indeed inherently differed by their memorability. This provided an appropriate context to analyze memorability at least in part by features of their melodic structure.

We used the package 'pls' (Mevik & Wehrens, 2007) from the statistical software environment R (R development core team, 2011) to compute two separate models for explicit and implicit memory tasks using explicit and implicit memory scores for each item as dependent variables. We used all first-order feature variables as well as all second-order features with testset as well as pop corpus as statistical context (134 predictor variables in total). We normalized all variables and we used the built-in cross-validation procedure to assess the fits of the different models. We chose the factor solution where the cross-validation of the root mean squared error of prediction (RMSEP) had its first minimum.

For explicit memory data this led to a PLSR model with two components (adjusted RMSEP = 0.1067 using the 80 leave-one-out cross-validation segments) and for the implicit memory data a model with four factors (adjusted cross-validated RMSEP = 0.3845) was optimal. The explicit model with two components explained 22.34% of the variance within the set of all predictor variables and had an R^2 value of 0.4948 for the explicit memory scores (i.e., it explained 49.48% of the variance). A lower R^2 value of 0.0933 resulted from the cross-validation.

For the implicit data the four-component model explained 30.71% of the variance among the predictor variables and 76.71% of the variance in the implicit scores on the dataset not using cross-validation. From cross-validation we obtained an R^2 value of 0.2534.

[6] We recognize that this is small difference may nonetheless have contributed some variance to the results. Judgment order and old/new were partly but not completely counterbalanced, which should be done in future studies.

TABLE 1. *Components and Feature: oadings of the Partial Least Squares (PLSR) Models for the Explicit and Implicit Memory Models.*

Feature name	Explicit Model		Implicit Memory Model			
	E1	E2	I1	I2	I3	I4
First-order features (No context)						
Summary Features						
i.abs.mean				−.203		
Step.cont.loc.var			.202			
d.mode						.223
d.entropy						.217
d.eq.trans						−.202
Int.cont.grad.std						−.216
M-Type Features						
Mean.entropy					.215	
Mean.productivity		−.205			.248	
Mean.Simpsons.D					−.214	
Mean.Yules.K					−.210	
Mean.Sichels.S		.202			−.214	
Second-order features (Testset as context)						
Summary Features						
d.median						−.209
int.contour.class						.263
int.cont.dir.change						.217
M-Type Features						
mtcf.TFDF.spearman					−.213	
mtcf.TFDF.kendall					−.212	
norm.log.dist					.207	
log.max.DF						−.234
mean.log.DF	−.238		−.254			
mean.g.weight	.239		.255			
mean.gl.weight	.237		.260			
Second-order features (Corpus as context)						
Summary Features						
d.eq.trans						.208
int.cont.grad.std						.201
d.mode						−.227
int.cont.dir.change						.227
M-Type Features						
TFDF.spearman		.226			−.231	
TFDF.kendall		.224			−.229	
mean.log.DF	−.235		−.261			
g.weight	.235		.261			
g.weight			.205			
mean.gl.weight	.238		.266			
TFIDF.m.D				.206		

Note. Loadings of individual feature variables on latent components of the PLSR model are displayed for explicit (E1 - E2), and implicit (I1 – I4) memory performance. For the sake of clarity only features with loadings > |0.2| are displayed and loadings below |0.2| are not shown in the table. For all feature definitions see Müllensiefen (2009).

In order to assess whether the same features predict explicit and implicit memory performance or to what extent the two feature sets overlap we extracted all features with an absolute loading value > 0.2 for any component of the latent factor model (see Table 1). All the features are listed for completeness; the main argument may be followed by attending to the feature categories as indicated by the subheadings.

All latent factors (or components) of both models were positively correlated with the dependent variables. Component E1 of the explicit model explains 13.9% of the variance among the predictor variables. It contains several high-loading feature variables using statistical information that reflect the frequency or uniqueness of the m-types of a melody with respect to the pop corpus and in the testset. Component E2 combines context-free complexity features with two features using statistical information from the pop corpus. Thus, a lower complexity combined with a match of m-type frequencies with the pop corpus seems to benefit explicit recognition.

The implicit model comprises four components. Its first component I1 is very similar to the first component of the explicit model. This set of features relating to the uniqueness of m-types might thus be interpreted as a general factor predicting higher memory accuracy.

Component I2 of the implicit model does not overlap with any other component of the explicit model and combines a feature reflecting contour variability (not using any context information) and variability in the m-type weightings with respect to the pop corpus, which is associated with higher implicit memory scores. Component I3 is interesting because it incorporates a very similar set of features as component E2 of the explicit model but with the sign of the loadings reversed. For implicit memory, a higher context-free complexity and lower repetition rate in combination with a low matching of m-type frequencies between statistical context and in the actual melody generate higher implicit memory accuracy. Thus, almost the same structural features can have differential relationships to explicit and implicit memory. The fourth component of the implicit model (I4) does not overlap with any other component from the explicit nor the implicit model. It mainly combines features reflecting the note durations and pitch contour.

