Journal of New Music Research
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/nnmr20

The Perception of Accents in Pop Music Melodies

Daniel Müllensiefen [a], Martin Pfleiderer [b] & Klaus Frieler [b]
[a] University of London, UK
[b] University of Hamburg, Germany

Version of record first published: 14 Oct 2009.

To cite this article: Daniel Müllensiefen , Martin Pfleiderer & Klaus Frieler (2009): The Perception of Accents in Pop Music Melodies, Journal of New Music Research, 38:1, 19-44

To link to this article: http://dx.doi.org/10.1080/09298210903085857

Journal of New Music Research
2009, Vol. 38, No. 1, pp. 19–44

Routledge
Taylor & Francis Group

The Perception of Accents in Pop Music Melodies

Daniel Müllensiefen[1], Martin Pfleiderer[2], and Klaus Frieler[2]

[1]University of London, UK; [2]University of Hamburg, Germany

Abstract

We examine several theoretical and empirical approaches to melodic accent perception and propose a heuristic classification system of formalized accent rules. To evaluate the validity of the accent rules a listening experiment was carried out. 29 participants had to rate every note of 15 pop music melodies presented as audio excerpts and as monophonic MIDI renditions for their perceived accent strength on a rating scale. The ratings were compared to accent predictions from 38 formalized, mainly binary accent rules. Two statistical procedures (logistic regression, and regression trees) were subsequently used in a data mining approach to determine a model consisting of an optimally weighted combination of smaller rule subset to predict the accents votes of the participants. Model evaluation on a set of unseen melodies indicates a very good predictive performance of both statistical models for the participants' votes obtained for the MIDI renditions. The two models derived for the audio data perform less well but still at an acceptable level. An analysis of the model components shows that Gestalt rules covering several different aspects of a monophonic melody are of importance for human accent perception. Among the aspects covered by both models are pitch interval structure, pitch contour, note duration, metrical position, as well as the position of a note within a phrase. In contrast, both audio models incorporate mainly rules relating to metre and syncopations. Potential applications of the presented accent models in automatic music analysis as well as options for future research following this computational approach are discussed.

1. Introduction

Exploring melodic accent structures and their underlying perceptual regularities is an important task in several areas of music research. Firstly, in structural and reductive approaches in music analysis, like the Schenkerian paradigm, it is important to know which notes have perceptually more weight than others. Secondly, the perception of melodic accent structure is probably closely related to the representation of melodies in memory. It can be hypothesized that the more salient notes of a melody (i.e. more accentuated notes) are better remembered whereas less important notes are likely to be forgotten or altered in memory in some way. Thirdly, accents play a fundamental role in our understanding of how metrical structures are constructed cognitively.

We define a melodic accent as the perceived property of a particular note of a melody, i.e. its perceptual salience or its (accent) strength. A melodic accent is thus a perceptual construct and does not relate to any particular source by which the percept was induced. As we will discuss in detail below, the sources that can induce melodic accents include the pitch domain, rhythm, harmony, metre, dynamics, agogic, timbre/instrumentation, lyrics, etc. as well as contextual sources such as simultaneously occurring events in the accompanying music.

The paper is organized as follows. In the first part, we start with an in-depth examination of theoretical and empirical approaches to accent perception. Then, a formalized system of 38 accentuation rules is given. Subsequently, we report a listening experiment yielding rule models for accent perception in pop music melodies. An evaluation of the models is given followed by a critical discussion and an outlook to further investigations.

Correspondence: Daniel Müllensiefen, Goldsmiths, University of London, New Cross Road, New Cross, London SE14 6NW, UK. E-mail: d.mullensiefen@gold.ac.uk

DOI: 10.1080/09298210903085857

2. Background

Following general approaches to perception and cognition (e.g. Neisser, 1976; Gibson, as cited in Reed & Jones, 1982), several theories in music research have emphasized the importance of selecting, structuring, and weighting external musical input to form expectations and facilitate cognitive processing (e.g. London, 2004; Huron, 2006). Cue abstraction (e.g. Deliège, 1996) and the ability to discriminate the important from the unimportant is believed to operate, at least in part, at a psychophysical level which involves semi-automatic processing (e.g. Balkwill & Thompson, 1999; Eerola et al., 2006) in addition to acquired semantic aspects and effects of enculturation. We believe that the perception and the processing of accents in melodies takes place along these general lines and serves several functions.

One important aspect of melodic accents is that they are efficient and robust cues that can be abstracted during listening and retained in memory. Abstracting melodic accents may therefore reduce memory load, while a sufficiently accurate representation of a melody can still be stored for comparison with future melodic events. Monahan et al. (1987), as well as Mari Riess Jones and collaborators (Jones et al., 1987; Jones & Ralston, 1992) were able to demonstrate that accent structures of short melodies were indeed kept accurately and robustly in memory, while surface information was forgotten more easily.

A second function of melodic accents is to provide temporal markers and establish a metrical grid, which supports the anticipation of musical events at certain time points (see e.g. Large & Jones, 1999). The expectations generated from these anticipations are believed to direct listeners' attention and can thus be considered as an important component in the generation of subjective musical structure, particularly in metre induction and rhythm perception.

Since Cooper's and Meyer's statement that 'an ultimate definition [of melodic accents] in terms of psychological causes does not seem possible with our present knowledge' (Cooper & Meyer, 1960, p. 7) various studies have focused on the cognitive induction of temporal structure through perception of rhythmic and melodic accents. Among the most prominent models in this field are the *internal clock* model by Povel and Essens (1985), the relation between melodic cues and metre perception as described by Hannon et al. (2004) and the *Joint Accent Structure* hypothesis by Jones and Boltz (1989) which has evolved into a somewhat general model of attention (Large & Jones, 1999; Jones et al., 2002) on one hand, and into a time and duration memory model for the auditory domain (Boltz, 1992, 1995, 1998) on the other.

However, the interplay between attention, expectation and accents is quite complex. Accented events can create expectations in the listener via establishing a metrical frame. Fulfilment of metrical expectations can give rise to events as being accented, e.g. on downbeats. But it is obvious that a violation of these expectations, i.e. surprise, will catch the listeners' attention and give rise to the perception of an accented event as well (see Huron, 2006). Viewing it from a melody composition perspective, these two opposing sources for accent perception, expectation fulfilment and expectation violation, need to be balanced in some way, otherwise neither of them can take place. One solution is to use different musical dimensions, instead of the temporal dimension alone. For example expectations with regard to metre can be confirmed by placing note onsets on strong metrical positions while expectations with regard to melodic contour can at the same time be violated by frequent reversals of the melodic direction or the use of large interval skips. Indeed, most approaches agree that there are several possible sources that contribute to the perceived accent strength of a particular note. Following Monahan and Carterette (1985, p. 3) and Parncutt (1994, p. 426) the perception of an accent can arise from note volume, duration, timbre (instrumentation), its relation to underlying tonality or harmony, its relation to a metric structure, and changes in the pitch sequence. As an additional dimension we would like to add the position of a note within a melodic structure.

In music theory and analysis one can find numerous explanations of why certain notes are more important than others (e.g. Piston, 1950a,b; de la Motte, 1993), and thus have greater salience. But an algorithmic systematization of procedures for determining the accent strength of any given melody has not been developed yet. Furthermore, explanations of note accent strengths are rarely backed up with empirical evidence from listener studies.

Turning to the psychological literature, there are several models that describe simple mechanisms or rules for determining the accent strength of melodic notes within one dimension. Thomassen (1982), for example, puts forward a classification of possible three-note patterns (trigrams) with respect to their melodic contour. The probability of an individual note being perceived as an accent is calculated in his model as the product of two accent probabilities determined by their position as second or third element of a three-notes group. Testing this model with an indirect paradigm using judgements for rhythmical regularity for isochronous tone sequences, Thomassen (1982) found a good correlation between the participants' responses and the predictions resulting from his contour based model. Other dimensions that might contribute to the perceived accent strengths of melodies were deliberately ignored by Thomassen.

Similarly, a number of studies by Dirk-Jan Povel and colleagues (Povel, 1981; Povel & Okkerman, 1981; Povel & Essens, 1985) investigate the influence of tone

durations on the perception of accent strength in monophonic sequences and on the induction of a cognitive metrical grid in the listener (see also Parncutt, 1994). Melodic movement or harmonic relations are again deliberately left out. In most empirical investigations on metre induction Eric F. Clarke noted a lack of the exploration '(...) of different kinds of accent and the ways in which accentual and temporal factors mutually influence each other (...). This is unfortunate both because purely temporal models of meter perception (...) are unrealistic in the demands that they make by deriving meter from temporal information alone and because such models tend to project a static and one-dimensional view of meter, rather than the more dynamic and fluid reality that is the consequence of the interplay of different sources of perceptual information (temporal and accentual)' (Clarke 1999, p. 489; see also Hannon et al., 2004).

There is a second kind of model in the literature, which explicitly regards several dimensions simultaneously that could be sources of accent perception in melodies. Huron and Royal (1996) showed that in real melodies (e.g. monophonic excerpts from Western art music, Gregorian chant) accents from selected dimensions significantly correlate with metrical position or the onset of text syllables. This notion of accent coincidence and its perceptual consequences is the subject of a large body of empirical psychological research. Two of the most prominent models in this respect are the model of melodic-rhythmic consonance by Caroline Monahan and colleagues (Monahan & Carterette, 1985; Monahan et al., 1987; Hirsh et al., 1990; Monahan, 1993) and the Joint Accent Structure hypothesis within the dynamic shape model developed by Mari Riess Jones and Marilyn Boltz (e.g. Boltz & Jones, 1986; Jones, 1987, 1993; Jones & Boltz, 1989; see also Pfordresher, 2003, for further empirical testing). Both models are quite similar in various aspects, which makes it possible to give a general description. We will mainly focus on the formulations as given in Monahan et al. (1987) and Jones (1987, 1993) for a brief summary of both models.

The main feature of the two models is that both assume accent coupling, i.e. stronger accents arise where accents from different dimensions coincide. For instance, they account for the intuitive notion that a long note that follows a large jump and reverses the contour of the melodic line, is likely to be perceived as an important note within a melodic sequence. Both models consider accents from intervallic movement, pitch contour, note duration, and specific note position. Monahan et al. (1987, p. 580) additionally take melodic grouping structure into account following work by Vos (1977), whereas Jones and Boltz incorporate the harmonic context of a melodic note (Boltz, 1991, 1993, 1999; see also Jones' 'tonal end accent', Jones, 1987, p. 623).

Both models employ summation for calculating a combined accent strength value from individual dimensions. Thus, accents from each dimension are usually assigned a weight of 1 while it is assumed that only one accent can arise from each dimension for a given note. An accent arising for example from pitch contour can be added to an accent from note duration and the resulting accent strength would thus be twice as high. Therefore, an explicit weighting scheme for accents from different dimensions is not introduced in either model. Inspired by Maury Yeston's work (1976), and in addition to the notion that accents can coincide on notes (*consonance*), Monahan and colleagues also proposed to model series of accents in different dimensions that are phase shifted, but can be characterized by regular time intervals forming ratios of small integer values. They term this phenomenon *out of phase consonance* and illustrate it by a series of notes in which pitch and rhythmic accents alternate regularly. The third concept to characterize the relation of accent series is *dissonance*, which they use to describe the unpredictable non-coincidence of accents from two different dimensions. Very similar concepts for the characterization of relative timing behaviour between accents from different dimensions have been proposed also within the Joint Accent Structure framework (e.g. Jones & Pfordresher, 1997; Pfordresher, 2003).

For estimating the importance of individual notes of a melody, Monahan and colleagues assume the consonance of accents as the appropriate modelling concept. From these considerations results the simple calculus of summing accents over all considered dimensions.

Both models assume the induction of a temporal or metrical structure that allows for event anticipation and expectations. Monahan et al. refer here to the concept of an internal best metrical clock developed by Povel and Essens (1985). Jones and Boltz postulate their own concept of dynamic rhythmic attending for listening to a melody, or more generally, for perceiving a series of temporal events (Jones & Boltz, 1989; Jones et al., 2002).

Finally, both models formulate very similarly specific rules for the determination of an accent for a specific note in a specific dimension that are largely in concordance with general knowledge from music theory and psychoacoustic research. Most of these rules are based on Gestalt laws, like the principles of proximity, similarity or good continuation, which are then individually applied to the musical information within the several dimensions.

But most rule formulations lack the precision to be unambiguously applicable in all analysis situations. The formulations of temporal structure accents due to long duration, might serve as an example. Jones (1987, p. 624) states that '[...] any relatively long or short tonal duration defines a temporal accent.' The same idea is expressed similarly by Monahan et al. (1987, p. 577): 'The strongest *natural* temporal accent tends to appear on tones initiating lengthened intervals (as measured from the attack point of one tone to the attack of the

next, denoted IOI for inter-onset interval).' Both formulations might serve well for the generation of artificial experimental stimuli where the experimenter chooses a minimum ratio that defines the relation between long and short tone durations. But in order to use this rule for determining temporal accents a decision must be made about the precise meaning of 'relatively long' and 'lengthened intervals'. Possible operational definitions range from 'longer than the previous note' over 'at least two times the duration of previous note' to 'longer than the mode of durations of the melodic phrase'.

Similarly, we find imprecise definitions for pitch and contour accents in the literature. For example Pfordresher (2003, p. 433) explains that '[m]elodic (m) accents were created by pairing changes in pitch direction with local increases in the semitone distance between successive pitches.' But unfortunately he does not make explicit how he defined 'local', nor what a sufficient criterion for an increase might be.

Therefore, it is one of the chief aims of the present study to test various operational definitions of Gestalt principles that might be suitable for accent perceptions. We formulated and implemented several rules differing in a previously specified parameter (e.g. relative duration length) when we could not find clear indications in the literature. In this aspect the present study can be seen as an exploratory search through the parameter space for certain rules. This search is facilitated by our attempt to impose a taxonomic structure onto the rule set that in turn helped with the definition of new rules.

