Chapter 9

Sloboda and Parker's recall paradigm for melodic memory: a new, computational perspective

Daniel Müllensiefen and Geraint A. Wiggins

Abstract

Sloboda and Parker (1985) proposed a new experimental paradigm for research on melodic memory in which participants are asked to listen to novel melodies and to sing back the parts they can recall from memory. In contrast to the many varieties of melodic recognition paradigms frequently used in memory research this sung recall paradigm can answer questions about how mental representations of a melody build up in memory over time, about the nature of memory errors, and about the interplay between different musical dimensions in memory. Although the paradigm has clear advantages with regard to ecological validity, Sloboda and Parker also note a number of difficulties inherent to the paradigm that mostly result from necessity to analyse 'dirty musical data' as sung by mostly untrained participants. This contribution reviews previous research done using the sung recall paradigm and proposes a computational approach for the analysis of dirty melodic data. This approach is applied to data from a new study using Sloboda and Parker's paradigm. This chapter discusses how this new approach not only enables researchers to handle large amounts of data but also make use of concepts from computational music analysis and music information retrieval that introduce a new level of analytic precision and conceptual clarity and thus provide a new interface which connects Sloboda's paradigm to rigorous quantitative data analysis.

The seminal study by John Sloboda and David Parker (1985) provided new insights into the construction of mental representations of melodic structures, and into mechanisms whereby learning of previously unknown melodies is accomplished over repeated attempts. In doing so, it introduced a new experimental paradigm for investigating melodic memory via sung recalls. The work is widely cited in subsequent literature on musical and melodic memory. However, despite the frequent citations and wide dissemination of the results of the study, the use of the recall paradigm has been limited to just a handful of subsequent studies. It has not achieved the level of usage and variation in the music cognition community achieved, for example, by paradigms based on recognition tasks.

In this chapter, we review the Sloboda and Parker paradigm from a methodological perspective, and hypothesize that one of the reasons why its uptake in the literature is at odds with the originality of result that it can produce are the relatively 'dirty' data it produces. A lack of standardized methods and the amount of manual work required to deal with these types of data seem to be major impediments to uptake of the paradigm. To ameliorate these drawbacks, we propose use of the computer as a tool for musical data analysis, and also as a means of model building and hypothesis generation. We demonstrate how it can help to make analysis procedures explicit and thus contribute to standardization. We reanalyse the original data from Sloboda and Parker's (1985) study, to demonstrate where computing technology can be applied and what additional value it can bring to the analysis of data from this very rich paradigm, whose full potential, we argue, has yet to be realized.

Paradigms for general memory research

In classic textbooks on experimental psychology (Anderson & Borkowski, 1978, 394–396; Kantowitz, Roediger, & Elmes, 1994, pp. 284–285; Kluwe, 1990), we find a well-developed canon of experimental paradigms that have been used to investigate verbal memory. The earliest go back to the very beginnings of psychology as a discipline and all of them have evolved and improved with time. For the purposes of the argument here a short description of each paradigm, in summary, will suffice.

Serial recall is the oldest experimental paradigm in the psychology of memory. It was first proposed by Hermann Ebbinghaus (1885) and has since been widely deployed in experimental studies, and considerably refined. Its basic task is to remember a list of items in order and recall them subsequently in the same order. Ebbinghaus and other well-known subsequent studies (e.g. Young, 1962) used meaningless syllables as experimental stimuli. Results generated by this paradigm inform us about the amount of repetition necessary to recall a list of items perfectly, the savings in effort and time when lists are relearned, and the decay of items in memory over time.

Paired-associate learning is another very early memory paradigm (Calkins, 1894) and can be directly linked to the stimulus–response concept of classical conditioning (Pavlov, 1927). Participants are given a list of pairs of items to memorize. They are then presented with a list containing one item of each pair, and asked to recall its match from the learned list before the match is shown. This process is repeated until the full list of pairs is memorized perfectly. Meaningless syllables and single words are used as items in verbal learning experiments using this paradigm. Central questions investigated concern item characteristics (e.g. the similarity between items, their imageability), the formation of association strengths, and the effects of cognitive mediators on recall performance.

Free recall, in contrast with serial recall, requires participants to learn a list of items, but leaves them free to recall the learned items in any order. If the study list is presented several times, then the position of the items is permuted, to avoid order associations. The most important observable in this paradigm is the effect of the position of an item in the study list on the probability of its recall. Regardless of the length of the study list, the initial items and the final items in the list are usually much better recalled than the items presented in middle positions (e.g. Murdock, 1962). These effects are known as *primacy* and *recency*; they can be reproduced very reliably with the *recency* effect usually being stronger than the *primacy* effect. One interpretation of these effects (Atkinson & Shiffrin, 1968) was that they suggest the existence of distinct memory stores where the first items of a list would have entered a long-term store and the last items would have still been present in a short-term store.

Recognition requires participants are to study a list of items, but, in contrast to the recall paradigms, described above, they are subsequently presented with a list containing items from the study list, and also new, previously unseen items. Participants are then asked to indicate which items were encountered previously, or one seen item can be presented along with several unseen items in a multiple-choice selection task. Recognition memory performance is generally much higher than recall memory performance (e.g. Shepard, 1967), though there are exceptions (Tulving, 1968). Not only can participants generally recognize high percentages (>80%) of long item lists (e.g. 500 items and more) but they are also able to maintain good recognition memories over long time intervals. This long-term performance is in clear contrast to the negative exponential forgetting curve that Ebbinghaus (1885) and others found with serial and other recall paradigms; it holds true for many different types of items (e.g. images, words).

Choosing a paradigm

There are good reasons for the existence and deployment of a wide range of experimental paradigms in memory research. First of all, different experimental procedures allow us to study different memory effects and to provide evidence for different types of hypothesis. For example, effects of list length and list position on correct recall from memory are typically studied using serial recall (Young, 1962), while ability to discriminate between stimuli has been investigated using the paired-associate learning paradigm (Underwood, Rundquist, & Schulz, 1959), and the effects of different rehearsal strategies can be revealed using the free recall procedure, for example with overt rehearsal (Rundus, 1974). Similarly, results from the various paradigms and their potential to falsify hypotheses of different nature have led to the proposal of correspondingly different memory models. For example, a specific class of memory trace model is primarily based on serial recall data (e.g. Nairne, 1990), while models using semantic hierarchies and clustering are often developed in connection with data from free recall studies (Tulving, 1968). Finally, data stemming from different experimental paradigms is often used to provide complementary evidence about the same model or hypothesis. This can help refute the claim that effects discovered with a certain paradigm are merely artefacts of that paradigm. Good examples are primacy and recency effects which reliably appear in both serial recall (Mueller, 1970) and free recall paradigms (Murdock, 1962). Complex, modern memory models that claim general applicability are usually based on data generated with diverse and very different paradigms. One such example is Baddeley's concept of working memory (1986, 2007) which is able to account for a large number of findings from recall studies using a variety of different tasks.

Paradigms in music memory research

Overview

Turning to research on memory for music, and specifically memory for *melodies*, we find a more limited range of basic experimental paradigms. If we naïvely suppose that the individual notes or pitch intervals of a melody are units comparable to words,[1] syllables, or digits in verbal memory research, then a free recall paradigm in a strict sense would be, by definition, melodically meaningless (though it might convey harmonic information): if notes are not recalled in a serial order, but freely, then it is hard to recognize the stimulus item and therefore to judge whether a participant has actually recalled it, and, if so, to what degree of accuracy. Some melodic recall studies include explicit instructions for the participants that the parts (sections, phrases, motives) of the melody items may be recalled in any order. Especially with longer melody items, this approach can make the experimental task much easier for the participants. Here, the paradigm comes closer to the verbal free recall paradigm, but it must still be considered a compromise variant of the basic melody recall procedure. We therefore conclude that there is essentially only one recall paradigm in music research, which is broadly comparable with verbal serial recall, *regardless of how the recall is actually performed in the experiment*. As we will see below, depending on the target group of participants, recall response modes can range from singing through playing on an instrument to using music notation or verbal labels for individual notes. In all instances though, the serial order of the recalled notes is of primary importance.

This said, it is possible, and not unreasonable, to define very abstract representations of melody which do not rely on strict note sequence—for example, by pitch class or interval counting. However, experience suggests that these are not good representations for musical memory: order does seem to be paramount, as one might expect. These representations can nevertheless be useful in applications such as music information retrieval (for example, harmonic interval statistics are used for this purpose by Pickens *et al.*, 2003).