As an interim summary, we note that both explicit and implicit memory are driven by a strong common factor (components E1 and I1) that indexes the uniqueness of short melodic motives. However, each type of memory retrieval has unique predictors or uses the same predictors but with opposite polarity (components E2, I2, I3, I4). For instance, shorter note durations and more variable pitch contours seem to predict better implicit memory but not explicit memory.

RESEARCH QUESTION 2: MELODIC FEATURES BIASING TRUE AND FALSE MEMORIES

We computed two different models, from the dataset of the explicit, non-aggregated memory judgements. The two datasets were based on responses only to those items that had been heard before (old items) and those items that were heard for the first time in the test phase (new items), respectively, each having 1,360 responses (34 participants x 40 melody items).

For both models we followed the same protocol that consisted of several analysis stages. In a first stage we identified a small selection of musical features as potentially powerful predictors, out of the full set of computational features that were used to analyze the melody items. This is because there is generally no built-in or generally accepted mechanism for variable selection in mixed-effects models. We made use of a combination of strategies traditionally employed in psycholinguistics looking at data on an item and participant level separately. For the item-level approach, we fit an ordinary linear regression model using mean-item ratings on the old-new 6-point confidence scale as dependent variable, with the data averaged over participants (Forster & Dickinson, 1976). We use backwards model selection based on the Bayesian Information Criterion (BIC) and arrived at a short list of candidate features with a significant predictive power on the participant level.

For the participant-level approach (Lorch & Myers, 1990) we fit 34 ordinary regression models to the data from individual participants (again with ratings on old-new scale as dependent variable) and then we tested for each feature whether its mean estimate across participants was significantly different from 0 by means of a *t*-test. We arrived at a list of candidate features with significant coefficients across participants. We then chose the features in the intersection of the item- and participant-level models as predictor variables. The result of this variable selection procedure gave us small sets of 14 and 6 features for the old item and new datasets respectively, out of the initial 58 features.

We then applied a mixed-model analysis where we centered participants' ratings around the grand mean and used *participant* and *melody item* as random effects affecting the intercept of the model. We added the musical features as selected by the variable selection procedure as fixed effects and inspected the *p*-values of their individual coefficient estimates after Markov Chain Monte Carlo (MCMC) sampling, and removed those features whose coefficient did not significantly differ from 0 (p values > .05). We compared the model only containing features with significant coefficients to the full model (including all features) as well as to the model where the feature with the next largest p value had been removed in a stepwise fashion. χ^2 tests were then used to compare data fits across models with different number of parameters. We chose the model with the fewest number of parameters (predictor variables) that did not significantly differ from the model with the best likelihood, as tested by the χ^2 tests evaluating likelihood-ratios.

The old-item model. This dataset comprised only the hits and misses, with higher values on the scale indicating more accurate performance. The old-item model only contained three predictor variables (log-likelihood = -2313; $R^2 = 0.276$) and compared very favorably to a null model with a random effect for participants and items only (log-likelihood = -2406; $R^2 = 0.103$) and a model with the best log-likelihood using additional predictors

TABLE 2. *Regression Models for Non-aggregated Memory Recognition Ratings from Old Items Only and New Items Only.*

Fixed effects	*B*	*SE*	*t*	95% CI
Old-item Model				
Constant	2.62	1.19	2.20	[0.4, 4.7]
int.cont.grad.std	0.07	0.03	2.53	[0.0, 0.1]
mtcf.norm.log.dist.testset	−12.88	7.08	−1.82	[-25.4, 0.1]
mtcf.std.g.weight.testset	−7.28	3.99	−1.82	[-14.4, -0.2]
R^2		.276		
New-item Model				
Constant	0.60	0.24	2.47	[0.1, 1.0]
dens.glob.duration.popcorpus	−0.85	0.43	−1.99	[-1.7, -0.1]
mtcf.TFDF.spearman.popcorpus	−1.13	0.50	−2.27	[-2.1, 0.2]
R^2		.229		

Note. CI = Confidence interval.

(log-likelihood = -2296; R^2 = 0.267). Table 2 gives the details of the chosen model for old item trials.

According to this model a (previously heard) melody is more accurately remembered if its contour is more variable (positive coefficient for `int.cont.grad.std`), which might be interpreted as a proxy for a higher surface complexity. The two remaining predictors in this model are based on m-types. `mtcf.norm.log.dist.testset` is a feature that is mainly driven by *document frequencies*, i.e., the number of melodies in the testset that contain the same m-type corrected by the number of occurrences of the same m-type in the given melody. The negative coefficient for `mtcf.norm.log.dist.testset` suggests that items with rare melodies or motives generate more hits. The third feature in this model, `mtcf.std.g.weight.testset`, makes a significant negative contribution, indicating that if all of the melody's motives have roughly the same level of distinctiveness with respect to the testset, the probability of a hit is increased. Summarizing this model we can say that a melody item with a fairly varied contour including lots of steep upward and downward movements, combined with motives that occur infrequently in the testset, increases ratings of "old" and thus makes a melody more memorable among items that had been presented.