Moreover, our study aims at selecting a set of rules corresponding to a set of dimensions and to combine the rules in a way that is optimal for explaining accent strength perceptions of notes in a melody. Previous studies have not rigorously tested the explanatory power of individual rules in the context of a model. In the case of the above delineated models by Monahan and Jones and Boltz, summative models were constructed on the basis of *a priori* considerations and then tested for their explanatory power. Likewise, Parncutt (1994) constructs a model for computing durational accents on the basis of several *a priori* assumptions. But none of these studies systematically experimented with adding and subtracting individual rules to or from the original model. Thus, we took very explicit care in the process of *variable selection* given a huge solution space defined by the number of possible rule combinations.

Not only should this study select some operational rule formulations that are superior to others in terms of explaining human accent perception, we also would like to address the question of how selected rules may be combined into an overall model with high predictive power. As mentioned above, most previous studies have chosen to sum accent values from different rules. An obvious improvement to this approach is to build weighted sums where weights might be interpreted as reflecting the relative importance of the individual rules within the model. An optimal rule weighting is the result of linear logistic regression models we applied in the analysis of the experimental data. We tried a second approach in the form of a classification tree model (e.g. Breiman et al., 1984). This type of model is very well suited for modelling interactions of rules in a hierarchical sequence. Tree models are generally quite efficient in situations where relations between variables are non-linear or data is missing. Often they yield simple graphical interpretations. As we only have little *a priori* knowledge of the nature of the cognitive processes which combine the rules for different musical dimensions into an overall model, regression models and tree models are two deliberately different attempts to explain experimental data, and thus are by comparison possibly illuminating the nature of underlying cognitive processes.

Another explicit goal of this study is the construction of an instrument for accent strength measurement. This is a slight difference from most of the aforementioned studies, which were primarily interested in testing the influence of certain Gestalt principles on melody or accent perception empirically (e.g. Eiting, 1984; Jones et al., 1987; Jones & Ralston, 1991; Boltz, 1993; Pfordresher, 2003). In many experiments the primary method of empirical validation has been to create artificial melodic sequences that are either in accordance with certain principles under scrutiny or not. The influence of that specific principle on melodic accent perception can then be assessed easily by statistical tests. At the same time the possibility to generalize the experimental results from artificially constructed sequences to real melodies of any style often remains unexplored and questionable. In contrast to artificially generated musical stimuli, we employ melodies from pop songs as experimental stimuli. They are presented as dead-pan MIDI files with uniform timbre as well as within their original audio context. The difference between the best fitting models for the MIDI melodies and the audio variants of the experiment should inform us about the ability to generalize models constructed on the basis of experiments with monophonic melodies to the perception of melodies embedded in a realistic song context. To our knowledge no experiment to date has tested the descriptive power of basic Gestalt-rules for melodic accent perception with excerpts from real songs.

3. Summary of goals

In summary, we define four primary goals of this study.

First, to identify the most important rules that are involved in human accent perception.

Second, to define models in which the perceptual rules can be combined. This follows the idea favoured by Boltz and Jones (1986), and by Monahan et al. (1987) that accent sources from different musical dimensions are evaluated simultaneously in human melody perception. But our concern is also the way of combining different accent rules.

Third, to compare results for monophonic melodies and melodies embedded in their original musical context. With respect to musical context, we limited ourselves to pop music melodies that are stylistically familiar to most listeners.

Fourth, to devise an analytic algorithm that assigns accent markers or accent weights to any monophonic melody in a cognitively valid way.

4. Layout of the study

With these goals in mind we designed a new experimental approach to study accent perception in melodies. Our approach encompasses several steps:

1. Precisely define a large set of mainly binary accent rules from the literature including possible variants.
2. Determine all possible accents algorithmically for all notes in a set of melodies according to the defined accent rules.
3. Let listeners judge the accent strengths of notes in a set of melodies—monophonic melodies as well as the same melodies in the original context.
4. Fit data from algorithmic accent rules to participants' ratings using different approaches (logistic regression, regression trees) of selecting accent rules with the highest explanatory power.
5. Evaluate the accuracy of the rule models using a different data test set.
6. Construct an accent assignment algorithm on the basis of the most accurate model.
7. Compare results for melodies in the original context and monophonic versions of the same melodies.

5. Accent rules

With regard to the selection of accent rules from the literature we limit ourselves mainly to simple rules that can be coded on the basis of individual note events and in a binary manner, i.e. a specific musical condition is fulfilled by a note in a given context or it is not. By limiting ourselves to mostly binary rules we exclude the more complex rules and accent functions that make use of internal parameters. A good example of a model that uses sophisticated perceptual functions, which might very well be incorporated into a model of accent perception, is Margulis' (2005) model of melodic expectancy. The several perceptual functions of this model (proximity, stability, expectancy, and tension functions) might be very relevant for explaining accent perception. However, by only considering simple binary accent rules we can estimate all the necessary parameters within the modelling procedure itself and do not have to rely on pre-existing parameters (e.g. a translations table of interval distance to perceived proximity) that might be conceived on rather music-theoretic grounds or have been developed empirically in the context of a different musical style (e.g. Western art music or folk music). Using only binary rules may at first seem a crude way to assess musical reality but it keeps the individual rules comparable and leaves the full control over the parameters (weights) associated with the individual rules to the development of a statistical model on this particular dataset. Using only binary rules has as well the advantage that they can be easily understood and implementations by other researchers can be realized immediately. We thus consider the simplicity of a model based on binary rules part of its attractiveness.

For the sake of exploring the space of binary accent rules gathered from the literature, we divide the accent rules tested here into six different categories: rules concerning pitch interval, pitch contour, note duration (inter-onset interval), position in phrase structure, metre and syncopation, and tonal context. These categories serve only as an initial heuristic for handling a large set of rules systematically and correspond to what seems to be common sense in accent research rather than to a particular psychological or musical theory. We ignore the dimensions of timbre and polyphonic coincidences (accompaniment) for the scope of this study, which focuses on monophonic melodies. Nonetheless, these dimensions might play an important role in the perception of melodies in the context of polyphonic music.

We define each rule unambiguously for all possible analytic situations in a way that can be implemented in a computer program straightforwardly. We do not claim that the following list of rules is exhaustive in any respect but we do think that it covers many of the ideas discussed in the accent literature and supplies alternative implementations where the transduction from an idea to an unambiguous rule is not directly evident. In this regard, the following list of rules can be viewed as a first and tentative attempt to explore the potentially large space of factors that have an influence on melodic accent perception.

We conceive melodies as sequences of pitches and onsets. Additionally, we make use of metre information that is provided with the test stimuli (see paragraph on metre related rules below). For most of the rules, accents are simple binary markers that indicate for every note if

the note receives an accent according to one accent rule. In this case, the accent value is 1, otherwise it is 0. The only exception is the contour rule by Thomassen (1982), which assigns a probability value between 0 and 1 to a note. We interpret this value as an accent strength value.

5.1 Pitch interval rules

From statistical surveys it is known that large intervals are much less common in most musical cultures than small intervals (Dowling & Harwood, 1986; Huron, 2006). Thus, large intervals can be seen as some kind of expectation violation forming possible events of interest within a melody. However, it is not clear what the critical interval size might be for an accent to be perceived. Beyond that, there are differing opinions in the literature on which of the two notes involved is accented. Therefore we formulated rules, where either the first, the second or both notes receive an accent. As a result we defined ten different accent rules using pitch interval information, which seem to cover most of the accent rules concerning interval size in the literature.

Rule name	Definition
jumpaft3	Accent on a note after a jump of 3 or more semitones
jumpaft4	Accent on a note after a jump of 4 or more semitones
jumpaft5	Accent on a note after a jump of 5 or more semitones
jumpbef3	Accent on a note before a jump of 3 or more semitones
jumpbef4	Accent on a note before a jump of 4 or more semitones
jumpbef5	Accent on a note before a jump of 5 or more semitones
jumpbea3	Accent on notes before and after a jump of 3 or more semitones
jumpbea4	Accent on notes before and after a jump of 4 or more semitones
jumpbea5	Accent on notes before and after a jump of 5 or more semitones
jumploc	Accent on the second note of an interval that is at least two semitones larger than its successor and predecessor interval and, thus, constitutes a change towards a class of a larger interval (see Monahan et al., 1987)

5.2 Pitch contour rules

For a long time pitch contour has been regarded as a very significant dimension of melody perception (e.g. Dowling & Fujitani, 1971; Dowling, 1978). The reversal of melodic direction seems to be an event that can trigger accent perception. But authors differ in their opinion on how to handle changing notes that are part of an ornament, like a trill. We formulated several variants that differ in the definition of a changing note. We also assigned the rules for contour based accents defined by Thomassen (1982) into this category. Unlike the rest of the contour based rules, Thomassen's rules result from taking into account a three-note context. As Thomassen postulates an algorithm, which is rather complex and since it is defined in detail in the original publication, we give only the reference here.

Rule name	Definition
pextrem	Accent on a contour extremum note, i.e. predecessor and successor notes both are lower or higher in pitch, as well as the first and last note of a melody
pextrst	Accent on contour extremum note excluding changing notes according to Steinbeck's (1982) definition, i.e. notes where the two preceding and succeeding notes are not all either lower or higher than the present note
pextrmf	Accent on contour extremum note excluding changing notes according to Müllensiefen and Frieler's (2004) definition, i.e. notes that are surrounded by two notes with the same pitch
pextrsta	Accent on note following note accented by **pextrst**
thom	Accent weight according to Thomassen's (1982) algorithm, which is based on the seven possible pitch direction patterns that can be formed by 2-interval chains (3-note patterns)
thomthr	Thresholded version of **thom**; all values <0.5 are assigned the value 0, all other values are set to 1

5.3 Inter-onset interval rules

In many previous studies duration and inter-onset intervals (IOI) have been identified as decisive factors for grouping, accent perception, and metre induction. We limit ourselves to inter-onset intervals and do not regard the true duration of notes and rests between notes. The accent rules based on inter-onset intervals differ in regard to time interval of reference, the deviation from that reference (short versus longer), and to the ratio of the inter-onset interval to the reference inter-onset interval that is sufficient to generate perceptual accent.

Rule name	Definition
longpr	Accent on a note starting an IOI longer than the IOI of the predecessor note
long2pr	Accent on a note starting an IOI at least two times as long as the IOI of the predecessor note
longmod	Accent on a note starting an IOI longer than the mode of IOIs in the melody
long2mod	Accent on a note starting an IOI at least two times as long as the mode of IOIs in melody
shortpr	Accent on a note starting an IOI shorter than IOI of predecessor note
short2pr	Accent on a note starting an IOI at most half as long as the IOI of the predecessor note
endloioi	Accent on a note that ends an IOI, which is at least two times as long as the mode of the IOIs in the melody

5.4 Position rules

Phrase ending and beginnings are widely believed to be places of high attention, and therefore induce accent perception. But what exactly is a musical phrase? We experimented with several existing models for melodic phrase segmentation (Cambouropoulos, 1998; Temperley, 2001). Informal tests with these models showed unsatisfactory results when applied to pop music melodies. We therefore used a 2-rule segmentation model, named SimpleSegmenter, that captures the factor which most of the current segmentation models regard as central for boundary perception, namely long inter-onset intervals. Thus, SimpleSegmenter can be seen as a common-sense model for melody segmentation, which is rather conservative because it recognizes only long temporal gaps as an indicator for phrase segmentation but disregards other sources of melodic discontinuity. Once an on-going evaluation of different segmentation approaches is completed, (preliminary results presented in Müllensiefen et al., 2007) SimpleSegmenter can easily be replaced by a more accurate (i.e. less conservative) segmentation model. According to the SimpleSegmenter model phrase endings are defined by notes with an inter-onset interval of at least 4 times the mode of IOIs found in the entire melody or by notes with an IOI of more than 1.5 s. The last note of a sequence is always counted as a phrase ending. In a way, the following position rules can also be defined in terms of inter-onset intervals, but for further extension to other segmentation algorithms and for conceptual clarity they are presented here in an extra section.

Rule name	Definition
phrasbeg	Accent on a note beginning a melodic phrase
phrasend	Accent on a note ending a melodic phrase
shortphr	Accent on the second note of melody phrases consisting of two notes only

5.5 Metre and syncopation rules

In the model by Monahan and colleagues accents arising from notes starting on a position on the metrical grid play an important role. In the audio excerpts used as experimental stimuli, beats and metre were clearly detectable due to the accompaniment. For the monophonic version of the melodies, participants had to induce the metre from the melody itself. For the computational model which needs metre information as input in order to apply binary metre and syncopation rules, we are left with two options: either we use a metre induction model from the cognitive and computational literature and have the metre induced for each melody. Candidate models for metre induction can be found for example in Eck (2001), Temperley (2001), Frieler (2004), and Volk (2008). Or we use the metre information as given in the MIDI files which stem from the transcriptions of the melodies in their song context by an expert transcriber and thus can be considered the *objective* metre. Either way we cannot be sure that all of our subjects deduced the metre of the monophonic melodies in the same way, either according to any cognitive-computational model or as the correct metre. Without explicitly testing or biasing their metre perception in a secondary task at the time of the experiment there is no way to determine what proportion of the listeners induced the objective metre or what alternative metre they represented cognitively. However, results of a small-scale pre-test with a few cognitive-computational models showed model-induced metre information was problematic both in terms of the quality of the results and also with regard to practical issues. In particular, it seemed that the tested computational models were developed for melodies from different repertoires (Western art music or folk songs) and thus performed less well on pop melodies most of which have inherently different rhythmic characteristics. We therefore chose to use the objective metre as given in the MIDI file. This also allows for a more general practical applicability of the model since it thus can be used on any monophonic MIDI (having the obligatory metre information) and does not require a pre-processing with foreign software. Nonetheless, to increase the cognitive validity of the proposed accent perception models, a necessary step in a subsequent study would be to evaluate the different existing metre induction models on a stimulus set taken from the pop

repertoire and to incorporate one or more metre induction models with acceptable performance as a pre-processing stage into the overall accent strength computation.

In summary, the effectiveness of the binary metre and syncopation rules in this category depends on two conditions, namely (a) that the objective metre found in the MIDI files is in some way related to the metre induced by the participants and (b) that the rules operating on the induced metre and the notes of the melody correspond to a cognitive process carried out by the participants. In a worst-case scenario where listeners would have consistently induced a metre different from the objective one, the metre related rules would show little predictive power at the time of evaluation against the experimental data. Thus, by including rules that make use of the correct metre information we test whether external metre information, if it is available, can help with predicting listeners' accent perception in pop melodies. Of course, if metre and syncopation rules fail to be effective predictors for accent perception it is difficult to determine whether themselves are badly constructed or whether the metre information in the MIDI files was incongruent with the listeners' perceptions. But if for our empirical dataset rules based on external metre information actually were effective for predicting the perception of melodic accents then external metre information would be very relevant for many practical applications dealing with melodic data from standard symbolic music formats like MIDI, **kern, EsAC, or MuseData.