[1] In fact, music is extremely context-sensitive, and, statistically, at least, there seems to be some comparison between sequences of three-or four-note chords and words (Mauch, Müllensiefen, Dixon, & Wiggins, 2008), though how deep this comparison runs is far from clear.

Nor is paired-associate learning often seen in music research. This may be again due to the sequential nature of music, and melodies in particular. It is hard to imagine a musically meaningful task where musical or melodic elements can be presented in pairs and to be encoded together. At least, one would have to consider passages of music that can be paired according to specific attributes (e.g. timbre, pitch range, tempo, specific rhythms). This paradigm would clearly make no musical sense at all if one took notes or intervals as the atomic elements to be remembered in pairs.

Recognition paradigms in music memory research

Recognition is the only other memory paradigm available for research on memory for melodies. Recognition has been used extensively over the past 40 years and has diversified into a few major subparadigms, the most important of which we now briefly describe.

Pitch comparison The **Deutsch paradigm** was first used by its eponymous designer in the early 1970s, in a tradition of many studies targeting short-term memory, its capacity, and the conditions for interference effects for verbal memory. Deutsch dedicated a series of studies to the limits of the auditory store holding pitch information and to the question of how memory representation for pitch could be eliminated by interference effects (Deutsch, 1970, 1972, 1974, 1975a, 1975b; Deutsch & Feroe, 1975). The paradigm used in these studies was later used by other researchers in follow-up studies. In its basic form, the Deutsch paradigm can be described thus. The participant is presented with a single target pitch which they are asked to keep in memory. Then, they listen to some intervening stimuli in an *interpolation* phase (called *retention* elsewhere): typically, the stimuli in the interpolation phase are varied as factors of the independent variable. These stimuli can be verbal auditory material or a varied number of pitches which can be more or less harmonically or melodically related to the target stimulus. Finally, a comparison pitch is played and the participant judges whether it is the same as the initial pitch or not.

The results generated by experiments using this paradigm seem to indicate that there may be a store for pitch information, separate from any verbal auditory store, and which is influenced only to a minor degree by other characteristics of a tone, such as loudness, timbre, or direction (i.e. the ear it is presented to). However, this pitch memory system, which has been hypothesized to be a separate subsystem in Baddeley's auditory working memory (Pechmann & Mohr, 1992), is fairly strongly influenced by pitches presented in the interpolation phase of the Deutsch paradigm. The interference with the memory representation of the target pitch is particularly strong if the interpolation pitches are close to the target in continuous pitch space, or if they are identical to the comparison pitch.

AB comparison is widely used in melodic memory research (e.g. DeWitt & Crowder, 1986; Dowling, 1972, 1978; Dowling & Fujitani, 1971; Idson & Massaro, 1978). It is similar to the Deutsch paradigm in that the participant is asked to compare two stimuli and indicate whether they are identical or not, and, sometimes, rate their confidence in their judgement. But with AB comparison, the target and comparison stimuli are longer melodic sequences, parts of real melodies, or excerpts from a polyphonic piece.

The retention phase between the presentation of the target stimulus, A, and the comparison stimulus, B, is less relevant than in the Deutsch paradigm: material presented in this phase often serves only as a distractor.

Variation in applications of AB comparison is often in the melodic attributes that are held the same or made different between A and B. The rationale is that, if two melodic sequences differ in a specific melodic attribute (e.g. contour) but participants nonetheless indicate that A and B are identical, then this particular attribute seems not to have been encoded in memory (Idson & Massaro, 1978, p. 554). Therefore, the paradigm is particularly well suited to uncovering which melodic attributes are encoded in memory, and how different melodic structures and contexts favour the encoding of particular attributes or combinations of attributes. The *scale and contour* theory of melodic memory (Dowling, 1978) is mostly based on data from AB comparison, and it has delivered much insight into how melodies of different lengths are encoded (Long, 1977), under which conditions interval versus contour information is encoded (Dowling & Bartlett, 1981; Dowling & Fujitani, 1971; Edworthy, 1983, 1985), and how the encoding process of melodic phrases proceeds during the course of a real listening experience (Dowling, Tillmann, & Ayers, 2002).

AB comparison is related to testing procedures from psychophysics, where two stimuli are to be compared with regard, for example, with difference in loudness or in pitch. Consequently, bias-free scoring procedures originally applied in psychophysical tests, such as *d′* (Green & Swets, 1966), are often employed in memory experiments of AB comparison to obtain values of the independent variable.

List-wise recognition is closer to the recognition procedure usually found in verbal memory experiments. In the study phase, short melodies are presented one after another forming a list. After a short retention interval, there is a test phase, in which another list of melodies is played to the participants. The test list contains melodies used in the study list as well as new melodies. Participants indicate whether each item on the test list was presented in the study phase, and rate their confidence in their judgement. In addition, or as an alternative, participants may be asked rate the pleasantness of the melodies. Pleasantness ratings haven been used as a measure of implicit memory (Halpern & Müllensiefen, 2008) on the basis that the previous exposure to the same melody increases its aesthetic appreciation (the mere exposure effect: Zajonc, 1968).

There are two major differences between list-wise recognition and AB comparison. First, in list-wise recognition, the significance of the retention phase is relatively very limited, and its length and the content are generally not part of the experimental design. Second, list-wise recognition admits greater distance in time and more intervening musical material between corresponding items on the study and test lists. If the lists are substantially long, then episodic memory representations of particular melody items are likely to be lost, due to interference or to decay of the original memory trace. This makes list-wise recognition well suited to implicit memory testing.

Dynamic recognition paradigm for familiar melodies was popularized by Matthew Schulkind and collaborators it in a series of publications (Schulkind, 2000, 2004; Schulkind, Posner, & Rubin, 2003) on the recognition of well-known tunes. In the first trial, participants listen to an incipit of a given length from a melody, and then are asked to indicate the title of the tune if they can recognize it. If a participant is unable to name the tune confidently, the incipit is played again, in repeated trials, with one more note from the melody included each time; there is a prompt for a recognition judgement after each trial. For each incipit length, the recognition rate across participants is recorded and is subsequently compared with melodic and rhythmic events happening or being completed at that point in the melody.

Recognition versus recall The greater diversification within the recognition paradigm corresponds with the far greater number of studies that have been published using recognition, rather than recall, as the participants' response. We will discuss some reasons for the dominance of the various forms of the recognition paradigm in research on melodic memory below. For now, we conclude by remarking that, in contrast to the few studies using a melodic recall paradigm, there is a plethora of melodic recognition studies in the literature and that recognition is significantly more prominent, more diversified and more developed in melodic memory research. Given the potential benefits of employing different memory paradigms in verbal research (finding different effects, coming up with different types of models, and corroborating the same effects/models from a different perspective, etc.), we suggest that there is a clear motivation to develop the recall paradigm for melodic memory. Furthermore, we will advocate the use of computers in the analysis and modelling of the melodic recalls generated from this paradigm.

Studies using recall paradigms for melodic memory

Studies using recall paradigms are much less common than those employing one of the variants of recognition surveyed above. The differences between recall studies lie mostly in whether novel or familiar melodies are used as experimental stimuli, in whether participants are musically trained or untrained (or perhaps the study compares performance between these two groups), and in the participants' response mode. It is worth briefly reviewing some recall studies here to provide a feel for the spectrum of existing methods, and to position the specific experimental paradigm of Sloboda and Parker (1985) therein.

Musical dictation was used by Deutsch (1980) to provide evidence for her hierarchical-generative system of mental representations (Deutsch & Feroe, 1981). Musically well-trained listeners listened to sequences of 12 sine tones, representing different degrees of structural difficulty. The tone sequences were produced from the Deutsch–Feroe model, applying different generative rules, to generate easier or more difficult sequences. Some sequences included leading tones, while others were mainly built around triads, and others were segmented on the basis of melodic versus temporal gaps. The participants' task was to write down, in staff notation, as much as they could remember from the stimuli. The notated responses were evaluated by calculating the relative number of correct pitches at the correct serial position. Deutsch found more correct recall for the easier melodies, that could be encoded more efficiently as predicted by the Deutsch–Feroe model. She therefore concluded that memorizing a melodic sequence consists of encoding the hierarchical structure of the sequence and its alphabet; a single chunk from the sequence (e.g. the incipit) would also have to be encoded, to allow for a subsequent reconstruction of the whole sequence from memory. The Deutsch–Feroe model is specified in a generative form, and producing sequences from it that can be used as experimental stimuli is straightforward. But it is not an analytic device capable of encoding an existing melody; to perform that more complex task, a fully functional parser would be needed. This memory model is therefore of limited use for experiments that use real, existing melodies as stimuli. The music dictation method employed in this experiment requires a good level of musical training among the participants and is therefore not suitable for exploring memory representation of untrained subjects.