Turning the perspective around, the same features are associated with the lack of recognition, or the illusion of novelty that is indexed by a miss: Old melodies with a very flat contour and that make use of motives that are very frequent in the testset as a whole, as well as having varying levels of distinctiveness, do not generate high ratings on the recognition scale and thus appear novel to listeners.

The new-item model. Finally, we computed a mixed-effects model on the basis of the new item trials only. Thus, the data comprise only the correct rejections and false alarms, with higher values on the scale indicating *less* accurate performance.The optimal model only makes use of two features as predictor variables and achieves very good indicators of fit (log-likelihood = -2178; R^2 = 0.229) when compared to a null model with only a random effect for participants (log-likelihood = -2221; R^2 = 0.146) and the full model with likelihood including additional predictors (log-likelihood = -2173; R^2 = 0.225). The details of this optimal model for new trials are given in Table 2. The model only comprises two features that are both derived from statistical context information using the entire pop corpus as context ("popcorpus"). This contrasts with the influence of the testset as a context for the old-item model. However, this is less surprising if one considers the fact that at the point in time where a participant was probed with a new item, this item was not part of the aggregate statistical representation of the corpus of test items that the participant had formed so far. Thus, the only context that participants might use for processing statistical information about features of a new item is their general knowledge of pop melodies.

Similar to the old-item model, high ratings towards the *old* end of the scale are generated by infrequent features. In particular, high recognition ratings (which are false alarms here) are driven by an uncommon global duration with respect to a typical phrase length in pop melodies in general (negative coefficient for `dens.glob.duration.popcorpus`) and if the repeated usage of the melody's motives is very different and atypical compared to how these motives are used in pop music in general (negative coefficient for `mtcf.TFDF.spearman.popcorpus`). In other words, a melodic phrase that is uncommonly short or long and has a high repetition of unusual motives makes listeners believe that they have heard this melodic phrase before, although the item was new.

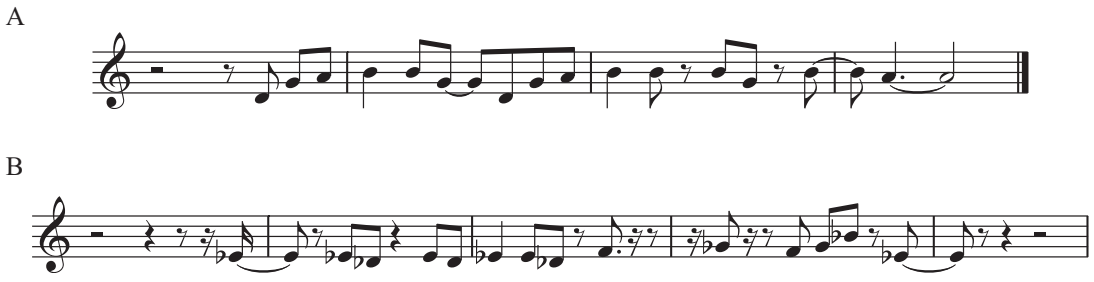

FIGURE 2. Items eliciting the highest average "old" ratings (A: *Under the Boardwalk*, The Drifters, 1964) and the lowest average "old" rating (B: *I Wanna Dance with Somebody*, Whitney Houston, 1986).

When comparing the two models we note that there is only limited overlap in the precise features selected as explanatory variables. However, "old" judgments seem to be driven by rather similar features in both models: Features that indicate the use of infrequent motives are associated with high ratings on the recognition scale. In fact, the features `mtcf.norm.log.dist` (from old-item model) and `mtcf.TFDF.spearman` (from new-item model) are conceptually related and measure the association between motivic frequencies in a corpus and a given melody item, albeit using different techniques (i.e., distance computation and rank-based correlation). Thus, the relationship between infrequent motives and "old" judgments holds true for old items and new items alike. And in fact, when we pooled hits and false alarms in our data across the 80 items we obtained a significant positive correlation of participants' explicit ratings over old and new items of $r = .34$, $t(78) = 3.16$, $p = .002$, 95% CI [.137, .518] confirming that certain items generate "old" judgments regardless of whether they were heard before or not.

Figure 2 shows the items that elicited the highest average "old" rating and lowest average "old" rating. *Under the Boardwalk* was judged as old most often, regardless of whether it had been heard or not. (As a reminder, all tunes were unfamiliar to our young adult participants.) This melody has several large interval jumps, which is uncommon in pop music. Conversely, the short excerpt from *I Wanna Dance with Somebody*, eliciting the fewest "old" judgments on average, has many small musical intervals that repeat frequently, which is similar to many other pop tunes. These most extreme examples from the item testset illustrate how rare melodic motives and the judgments on the "old-new" ratings scale are related.