While we are aware of a number of complex models of rhythm and syncopation perception that generate differentiated parametric profiles of metrical and syncopation strengths (Longuet-Higgins & Lee, 1984; Smith & Honing, 2006; Fitch & Rosenfeld, 2007), we limit ourselves again to binary rules for reasons of comparison and parameterization outlined above. As said before, it would be very interesting to incorporate some of these parametric models of syncopation strength into an accent perception model but a different framework for combining and weighting output from different pre-existing models would have to be established.

Syncopation is generally assumed to be important in Western popular music. There are several algorithmic definitions in the literature and we used different methods to determine syncopations. We adopted one view on syncopation as events being shifted forward from a beat position to a weaker position directly before, leaving the following beat position unoccupied. This phenomenon is known by musicians as 'anticipated beats' or 'offbeats' (Temperley, 1999). According to this view no distinction between different kinds of subdivisions of a beat IOI are necessary.

In a slightly alternate definition, syncopation can occur on any level of a metric hierarchy (Temperley, 2001). Higher levels of a metrical hierarchy are obtained by grouping and lower levels are achieved by dividing of inter-onset intervals. Each inter-onset interval of a metric level can be divided (or grouped) individually in either a binary or ternary pattern, which again can be divided (grouped) individually, thus giving a metric tree in a bi-fold recursive way. The reference for building a metrical hierarchy is the beat level, which is thought to be most stable, so no subdivision of a higher level is admissible which doesn't coincide with the beat level. A strong position of one level is defined as being part of the next higher level. Then, syncopation is defined as the occupation of a weak position within a metrical level while the next following strong position is left free; in the case of ternary subdivisions ('strong–weak–weak') according to this view, a note on the first weak position followed by a pause on the second weak position and followed by a note on the next strong position is a syncopation too. However, all melodies used in this experiment are in 4/4 metre with largely binary subdivisions of the beat.

Rule name	Definition
beat1	Accent on a note on the first beat of a bar
beat13	Accent on a note on the first or second metrically strong position of a bar, i.e. in 4/4 metre on the first or third beat
beat1234	Accent on a note on any beat
sync1	Accent on a note before the first beat of a bar and with an IOI extending over the next beat position
sync13	Accent on a note before beat 1 or 3 of a bar and with an IOI extending over the next beat position
sync1234	Accent on a note not on any beat and with an IOI extending over the next beat position
synchalf	Accent on a note with an onset on beat 2 or 4 of a bar and an IOI extending over next beat position 3 or 1
sync0	Accent on a note with an onset on the first subdivision level of the beat level (quaver or quaver triplet) with IOI longer than the time span of the subdivision
sync8s	Accent on a note with an onset on a second subdivision level of the beat level (semiquaver or semiquaver sextuplet) with an IOI longer than the time span of the subdivision
sync16s	Accent on a note with an onset on a third subdivision level of the beat level with inter-onset interval longer than the IOI of the subdivision

5.6 Tonal context rules

Tonal context has been introduced comparatively recently into the above cited multi-accent models. Particularly Marilyn Boltz (1991, 1993, 1999) explored the effect of a note being part of an underlying harmony and simultaneously falling into a specific structural position (phrase ending) or metrical position. In order to decide whether a note is part of a tonal context the underlying tonality has first to be induced from the monophonic melody. The Krumhansl–Schmuckler algorithm (Krumhansl, 1990) has generally proven to yield quite reliable results for tonal music. We used the Krumhansl–Schmuckler algorithm on a bar-wise basis like suggested in Krumhansl (1990). For determining the underlying harmonies from the audio examples, a manual transcription of the chord sequences was carried out and the two rules of tonal context were applied on the basis of the transcribed chords from the accompaniment and not using the Krumhansl–Schmuckler algorithm.

Rule name	Definition
triad	Accent on a note that is part of implied tonality of the bar (as determined by the Krumhansl–Schmuckler algorithm)
triadphen	Accent on a note that is part of implied tonality (Krumhansl–Schmuckler) of the bar and that ends a phrase (according to SimpleSegmenter, see above)

In addition to the perception of accents which might arise from the above-defined 38 rules, of course, supplementary accentuations of notes are perceived in actual performances of songs according to relative loudness, articulation (e.g. slides and slurs, micro-rhythmic deviations) and lyrics. We deliberately did not account for these possible sources of accentuation, firstly because we aimed for a general comparability between monophonic MIDI-melodies and the original audio excerpts. Secondly, it is a very challenging task in itself to measure perceived loudness, articulation and semantic content. These features can correlate with the structural accents or a performer can deliberately decide to use them against the melodic structure. In any case, these supplementary accentuations are not catered for in our models and thus constitute uncontrolled variables which might contribute to the overall variance.

The 38 rules were implemented in the MELFEATURE software toolbox for melodic analysis. The output is a set of 38 binary vectors (with the exception of the original formulation of Thomassen's contour rule), each vector having as many elements as there are notes in the melody. Each binary vector element represents the existence or the absence of a given accent feature at a particular note in the melody. Figure 1 shows a melody example in staff notation with the binary values resulting from the application of six arbitrarily chosen accent rules written underneath.

As can be seen from Figure 1, according to the rule **longpr** (i.e. give an accent to a relatively long note, i.e. a note starting an inter-onset interval longer than the preceding inter-onset interval) the second and the fourth note receive an accent. As can be seen, the frequencies of assigned accents vary considerably among the rules. For example, since the melody contains only three phrases, the rule **phrasend** gives only three accents, whereas **longmod** assigns an accent to 15 out of the 31 notes.

6. Experiment

6.1 Method

Given the opportunity to recruit musicology students with a high musical background as participants, we chose an explicit experimental paradigm, which involved reading pseudo-staff notation and marking the accent strength of individual notes.

6.2 Procedure

The participants were asked to listen to a short melody and to mark each note within a graphic representation of the melody indicating the perceived degree of accentuation. A pre-test showed that a three-point scale was suitable, because a finer scaling made the decision process considerably slower and more difficult. The three scale steps were termed 'not accentuated' (no mark, coded as 1), 'accentuated' (circle around note, coded as 2), and 'strongly accentuated' (circle and additional cross, coded as 3). The graphical representation of each melody (see Figure 2) was intended to be easy to follow along while listening to the melody excerpt. It contains all the necessary elements to accomplish the task and is still quite close to the original transcription from which it was derived. It shows only the approximate lengths (mainly quarter and eighth notes) and the relative pitch heights of the individual notes. There are no rests, bar lines, or staff lines included. For all audio examples that included vocals, the lyrics were printed below each note, to make it easier to follow the notation.

6.3 Participants

29 students of an introductory course in music psychology at the University of Hamburg took part in the experiment (12 female and 17 male students from 20 to 29 years old, mean age 23.2). All participants showed strong preferences for popular music styles, which

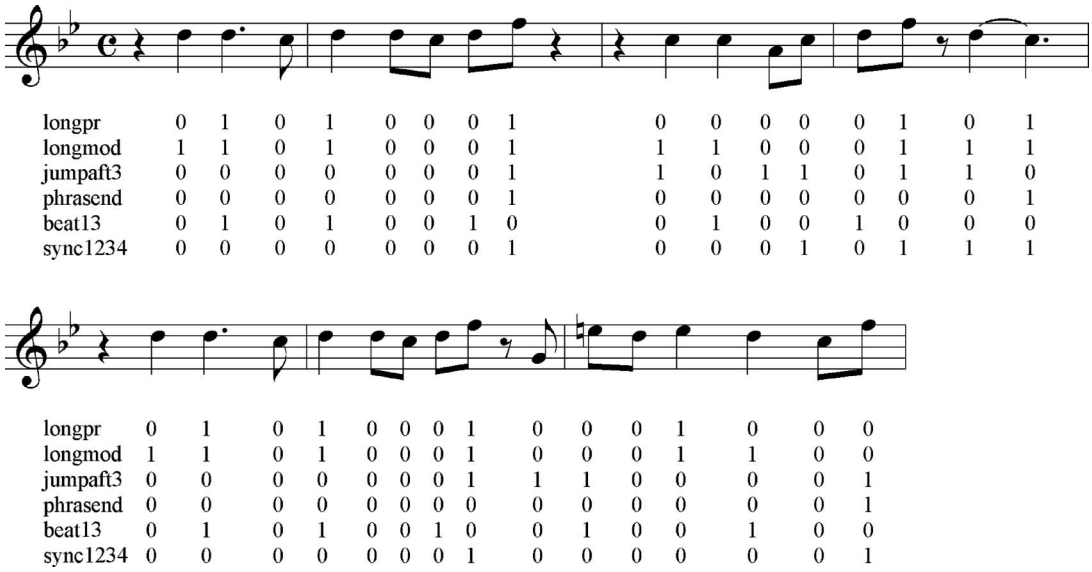

Fig. 1. 'Climb up (Stairway to heaven)' by Neil Sedaka; the values for each note of six arbitrarily chosen accent rules are depicted below staff notation.

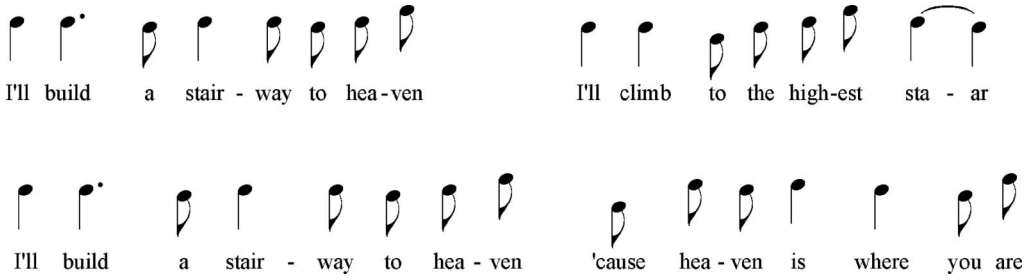

Fig. 2. Graphical representation of a melody ('Climb up', see Figure 1) on the test sheet.

suggests that they were familiar with the musical idiom of the test items (see below). Most of the participants had been playing a musical instrument for a long time (mean 12.1 years) and are still practicing (mean 4.1 h a week) and playing music (8.9 h a week) on a regular basis.

Two experimental sessions were conducted separated by a two-week time interval. Each session lasted for about 25 min. The participants were instructed that the experiment was about the perception of melodies and that the degree of accentuation was assumed to differ between the notes of a melody. Since each test melody was presented four times the participants were advised to just listen to the melody first, to circle the important notes the second time, to indicate at the very important notes the third time on the response sheet, and finally to examine their own ratings the last time the melody was played. Following the instructions, a training example was played to ensure that the graphical melody representation and the marking scheme were well understood. Participants could indicate whether they had problems understanding the concept of melodic accents or the task itself.

Both test sessions comprised 15 melodies. Additionally, two melodies were used exclusively to check for within-subject consistency (see below). Silent intervals of approximately 4 s were inserted between the four repetitions of each melody while different melodies were separated by 10 s of silence. In a pre-test, participants reported that it was easier for them to rate the monophonic melodies than the original excerpts. In order to maintain a high concentration level throughout the experimental sessions, we presented in both sessions the original excerpts first and thereafter the monophonic examples in a randomized order. After the first experimental session participants answered a questionnaire on personal data, musical training, music preferences, and feedback concerning the task. The second experimental session two weeks later involved the same participants, but an alternate version of each of the 15 melodies.

6.4 Stimuli (the melodies)

We prepared two versions of each of the 17 melodies: an excerpt from the audio recording of a popular music song

('original melody'), and a single-voice melody extract ('MIDI melody') played with the 'grand piano'-sound (MIDI patch 1) of the general MIDI device of a PC (Microsoft Software Wavetable Synthesizer with Roland Sound Canvas digital samples). Most of the excerpts were chosen from successful and typical but not very well-known pop tunes. In addition, there were three excerpts of contemporary R'n'B songs, one example of Jamaican Ska, one Reggae song and one Trinidadian Calypso tune (see Appendix for details of the songs used as experimental stimuli). The MIDI melodies were based on the vocal melodies of the recording excerpts as transcribed by the authors. In the first experimental session, eight original melodies and nine MIDI melodies were presented. In the second session, the participants listened to the original melodies corresponding to the MIDI melodies they had heard two weeks before and vice versa, except for the two consistency items. To test the consistency of the participants' ratings, additionally one original melody and one MIDI melody were played in both parts of the experiment. Each excerpt had a duration between 10 and 20 s and a 2- or 4-beat metre with binary eighth notes (quavers, i.e. no 'swing' eighths). The tempo of the melodies varied between 75 and 155 bpm.

7. Results

7.1 Task difficulty

The majority of the 29 participants reported that they seldom (17 participants) or never (11 participants) had difficulties relating the note symbols to the notes of the aurally presented melodies. They were able to maintain a good level of concentration during the task (with a median of 7 on a 10-step rating scale with '10' indicating perfect concentration) and they rated the task to be of medium difficulty (median of 5). Furthermore, contrary to the participants in our pre-test a majority of participants (24) reported the MIDI task to be more difficult than the audio task. Possibly, this might be due to the missing cues regarding metre and melodic phrasing in the MIDI condition with unaccompanied monophonic melodies. Missing these cues, the induction of the metrical structure can become more difficult for the participants, particularly for the often highly syncopated pop melodies. However, overall the experimental task was of medium difficulty to the participants and the explicit rating paradigm appears to be apt for the collection of empirical data on accent perception.

7.2 Consistency tests

The question of between-subject consistency concerns the possibility of whether modelling all participants' answers with one algorithm is feasible or whether any important or interesting of information would be lost when only some kind of average is modelled.

7.2.1 Within-subject consistency

As one of the primary aims of this study is to model human perception of melodic accents with algorithms we have to first answer the question whether the human behaviour as evidenced by the data from this accent task satisfies the basic conditions to be modelled by an algorithm. If most of our participants gave wildly diverging answers to identical stimuli in two different testing sessions then modelling the participants behaviour with an algorithm that always returns exactly the same answer when given the same stimulus would not be adequate.