To overcome this last limitation, Davies and Yelland (1977) asked their participants to draw from memory a representation of the contours of short, song-like, newly composed sequences, to which they had listened. Contour was represented as rising, falling, or horizontal lines between successive notes. The resulting drawings were scored according to the number of lines with correct inclination at the correct serial position. Davies and Yelland were not concerned with effects of melodic structure, but varied the number of repeated listenings and the type of training procedure used to familiarize participants with the contour drawing method as independent variables. As expected, participants drew increasingly more accurate contours over repeated listenings to the same melody. This corroborates the assumption that mental representations of melodies become more accurate with repeated exposure. With regard to the training procedure comparison, the best results came from participants who had practised contour drawing with well-known melodies from long-term memory in silence as opposed to another group that practised drawing novel melodies that were presented aurally and whose members received feedback on their drawings. Davies and Yelland interpret this group difference as pointing to the importance of comparing novel stimuli with settled internal representations. Also, it seems that contour information is not explicitly abstracted during listening, but that melodies are represented as tonal analogies from which abstractions such as contour can be derived in subsequent serial scans through the representation.

While Davies and Yelland's experimental procedure is suitable for participants with little or no musical training, it is limited to reflecting the accuracy of melodic contour encoding and it ignores other parameters such as interval size, harmonic implication or rhythmic duration. It also requires a transfer of 'up-and-down' information from the auditory domain into the visual domain, and so it is not possible to separately distinguish if participants with low scores had an inaccurate representation of the melody as an auditory object, or if they were less good at transferring information from the auditory domain to the visual response domain. The different training procedures could potentially have trained these two different processes to different degrees.

Contour label recall was used by Williamson, Baddeley, and Hitch (2006) (and similarly Keller, Cowen, & Saults, 1995) to test recall memory for short pitch sequences. It asks for recall of the contour of four-note isochronic pitch sequences from working memory. Each note of a sequence could be either 'high', 'medium', or 'low', in relation to the other notes. In each trial, after hearing the pitch sequence, participants were asked to fill out a matrix in a paper booklet indicating the contour category of each of the four notes. Since participants' responses are expressed in a written notation requiring no musical knowledge, they need no musical training or any singing skills. The resulting data are very clean compared with sung recalls from the paradigms, described below, which require skilled transcription and expert analysis. In principle, the paradigm is not limited to the evaluation of memory representation of melodic contour (i.e. higher-than and lower-than judgements with respect to pitch height): judging duration or rhythmic categories as well as intervallic distance should, in theory, also be possible. However, it is limited in that, with musically untrained listeners, memory load issues can mean that it is possible test only a single musical dimension (contour or rhythm or intervals) at a time, so interactions between dimensions cannot be studied.

Also, experimental stimuli must be relatively simple and the number of different stimulus class labels is limited; Miller's (1956) rule of thumb of 7 ± 2 different magnitude values or categories that are simultaneously manageable on the same perceptual dimension may possibly hold as an upper limit (for a more recent affirmation of Miller's original results, see Shiffrin & Nosofsky, 1994). However, many popular melodies make use of more than seven different pitches, and processing and memory constraints are aggravated by the requirement to report the category label of each note of a series in order. Determining the limits of the paradigm before floor effects are reached in terms of the number of different category classes, as well as with respect to number of musical dimensions that can be combined and in terms of the length of the note sequence would give valuable insight into to which degree this recall paradigm can be used with non-artificial melodic stimuli.

Vocal recall of known songs was used by Halpern (1989) and Levitin (1994) to determine the accuracy for the absolute pitch of melodies from well-known folk or pop songs. Here, the accuracy of the memory representation of the relative structure of the melodies was of less importance because only the first note (Halpern) or the first three notes (Levitin) sung by the participants were used for analysis; participants that failed to produce a stable pitch were excluded from the sample. The analysis of the sung recalls must have been quite laborious in both cases. In Halpern's study, the starting pitches of the folk songs were transcribed by an expert musician with reference to a keyboard. Levitin used a computer program to estimate fundamental frequency using the Fast Fourier Transform. He determined the pitch of each of the three starting notes of the participants' recalls by selecting just the steady state portion of the sung notes with an audio editor and rounding the fundamental frequency to the nearest semitone. This paradigm is, in principle, quite similar to Sloboda and Parker's paradigm in that sung responses from memory are recorded and transcribed, and, in consequence, researchers using vocal recall face the same problems as Sloboda and Parker, which we outline below. A lot of expert work is required to transform the sung audio recordings into score notation and finally into a numerical representation of pitch. However, because Halpern (1989) and Levitin (1994) are only concerned with long-term memory for absolute pitch, and melodic structure, rhythmic, metric, phrasing, and interval information in the sung recalls need not be transcribed, the complexity of the task is greatly reduced.

The Sloboda and Parker recall paradigm

The core component of the Sloboda and Parker's (1985) procedure is the request to participants to sing the stimulus melody item back from memory. The experimental procedure, as used in their original study, can be split into four stages:

1. A melody is played to the participant up to six times. After each repetition, the participant is asked to sing into a microphone and recording device those parts of the melody that they can remember.
2. The audio recordings are transcribed into music notation by a human expert.
3. From the notation, each recall is analysed with regard to different aspects such as metre, breathing breaks, melodic contour, and phrase structure.
4. The dependent variables are compared over the values of the independent variables of interest: trial number, participants' musical background, melody type, and presentation mode (e.g. with or without lyrics: Ginsborg & Sloboda, 2007).

To obtain accuracy scores for different types of errors, in step 3, and ultimately to enable quantitative analyses, Oura and Hatano (1988) developed a more formal scoring method which was subsequently used by Zielinska and Miklaszewski (1992), Drake, Dowling, and Palmer (1991) and Ogawa *et al.* (1995) with slight variations. In this scoring procedure, each phrase in the recall where a tonality can be identified is assigned to the best matching phrase in the original melody. Each phrase is then compared, in half-bar windows, to its counterpart in the original. If the half-bar window is identical to the original, it is labelled as *correct*. If it differs in either rhythm or contour, it is labelled as *modified* and if it differs in both parameters from the original it is declared *non-identified*. The windows are summed together by category (and error type) and their proportions with respect to the overall number of half-bars are taken as dependent variables for that particular trial.

While the method seems superficially straightforward, and the results that can be produced from it are of great interest, much cited, and truly complementary to what can be obtained from a recognition paradigm, it has been used only a few times in the 25 years since it was published. This disappointing uptake is due to some inherent difficulties with the method, most of which were discussed in the original article (Sloboda & Parker, 1985); we summarize them as follows.

1. The raw data recorded in these studies are extremely 'dirty', requiring expert interpretation. Participants are required to sing, and their singing may be inaccurate; in some places, it is necessary to infer which note(s) they *meant* to sing. The participants' singing is recorded, and it is possible that the recording may be imperfect; it is impossible to prevent this without breaking the paradigm, for obvious reasons.
2. There is 'no theory of melodic identity' (Sloboda & Parker, 1985, p. 159).Therefore, subjective judgements of the expert transcribers are crucial in doing the analysis. This subjectivity can clearly affect the results, and cannot be controlled.
3. It follows that there can be no standard, verified scoring method for these complex response data; that is, there is no theory of melodic *similarity* to give a

Table 9.1 Numbers of recalled fragments reported in studies applying Sloboda and Parker's recall paradigm

Study	Recalled fragments
Sloboda and Parker (1985)	48
Oura and Hatano (1988)	320
Zielinska and Miklaszewski (1992)	310
Ogawa *et al.* (1995)	80
Ginsborg and Sloboda (2007)	60
Müllensiefen and Wiggins (manuscript in preparation)	1900

measure of how close the attempts are to the original. Oura and Hatano (1988) aimed to address this with their procedure, which was described above.