Discussion

The first analysis set out to investigate the structural features of unfamiliar melodies that facilitate or inhibit implicit and explicit memory performance. In summary, we found that melodies with more unusual melodic motives are recognized better in both explicit and implicit tasks. However, other components clearly distinguish between accuracy in explicit and implicit tasks. Whereas lower motivic complexity combined with a closer match of relative motive frequencies to the corpus help explicit memory accuracy, the reverse is true for implicit memory accuracy. In addition, only implicit retrieval (component I4) seems to make use of contour and rhythm information in a summarized global form and exploits comparisons to the statistical contexts of item testset and pop corpus. A simple and familiar contour shape combined with an unusual and complex rhythm makes a melody easier to retrieve implicitly.

In contrast to our previous work (Halpern & Müllensiefen, 2008) we found explicit and implicit memory scores to be moderately correlated across items. However, with the PLSR analysis we were able to understand this correlation as resulting from a general factor capturing features related to the uniqueness of individual melodic motives with respect to a context. Beyond this factor that is common to both types of memory, we find striking differences between the features explaining explicit and implicit memory performance. Thus, it seems that memorability for melodies is partially an intrinsic and task-independent characteristic of individual items, but different features can also make a short tune more easily encodable and retrievable in either explicit or implicit tasks.

It is interesting to note that the implicit memory models explain a larger proportion of variance than their explicit counterpart. This suggests that explicit memory tasks might be more dependent on individual encoding strategies and incidental associations than implicit memory. Learning novel melodic items in a sequential paradigm is a difficult task (see Halpern & Bartlett, 2010, for a review) often resulting in modest

memory scores. It would thus be interesting to see whether feature models for explicit and implicit memory would substantially differ for data from an experimental paradigm with multiple exposures.

The role of referential context was one of the important motivations of this study, given the explanatory power of features defined with reference to linguistic (see Howard, Jing, Addis, & Kahana (2007) for a review) or music corpora (Bailes, 2010). We note a number of interesting differences in the two models regarding the different classes of features (summary vs. m-type features) and the three different contexts (no context, testset, pop corpus). The general memory component in neither PLSR model makes much use of first-order features but does draw heavily upon features that make use of statistical information from testset and pop corpus. And in fact, neither component of the explicit PLSR model has any summary feature among their highly loading variables. Explicit memory performance seems to be mainly driven by the recognition of unique motivic elements and "uniqueness" is only meaningful with reference to a corpus. However, whether the features used statistical information from testset or the full pop corpus did not much matter: in both models we found that the testset and corpus variants of the same second-order feature loaded highly on the same component. This might be a result of our attempt to select a testset of items that represent a fair sample of the corpus.

In contrast to the explicit memory model, the implicit model makes more use of first-order features. This suggests that for implicit memory processing, participants rely on surface characteristics of the stimulus items over and above statistical information from prior musical listening. This greater dependence on first-order item features might explain why models of implicit memory explain a greater proportion of the variance in accuracy as shown above. Here, many surface and sensory characteristics of the melody items were controlled: isolated melodies without any accompaniment, neutral piano timbre, controlled expressivity with respect to microtiming and dynamics, rapid presentation of many unrelated melodies as a list). This situation may be less favorable for the implicit use of prior musical knowledge acquired from real-world listening episodes. Thus, the characteristics of the individual items and the testset as a context become relatively more important in the implicit task. It is impressive that these computations of complex musical relationships take place presumably automatically, and in individuals without much formal training in music.

Interestingly, the amount of music training, from none to moderate, did not have a measurable relationship to memory performance. Thus, item characteristics as modeled by the computational features predicted implicit and explicit memory performance in people both with absolutely no training and those with some training. This finding is in line with a number of prior studies (summarized in Halpern & Bartlett, 2010) showing that performance in standard memory paradigms for melodies are relatively insensitive to specific training (or the innate skills) of musicians.

The results supporting at least partial separation of mechanisms subserving the two ways of querying memory constitute working hypotheses for future studies aimed at testing our models of memory in general, and for musical memory in particular. Melodies could be selected that have high and low memorability scores according to the models derived in this study. If the models are valid, then memory performance should significantly differ for the two groups of items. Even more interesting would be to identify melody items with high explicit but low implicit memorability scores (and vice versa) for which we would hypothesize differing explicit and implicit memory performance.

Concerning our second research question using a feature-based analysis approach for raw memory judgments in a melodic recognition task, we found that relatively few features were required to significantly predict which melodies are more likely to elicit judgments of "old" in recognition memory. Furthermore, although the features selected in each model were not identical, a large proportion of them captured, in various ways, a tendency to call "old" those melodies containing relatively rare motivic patterns. This occurred whether the melody was, in fact, old (hits) or new (false alarms), so that both true and false feelings of oldness were driven by these rare motives. This relationship went contrary to our prediction that relatively *common* features would have increased the tendency to call an item old, based on generalized feelings of feature familiarity.