We assessed within-subject consistency on the basis of the two melody items that were presented identically in both test sessions using Cohen's (1968) kappa as a standard procedure. The kappa coefficient ranges from $-\infty$ to 1 and as a rule of thumb Landis and Koch (1977) associate ranges of the κ-scale with categories of agreement. According to their heuristic scheme κ-values ≤ 0 indicate no agreement while values of 0.21–0.4 denote low agreement, and values of 0.41–0.6 mean moderate agreement and so on. We computed two κ-values from the 3-point accent ratings for all the notes of the two repeated melody items (1 original and 1 audio item) for each participant. We then screened the 28 participants from which we obtained ratings for both repeated items for low κ-values (1 of the 29 original participants gave missing values for one of the repeated items). Figure 3

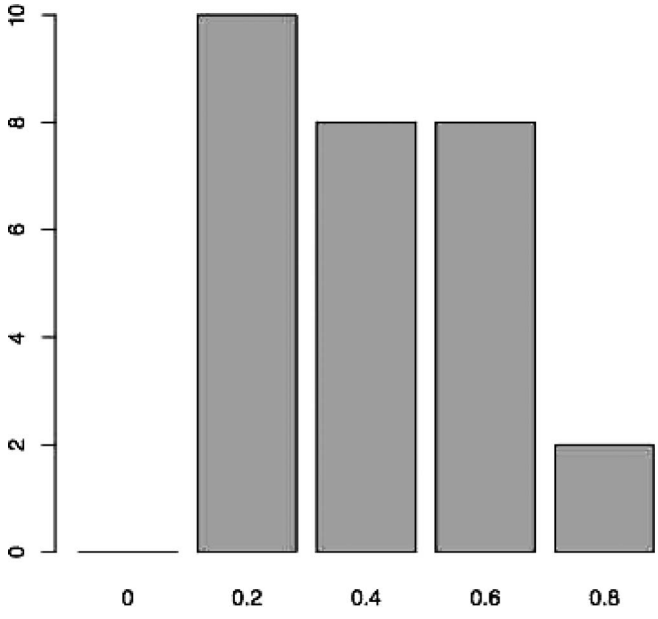

Fig. 3. Distribution of κ-values reflecting within-subject consistency for 3-point rating data of the test items (each participant is represented by the higher of his/her two κ-values).

shows the distribution of κ-values binned to Landis and Koch's scheme for the original 3-point rating data (each participant is represented by the higher of the two κ-values): none of the 28 participants had κ-values ≤ 0 for either melody item and only two did not reach the 0.2 level for one of the two items. 18 participants showed an agreement of 0.41 (moderate agreement) or higher. Mean κ over all 56 values (28 participants $\times$ 2 items) was 0.36.

To check how within-subject consistency depends on the use of the rating scale we collapsed all ratings for normal and strong accents to one category and computed the same statistics. The distribution of the κ-values from the binary data is depicted in Figure 4.

It is very clearly skewed towards the upper end of the scale. Mean κ for the binary values is 0.45. A paired Wilcoxon test (test statistic: 251, $p < 0.001$) confirmed the greater agreement as measured by Cohen's κ for the binary data (accent versus non-accent). Taken together, the data of the participant group seems to have an acceptable level of within-subject consistency already for the original 3-point ratings and we therefore decided to include the data of all participants in the subsequent analysis. A significantly higher agreement is found in the experimental data when only the participants' binary choices of whether a note is an accent or not are considered. For this reason, we chose to use only the binary information from the participants' ratings data in the subsequent analysis and collapsed all strong and normal accent ratings to one category. Additionally, it appears to be conceptually easier to model a data set with a binary response variable instead of trying to capture the additional information provided by the ratings on a 3-point scale. The 3-point scale ratings can at best reflect an ordinal scale level in the participants' judgments but they do not really allow for models that are designed for interval-scale dependent variables and that make assumptions like normality in the error distribution which are difficult to obtain with data from such a restricted measurement range.

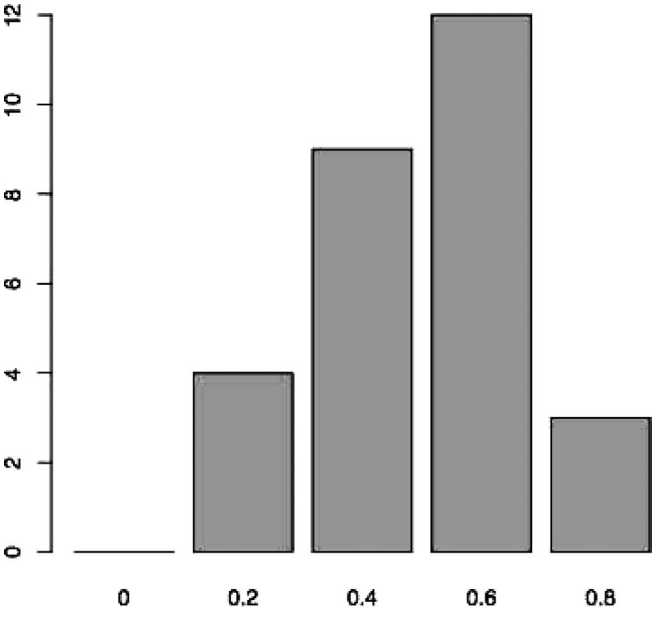

Fig. 4. Distribution of κ-values reflecting within-subject consistency for rating data of the test items collapsed to binary accent votes (each participant is represented by the higher of the two κ-values).

7.3 Layout of the data analysis

In this section we will report the results of the variable selection, modelling procedures, and the model evaluation. Due to space limitations, we will focus primarily on the treatment of the data obtained for the MIDI melodies. We processed the audio data almost identically and we will therefore only report the main results derived from the original audio items. The main goal is here to provide sufficient data and results for the reader to judge whether the models generated from MIDI and audio stimuli are indeed similar. Thus, unless stated explicitly otherwise, all numbers presented in this section refer to the data from the MIDI experiment.

In order to obtain one dependent response variable we added up all of the binary ratings and divided by the total number of participants ($n = 29$) for each note of each melody. This response variable, thus, represents proportion data in the sense that it indicates the proportion of participants in our sample who consider a particular note to be an accent. We aim at modelling, i.e. predicting, this response variable with the 38 accent rules. With the exception of the untransformed Thomassen rule (**thom**) these 38 predictor variables contain binary data. As it is unclear from *a priori* knowledge how human listeners derive accent perceptions and judgments from the various musical parameters we decided to model the influence of the predictors on the response variable in two different ways. The first model comes from the family of General Linear Models and is additive by nature. This notion of additivity corresponds to previous accent models e.g. by Monahan et al. (1987), Boltz and Jones (1986) and Parncutt (1994). The response variable in this model is modelled by a binomial model, which is an obvious choice for proportion data (see e.g. Collet, 2003; Crawley, 2007, Ch. 16).[1] In contrast, the second model is conditional in nature and primarily models the interaction between predictors and not their sum. We chose a regression tree model as proposed by Breiman et al. (1984). The response

[1]The term *logistic regression* is also commonly used to refer to this type of model emphasizing the type of transform (logit) applied to the response variable. However, we like to refer to this model as a *binomial model* emphasizing the type of error distribution of the response variable that we assume, leaving some leeway to explore other options of transformation (e.g. probit transformation) of the response variable.

variable in this model is understood as a continuous variable bounded within the limits of 0 and 1. Since our aim is to generalize the resulting models to unseen pop melodies beyond our experimental stimuli we took special care not to overfit the models on the basis of this particular dataset. We therefore randomly split the 15 tested melodies into three subsets of five melodies each. The subsets were used for the different steps in the data analysis. We performed variable selection and model fitting on the first and second subset respectively and reserved the data of the third one for evaluation of the two models. The evaluation results should, thus, reflect the model performance in a more honest way, as the melodies of the evaluation subset were not used in the model construction.

7.4 The binomial model

Many of the 38 accent rules described above were constructed in a similar way and only differ in a parameter value. As mentioned before this was done to search the parameter space within which accent rules reflecting perceptual principles can be implemented. As a result variable selection is a necessary and important first stage in the data analysis. The variable selection was carried out in three consecutive stages: variable clustering, single variable selection from clusters, and variable selection within model fitting. The first two stages were run on data subset 1 ($n = 117$ note events) while the second one was done in combination with the parameter estimation for the predictors on subset 2 ($n = 118$ note events).

Instead of relying on the heuristic preconceptions about how rules should be grouped that were rather based on their algorithmic construction principles, variable clustering provides a way to check empirically the closeness of the relationship that exists between variables given our set of pop melodies. We first computed the pair-wise distances between all 36 rules, excluding the non-binary **thom** rule and the **sync16** rule which did not indicate any accent for the melodic data of subset 1. We chose the widely used *Jaccard distance* (Jaccard, 1901) as a distance measure. The Jaccard distance considers only those cases (notes) where either of the two accent rules to be compared predicts an accent. This is a desired property of the distance measure given that the distribution of accents versus non-accented notes is heavily biased towards the non-accent category for almost all accent rules. We then used hierarchical agglomerative complete linkage clustering (see e.g. Everitt, 1974) to construct the tree of clusters that is visualized in Figure 5 in a so-called cluster dendrogram.

It is easy to see that the cluster tree is closely related to our initial heuristic grouping of accent rules. For

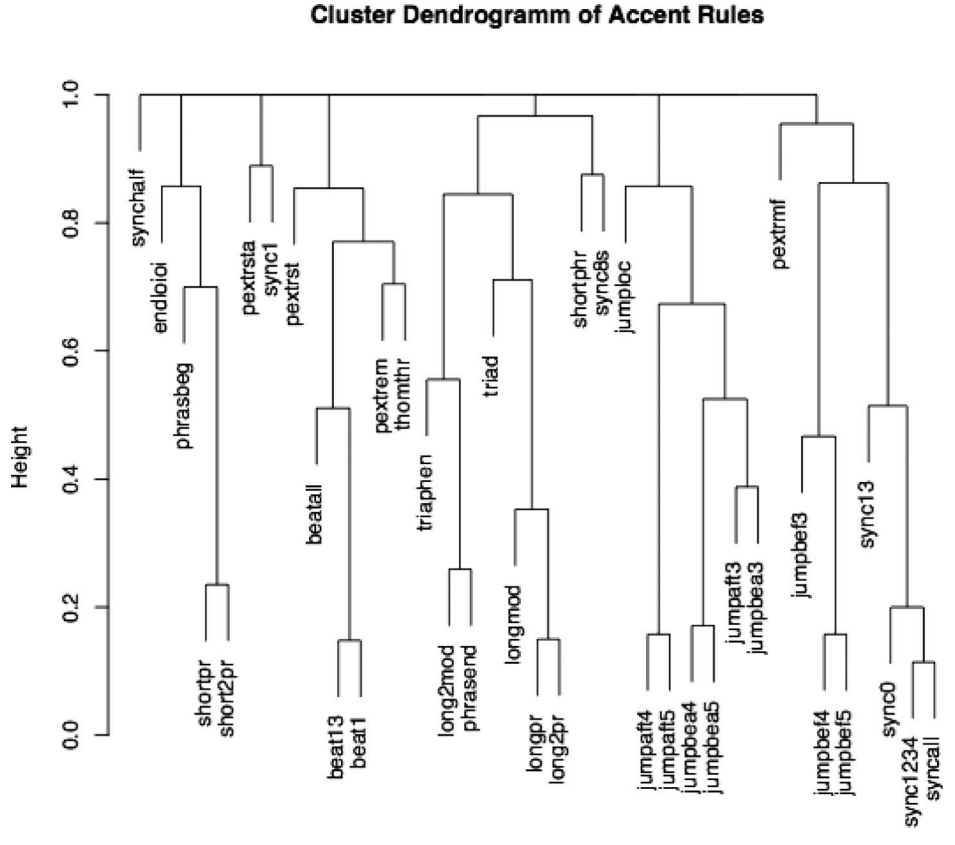

Fig. 5. Dendrogram of agglomerative complete linkage clustering for 37 binary accent rules on melodies of data subset 1.

example all rules that indicate accents after interval jumps of varying size are grouped together very early into one cluster. Similarly other rule sets that differed in a parameter value like the rules giving accents to longer notes and the rules giving accents on metrical beats also form close and early clusters. But also some rules that were previously not conceived as similar reveal correlations according to this clustering solution. As an example, the cluster formed by **long2mod**, **phrasend**, and later **triadphen** can be explained by frequent coincidences of very long notes coming from the triad of the underlying chord on phrase endings.

The cluster tree does not imply a single clustering solution with a fixed set of numbers but encompasses all possible clusterings of a dataset. A decision has to be made by the researcher where to cut the tree or, looking at it from a different perspective, how many clusters to choose. At this first stage of the variable selection process we intended to be fairly inclusive, i.e. to aim for a high number of clusters, given that there are two further stages of variable selection. The cut of the tree was based on the knowledge about which aspects of the melodic data the rules reflected and on the increase of the cluster height, i.e. the between-cluster distance according to the complete linkage criterion. The increase in cluster height is depicted in Figure 6.