4 A simple but very limiting issue is the amount of time it takes to transcribe the data and perform these analyses. No really large studies using the method have yet been published; those studies which have been carried out are listed, with their size, in Table 9.1. This last issue, superficially, is the kind of problem that one would like to solve by using a computer, on the grounds that computers can analyse data very quickly and very accurately. However, because music is fundamentally a perceptual and cognitive construct, any program used for these tasks needs to encode *human expert* levels of musical understanding. In the next section, we explore the extent to which this is currently possible, show what can be done, and examine prospects for the future.

Computational methods for researching melodic memory

Towards automatic support of Sloboda and Parker's paradigm

Our central idea in this paper is that the experimental paradigm proposed by Sloboda and Parker (1985) may be greatly facilitated by appropriate application of computational technology. Broadly, it should be possible to use computational methods to process, analyse, and compare recall data, thus allowing the method to be applied more widely.

We summarized the drawbacks with the method in the previous section, above; a computational approach can, at least in principle, help with most of them. Digital recording and post-hoc computational audio analysis can greatly help with problem number 1; this technology has been readily and cheaply available for more than a decade, though it is becoming established in music and music psychology research only recently (e.g. Cook, 2008). In particular, as noted above, the speed and repeatability of computer processing could help in principle with problem number 4; however, this is dependent on solutions to the more musical aspects in problems 2 and 3, which are very difficult. Sloboda and Parker's statement that there is 'no theory of melodic identity' (1985, p. 159) holds just as true 25 years on. However, there is now a substantial body of research leading towards computational theories of melodic similarity which are applicable in an analytical context. Progress towards a computational theory brings concomitant benefits beyond the immediate application: in particular, to build a program that embodies a theory, one must specify that theory to an extreme—indeed *absolute*—level of detail. Once such a specification is given and a program written, it can be tested to destruction against all the data available from studies involving humans, without problems of fatigue, priming, or even ethics. In this way, the computational approach not only facilitates the empirical work, but can make its analysis more rigorous and objective; new knowledge can be created in terms of novel hypotheses generated by the models.

Extant computational methods developed for various musical purposes can be applied in this context. Since musical memory, and hence musical similarity, de facto underpins the vast majority of musical behaviour (Wiggins, 2007), there is a rich seam of more general work to be mined and repurposed to help solve the current problem. There is not space here to review or even to list all the work that could contribute in this way, but some examples follow. Temperley (2001) and Huron (1995, 2006) have developed cognitively informed methods intended for computational music analysis; in particular, attempts to understand structural reference in this context help in our understanding of memory, for it is this which enables such reference to be understood. These essentially musicological approaches can be enhanced by models from mathematical music theory (e.g. RUBATO: Milmeister, Mazzola, & Thalmann, 2009), which generally aim to find mathematical systems underpinning perception and cognition, with a view to explaining why either the music or the perception is the way it is.

The study of music cognition in general (e.g. Eerola, Järvinen, Louhivuori, & Toiviainen, 2002; Krumhansl, 1990; Thomassen, 1982), from an empirical standpoint, allows different kinds of insight into what may be expected in perception of music, and can inform the construction of heuristics (see below) for helping with analysis of musical data, as well as supplying models of musical competence, such as memory.

In a practically motivated context, a significant amount of work in the field of music information retrieval is concerned with understanding of perceived similarity of various kinds, such as harmonic similarity, usually at the level of whole pieces, (e.g. Pickens *et al.*, 2003) and melodic similarity (e.g. Crawford, Iliopoulos, & Raman, 1998; Müllensiefen & Frieler, 2004a). Finally, since there seem to be shared effects and behaviours between music and language, it has proven useful to consider techniques from computational linguistics in the musical context (e.g. Downie, 2003).

Taken together, models and methods developed in these areas can already analyse many aspects of melodies that are of core interest in many psychological studies. These include: melodic identity and similarity (Müllensiefen & Pendzich, 2009); metrical structure, metre induction (Müllensiefen and Frieler, 2004b); (Eck, 2002; Volk, 2008); phrase structure (Pearce, Müllensiefen, & Wiggins, 2010; Temperley, 2001); rhythmic structure (Weyde, 2004); harmonic structure: tonality induction (Krumhansl, 1990; Longuet-Higgins & Steedman, 1971); accent strength (Müllensiefen, Pfleiderer, and Frieler, 2009), complexity (North & Hargreaves, 1995), expectedness (Pearce & Wiggins, 2006), high-level structure identification (Abdallah, Sandler, Rhodes, & Casey, 2006).

While computational methods are nowhere near the position of simulating a human listener as a whole, we suggest that the time is ripe to begin applying these methods to focused problems such as those provided by the Sloboda and Parker paradigm. Through feedback from such application, the methods will themselves be improved.

Computational modelling of perceived melodic similarity

A key issue in executing the Sloboda and Parker paradigm is the comparison of recalled melodies with the original stimulus. This is a hard task for human experts, not simply because it is a difficult task in any case, but in particular because the judgements required can sometimes be very subjective, based on the prior musical bias of the expert. One way one might increase objectivity would be to do the analysis automatically: this has the advantage of removing subjectivity based on purely human judgement. However, such judgements cannot, in the current state of the art, be made entirely reliably by computer. In this section, we briefly survey the state of that art, with a view to identifying what can be done and what needs further work. We also exemplify this concept with a short description of the *edit distance* (or *Levenshtein distance*), an algorithm that is very simple but which has proven to be effective for similarity computations in automatic text processing, computational biology (Gusfield, 1997), and also audio and music computing (e.g. Crawford *et al.*, 1998; Unal, Chew, Georgiou, & Narayanan, 2008).

We described the scoring procedure developed by Ogawa *et al.* (1995) in an earlier section. In this method, melodic similarity is measured by (manually) determining the identity of half-bar segments from the original melody and the sung recalls. This approach, while not computerized, is archetypal of computerized approaches. Because musical similarity is a multidimensional thing, it is necessary to be clear about which dimensions one is interested in (and, of course, this may be a parameter of one's study). It follows that, if we are interested in more than one dimension, but we need a linear scale for comparison, we also need a mathematical means of mapping vector distances in a multidimensional space on to scalar distances; fortunately, the mathematics required is undaunting to psychologists. Both the dimensions and mapping methods must correspond properly with perceived similarity; where there is ambiguity in human response, this should ideally be detectable in the measurement system (for example, two different multidimensional distances might map on to the same scalar distance). To produce a similarity measure which meets this specification in full is a very tall order, and much more research is required to do so.

However, we are in a position to begin. Assuming that we have transcriptions of all melodies in a symbolic format to start from, there are two essential steps in measuring melodic similarity: first, the melodies are transformed into one (or several) representation(s) which encode(s) all the dimensions of the music in which one is interested; second, the abstract representations are compared and a numerical value (usually normalized between 0 and 1) is derived to indicate the distance between them.

In the first step, the raw melodies (which might be represented, for example, only by a sequence of tuples of pitch and onset values) are transformed into one or more abstract representations that are cognitively more meaningful and of interest with regard to the similarity comparison. Melodic representations that can be usefully employed for similarity measurement include melodic intervals, melody contour, a sequence of durational or rhythmic values, and a sequence of tonality values that are implied by the melodic pitches. The important point about this step is that the raw melodic information is *abstracted* and melodies are represented as sequences of numerical or digit symbols. The transformation is in principal independent of the algorithm that is used to compare symbol sequences subsequently. The choice of transformation depends rather on whether the researcher deems it important with respect to the similarity measurement or whether it contains an optimal amount of information in an information-theoretical sense (Pearce & Wiggins, 2004). Of course, various different representations of the same melody (e.g. one containing pitch intervals and one containing rhythmic values) can be combined for subsequent similarity measurement.

In the second step, from the abstract representations of two melodies a similarity value from the interval $\in [0, ..., 1]$ is computed where 1 indicates maximal similarity or identity and 0 indicates no similarity between the two melodies. In recent years, several different measures, and algorithms to implement them, have been proposed and tested in various applications needing to determine the similarity between melodies. These include geometric measures (Aloupis *et al.*, 2003; O Maidin, 1998), string matching techniques such as edit distance (Crawford *et al.*, 1998; Mongeau & Sankoff, 1990), n-gram measures (Downie, 2003; Uitdenbogerd, 2002), and hidden Markov models (Meek & Birmingham, 2002) from speech recognition, as well as the Earth Movers Distance algorithm (Typke, Wiering, & Veltkamp, 2007) from computer vision, and hybrid algorithms that combine the output of several different similarity measurement procedures (Müllensiefen & Frieler, 2004a). Here, by way of example, we briefly explain the edit distance measure, which, in recent years has proven to be a surprisingly effective comparison or benchmark algorithm that appears to be similar to informal and rather intuition-guided music-analytic applications of melodic similarity concepts (for a comparative discussion of quantitative and qualitative approaches to melodic similarity, see Müllensiefen & Pendzich, 2009).