This relationship between infrequent features and "old" judgments stands in contrast with the often observed *mirror effect* in memory, in which factors that increase high hit rates also tend to reduce false alarm rates (Dobbins & Kroll, 2005). Interestingly though, a few recent studies examining recognition memory for verbal material on an item level failed to find the typical mirror effect for word frequency (Cortese et al., 2010; Kang, Balota, & Yap, 2009). In the latter study, hit and false alarm rates were even found to be significantly correlated over items. As an explanation, Cortese et al. point to sublexical processes (e.g., orthographic and phonological processing) that could gain importance especially when semantic processing is not possible for

all items, e.g., if the context contains non-words. Given that unfamiliar nonverbal music largely lacks a semantic dimension, this argument might apply even more strongly in our case and can explain why similar features drive old judgments for old and new items, i.e., why hits and false alarms are both largely driven by the use of infrequent motives.

A possible interpretation of the empirical relationship between items with infrequent short motives and higher "old" ratings could be a single-process model for explaining subjective melody recognition: Listeners, even musically untrained ones, seem to be quite sensitive to uncommon short motives and seem to build up a frequency distribution of them from the experimental dataset, after only one exposure to each item. We hypothesize that an item containing an unusual melodic motive is registered by the listener as a special event either at encoding or test. This registration of a special event in the test phase might then be compared to an existing (and probably relatively strong) memory trace for this particular motive from the study phase, and generates a hit. This assumption is supported by the fact that in the old-item model, the features reflecting commonness of motives are defined with reference to the testset. In contrast, if no memory trace exists because the item is novel, the mere fact that the unusual motive is registered is then misattributed to recognition of that item and leads to a false alarm. In this case the uncommonness of the motive is judged according to the listener's general knowledge about pop music. This is supported by the fact the features selected as significant predictors in the new-item model are defined with reference to the entire corpus of pop melodies.

This misattribution may be related to the illusions of memory engendered by perceptual salience. Items that are easier to process, such as a loudly spoken word or a visually presented word in a large font, are given higher Judgment of Learning (JOL) ratings at encoding, but these items are not necessarily remembered better (Rhodes & Castel, 2009). In a list of unfamiliar tunes, which have no semantic content, statistical rarity of patterns may also be salient and receive priority in processing. Our response scale was implemented as a set of confidence ratings. Thus the significant features in the models predict increasing confidence in an "old" judgment, which is a type of meta-cognitive self-assessment. Increased salience was apparently interpreted as an increased subjective probability of the item being old. For detecting old items, this is a helpful memory mechanism because attention is directed to the rare motive in study and test phase. However, if the item is novel, the listener may confuse the detection of a rare motive with an actual match in memory. The misattribution of fluent processing due to *salience* to fluent processing due to *prior presentation*, causes false alarms.

Both analyses demonstrated that determining the influence the structural features of musical items can be a powerful analytic approach for music cognition research. This approach is largely unsupervised and makes few assumptions about participants (other than a shared musical culture) or the immediate learning context. However, as indicated by the moderate fit indices of all models, the present analyses do not explain everything about memory for melodies. At an item-level, some important aspects of musical structure likely were not covered by the features in the present feature set and thus may not have been optimal. At the participant level, we could not capture variance in the data due to individual differences in general or musical memory performance. On the other hand, we have been very cautious in assessing the fits of all models by using cross-validation and random effects models that use unbiased estimators and include a shrinkage correction for parameter estimates. Thus, the R^2 values reported above are much more conservative than they would have been for ordinary least squares regression models.

The results from both feature analyses indicates that distinctiveness in the short motives that constitute the melody items is very important for explaining good explicit and implicit retrieval and also for explaining high recognition ratings even if recognition is false. This ties in with the fact that most high-loading features of the two PLSR models came from the m-type category (based on short motives) and in fact, the results from all four models from both analyses seem to suggest that short melodic motives are the main building blocks in memory for melodies.

The results from all models also point to the importance of the listening context and implicit structural knowledge. Four out of the five features from the two mixed effects models made use of statistical context information and almost all high-loading features of the PLSR model for explicit recognition also made use of statistical information from the testset or pop corpus. Context-free features not using any statistical information about music are mainly found in the PLSR model for implicit memory. The two mixed effects models as well as the explicit PLSR model suggest that two types of statistical learning of musical context information are employed for solving the memory task: immediate learning of local context and long-term statistical learning of regularities in a familiar genre. Neither type of learning is deliberate and both take place in nonexpert participants.

A simplified account of these results can serve as advice to a hit song composer: A tune that aims to be highly memorable for explicit recognition ("I know that tune!") needs to make use of melodic motives that are rare in terms of their occurrence in a corpus. It also needs to repeat all motives frequently while relative frequency for using the motives should resemble their relative frequencies in the corpus. For a tune that aims to be perceived as more pleasurable on repeated listenings (and thus be remembered well implicitly) the usage of unique motives is also important. In addition, less repetition in its motives and a smaller average interval size as well as simple contour but complex rhythm is also beneficial for implicit memorability.