A sharp rise in cluster height is clearly visible between 21 and 20 clusters (reading from right to left). At this stage, the **thomthr** rule is still separated from **pextrem** and **phrasebeg** and has not yet joined the small cluster of rules assigning accents to relatively short notes (**shortpr** and **short2pr**). We subsequently selected the best-performing rule from each of the 21 clusters. This was carried out by determining the correlation between the binary predictions of each rule and the proportion of participants that indicated an accent at each note from the melodies of subset 1. Previous studies used the Pearson correlation coefficient to quantify the relation between a binary

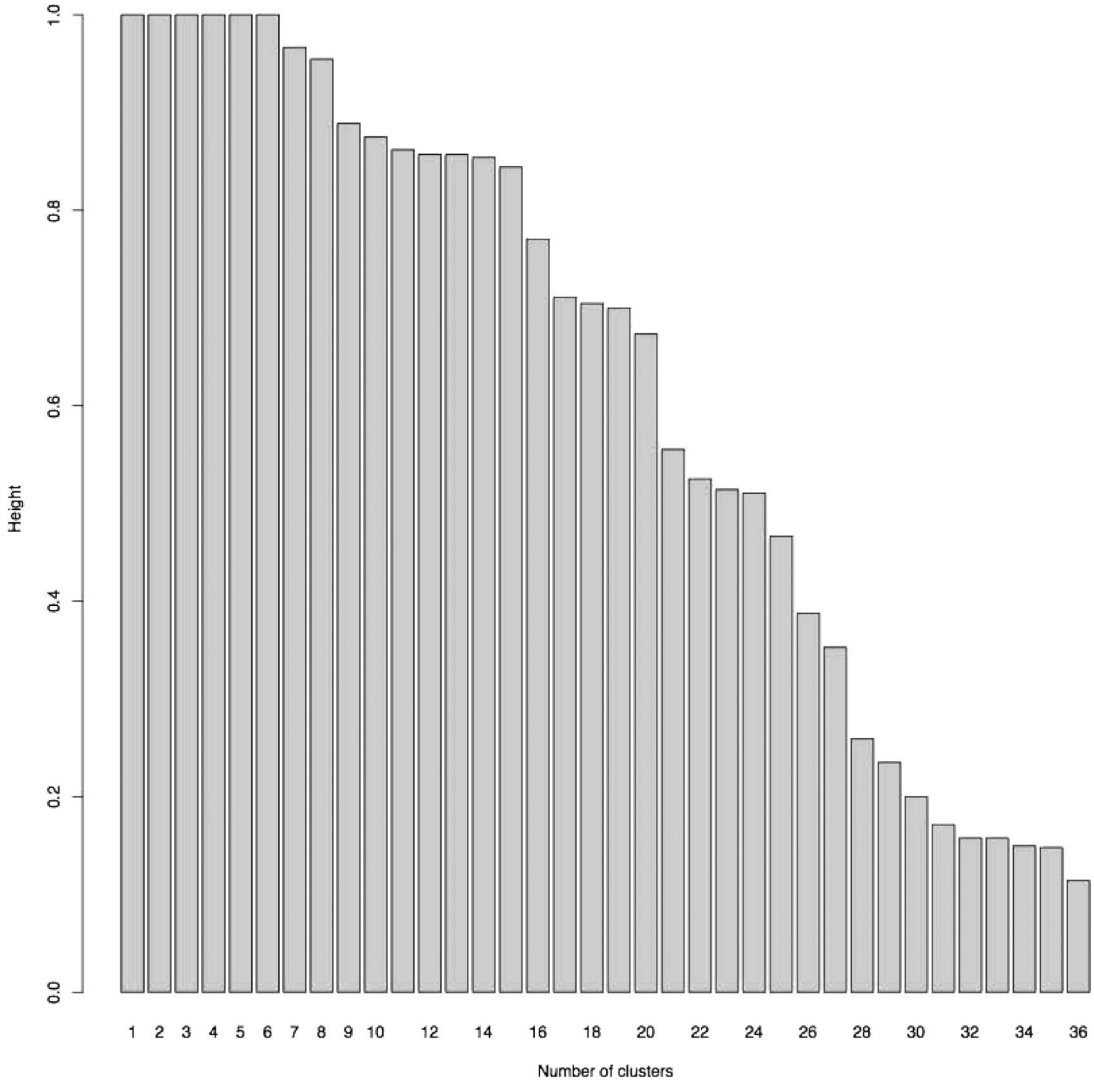

Fig. 6. Decrease in overall cluster height by increasing number of clusters of accent rules.

variable and a variable reflecting proportions. But this coefficient is based on the deviation of the individual values from the mean and we argue that the mean is not a very meaningful concept for vectors of binary values and with strongly unbalanced categories (on average, all accent rules give an accent on every fifth note, and the mean is about 0.2). In this situation, false negative predictions by an accent rule (i.e. accent rule indicating 0 on specific note and more than average participants indicate 1) are much less strongly penalized than false positive predictions where the difference between the mean (0.2) and the binary prediction (1) is greater. As an alternative to Pearson correlation for relating the binary predictions of the accent rules to the interval-scaled values of count variable we chose the *area under the receiver operating characteristic curve* (AUC). The AUC is a well-known measure from signal detection theory (e.g. Swets, 1973) and a widely used evaluation metric rating scale paradigms. In theory, the AUC values range from 0 to 1, however values below 0.5 are considered to indicate a negative correlation. The AUC is not equivalent to the Pearson correlation between a continuous and a binary variable and in contrast to the Pearson correlation the AUC is bias-free which means that it takes the distribution of 0s and 1s in the binary variable into account.[2] For each of the 21 accent rule clusters, we selected the one with the highest AUC value measured in comparison to the proportions data of subset 1. Table 1 shows the AUC values of all 37 binary accent rules along with their cluster memberships as well as indications whether the rule represents the best rule form its respective cluster.

These 21 selected accent rules and the non-binary **thom** rule then entered a combined model fitting and another round of variable selection on data subset 2: we fitted a full model of all 21 predictors from the binomial family to the proportions response variable. Then we excluded predictor variables in a step-wise backwards selection using the Bayes' Information Criterion (BIC) that combines an indicator of model fitness (log-likelihood) and a penalty for model complexity. In a final step we eliminated all model terms where the probability of the coefficient value did not reach a significance level of 0.05.[3]

The resulting binomial model is stated in terms of the so-called log odds ratio, i.e. the logarithm of the ratio between successes (i.e. p, number of participants indicating an accent) and failures (i.e. q, number of participants not indicating an accent) and involves the following nine predictors and corresponding linear weights:

$$\begin{aligned}\ln\left(\frac{p}{q}\right) = &-1.55 - 0.52 \cdot jumpbea4 + 0.53\ pextrmf \\ &+ 0.55 \cdot sync1 + 0.57 \cdot pextrem \\ &+ 0.68 \cdot longpr + 0.7 \cdot beat1 + 0.71 \cdot jumpaft4 \\ &+ 0.94 \cdot jumploc + 1.44 \cdot phrasend.\end{aligned}$$

Table 1. AUC values of the 37 binary accent rules indicating closeness of relation to participants' accent votes for the melodic data of subset 1. Cluster membership and an indication whether the rule is the best performing rule in its cluster are also given.

Rule name	AUC	Cluster no.	Best in Cluster
pextrsta	0.299	9	*
short2pr	0.388	12	
shortpr	0.404	12	*
sync8s	0.407	18	*
phrasbeg	0.462	20	*
triad	0.528	19	*
jumpbef3	0.551	3	
beat1234	0.573	14	
pextrmf	0.576	8	*
jumpbef5	0.583	3	
jumpbef4	0.595	3	*
pextrem	0.615	6	*
jumploc	0.631	4	*
jumpbea3	0.633	1	
pextrst	0.635	7	*
jumpbea5	0.641	1	
jumpaft3	0.642	1	
endloioi	0.658	13	*
syncall	0.673	15	
sync13	0.675	15	
jumpbea4	0.682	1	*
sync1234	0.691	15	
jumpaft5	0.699	2	
sync0	0.705	15	*
jumpaft4	0.736	2	*
sync1	0.743	16	*
shortphr	0.784	5	*
longmod	0.798	10	
long2pr	0.8	10	
beat13	0.809	14	
beat1	0.809	14	*
thomthr	0.812	21	*
triaphen	0.819	11	
longpr	0.853	10	*
long2mod	0.86	11	
phrasend	0.867	11	*
synchalf	0.897	17	*

[2]In fact, a rank correlation between the Pearson values and the AUC values comparing all accent rules and the proportion variable for the data of subset 1 yields a correlation value of 0.85.

[3]The parameter estimation was done using an approximation to the maximum likelihood criterion as obtained by the *iteratively weighted least squares* (IWLS) procedure (e.g. Venables & Ripley, 2002, p. 185). We used the functions `glm`, `stepAIC`, and `update` as implemented in the statistical software R.

The model has a relatively high residual deviance of 293 on 108 degrees of freedom. But except for **jumbea4**—which acts as a modifier on the effect of **jumaft4**—all model terms have a positive coefficient value. The largest contribution comes from **phrasend** while the local jump rule (**jumploc**), the metrical rule **beat1**, and **longpr**, which accents relatively long notes, also play an important role. Taken together, the model combines rules from almost all rule families that we used for the initial categorization (rules from the categories of pitch intervals, pitch contour, inter-onset intervals, metre and syncopation, and phrase position) and we therefore consider it a comprehensive model.

A simpler model with just four rules was constructed from the original audio data. The model formula including the relative weights of the individual rules reads as follows:

$$\ln\left(\frac{p}{q}\right) = -1.11 + 0.32 \cdot jumpaft4 + 1.25 \cdot sync8s + 1.6 \cdot sync1 + 2.02 \cdot beat1.$$

The model features the metrical rules **beat1**, **sync1**, and **sync8s** strongly and gives some weight also to notes after large pitch jumps (**jumpaft4**). It appears to be much more a rhythmical model than the binomial model constructed from the MIDI data. The residual deviance is 640 over 113 degrees of freedom, which indicates a worse model fit.

7.5 The tree model

As an alternative model to combine the predictions from the various accent rules we used a regression tree model. In contrast to the additive nature of regression models, tree models rather model the interaction between predictor variables. The modelling approach is in this sense complementary to the binomial model described above. Although regression and classification trees have been a standard technique in machine learning for almost 30 years they are rarely used in modelling data from music perception experiments (exceptions are Müllensiefen, 2004; Kopiez et al., 2006; Müllensiefen & Hennig, 2006). The idea behind regression trees is to partition cases into a few categories of the dependent variable by a set of independent variables. The process of partitioning is hierarchical and recursive such that the resulting output lends itself very easily to be displayed in a graphical tree structure. Amongst others, the built-in mechanism of variable selection and the ability to deal with missing cases by so-called surrogate variables are the most convenient features of tree models. We used the CART algorithm as proposed by Breiman et al. (1984) and as implemented in the *rpart* package of the statistical software environment R.

As dependent variable we used again the proportion of participants indicating a note as accented and we included all 38 accent rules in the set of potential predictors. We then constructed the regression tree model using the default parameters as given in the R implementation with the exception of the complexity parameter (cp) that defines the pruning of the regression tree to a certain level. We chose cp according to Therneau's and Atkinson's (1997) rule of thumb that looks for a minimal sized tree within one standard deviation of the lowest cross-validation error. As variable selection and parameter estimation is performed as part of the same process in the construction and pruning of the tree model we used data subsets 1 and 2 ($n = 235$ note events) to arrive at the final tree model. The model is shown graphically in Figure 7.

The tree graph reads as follows: if the condition at the node is satisfied one descends to the left otherwise one goes to the right. For example, at the first node the model splits at notes according to whether they are longer than the previous note (**longpr**). Shorter notes then get split into notes syncopated at any level and non-syncopated notes (**sync1234**). The non-syncopated notes mark a terminal node at this branch and are associated with the low value of the response variable of 0.211. The syncopated notes are eventually split into those that are part of a large interval (**jumpbea5**) and those that aren't. The former terminal node receives a higher value of 0.537 while the latter one gets a lower value of 0.321. Relatively long notes on the other hand get subsequently split according to their Thomassen contour value, whether they fall on any beat in the bar, and according to whether they constitute a phrase end or not. The highest proportion value receives those notes that are longer than their predecessor note and have a Thomassen contour value of at least 0.108 and have their onset coinciding with a beat in the bar. This model could be considered simpler than the binomial model in that it only makes use of five different accent rules from four different conceptual categories. It explains 57.4% of the variance in the data (equivalent to the R^2 value of a linear regression model) and a cross-validation error of 58%. It is interesting to note that, like the binomial model, it also features **longpr** and **phrasend** as well a rule associated with jumps (**jumpbea5**) from the pitch interval category.

For the original audio data we obtain a tree model that is similar in some aspects (see Figure 8): **longpr** is the rule at the first node by which the cross-validation error is reduced most. It is followed on either branch of this tree by rules giving weights to the strong beats of the bar (**beat1**, **beat13**). The model contains also the rules **phrasend** and **sync13** which corresponds to the MIDI tree model. The differences to the MIDI model consist mainly in the exclusive use of the rule that uses metrical

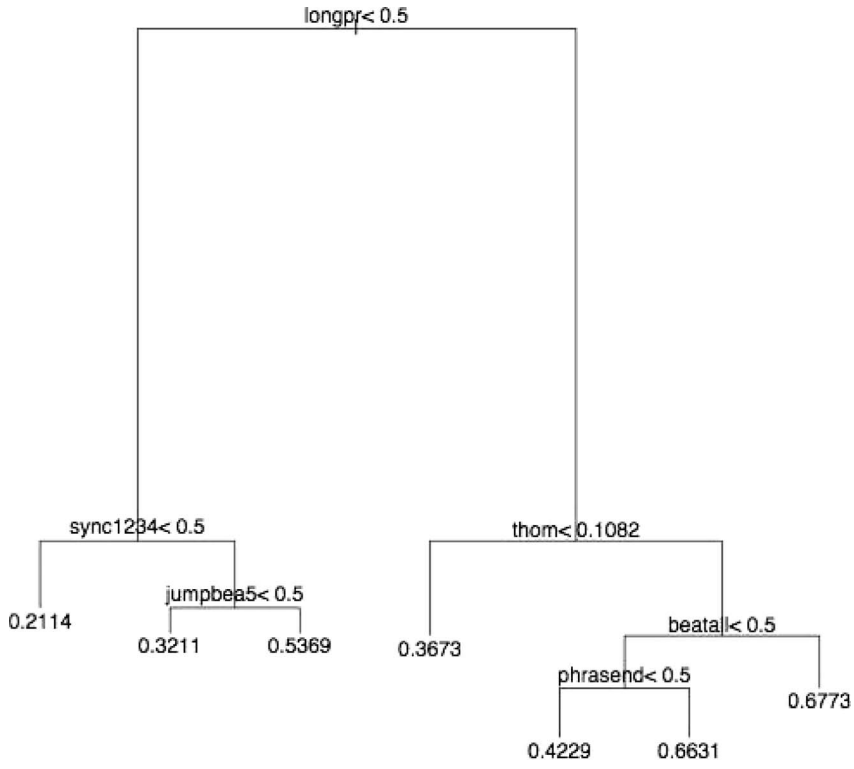

Fig. 7. Regression tree model derived from accent ratings of MIDI melodies.