The main idea behind the edit distance, or Levenshtein distance, is to treat the minimum number of operations ('edits') needed to transform one string into another as a measure of the distance between them. The permitted operations are insertion, deletion, and substitution. Since the operation of the algorithm is independent of the denotation of the symbols in the strings compared, we use letter strings to demonstrate the edit distance in our example. Consider the two letter strings SCHOENBERG and SCHUBERT. These strings can be aligned optimally by applying the series of operations applied to each letter that is recorded in the *edit transcript* shown. The symbol D denotes the deletion of a letter, I stands for insertion, S for a substitution, and M for a match of two letters, i.e. no operation is carried out.

$$\begin{array}{l}
\text{String 1} \quad \text{S C C H O E N B E R G} \\
\text{String 2} \quad \text{S C H _ _ U B E R T} \\
\hline
\text{Edit Transcript M M M D D S M M M S}
\end{array} \tag{9.1}$$

Since there are two deletions and two substitutions, the edit distance in this case is 4—if deletions and substitutions are given the same weight, though these amounts may be varied to achieve different matching behaviours. To arrive at a similarity value, rather than a distance metric, and to confine that value to the interval (0,1), we divide the distance value by the maximal distance (the length of the longer string) and subtract the result from one:

$$\sigma(s_1, s_2) = 1 - \frac{d_{edit}(s_1, s_2)}{\max(|s_1|, |s_2|)} \tag{9.2}$$

For our example, we thus obtain an edit distance similarity value of 1 – 4/10=0.6. Of course, this same measure works with symbols representing pitch, intervals or rhythmic categories; all that we need is the ability to decide whether two symbols are the same, regardless of their meaning. What makes it attractive for music researchers is the fact that the quality of the operations (insertions, deletions, and substitutions) have parallels in musical composition, where it is common practice to produce variations or related pieces of music by, for example, inserting ornamental notes or by replacing certain structurally important notes by others that may serve a similar—or different—harmonic purpose. This neat simplicity is necessarily lost if we wish to consider more than one musical dimension at once, because of the intimate relationship between the dimensions of music, but Mongeau and Sankoff (1990) have adapted the measure to work better in these circumstances.

The description above, however, covers only the measure, and not the algorithm required to compute it. For edit distance to be useful, two strings must be aligned *optimally*—that is, with the smallest possible distance. This is often achieved, in $O(n^2)$ time and memory, using *dynamic programming* (Gusfield, 1997), whose implementation is straightforward. This, and the fact that the measure itself and its outputs are mostly easy to understand, have made edit distance a popular tool in computational music analysis in recent times.

Revisiting Sloboda's and Parker's results

Motivation

We now revisit the results of Sloboda's and Parker's paper as discussed in their nine conclusions (Sloboda & Parker, 1985, pp. 159–160). The primary purpose of reanalysing the musical data from the original study is not to refute or support the authors' findings, but rather to show how analytical algorithms can be usefully employed in cognitive studies, how they might help to cope with large amounts of difficult-to-analyse musical data, and how they might shed a different light on some analytical procedures that are commonly executed by human (as opposed to computational) effort.

As musical data, we used the 48 transcribed recalls that are given in appendix by Sloboda and Parker (1985). We applied similarity measurements and other algorithmic analysis methods to the research questions and results summarized at the beginning of their conclusion section. We limit ourselves to discussion of six out of the nine results, where the application of computational methods seem most straightforward and productive, and omit Sloboda and Parker's main findings 4, 5, and 7.

Main finding 1: 'No recall is perfect'

We measured the similarity between the target melody and each recall using the edit distance algorithm (see above). To have a 'second opinion' from a similarity algorithm that makes use of musical background information from a corpus, we also applied an asymmetric similarity algorithm proposed by Müllensiefen and Pendzich (2009), which is based on a conception of feature similarity proposed by Tversky (1977), the *Tversky.target.only* algorithm. We applied both algorithms only to the *pitch sequences* of the melodies. The maximal similarity values were 0.93, from edit distance, and 0.98, from the Tversky similarity algorithm. Both were measured for participant 5 on her sixth trial. Sloboda and Parker's assertion that, even for this simple folk tune and with some musically trained participants, there is no single 'note-for-note perfect' recall among the 48 attempts. The computational analysis corroborates this surprising limitation of human melody recall.

Main finding 2: 'Recalls are highly related to the original in many respects'

We measured the similarity between recalls and the target melody using our simple edit distance algorithm, but feeding it different types of musical information, viz., sequences of pitches, sequences of interpolated contour values, sequences of categorized rhythmic (duration) values, and sequences of implied tonalities at the bar level as derived from the Krumhansl–Schmuckler algorithm for tonality induction (Krumhansl, 1990). How these different types of musical information are obtained, through appropriate transformations from the raw melodic pitch and onset data, is documented by Müllensiefen and Frieler (2004a). We obtained mean similarity values, averaged over all 48 recalls, of 0.37 for pitch similarity, 0.21 for contour similarity, 0.56 rhythmic similarity, and 0.20 for similarity of implied tonalities. Taking into account that these values are averaged over all six trials and therefore include low similarity values from initial (and so mostly incomplete) trials, it seems fair to say that the recorded recalls were related to the target melody in many different respects. In particular, the similarity is found to some extent in all the dimensions tested. Thus, the original findings are supported.

Main finding 3: 'Metrical structure is preserved in almost all recalls'

We used the beat and metre extraction model of Frieler (2004) to induce the metre of the target melody and the recalls. The model uses temporal smoothing with Gaussian kernels, accent rules based on note durations and autocorrelation to determine the beat level, metre, and metrical phase from monophonic input. The model induced a 2/4 metre for the target melody, which is not unreasonable; the melody being actually notated in 4/4 may simply be because 4/4 metres are more common in Western musical styles. Eighty one per cent of the recalls were classified as having a 2/4 metre and 19% were classified as 4/4. So all trials had a simple duple metre, like the target melody. However, since 4/4 is the default for most (popular) Western music, this result is not so surprising; it would be interesting to apply the approach to recalls where the target was in triple and/or compound meter.

Main finding 6: 'Subjects vary significantly regarding their melodic and harmonic accuracy'

We compared the recalls of musical novices (participants 1–4) and musicians (participants 5–8) with regard to their melodic and harmonic accuracy (i.e. similarity to the target melody), as did Sloboda and Parker. We measured pitch edit distance to estimate pitch accuracy. Novices achieved a mean similarity of 0.30 over all trials and musicians reached a mean of 0.45, which proved to be a significant difference ($t(30.4) = 2.86$, $p = 0.004$). Similarly, mean harmonic similarity as measured by the edit distance over implied tonalities differed significantly between novices ($\bar{x}_{novices} = 0.09$) and musicians ($\bar{x}_{musicians} = 0.31$; $t(28.5) = 3.17$, $p = 0.002$). Although the difference between novices and musicians is significant for both measures, the it is much larger for the harmonic comparison. This supports Sloboda and Parker's conclusion that 'memory for harmonic structure seems to be related to musical expertise' (1985, p. 160).

We also compared rhythmic similarity values between the two groups. The resulting t-test shows no significant difference between the two groups of participants ($\bar{x}_{novices} = 0.59$; $\bar{x}_{musicians} = 0.54$; $t(45.5) = -0.75$, $p = 0.77$). Sloboda and Parker observed that while there are considerable differences between listeners regarding the memory retention of harmonic structure, metrical structure was abstracted well by all participants. This finding seems now to extend to rhythmic structure as well.

Main finding 8: 'Musicians and non-musicians differ in retention of harmony'

We looked at the development of the representation of harmonic structure in both groups of participants over the six trials, with main finding 6 (above) in mind. We averaged the harmonic similarity values, as measured by the edit distance on the implied tonalities, for each trial number over each participant group. The development is illustrated in Figure 9.1.

Figure 9.1 shows a clear increase in harmonic similarity between the musicians' recalls and the target melody from trial 4 on while the harmonic similarity between recalls of novices and the target melody stay at the same low level of similarity over all six trials. In contrast, no learning effect or improvement of accuracy in the mental representations seems to take place for the novices. Accordingly, a dependent sample t-test between novices' and musicians' mean recall similarity values indicated a significant difference ($\bar{x}_{differences} = 0.22$; $t(5) = 3.16$; $p = 0.01$). Therefore, it seems evident that the perceptual and cognitive abstraction and refinement of harmonic structure is achieved only by participants with some musical training.