Author Note

Correspondence concerning this article should be addressed to Daniel Müllensiefen, Department of Psychology, Goldsmiths, University of London, New Cross Road, New Cross London SE14 6NW. E-mail: d.mullensiefen@gold.ac.uk

References

Baayen, H. (1992). Statistical models for word frequency distributions: A linguistic evaluation. *Computers and the Humanities, 26(5-6)*, 347-363.

Baayen, H. (2001). *Word frequency distributions*. Dordrecht: Kluwer Academic Publishers.

Baeza-Yates, R., & Ribeiro-Neto, B. (Eds.). (2011). *Modern information retrieval: The concepts and technology behind search* (2nd ed.). Harlow: Pearson Education.

Bailes, F. (2010). Dynamic melody recognition: Distinctiveness and the role of musical expertise. *Memory and Cognition, 38*, 641-650.

Bartlett, J. C., Halpern, A. R., & Dowling, W. J. (1995). Recognition of familiar and unfamiliar music in normal aging and Alzheimer's disease. *Memory and Cognition, 23,* 531-546.

Bartlett, J. C., Hurry, S., & Thorley, W. (1984). Typicality and familiarity of faces. *Memory and Cognition, 12,* 219-228.

Bartlett, J. C., & Snelus(1981). Lifespan memory for popular songs. *American Journal of Psychology, 93*, 551-560.

Cabeza, R., & Kato, T. (2000). Features are also important: Contributions of featural and configural processing to face recognition. *Psychological Science, 11*, 429-433.

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research, 45*, 891-900.

Cortese, M. J., Khanna, M. M., & Hacker, S. (2010). Recognition memory for 2,578 monosyllabic words. *Memory, 18*, 595-605.

Daselaar, S. M., Rombouts, S. A., Veltman, D. J., Raaijmakers, J. G., & Jonker, C. (2003). Similar networks activated by young and old adults during the acquisition of a motor sequence. *Neurobiology of Aging, 24,* 1013-1019.

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

Demorest, S. M., Morrison, S. J., Beken, M. B., & Jungbluth, D. (2008). Lost in translation: An enculturation effect in music memory performance. *Music Perception, 25*, 213-223.

Dobbins, I. G., & Kroll, N. E. (2005). Distinctiveness and the recognition mirror effect: Evidence for an item-based criterion placement heuristic. *Journal of Experimental Psychology-Learning Memory and Cognition, 31*, 1186-1198.

Dowling, W. J., Bartlett, J. C., Halpern, A. R., & Andrews, M. W. (2008). Melody recognition at fast and slow tempos: Effects of age, experience, and familiarity. *Perception and Psychophysics, 70,* 496-502.

Eerola, T., Himberg, T., Toivianen, P., & Louhivuori, J. (2006). Perceived complexity of western and African folk melodies by western and African listeners. *Psychology of Music, 34,* 337-371.

Eerola, T., Järvinen, T., & Toiviainen, P. (2001). Features and perceived similarity of folk melodies. *Music Perception, 18*, 275-296.

Forster, K. I., & Dickinson, R. G. (1976). More on the language-as-fixed-effect fallacy: Monte Carlo estimates of error rates for F1, F2, F′, and min F′. *Journal of Verbal Learning and Verbal Behavior, 15*, 135-142.

Garthwaite, P. H. (1994). An interpretation of partial least squares. *Journal of the American Statistical Association*, *89*, 122-127.

Graf, P., & Schacter, D. L. (1985). Implicit and explicit memory for new associations in normal and amnesic patients. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 501-518.

Haenlein, M., & Kaplan, A. M. (2004). A beginner's guide to partial least squares (PLS) analysis. *Understanding statistics–statistical issues in psychology, education and the social sciences, 3*, 283-297.

Halpern, A. R., & Bartlett, J. C. (2010). Memory for melodies. In M. Jones, A. Popper, & R. Fay (Eds.), *Music perception* (pp. 234-258). New York: Springer-Verlag.

Halpern, A. R., Bartlett, J. C., & Dowling, W. J. (1995). Aging and experience in the recognition of musical transpositions. *Psychology and Aging*, *10*, 325-342.

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

Halpern, A. R., & O'Connor, M. G. (2000). Implicit memory for music in Alzheimer's disease. *Neuropsychology, 14,* 391–397.

Henson, R. N., Hornberger, M., & Rugg, M. D. (2005). Further dissociating the processes involved in recognition memory: An fMRI study. *Journal of Cognitive Neuroscience, 17*, 1058-1073.

Howard, M. W., Jing, B., Addis, K., & Kahana, M. K. (2007). Semantic structure and episodic memory. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 121-142). Mahwah, NJ: Erlbaum.

Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation*. Cambridge, MA: MIT Press.

Kang, S. H., Balota, D. A., & Yap, M. J. (2009). Pathway control in visual word processing: Converging evidence from recognition memory. *Psychonomic Bulletin and Review, 16*, 692-698.