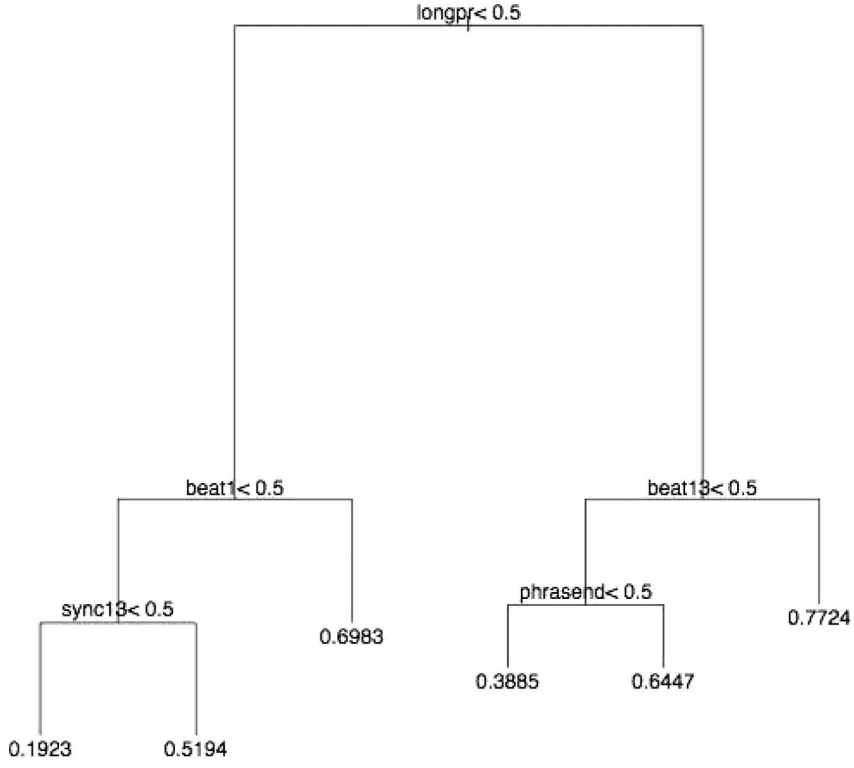

Fig. 8. Regression tree model derived from accent ratings of original audio melodies.

or note duration information. None of the pitch contour or pitch interval rules has any significance in this tree model for the audio data. The model fit is slightly higher compared to the MIDI model featuring an R^2 value of 61.9 and a cross-validation error of 43.8%.

7.6 Model evaluation

To compare the two different models and to obtain an honest estimate of their predictive power on unbiased data, we tested them on the third data subset reserved for evaluation that comprises 147 note events. The proportion of participants that would assign an accent value was predicted for each note from the model formulae of the binomial (after converting back from log-odd ratios to proportion values) and the tree model. As evaluation measures we computed (a) the Pearson correlation and (b) the mean accuracy which is defined as 1—the difference between the model's predictions and the actual proportional value from the empirical data over all notes in the evaluation set:

$$a = 1 - \frac{\sum_{i=1}^{N} |\hat{y}_i - y_i|}{N}.$$

To get a feel for the absolute degree of performance of the models we also included a trivial model in the comparison. The trivial model simply predicts for each note the mean proportional value of the subsets 1 and 2. Table 2 summarizes the results of the evaluation performance for both midi and audio data.

The best performance is achieved by the binomial model on the MIDI data. On average the binomial model is accurate by 87% in its predictions of the number of participants that will classify a note and accent. This is approximately equivalent to a deviation of 4 (3.73) out of the 29 participants. The tree model performs almost as well on the MIDI data (86%), and in fact when rounded, its average deviation is also 4 (4.16) participants. We tested the significance of this difference in model performance with a two-sided t-test corrected for unequal variances. The t-test did not indicate a significant difference between the two models ($t(289) = -1.17$, $p = 0.24$) but both models proved to have a significantly different performance from the trivial model (binomial versus trivial: $t(277) = -4.28$, $p < 0.01$; trivial versus tree: $t(287) = 3.06$, $p < 0.01$) which has an average deviation of 5 (5.42) participants.

Table 2. Evaluation results for binomial and tree models derived from MIDI and audio data.

Model	Mean accuracy	Correlation
MIDI binomial model	0.87	0.7
MIDI tree model	0.86	0.62
MIDI trivial model	0.81	0
Audio binomial model	0.82	0.57
Audio tree model	0.84	0.69
Audio trivial model	0.76	0

Note: Since Pearson correlation is not defined between a variable and a constant value we computed the covariance instead only for the trivial models.

A similar result comes from the models constructed from and tested on the audio data. Here, the tree model shows a superior performance (mean accuracy 84%, meaning an average deviation of 5 participants). But the superiority over the binomial model turns out to be insignificant ($t(291) = 1.4$, $p = 0.16$) while the difference between each model and the trivial model is highly significant (binomial versus trivial: $t(291) = -3.69$, $p < 0.01$; trivial versus tree: $t(291.91) = 5.25$, $p < 0.01$) with the trivial model exhibiting a high average deviation of 7 (6.95) participants.

With regard to the audio example it cannot be decided from this data, of course, to which degree the perceived accents were induced by melodic structure alone that is modelled by the accent model and to which degree accents were evoked from the interplay between melody and events in the accompaniment. All the evaluation figures tell us is that when we apply a model working on purely melodic information we obtain a prediction accuracy that is worse than if the model is tested against listening data from monophonic melodies but still at an acceptable level and still significantly better than a trivial model. Epistemologically, it cannot be claimed that this model gives a comprehensive account of melodic accent perception in listening to full pop songs. But nonetheless, this evaluation indicates to which degree a purely melodic model might be useful in applications dealing with full songs.

A look at Figures 9 and 10 that compare empirical and model predicted proportional accent values for two

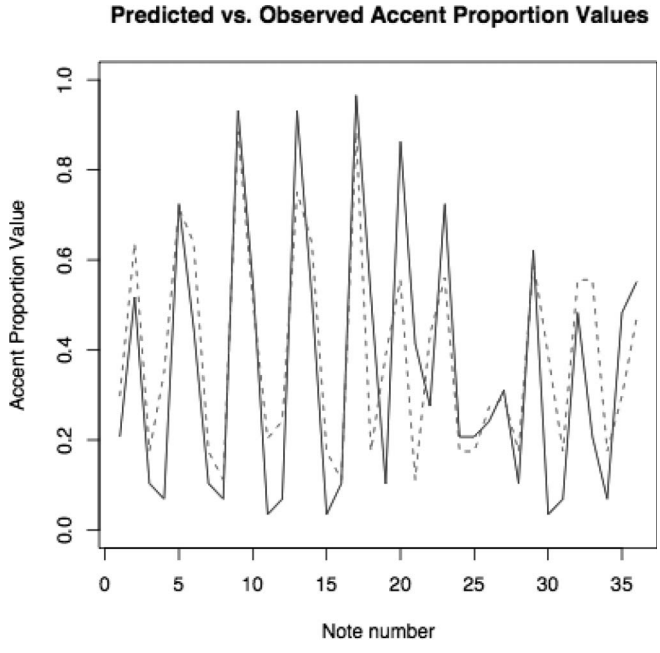

Fig. 9. Observed (solid line) and predicted (dashed line) proportional accent values for MIDI version of melody no. 15 'Mas in Madison Square Gardens' from binomial model.

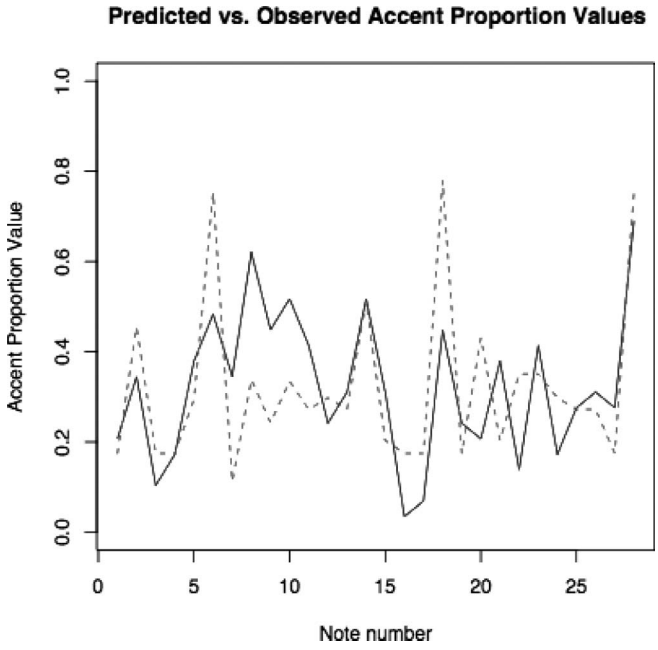

Fig. 10. Observed (solid line) and predicted (dashed line) proportional accent values for MIDI version of melody no. 4 'Take good care of my baby' from binomial model.

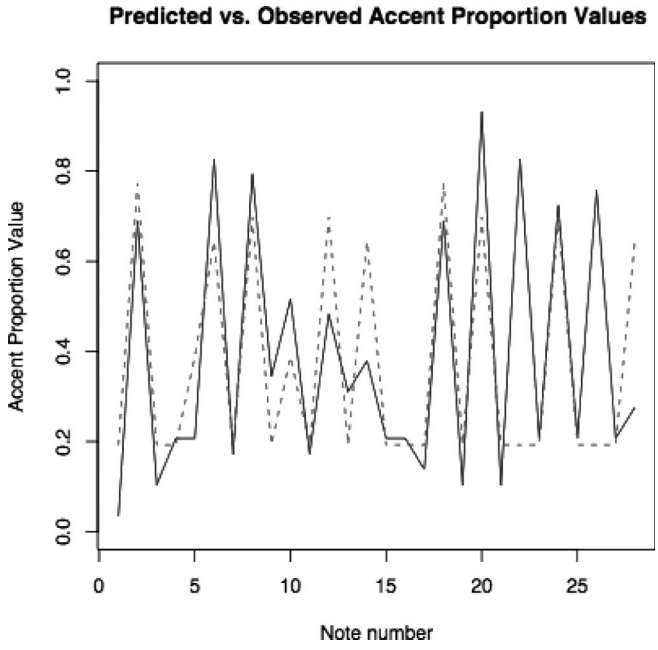

Fig. 11. Observed (solid line) and predicted (dashed line) proportional accent values for audio version of melody no. 4 'Take good care of my baby' from tree model.

example melodies from the data test set illuminates the quality of the predictions a little further.

For melody no. 15 we see an almost perfect fit (apart from a difference in scale on some notes), which indicates clearly that the model follows closely the accent structure as derived from the experimental data. In contrast, the predictions for melody no. 4 seem to miss out on some clear perceived accents, e.g. note no. 8, 10, 21, and also erroneously predict a higher value for some non-accented notes, e.g. note no. 6, 17, 20. But still parts of the accent structure seem to be captured even in this example melody where the binomial model performs rather badly. Also, we find good and bad examples of accent predictions by the tree model for the audio data shown in Figures 11 and 12.

The audio version of melody no. 4 is predicted rather well by the tree model despite a few missed accents, e.g. notes 22, and 26. However, the tree predictions for the audio version of melody no. 7 has comparatively more misses, notes 10, 12, and 14, and false alarms, e.g. notes 7 and 16.

Taken together, the overall performance of the two models on the test set and the qualitative look at badly and well-predicted melodies yields a satisfying result. Both models do very well on some of the test melodies while for the melodies where they do less well there is no obvious trend indicating systematic failure, i.e. misses and false alarms have an equal occurrence and even with the difficult melodies the models pick up parts of the accent structure correctly.

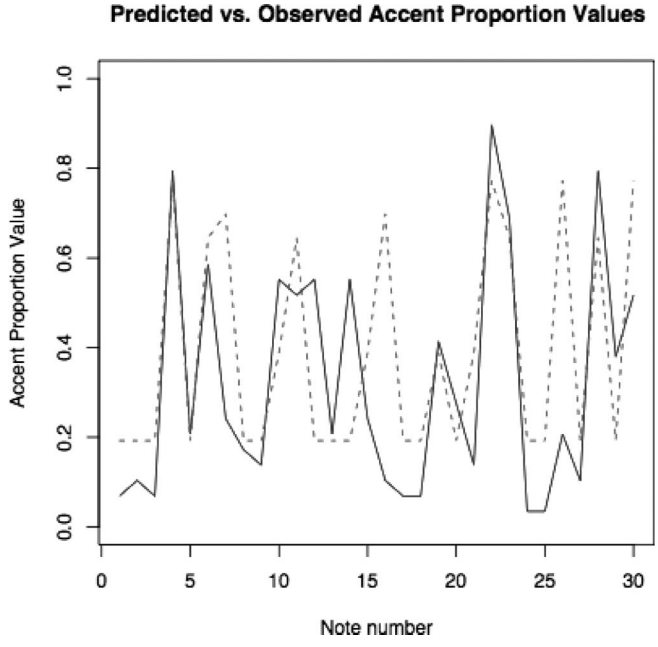

Fig. 12. Observed (solid line) and predicted (dashed line) proportional accent values for audio version of melody no. 7 'Do you want to dance' from tree model.

7.7 Predicting binary responses

Depending on the use or application of a melodic accent perception model it can be desirable to predict only the binary response whether or not a note is

perceived as having a melodic accent. A naive approach would be to consider all notes that get an accent voting denotes at least 50% of the participants to be true accents in a binary scheme. Unfortunately, a fixed threshold like 50% implicitly assumes a certain level of agreement between participants but if the participants' votes are rather noisy or follow multiple but diverging strategies for accent a particular melody then reaching a 50% agreement on a particular note is rather unlikely. Figure 10 (MIDI data for melody no. 7) is a good example of a melody where votes are split between participants. Out of the 28 notes of this melody there are only two where the accent votes clearly sum up to more than 50%, and there are about four notes with proportional accent values that hover around 50%. A fixed threshold of 50% seems to be quite arbitrary in this case. And in fact, the participants marked 7.84 accents on average on this melody (SD = 2.8, range: 3 to 12). A binary accent solution for this melody using a fixed threshold of 50% would therefore result in an accent profile that has too few accents, already for the participants' data but even more so for the model predictions which contain only four notes with proportional accent values of $>50\%$.