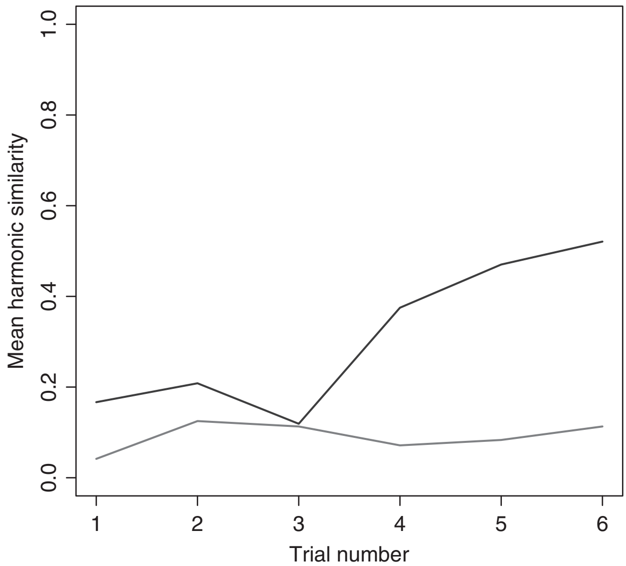

Fig. 9.1 Mean similarity between participants' recalls and target melody over trials. Blue: musicians; red: novices.

Main finding 9: 'Subjects do not show improvement on any measure over the six trials'

This is one of Sloboda and Parker's most interesting conclusions, since it seems to imply that human memory for melodies does not increase in accuracy with repeated exposure. To test whether there is actually no such learning effect, we averaged for each trial similarity for pitch, harmonic, rhythmic, and overall similarity (measure *opti2* combining duration weighted pitch values and a similarity measure on short pitch motives, see Müllensiefen & Frieler, 2004a, for details) over all participants and computed Pearson's correlation, *r*, between trials number and means similarity for each similarity measure. We found high and significant values of *r* for all similarity measures as given in Table 9.2. Figure 9.2 illustrates these correlations.

These correlations do suggest that a learning effect is taking place over repeated trials. The effect seems to be less strong for the rhythmic aspect of the memory representation than for the pitch and harmonic aspects. So, while Sloboda and Parker assert that recalls get longer over trials (starting at a mean of 16.5 notes for trial 1 and monotonically increasing to 26.625 notes for trial 6), we found that recalls also get more accurate and that participants also seem to refine their memory representations.

This is in line with results of Zielinska and Miklaszewski (1992), who found (using 10 instead of 6 trials) an increase in the number of correctly recalled half-measures.

Discussion

The translation of the six main analyses discussed in the previous section into computational methods proved to be very simple and straightforward. Translating main

Table 9.2 Values of Pearson's correlation coefficient *r* with corresponding *p* values for different similarity measures measuring the correlation between trial number and mean similarity with the target melody

Similarity measure	*r*	p Value
Edit distance, pitch	0.93	0.008
Edit distance, implied tonality	0.93	0.008
Edit distance, rhythmic category	0.77	0.08
opti2	0.91	0.01

findings 4, 5, and 7 into computational form requires some more detailed argumentation; they can probably be realized in more than one way.[2] Therefore, we defer these for future work, because the purpose of this chapter is merely to support our claim that analytical tasks in music cognition research can be solved computationally.

For main finding 4, 'subjects preserve phrase structure', one possible analysis would have been to compute, for each recall, the similarity between all possible pairs that can be created from the three phrases, *A*1, *A*2, and *B*. The overall similarity between pair (*A*1, *A*2) should then be higher than the similarity between pairs (*A*1, *B*) and (*A*2, *B*).

The translation of main finding 5, 'original rhythms are often replaced by metrical equivalents', would require the measurement of rhythmic similarity at the phrase

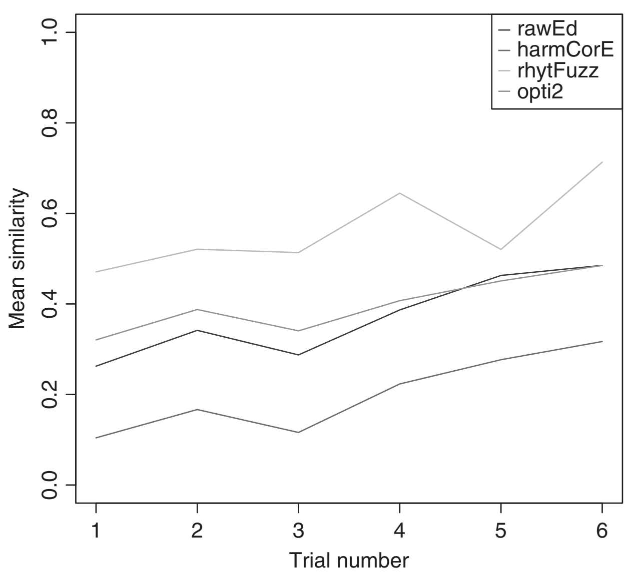

Fig. 9.2 Mean similarity over trials according to different similarity measures; colours correspond with the different measures.

[2] Specifically, for main findings 4 and 5 the phrases all recalls need to be identified initially to enable the subsequent phrase-level analyses.

level, between original and recalled phrase, and a metre induction procedure giving both metre and phase information. The test would then be that, for most phrases, the induced metre and phase is identical to the original but rhythmic similarity to the original varies to a greater extent.

The claim from main finding 7, 'harmonic structure is preserved even when exact melodic structure is lost', is similar to main finding 5: Sloboda and Parker observe that while the recall of melodic details shows a greater variation, the more abstract harmonic structure, having fewer degrees of freedom, is less likely to vary and is often close to the original. This might be demonstrated by comparing the variation in harmonic with melodic similarity measurements, within a series of repeated measurement.

Conclusion and outlook

Several central questions motivated Sloboda and Parker's original study, including:

- How do memory representations build up over repeated exposure to the same novel melody?
- What melodic parameters or dimensions are easier to grasp and to encode and which ones or more difficult?
- How do people with and without musical training differ with respect to building up representations of melodies in memory?

In this chapter, we have demonstrated how a computational approach can help in answering these questions, and how the nature of the answers changes when a computer is involved as a tool for analysis and modelling. We have seen answers become necessarily very precise and quantitative; this we consider to be an advantage, especially if experiments are repeated with the same method but on different data or with different stimuli. However, one must take care not to confuse this necessary precision with greater validity—indeed, the converse can easily be the case. Therefore, in order to obtain valid results, the questions that the computer is supposed to answer must also be asked very precisely. This is no bad thing: the necessity to make explicit every tiny step from the formulation of the question (via the choice and implementation of a formal procedure) to the interpretation of the results can often win us additional insight into the phenomena under study. As can be seen from our discussion of melodic similarity algorithms, the use of computational models and algorithms offers the researcher a huge choice of analytic procedures—but, in consequence, that researcher must take responsibility for their decision and argue it. Though unavoidably painful at the beginning, this extra rigour in several aspects of methodology, can ultimately only be good for the field.

While we certainly do not suggest that the edit distance measure is necessarily the best measure of melodic similarity in the context of this experimental data, we chose it deliberately because it is easy to understand, and it is at least loosely connected to semi-formal procedures that musicologists use to determine similarity relations and derivative procedures between melodies. However, it would be interesting to reanalyse this experimental dataset with more complex similarity models that are claimed to have more cognitive validity (e.g. Müllensiefen and Frieler, 2004a; Müllensiefen and Pendzich, 2009). Then, employment of a computational procedure for determining melodic similarity would change status, from merely using an adequate tool, to modelling human behaviour with a cognitive-computational model—an epistemic transition for which we have argued in detail elsewhere (Pearce, Müllensiefen, & Wiggins, 2009; Wiggins, Pearce, & Müllensiefen, 2009).