Kopiez, R., & Müllensiefen, D. (2011). Auf der Suche nach den 'Popularitätsfaktoren'in den Song-Melodien des Beatles-albums Revolver[In search of popularity factors in the tunes from the album Revolver by the Beatles]. In S. Meine & N. Noeske (Eds.), *Music und Popularität: Aspekte zu einer Kulturgeschichte zwischen 1500 und heute* (pp. 207-225). Münster: Waxmann.

Johnson, M. K., Kim, J. K., & Risse, G. (1985). Do alcoholic Korsakoff syndrome patients acquire affective responses? *Journal of Experimental Psychology: Learning Memory and Cognition, 11*, 3-11.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.) (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Lawrence Erlbaum.

Lorch, R. F., & Myers, J. L. (1990). Regression-analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning Memory and Cognition, 16*, 149-157.

Loui, P., & Wessel, D. L. (2006). Acquiring new musical grammars: A statistical learning approach. In M. Baroni, A. R. Addessi, R. Caterina, M. Costa (Eds.), *Proceedings of the 9th International Conference on Music Perception and Cognition* (pp. 1009-1017). Bologna, Italy: ICMPC-ESCOM.

Mevik, B. H., & Wehrens, R. (2007). The pls package: Principal component and partial least squares regression in R. *Journal of Statistical Software*, *18*, 1-24.

Müllensiefen, D. (2009). *FANTASTIC: Feature ANalysis Technology Accessing STatistics (In a Corpus)* [Technical report v1.5]. London: Goldsmiths, University of London. Retrieved July 14, 2010, from http://www.doc.gold.ac.uk/isms/m4s/FANTASTIC_docs.pdf

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PLoS ONE*, 9, e89642.

Müllensiefen, D., Wiggins, G., & Lewis, D. (2008).High-level feature descriptors and corpus-based musicology: Techniques for modelling music cognition. In A. Schneider (Ed.), *Hamburger Jahrbuch für Musikwissenschaft, 24* (pp. 133-155). Frankfurt: Peter Lang.

Pearce, M. T., Müllensiefen, D., & Wiggins, G. A. (2010). The role of expectation and probabilistic reasoning in auditory boundary perception: A model comparison. *Perception, 39*, 1367-1391.

Pearce, M. T., & Wiggins, G. A. (2006). Expectation in melody: The influence of context and learning. *Music Perception, 23,* 377-405.

Pelucchi, B., Hay, J. F. & Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development, 80,* 674-685.

Peretz, I., Gaudreau, D., & Bonnel, A.-M. (1998). Exposure effects on music preference and recognition. *Memory and Cognition*, *26*, 884-902.

Quesada, J. (2007). Creating your own LSA spaces. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch(Eds.), *Handbook of latent semantic analysis* (pp. 71-85). Mahwah, NJ: Erlbaum.

R development core team (2011). *R: A language and environment for statistical computing*. Vienna: R foundation for Statistical Computing.

Reder, L. M., Park, H., & Kieffaber, P. D. (2009). Memory systems do not divide on consciousness: Reinterpreting memory in terms of activation and binding. *Psychological Bulletin, 13*, 23-49.

Rhodes, M. G., & Castel, A. D. (2009). Metacognitive illusions for auditory information: Effects on monitoring and control. *Psychonomic Bulletin and Review, 16*, 550-554.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*, 27-52.

Schacter, D. L., & Buckner, R. L. (1998). On the relations among priming, conscious recollection, and intentional retrieval: Evidence from neuroimaging research. *Neurobiology of Learning and Memory, 70*, 284-303.

Schacter, D. L., Chiu, C-Y. P., & Ochsner, K. N. (1993). Implicit memory: A selective review. *Annual Review of Neuroscience, 16*, 159-182.

Schulkind, M. D., Posner, R. J., & Rubin, D. C. (2003). Musical features that facilitate melody identification: How do you know it's "your song" when they finally play it? *Music Perception, 21,* 217-249.

Swets, J. A. (1973). The relative operating characteristic in psychology. *Science, 182,* 990-1000.

Temperley, D. (2001). *The cognition of basic musical structures.* Cambridge, MA: MIT Press.

Temperley, D. (2007). *Music and probability.* Cambridge, MA: MIT Press.

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

Wiggins, G. A., Pearce, M. T., & Müllensiefen, D. (2009). Computational modelling of music cognition and musical creativity. In R. Dean (Ed.), *The Oxford handbook of computer music* (pp. 383-420). Oxford, UK: Oxford University Press.

Wold, H. (1975). Soft modelling by latent variables: The non-linear iterative partial least squares (NIPALS) approach. In J. Gani (Ed.), *Perspectives in probability and statistics: Papers in honour of M. S. Bartlett* (pp. 117-142). London: Academic Press

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441-517.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist, 35,* 151-135.

Appendix

DEFINITION OF ANALYTICAL FEATURES

We represent melodies in the most basic form as a sequence of notes n_i where each note is a tuple of an onset time value t_i measured in seconds and a pitch value p_i. Because we are dealing with melodies from Western cultures using an equal-tempered tuning system, pitch values are represented in MIDI.

In the following all features discussed in the text are defined using formula notation.