Therefore, instead of cutting off the accent profiles at a fixed threshold and in order to avoid under- or over-accented melodies we are aiming at binary accent solutions that have a reasonable number of accented notes per melody.

For the two data subset we used for training each participant marked an average of 8.02 accents per melody with melodies having a length of 15 to 31 notes. Or to express this differently, over the ten melodies in the training datasets, the rate of notes per accented note has a mean of 2.9 for the MIDI data. This means that approximately every third note has an accent. In fact, the variation of this rate between melodies seems rather small as evidenced by its standard deviation of 0.46. For the audio data the average is 11.67 accented notes per melody, which results in an accent rate of 2.2 notes per accent (SD = 0.45). In contrast, when we look at the average number of accents between participants we obtain the same mean (8.02) for the MIDI data but with a much higher standard deviation of 2.68. The range is marked by a low 3.5 accents per melody assigned by participant number 28 and a high 12.5 accents per melody as given by participant number 18. Given these different susceptibilities or thresholds to accent perception, which we cannot address in this study due to space limitations we stick to the mean accent rates of about 2.9 (MIDI) and 2.2 (audio) for determining how many notes of an unseen melody should be accented.

To measure the relation between the proportional accent value and the binary model predictions we use again the AUC measure. Table 3 summarizes the results of the evaluation of the binary measures.

A *t*-test for the difference in the mean AUC values did not reach significance for the MIDI data ($t(7) = 1.48$, $p = 0.18$) nor for the audio data ($t(6) = -1.47$, $p = 0.19$). The mean AUC values of 0.89 and 0.82 reached for the MIDI models are comparatively good and indicate a very satisfying model performance. AUC values of between 0.8 and 0.9 are considered indicating 'excellent classification' according to Hosmer and Lemeshow (2000). Still for the audio models we obtain 'excellent' to 'acceptable' classification results with AUC values above of 0.8 and 0.71.

This good performance is evidenced by Figures 13 and 14, which show the proportional accent values against the binary accent prediction by the binomial and the tree model for two example melodies with a model performance near the average.

Table 3. Evaluation results for binomial and tree models making binary predictions for MIDI and audio data.

Model	Mean AUC	Std. of mean difference
MIDI binomial model	0.89	0.06
MIDI tree model	0.82	0.09
Audio binomial model	0.71	0.12
Audio tree model	0.8	0.06

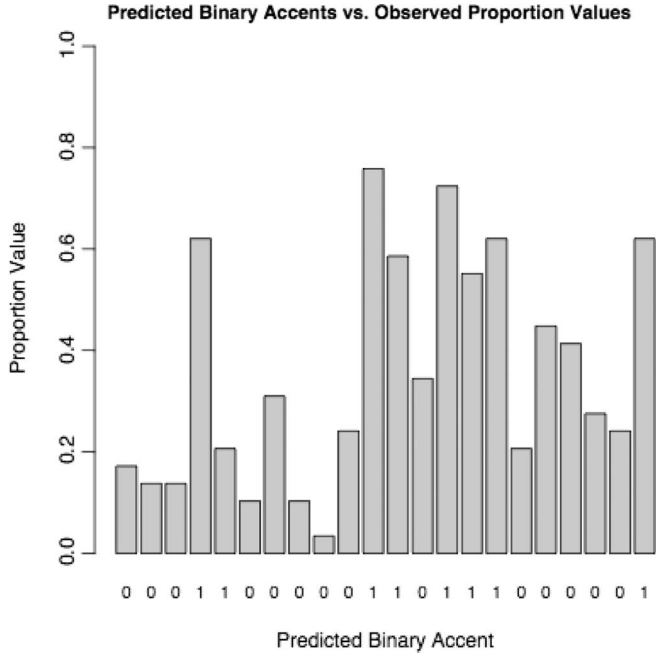

Fig. 13. Observed proportional accent values and predicted binary accents for MIDI version of melody no. 10 'Let me be your only one' from binomial model.

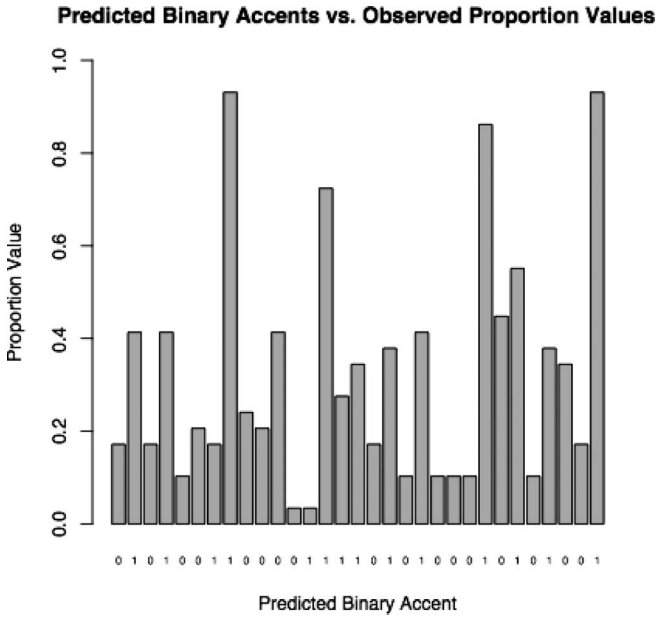

Fig. 14. Observed proportional accent values and predicted binary accents for Audio version of melody no. 8 'Climb up' from tree model.

8. Discussion

We presented a new approach and framework for investigating the perception of melodic accents. This approach consists of (a) defining a large set of possible accent rules in algorithmic terms, (b) selecting a smaller subset of rules from the initial set and modelling the sum of the accents votes of the participants by prediction models, and (c) evaluating the model predictions on an unseen dataset that had not been used for variable selection or modelling. We tested two different models, a binomial model that combines the accent rules as linear predictors and a tree model, which combines accent rules in a hierarchical way. The resulting models show an acceptable to excellent performance on the data test set for both predicting proportions and binary accent decisions. Of course, both models will need further testing for robustness and the parameter estimates should rather be viewed as indicating means of confidence intervals in that region. However, there are a number of results which we are confident about already from the results with the present data and which are, thus, worth summarizing.

- Both models perform better on the MIDI data than on the audio data. This is easily explained by the fact that the original version of the melodies naturally contain much more information that might be relevant for melodic accent perception and that is not captured by the 38 melodic accent rules, e.g. the rhythmic and harmonic accompaniment, timbral information, or cues in the lyrics or the singer's phrasing.
- The models derived from the audio data feature much more prominently rules related to temporal information, i.e. inter-onset intervals and metrical position.
- The models derived from MIDI data feature rules from a variety of different categories, i.e. reflecting different aspects of the melodic data.
- The rules incorporated in the binomial and the tree model exhibit a certain overlap, or at least reflect similar aspects of the melodic data. This seems to apply more to the MIDI than to the audio data.
- There are no significant performance differences between the binomial and the tree models for both data sets and although there is a tendency for the binomial model to perform better on the MIDI data but worse on the audio data one could choose either model for application purposes in light of the present data.
- The rates of number of notes per accent are relatively constant between melodies when averaged over participants and have means of 2.9 (MIDI) and 2.2 (audio). They are less constant when compared between participants and averaged over melodies. Thus, different listeners seem to have different sensitivities to accent cues in the melodies but these differences seem to level out when looking at the entire sample.
- The performance of both models on the unseen MIDI data is indeed very satisfying.

This holds true for the model predictions in the form of the original response variable (proportions of accent votes) as well as for the reduction to binary accent predictions.

Looking at the commonalities of the models derived from the MIDI data, it has to be emphasized that both models highlight Gestalt-based rules for accent perception concerning pitch interval, pitch contour and inter-onset intervals as well as metre and syncopation accent rules. In conclusion, it is safe to state from the present results that human perception of melodic accents uses information about several different aspects of the melodic data. Therefore, any model that is limited to only one or a few melodic dimensions can only be considered to account partially for listeners' perception of melodic accents. Due to the different nature of the binomial and the tree model the relative importance (as measured by the beta weights in the binomial model and the decrease of the relative error in the tree model) of the individual rules differs between the two models. The largest decrease in the tree model is obtained by splitting notes according to their relative inter-onset interval length (**longpr**). In contrast, **longpr** has a beta-weight of only medium magnitude when compared to the parameter values of the other rules in the binomial model.

Therefore, it seems that despite their comparable performance on the evaluation dataset the two models are similar in their choice of accent rules but not reducible to each other. We view them as two different but equally valid ways to model the empirical data from this experiment.

Both models derived from the audio data employ a clearly different selection of accent rules. This indicates that models using only rules based on melodic structure and external metre information are less adequate to describe melodic accent perception when the melody is presented in its full musical context. Both audio models make heavy use of metrical and syncopation rules, where accents on beats seem to be balanced or even emphasized by syncopations. These findings are consistent with the theory of a metric hierarchy of accentuated notes in each measure along with modifications for pop and rock music regarding anticipating syncopations which can be a highly effective aesthetic device when employed in combination with a strong metrical grid (see Temperley 1999, 2001; Pfleiderer, 2006).

Employing a staged variable selection technique for the binomial model involving hierarchical clustering and step-wise variable elimination we arrived at compact models for the two datasets. The models comprise respectively nine and four rules out of the initial set of 38 rules. A qualitative look at the selected rules showed that they contribute largely complementary information to the resulting model. Searching the solution space in terms of rules and parameters as being one of the motivations of this study has, thus, been successful. The largely differing AUC values for the various rules pertaining to the same pre-defined category make it very obvious that certain formulations of a high-level influential factor are more appropriate than others (see Table 1). For example, rules predicting accents only before a large pitch jump perform very badly compared to both, rules predicting accents after and before a jump and also rules predicting an accent only after a jump. Regarding pitch contour accentuation we found that accenting every change in melodic direction (**pexstra**) is a very bad accent predictor. Filtering for trivial contour changes is certainly needed to enhance the predictive power of the rules from this category. However, the threshholded version of the Thomassen-algorithm which takes three-note patterns into account gives still a far better performance. Phrase endings and inter-onset intervals in relation to the predecessor inter-onset interval or to the mode of inter-onset intervals in a melody are very important for accent perception, whereas phrase beginnings and short notes do not seem to correspond to accent perception at all. For metre rules only beat one and beat three of a 4/4 bar are perceived as being strongly accented. These findings might feed into a follow-up study where accent rules with a consistently bad performance can be eliminated from the initial rule set making variable selection easier and quicker.

Taken together, the results reported in this study and summarized in the previous paragraphs can be viewed as novel contributions to research the topic of melodic accent perception. While aiming at the general applicability of rule-based models, we introduced precision in the formulation of the individual rules as well as for the method of rule combination. As a result we propose sufficiently specified models that we claim are adequate at least for the prediction of perceived accents in monophonic Western pop melodies when objective metre information is provided by an external source. The models are ready for testing and we would like to invite researchers in the area to investigate their performance and usefulness with different melody repertoires and in different experimental designs or application environments.

9. Outlook

The design of this study was partly set out as a statistical modelling paradigm to handle a typical data mining situation where we have a large set of potential predictor variables, a limited number of observations, noisy empirical data, and no firm hypothesis of the type of model that would fit the data best. Given the recent growth in techniques and algorithms developed for similar data analysis problems, we could naturally only explore a tiny fraction of the conceptually applicable options that have implemented versions available. A list of options to follow up in a subsequent study would include the following ideas.

It is very straightforward within the present framework to apply different transformations to the response variable in the binomial model (e.g. probit transformation) and to employ different parameter optimization techniques for the maximum likelihood estimation (Newton–Raphson type algorithms, the Nelder–Mead method or *simulated annealing*). This may result in better prediction results from the binomial model. Similarly, other state-of-the-art data mining techniques like *boosting* or *random forests* would probably be well suited to explore the possibly non-linear relation between predictors and the response variable as well as the non-additive relation among the predictors themselves. However, one has to keep in mind that these more sophisticated data mining models might gain some additional predictive power but at the same time are usually more difficult to understand and interpret than the very clear model structures that result from a general linear model or a tree model.

Apart from the statistical power of these techniques that might lead to better prediction a qualitative inspection of the types of errors that both models make

would be a useful exercise. Given that most rules are based on Gestalt laws and have simple algorithmic mechanisms, the analytical eye should be able to spot missing rules in the models or musical situations that the models in the present state can't accommodate. A qualitative analysis such as this could then lead to a formulation of additional rules that cover musical aspects that have not be been catered for so far. A quick qualitative look at model predictions and scores of the examples mentioned above suggests that at least rules for identifying global contour extrema, parallelism, and suspension notes could be a useful addition. For modelling accent perception from audio data complementary rules are certainly needed as well, since the performance of the present models indicates that they remain incomplete. But extracting relevant predictors (e.g. harmony, metre, phrasing of melody line) automatically from the raw audio data with a precision comparable to the Gestalt rules presented here remains a very difficult audio engineering task.

A sensible extension of the present approach would certainly be the introduction of the concept of time-dependency within a sequence of accents. So far, we have modelled accents only on the basis of the musical events happening on a particular note or within a very limited time-window (2 notes for the jump-before-and-after rules and 3 notes for the Thomassen rule). The tacit assumption of most rules is that subsequent notes are independent in their accent values. But repetitions, metrical parallelism, and the accent structure of the immediately preceding notes could all very well be valuable sources of information that our human participants used in making accent judgements and that would need to be modelled as well. Time-series analysis and Markov chains are two models that would allow for the inclusion of the notion of dependency between subsequent notes with regard to accentedness.

Just as Markov chains introduce dependency between subsequent events so could general global characteristics be introduced on which local accent rules are dependent. For example the complexity, the syncopation level or the style/genre of a piece or a monophonic melody could be global predictors that may have strong influences on how local Gestalt rules trigger accent perceptions.