Apart from answering Sloboda and Parker's questions in a computational way that is of a different quality, the use of computers to analyse and model data resulting from their experimental paradigm also enables us to ask some additional questions. Because these questions entail comparison across different melodic stimuli, we conducted a new experiment with Sloboda and Parker's design but using 14 different melodies and 30 subjects with acceptable singing abilities. As a result, we obtained about 1900 usable sung recalls which have been transcribed by a professional human transcriber. These data are about to be analysed using the computational approach proposed here. The additional questions that we hope to answer from the analysis of this new dataset concern mainly features of melodic structure, and they aim at being generalizable to the memory processing of Western popular melodies in general:

- How do musical features affect the recall of melodies?
- What makes a melody easy or difficult to recall?
- Which parts of a melody are represented first and most accurately?
- Does commonness or rarity of melodic features play a role?
- Are melodies in their song context and as audio excerpts recalled better than single-line melodies or is this vice versa?

To facilitate this type of feature analysis, we have developed an open source software toolbox, called FANTASTIC,[3] which computes summary and sequence-based features of monophonic melodies. FANTASTIC also enables researchers to model melody perception and cognitive processing in the context of a corpus of melodies and, therefore, it can take into account previous listening experience given a suitable corpus of music. The software has already proven useful in the analysis of experimental data from a recognition paradigm similar to the one used by Halpern and Müllensiefen (2008), and preliminary results suggest that objective measures of implicit and explicit memory performance, as well as subjective memory measures including false alarms and misses, can be explained to a certain degree by features of the melodic structure (Halpern, Müllensiefen, & Wiggins, 2008). Once our recognition dataset is fully analysed, we aim to model our recall data with the same feature approach, to see whether recall and recognition memory can be explained by the same structural features. This would supply strong evidence for the hypothesis that the performance of musical memory is indeed dependent on musical structure, and that we can explain and predict cognitive behaviour, at least partially, from the structure of the music itself.

[3] Feature ANalysis Technology Accessing STatistics (In a Corpus); download from our project website: http://www.doc.gold.ac.uk/isms/mmm/?page=Software\%20and\%20Documentation.

Acknowledgements

This research was supported by EPSRC project EP/D0388551, *Modelling Musical Memory and the Perception of Melodic Similarity*. It was inspired by the work of John Sloboda.

References

Abdallah, S., Sandler, M., Rhodes, C., & Casey, M. (2006). Using duration models to reduce fragmentation in audio segmentation. *Machine Learning, Special Issue on Machine Learning for Music*, *65*(2–3), 485–515.

Aloupis, G., Fevens, T., Langerman, S., Matsui, T., Mesa, A., Rappaport, D., & Toussaint, G. (2003). Computing the similarity of two melodies. In *Proceedings of the 15th Canadian Conference on Computational Geometry*. Halifax, Canada: Dalhousie University.

Anderson, C., & Borkowski, J. G. (1978). *Experimental psychology*. Glenview, IL: Scott, Foresman and Company.

Atkinson, R., & Shiffrin, R. (1968). Human memory: a proposed system and its control processes. In K. Spence, & J. Spence (Eds.), *The psychology of learning and motivation*. New York, NY: Academic Press.

Baddeley, A. (1986). *Working memory*. Oxford: Oxford University Press.

Baddeley, A. (2007). *Working memory, thought and action*. Oxford: Oxford University Press.

Calkins, M. (1894). Association. *Psychological Review*, *1*, 476–483.

Cook, N. (2008). Techniques for analysing recordings. In N. Cook, E. F. Clarke, D. Leech-Wilkinson, & J. Rink (Eds.), *The Cambridge companion to recorded music*. Cambridge, UK: Cambridge University Press.

Crawford, T., Iliopoulos, C., & Raman, R. (1998). String-matching techniques for musical similarity and melodic recognition. *Computing in Musicology*, *11*, 73–100.

Davies, J., & Yelland, A. (1977). Effects of two training procedures on the production of melodic contour, in short-term memory for tonal sequences. *Psychology of Music*, *5*(2), 3–9.

Deutsch, D. (1970). Tones and numbers: Specificity of interference in immediate memory. *Science*, *168*, 1604–1605.

Deutsch, D. (1972). Octave generalization and tune recognition. *Perception & Psychophysics*, *11*(6), 411–412.

Deutsch, D. (1974). Generality of inference by tonal stimuli in recognition memory for pitch. *Quarterly Journal of Experimental Psychology*, *26*, 229–234.

Deutsch, D. (1975a). The organization of short-term memory for a single acoustic attribute. In D. Deutsch, & J. Deutsch (Eds.), *Short term memory*. New York, NY: Academic Press.

Deutsch, D. (1975b). Two-channel listening to musical scales. *Journal of the Acoustical Society of America*, *57*, 1156–1160.

Deutsch, D. (1980). The processing of structured and unstructured tonal sequences. *Perception & Psychophysics*, *28*(5), 381–389.

Deutsch, D., & Feroe, J. (1975). Disinhibition in pitch memory. *Perception & Psychophysics*, *17*, 40–51.

Deutsch, D., & Feroe, J. (1981). The internal representation of pitch sequences in tonal music. *Psychological Review*, *88*, 503–522.

DeWitt, L., & Crowder, R. (1986). Recognition of novel melodies after brief delays. *Music Perception*, *3*(3), 259–274.

Dowling, W., Tillmann, B., & Ayers, D. (2002). Memory and the experience of hearing music. *Music Perception*, *19*(2), 136–149.

Dowling, W. J. (1972). Recognition of melodic transformations: Inversion, retrograde, and retrograde inversion. *Perception & Psychophysics*, *12*(5), 417–421.

Dowling, W. J. (1978). Scale and contour: Two components of a theory of memory for melodies. *Psychological Review*, *85*(4), 341–354.

Dowling, W. J., & Bartlett, J. C. (1981). The importance of interval information in long-term memory for melodies. *Psychomusicology*, *1*, 30–49.

Dowling, W. J., & Fujitani, D. S. (1971). Contour, interval, and pitch recognition in memory for melodies. *Journal of the Acoustical Society of America*, *49*(2, Part 2), 524–531.

Downie, J. S. (2003). *Evaluating a simple approach to music information retrieval. Evaluating a simple approach to music information retrieval. Conceiving melodic n-grams as text*. Ph.D. thesis. London (Ontario), Canada: Faculty of Information and Media Studies, University of Western Ontario.

Drake, C., Dowling, W., & Palmer, C. (1991). Accent structures in the reproduction of simple tunes by children and adult pianists. *Music Perception*, *8*(3), 315–334.

Ebbinghaus, H. (1885). *U das Ged¨achnis. Untersuchungen zur experimentellen Psychologie*. Amsterdam: Bonset.

Eck, D. (2002). Finding downbeats with a relaxation oscillator. *Psychological Research*, *66*(1), 18–25.

Edworthy, J. (1983). Towards a contour-pitch continuum theory of memory for melodie. In D. Rogers, & J. Sloboda (Eds.), *The acquisition of symbolic skills* (pp. 263–271). New York, NY: Plenum.

Edworthy, J. (1985). Interval and contour in melody processing. *Music Perception*, *2*(3), 375–388.

Eerola, T., J¨arvinen, T., Louhivuori, J., & Toiviainen, P. (2002). Statistical features and perceived similarity of folk melodies. *Music Perception*, *18*(3), 275–296.

Frieler, K. (2004). Beat and meter extraction using gaussified onsets. In C. Lomeli Buyoli, & R. Loureiro (Eds.), *Proceedings of the 5th International Conference on Music Information Retrieval* (pp. 178–183). Barcelona: Universitat Pompeu Fabra.

Ginsborg, J., & Sloboda, J. A. (2007). Singers' recall for the words and melody of a new, unaccompanied song. *Psychology of Music*, *35*(3), 421–440.

Green, D., & Swets, J. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Gusfield, D. (1997). *Algorithms on strings, trees, and sequences: Computer science and computational biology*. Cambridge: Cambridge University Press.

Halpern, A. (1989). Memory for the absolute pitch of familiar songs. *Memory & Cognition*, *17*(5), 572–581.

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

Halpern, A., Müllensiefen, D., & Wiggins, G. (2008). *Modelling memory responses in a memory recognition task*. Talk at the 10th International Conference for Music Perception and Cognition (ICMPC10), Sapporo (Japan).

Huron, D. (1995). *The humdrum toolkit: Reference manual*. Menlo Park, CA: Center for Computer Assisted Research in the Humanities.

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

Idson, W., & Massaro, D. (1978). A bidimensional model of pitch in the recognition of melodies. *Perception & Psychophysics*, *24*(6), 551–565.

Kantowitz, B. H., Roediger, H. L. I., & Elmes, D. G. (1994). *Experimental psychology: understanding psychological research*. Minneapolis: West.