Global duration (`glob.duration`):

This is the difference between the onset of the last and the first note measured in milliseconds.

$$glob.duration = t_n - t_1$$

t_1 and t_n designate the onset time value of its first and last note, respectively.

Note density (`note.dens`):

$$note.dens = \frac{\#notes}{t_n - t_1}$$

#notes represents the number of notes in a melody.

Interpolation Contour Gradients Standard Deviation (`int.cont.grad.std`):

This feature is based on a representation of melodic contour as an interpolation between the high and low points (i.e., contour turning points) of a melody using straight lines.

We substitute the pitch values of a melody with the sequence of gradients that represent the direction and steepness of the melodic motion at evenly spaced points in time. The gradient values are represented by the interpolation contour vector x and the variability of the contour of a melodic phrase can be measured as the standard deviation of the interpolation contour vector x:

$$\text{int.cont.grad.std} = \sqrt{\frac{\sum_i (x_i - \bar{x})}{N - 1}}$$

M-type features and second-order m-type features:

M-types operate on pitch intervals and duration ratios of consecutive notes and are created from a melody in several steps: Firstly, a melody is segmented into melodic phrases by means of the standard algorithm Grouper (Temperley, 2001). Then for each phrase, all pitch intervals are classified into 19 different categories and all duration ratios are classified into three different categories (corresponding roughly to whether the second note is shorter, equal, or longer than the first one). After this, for each pair of adjacent notes the pitch interval class and the duration ratio class are represented as a 3-digit symbol. As a next step, a window of length n is slid over the string of symbols. The content of the window at any position in the melodic phrase is called an m-type. All m-types encountered in all melodic phrases are tabulated and for each m-type its frequency of occurrence in the entire melody is recorded. This procedure is repeated for different window lengths (currently $n = 1, \ldots, 5$, i.e., spanning two to six notes in the original melody) and each time a separate table of m-types (of a certain length n) is produced.

In this study only second-order variants of m-type features proved to be of explanatory power that make use of frequency statistics from a music corpus. In analogy to the terminology used in computational text retrieval we use the names *term frequency* ($TF(\tau_i)$) to denote the relative frequency of m-type τ_i in the given melody and *document frequency* ($DF(\tau_i)$) to mean the relative number of melodies (i.e., documents) in the corpus where m-type τ_i occurs at least once.

Spearman correlation of term and document frequencies (`mtcf.TFDF.spearman`):

For this feature term and document frequencies are ranked separately and are then compared using Spearman's rank correlation. For ties the minimum rank is assigned to all the values of the same rank.

Normalized distance of term and document frequencies (`mtcf.norm.log.dist`)**:** This feature uses document frequencies to index rare m-types and adjusts the document frequencies by the term frequency of the m-type. $\tau \in m$ denotes an m-type occurring in melody m. The logarithm of the raw frequencies is taken and TF and DF are both normalized according to

$$TF'_{\tau_{i \in m}} = \frac{\log_2 TF_{\tau_i \in m}}{\sum_i \log_2 TF_{\tau_i \in m}}$$

and then the absolute difference between the two vectors is calculated element-wise and normalized by the number of m-types occurring in melody m ($|\tau \in m|$).

$$mtcf.norm.\log.dist = \frac{\sum_{\tau_i \in m} |TF'_{\tau_i}| - DF'_{\tau_i}}{|\tau \in m|}$$

Mean product of term and document frequencies (`mtcf.mean.log. TFDF`)**:**

This feature is very similar to the previous one except that term and document frequencies are combined by vector multiplication.

$$mtcf.mean.\log.TFDF = \frac{TF'_{\tau \in m} DF'_{\tau \in m}}{|\tau \in m|}$$

Mean document frequency productivity (`mtcf.mean.productivity`)**:**

This is the number of m-types only occurring once in the corpus (in linguistics these are known a hapax legommena) divided by N, the number of all m-tokens, in a melody.

$$mean.productivity = \frac{\sum_n \frac{V(1,N)}{N}}{|n|}$$

Here $V(1, N)$ is a function denoting the number of m-types occurring once among m-tokens of a set. n represents the lengths of the m-type and $|n|$ is the number of different m-type lengths considered.

Standard deviation of global frequency weights (`mtcf.std.g.weight`)**:**

For this feature, each m-type is weighted according to its frequency in a given melody as well as in a corpus using a weighting scheme suggested by Quesada (2007). The weight is based on the ratio $P_{m,\,C}(\tau)$, i.e., the ratio of its local frequency in a melody m of corpus C to the overall frequency in C:

$$P_{m,C}(\tau) = \frac{f_m(\tau)}{f_C(\tau)}$$

Then, the global weight of this m-type τ is calculated using entropy-based weighting:

$$glob.w = 1 + \frac{\sum_{m \in C} P_{m,C}(\tau).\log_2(P_{m,C}(\tau))}{\log_2(|C|)}$$

The final feature is then the standard deviation of the global weights of all m-types τ of a melody m.