Another tacit assumption with regard to the data analysis was to assume that there is a somehow 'correct' or 'true' accent strength associated with each note and that all participants respond to this true accent strength with different internal thresholds. These hypothesized different internal thresholds would explain why some participants judge a note to be accented and others don't. The binomial model assumes that by summing up the binary decisions we obtain a measurement that reflects the true accent strength. Contrary to this assumption, it might be perfectly possible that different participants perceive the accent structure of a melody very differently from each other with their accent having little overlap. Then, by adding together binary accent votes from participants that differ substantially in their perceptions and judgements we would obtain meaningless averages that would not reflect any accent strength profile as perceived by one of the participants. From a qualitative look at the present data this divergence between participants certainly isn't the case for some melodies. For example, the very regular patterns and rather unanimous votes for melody no. 15 'Mas in Madison Square Garden' (Figure 9) indicate a rather large agreement between participants. In contrast, the longer passage (notes 5 to 11) of the MIDI version of melody no. 4, 'Take good' (Figure 10), could potentially be an indicator of accent votes getting split between two or more perceptive strategies. This would have to be confirmed by inter-subject agreement measures applied to each melody item individually. One way of dealing with diverging perception patterns is to cluster participants into groups with largely similar judgements, which then allows one to model each group of participants separately.

In either case, the models will certainly gain in prediction performance and robustness if they have more data to learn from. Therefore, the collection of more empirical data is necessary for the continuation of the approach proposed in this study. It would certainly make for interesting comparisons if we could train and test accent models of melodies from different styles (e.g. jazz melodies and improvised solo lines or folk songs). Similarly, it would be highly interesting to compare the here presented empirical data to experimental results where the perceived accent strength was measured using other experimental paradigms, e.g. paradigms that do not require conscious and overt accent strength judgements but make use of implicit techniques.

Above all, the models constructed and tested in this study need to be challenged by an application with melodic data outside the experimentally collected accent ratings. We expect that using accent profiles as a melodic salience function actually improves the performance in musicological and music retrieval applications, like query-by-humming applications (e.g. Unal et al., 2008), tune retrieval from large databases (e.g. Kornstädt, 1998), research on melodic variations (e.g. Hörnel, 1998) or stylistic clustering in folk-song archives (e.g. Sagrillo, 1999; Müllensiefen & Frieler, 2007).

Acknowledgements

We thank the 29 participants of the experiment, Steffen Just for entering the experimental data and Marcus Pearce for revising several drafts of this paper. Daniel Müllensiefen is supported by EPSRC grant EP/D038855/1.

References

Balkwill, L. & Thompson, W.F. (1999). A cross-cultural investigation of the perception of emotion in music: psychophysical and cultural cues. *Music Perception*, *17*, 43–64.

Boltz, M. (1991). Some structural determinants of melody recall. *Memory & Cognition*, *19*(3), 239–251.

Boltz, M. (1992). The remembering of auditory event durations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 938–956.

Boltz, M. (1993). The generation of temporal and melodic expectancies during musical listening. *Perception & Psychophysics*, *53*(6), 585–600.

Boltz, M. (1995). Effects of event structure on retrospective duration judgements. *Perception & Psychophysics*, *57*, 1080–1096.

Boltz, M. (1998). The processing of temporal and non-temporal information in the remembering of event durations and musical structure. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(4), 1087–1104.

Boltz, M. (1999). The processing of melodic and temporal information: Independent or unified dimensions? *Journal of New Music Research*, *28*(1), 67–79.

Boltz, M. & Jones, M.R. (1986). Does rule recursion make melodies easier to reproduce? If not, what does? *Cognitive Psychology*, *18*, 389–431.

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.

Cambouropoulos, E. (1998). Towards a general computational theory of musical structure. PhD thesis, University of Edinburgh, UK.

Clarke, E.F. (1999). Rhythm and timing in music. In D. Deutsch (Ed.), *The Psychology of Music* (2nd ed.). (pp. 473–500). San Diego: Academic Press.

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213–220.

Collett, D. (2003). *Modelling Binary Data* (2nd ed.). Boca Raton: Chapman & Hall/CRC.

Cooper, G. & Meyer, L.B. (1960). *The Rhythmic Structure of Music*. Chicago: University Press.

Crawley, M.J. (2007). *The R Book*. Chichester: John Wiley & Sons.

de la Motte, D. (1993). *Melodie: Ein Lese- und Arbeitsbuch*. Kassel: dtv/Bärenreiter.

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

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

Dowling, W.J. & Harwood, D.L. (1986). Melody: Attention and memory. In W.J. Dowling & D.L. Harwood (Eds.), *Music Cognition* (pp. 124–152). Orlando: Academic Press.

Eck, D. (2001). A positive-evidence model for rhythmical beat induction. *Journal of New Music Research*, *30*, 187–200.

Eerola, T., Himberg, T., Toiviainen, P. & Louhivuori, J. (2006). Perceived complexity of Western and African folk melodies by Western and African listeners. *Psychology of Music*, *34*(3), 341–375.

Eiting, M.H. (1984). Perceptual similarities between musical motifs. *Music Perception*, *2*(1), 78–94.

Everitt, B. (1974). *Cluster Analysis*. London: Heinemann.

Fitch, W. & Rosenfeld, A. (2007). Perception and production of syncopated rhythms. *Music Perception*, *25*, 43–58.

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.

Hannon, E.E., Snyder, J.S., Eerola, T. & Krumhansl, C.L. (2004). The role of melodic and temporal cues in perceiving musical meter. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(5), 956–974.

Hirsh, I.J., Monahan, C.B., Grant, K.W. & Singh, P.G. (1990). Studies in auditory timing: 1. Simple patterns. *Perception & Psychophysics*, *47*(3), 215–226.

Hörnel, D. (1998). A multi-scale neural-network model for learning and reproducing chorale variations. In E. Selfridge-Field & W.B. Hewlett (Eds.), *Computing in Musicology 11* (pp. 141–157). Cambridge, MA: MIT Press.

Hosmer, D.W. & Lemeshow, S. (2000). *Applied Logistic Regression*. New York: Wiley.

Huron, D. (2006). *Sweet Anticipation: Music and the Psychology of Expectation*. Cambridge, MA: MIT Press.

Huron, D. & Royal, M. (1996). What is melodic accent? Converging evidence from musical practice. *Music Perception*, *13*(4), 489–516.

Jaccard. P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, *37*, 547–579.

Jones, M.R. (1987). Dynamic pattern structure in music: recent theory and research. *Perception & Psychophysics*, *41*, 621–634.

Jones, M.R. (1993). Dynamics of musical patterns: How do melody and rhythm fit together? In Th. Tighe & W.J. Dowling (Eds.). *Psychology and Music: The Understanding of Melody and Rhythm* (pp. 67–92). Hillsdale (NJ): Lawrence Erlbaum.

Jones, M.R. & Boltz, M. (1989). Dynamic attending and responses to time. *Psychological Review*, *96*(3), 459–491.

Jones, M.R., Moynihan, H., MacKenzie, N. & Puente, J. (2002). Temporal aspects of stimulus-driven attending in dynamic arrays. *Psychological Science*, *14*(4), 313–319.

Jones, M.R. & Pfordresher, P.Q. (1997). Tracking musical patterns using Joint Accent Structure. *Canadian Journal of Experimental Psychology*, *51*, 271–290.

Jones, M.R. & Ralston J.T. (1991). Some influences of accent structure on melody recognition. *Memory & Cognition*, *19*, 8–20.

Jones, M.R., Summerell, L. & Marshburn, E. (1987). Recognizing melodies: A dynamic interpretation. *Quaterly Journal of Experimental Psychology*, *39A*, 89–121.

Kaiser, H. (1974). An index of factorial simplicity. *Psychometrika*, *39*, 31–36.

Kopiez, R., Weihs, C., Ligges, U. & Lee, J.I. (2006). Classification of high and low achievers in a music sight reading task. *Psychology of Music*, *34*(1), 5–26.

Kornstädt, A. (1998). Themefinder: A web-based melodic search tool. In E. Selfridge-Field & W.B. Hewlett (Eds.), *Computing in Musicology 11* (pp. 231–236). Cambridge, MA: MIT Press.

Krumhansl, C. (1990). *Cognitive Foundations of Musical Pitch*. New York: Oxford University Press.

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Large, E.W. & Jones, M.R. (1999). The dynamics of attending: How people track time varying events. *Psychological Review*, *106*, 119–159.

London, J. (2004). *Hearing in Time. Psychological Aspects of Musical Meter*. Oxford: Oxford University Press.

Longuet-Higgins, H. & Lee, C. (1984). The rhythmic interpretation of monophonic music. *Music Perception*, *1*, 424–441.

Margulis, E.H. (2005). A model of melodic expectation. *Music Perception*, *22*(4), 663–714.

Monahan, C.B. (1993). Parallels between pitch and time and how they go together. In T.J. Tighe & W.J. Dowling (Eds.), *Psychology and Music: The Understanding of Melody and Rhythm* (pp. 121–154). Hillsdale, NJ: Lawrence Erlbaum.

Monahan, C.B. & Carterette, E.C. (1985). Pitch and duration as determinants of musical space. *Music Perception*, *3*(1), 1–32.

Monahan, C.B., Kendall, R.A. & Carterette, E.C. (1987). The effect of melodic and temporal contour on recognition memory for pitch change. *Perception & Psychophysics*, *41*(6), 576–600.

Müllensiefen. D. (2004). Variabilität und Konstanz von Melodien in der Erinnerung: Ein Beitrag zur musikpsychologischen Gedächtnisforschung. PhD thesis, University of Hamburg, Germany.

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

Müllensiefen, D. & Frieler, K. (2007). Modelling experts' notion of melodic similarity. *Musicae Scientiae, Discussion Forum*, *4A*, 183–210.

Müllensiefen, D. & Hennig. C. (2006). Modeling memory for melodies. In M. Siliopoulou, R. Kruse, C. Bongelt, A. Nürnberger & W. Gaul (Eds.), *From Data and Information Analysis to Knowledge Engineering* (pp. 732–739). Berlin: Springer.

Müllensiefen, D., Pearce, M., Wiggins, G. & Frieler, K. (2007). Segmenting pop melodies: A model comparison approach. Paper held at *SMPC07*, Montreal, Canada.

Neisser, U. (1976). *Cognition and Reality: Principles and Implications of Cognitive Psychology*. San Francisco: W.H. Freeman.

Parncutt, R. (1994). A perceptual model of pulse salience and metrical accent in musical rhythms. *Music Perception*, *11*, 409–464.

Piston, W. (1950a). *Counterpoint*. London: Victor Gollancz.

Piston, W. (1950b). *Harmony*. London: Victor Gollancz.

Pfleiderer, M. (2006). *Rhythmus. Psychologische, theoretische und stilanalytische Aspekte populärer Musik*. Bielefeld: Transcript.

Pfordresher, P.Q. (2003). The role of melodic and rhythmic accents in musical structure. *Music Perception*, *20*(4), 431–464.

Povel, D.J. (1981). Internal representations of simple temporal patterns. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(1), 3–18.

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

Povel, D.J. & Okkerman, H. (1981). Accents in equitone sequences. *Perception & Psychophysics*, *30*(6), 565–572.

Reed, E.S. & Jones, R. (Eds.). *Reasons for Realism. Selected Essays of James J. Gibson*. New Jersey: Lawrence Erlbaum Associates.

Sagrillo, D. (1999). *Melodiegestalten im luxemburgischen Volkslied: Zur Anwendung computergestützter Verfahren bei der Klassifikation von Volksliedabschnitten*. Bonn: Holos.

Smith, L.M. & Honing, H. (2006). Evaluating and extending computational models of rhythmic syncopation in music. In *Proceedings of the International Computer Music Conference*, New Orleans, USA, pp. 688–691.

Steinbeck, W. (1982). Kieler Schriften zur Musikwissenschaft XXV. *Struktur und Ähnlichkeit. Methoden automatisierter Melodieanalyse*. Kassel: Bärenreiter.

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

Temperley, D. (1999). Syncopation in rock. A perceptual perspective. *Popular Music*, *18*, 19–40.

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

Therneau, T.M. & Atkinson, E.J. (1997). An introduction to recursive partitioning using the RPART routines. *Technical Report Series No. 61*. Department of Health Science Research, Mayo Clinic, Rochester, Minnesota. Retrieved August 25, 2009, from http://mayoresearch.mayo.edu/mayo/research/biostat/upload/61.pdf

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

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

Venables, W.N. & Ripley, B.D. (2002). *Modern Applied Statistics with S*. New York: Springer.

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.

Yeston, M. (1976). *The Stratification of Musical Rhythm*. New Haven: Yale University Press.

Appendix. Song excerpts used as experimental stimuli.

No.	Title	Artist	Genre	Released	Source Recording
1	Children Of The Night	Richard Marx	Pop	1989	Richard Marx: Repeat Offender
2	Goodbye My Love Goodbye	Demis Roussos	Pop (Schlager)	1974	Demis Roussos: Demis Roussos
3	Longer	Dan Fogelberg	Pop ballad	1980	Dan Fogelberg: Phoenix
4	Take Good Care of My Baby	Smokie	Pop ballad	1981	Smokie: Solid Ground
5	Enjoy Your Life	Risquee	Funk/Rock	1999	Risquee: Upside Down
6	Cold Cold Heart	Wet Wet Wet	Pop	1994	Wet Wet Wet: Part One
7	Do You Want To Dance	Puhdys	RocknRoll	1977	Phudys: Rock'n'Roll Music
8	Climb Up (Stairway to heaven)	Neil Sedaka	Entertainer	1962	Neil Sedaka: Sings His Greatest Hits
9	Love Is Like A Rainbow	Modern Talking	Disco/Pop	1999	Modern Talking: Alone
10	Let Me Be Your Only One	Risquee	Funk/Rock	1999	Let Me Be Your Only One
11	Don't Stay Away	Phyllis Dillon	Reggae	1967	Tougher Than Tough. The Story of Jamaican Music
12	She	Will Downing	R'n'B	1991	Will Downing: A Dream Fulfilled
13	No No No	K. C. White	Reggae	early 1970s	Holy Ground: Alvin Ranglin's GG Records
14	Foolish	Ashanti	R'n'B	2002	Ashanti: Ashanti
15	Mas in Madison Square Garden	Lord Kitchener	Calypso	1971	Mighty Sparrow & Lord Kitchener: Carnival Hits
16	Rim Shot	Erykah Badu	R'n'B	1997	Erykah Badu: Baduism
17	Bitter Blue	Kind of Blue	Rock	2000	Kind of Blue: In Sight