Keller, T., Cowen, N., & Saults, J. (1995). Can auditory memory for tone pitch be rehearsed? *Journal of Experimental Psychology: Learning, Memory and Cognition*, *21*(3), 635–645.

Kluwe, R. (1990). Gedächtnis und Wissen. In H. Spada (Ed.), *Allgemeine psychologie* (pp. 115–187). Stuttgart: Hans Huber.

Krumhansl, C. L. (1990). *Cognitive foundations of musical pitch*. Oxford Psychology Series 17. Oxford: Oxford University Press.

Levitin, D. J. (1994). Absolute memory for musical pitch: Evidence from the production of learned melodies. *Perception & Psychophysics*, *56*(4), 414–423.

Long, P. (1977). Relationships between pitch memory in short melodies and selected factors. *Journal of Research in Music Education*, *25*(4), 272–282.

Longuet-Higgins, C., & Steedman, M. (1971). On interpreting Bach. *Machine Intelligence*, *6*, 221–242.

Mauch, M., Müllensiefen, D., Dixon, S., & Wiggins, G. (2008). What's a work in harmony? In: K. Miyazaki, Y. Hiraga, M. Adachi, Y. Nakajima, and M. Tsuzaki (Eds.), *Proceedings of the 10th International Conference for Music Perception and Cognition* (pp. 683–689), 25–29 August 2008, Sapporo, Japan (CD). Adelaide: Causal Productions.

Meek, C., & Birmingham, W. (2002). Johnny can't sing. a comprehensive error model for sung music queries. In M. Fingerhut (Ed.), *Proceedings of the Third International Conference on Music Information Retrieval* (pp. 124–132). Paris: IRCAM.

Miller, G. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Milmeister, G., Mazzola, G. B., & Thalmann, F. (2009). *The Rubato composer musicsoftware*. Berlin: Springer.

Mongeau, M., & Sankoff, D. (1990). Comparison of musical sequences. *Computers and the Humanities*, *24*, 161–175.

Mueller, J. (1970). Response properties of the position indicant in serial learning. *Journal of Experimental Psychology*, *84*, 35–39.

Müllensiefen, D., & Frieler, K. (2004a). Cognitive adequacy in the measurement of melodic similarity: Algorithmic vs. human judgments. *Computing in Musicology*, *13*, 147–176.

Müllensiefen, D., & Frieler, K. (2004b). Optimizing measures of melodic similarity for the exploration of a large folk-song database. In *Proceedings of the 5th International Conference on Music Information Retrieval* (pp. 274–280). Barcelona: Universitat Pompeu Fabra.

Müllensiefen, D., & Pendzich, M. (2009). Court decisions on music plagiarism and the predictive value of similarity algorithms. *Musicæ Scientiæ, Discussion Forum 4B*, 257–295.

Müllensiefen, D., Pfleiderer, M., & Frieler, K. (in press). The perception of accents in pop music melodies. *Journal of New Music Research*, *38*(1), 19–44.

Murdock, B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*, 482–488.

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, *18*, 251–269.

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

O Maidin, D. (1998). A geometrical algorithm for melodic difference in melodic similarity. *Computing in Musicology*, *11*, 65–72.

Ogawa, Y. T. K., & Mito, H. (1995). Modification of musical schema for Japanese melody: A study of comprehensible and memorable melody. *Bulletin of the Council for Research in Music Education*, *127*, 136–141.

Oura, Y., & Hatano, G. (1988). Memory for melodies among subjects differing in age and experience in music. *Psychology of Music*, *16*, 91–109.

Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex* (G. V. Anrep, Trans.). London: Oxford University Press.

Pearce, M. T., Müllensiefen, D., & Wiggins, G. (2009). ¨Melodic grouping in music information retrieval: New methods and applications. In R. Zbigniew, & A. Wieczorkowska (Eds.), *Advances in music information retrieval*, Studies in Computational Intelligence. Berlin: Springer.

Pearce, M. T., Müllensiefen, D., & Wiggins, G. A. (2010). Melodic grouping in music information retrieval: new methods and applications. In: R. Zbigniew & A. Wieczorkowska (Eds.), *Advances in music information retrieval. studies in computational intelligence*, vol. 274 (pp. 365–390). Berlin: Springer.

Pearce, M. T., & Wiggins, G. A. (2004). Improved methods for statistical modelling of monophonic music. *Journal of New Music Research*, *33*(4), 367–385.

Pearce, M. T., & Wiggins, G. A. (2006). Expectation in melody: The influence of context and learning. *Music Perception*, *23*(5), 377–406.

Pechmann, T., & Mohr, G. (1992). Interference in memory for tonal pitch: Implications for a working-memory model. *Memory & Cognition*, *20*(3), 314–320.

Pickens, J., Bello, J. P., Monti, G., Sandler, M. B., Crawford, T., Dovey, M., & Byrd, D. (2003). Polyphonic score retrieval using polyphonic audio queries: A harmonic modeling approach. *Journal of New Music Research*, *32*(2), 223–236.

Rundus, D. (1974). Output order and rehearsal in multi-trial free recall. *Journal of Verbal Learning and Verbal Behaviour*, *13*, 656–663.

Schulkind, M. D. (2000). Perceptual interference decays over short unfilled intervals. *Memory & Cognition*, *28*(6), 949–956.

Schulkind, M. D. (2004). Serial processing in melody identification and the organization of musical semantic memory. *Perception & Psychophysics*, *66*(8), 1351–1362.

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*(2), 217–249.

Shepard, R. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behaviour*, *6*, 156–163.

Shiffrin, R., & Nosofsky, R. (1994). Seven plus or minus two: A commentary on capacity limitations. *Psychological Review*, *101*(2), 357–361.

Sloboda, J. A., & Parker, D. H. (1985). Immediate recall of melodies. In P. I. C. Howell, & R. West (Eds.), *Musical structure and cognition* (pp. 143–167). London: Academic Press.

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

Thomassen, J. (1982). Melodic accent: Experiments and a tentative model. *Journal of the Acoustical Society of America*, *71*, 1596–1605.

Tulving, E. (1968). When is recall higher than recognition? *Psychonimic Science*, *10*, 53–54.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.

Typke, R., Wiering, F., & Veltkamp, R. (2007). Transportation distances in human perception of melodic similarity. *Musicæ Scientiæ, Discussion Forum 4A*, 153–181.

Uitdenbogerd, A. L. (2002). *Music information retrieval rechnology*. Ph.D. thesis, RMIT University of Melbourne, Australia.

Unal, E., Chew, E., Georgiou, P., & Narayanan, S. (2008). Challenging uncertainty in query by humming systems: A fingerprinting approachenging. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(2), 359–371.

Underwood, B., Rundquist, W., & Schulz, R. (1959). Response learning in paired-associate lists as a function of intralist similarity. *Journal of Experimental Psychology*, *58*, 70–78.

Volk, A. (2008). Persistence and change: Local and global components of metre induction using inner metric analysis. *Journal of Mathematics and Computation in Music*, *2*(2), 99–115.

Weyde, T. (2004). Modeling rhythmic motif structure with fuzzy logic and machine learning in music query: Methods, models, and user studies. *Computing in Musicology*, *13*, 35–50.

Wiggins, G., Pearce, M., & M¨ullensiefen, D. (2009). Computational modelling of music cognition and musical creativity. In R. Dean (Ed.), *The Oxford handbook of computer music and digital sound culture*. Oxford: Oxford University Press.

Wiggins, G. A. (2007). Models of musical similarity. *Musicæ Scientiæ, Discussion Forum 4A*, 315–337.

Williamson, V. J., Baddeley, A. D., & Hitch, G. J. (2006). Music in working memory? examining the effect of pitch proximity an the recall performance of nonmusicians. In M. Baroni, R. C. Addessi, & M. Costa (Eds.), *Proceedings of the 9th International Conference of Music Perception and Cognition (ICMPC9)* (pp. 1581–1589). Bologna: Society of Music Perception and Cognition (SMPC) AND European Society for the Cognitive Sciences of Music (ESCOM).

Young, R. (1962). Tests of three hypotheses of the effective stimulus in serial learning. *Journal of Experimental Psychology*, *63*, 307–313.

Zajonc, R. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology Monograph Supplement*, 9, 1–27.

Zielinska, H., & Miklaszewski, K. (1992). Memorising two melodies of different style. *Psychology of Music*, *20*, 95–111.