Journal of New Music Research
2001, Vol. 30, No. 3, pp. ••–••

0271-3683/01/2205-367$16.00

Memory-Based Models of Melodic Analysis: Challenging the Gestalt Principles

Rens Bod

School of Computing, University of Leeds, Leeds LS2 9JT, UK, and Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, Netherlands

Abstract

We argue for a memory-based approach to music analysis which works with concrete musical experiences rather than with abstract rules or principles. New pieces of music are analyzed by combining fragments from structures of previously encountered pieces. The occurrence-frequencies of the fragments are used to determine the preferred analysis of a piece. We test some instances of this approach against a set of 1,000 manually annotated folksongs from the Essen Folksong Collection, yielding up to 85.9% phrase accuracy. A qualitative analysis of our results indicates that there are grouping phenomena that challenge the commonly accepted Gestalt principles of proximity, similarity and parallelism. These grouping phenomena can neither be explained by other musical factors, such as meter and harmony. We argue that music perception may be much more memory-based than previously assumed.

1. Introduction

In listening to a piece of music, the human perceptual system segments the sequence of notes into groups or phrases that form a grouping structure for the whole piece (cf. Longuet-Higgins, 1976; Tenney & Polansky, 1980; Lerdahl & Jackendoff, 1983; Stoffer, 1985). One of the main challenges in modeling musical segmentation is the problem of ambiguity: several different grouping structures may be compatible with a sequence of notes while a listener usually perceives only one particular structure. It is widely assumed that the preferred grouping structure of a piece depends on a combination of low-level phenomena, such as local discontinuities and intervallic distances, and high-level phenomena, such as melodic parallelism and internal harmony.

Most models of musical segmentation use the Gestalt principles of proximity and similarity (Wertheimer, 1923) to predict the low-level grouping structure of a piece: grouping boundaries preferably fall on larger inter-onset-intervals, larger pitch intervals, etc. (see Tenney & Polansky, 1980; Lerdahl & Jackendoff, 1983; Cambouropoulos, 1996, 1997). While most models also incorporate higher-level grouping phenomena, such as melodic parallelism and harmony, these phenomena remain often unformalized. For example, Lerdahl & Jackendoff (1983) do not provide any systematic description of higher-level musical parallelism, and Narmour's Implication-Realization model (Narmour, 1990, 1992) relies on factors such as meter, harmony and similarity which are not fully described by the model. As a result, these models have not been evaluated against large sets of musical data, such as the Essen Folksong Collection (Schaffrath, 1995; Huron, 1996). Only a few, hand-selected passages are typically used to evaluate these models, which questions the objectivity of the results.

The current paper investigates a rather different approach to music analysis. Instead of using a predefined set of rules or principles, we present a model which works with a corpus of grouping structures of previously encountered musical pieces. New pieces are analyzed by combining fragments from the corpus-structures; the frequencies of the fragments are used to determine the preferred analysis. We thus propose a supervised, memory-based approach to music analysis which works with concrete musical fragments rather than with abstract formalizations of intervallic distances, paral-

Accepted: 22 August, 2001

Correspondence: Rens Bod, Institute for Logic, Language and Computation, University of Amsterdam, Spuistreat 134, 1012 VB, Amsterdam, Netherlands. E-mail: rens@science.uva.nl

lelism, meter, harmony or other musical phenomena. In other fields of cognitive science, such as natural language processing and machine learning, memory-based models have become increasingly influential (cf. Mitchell, 1997; Bod, 1998; Manning & Schütze, 1999). Moreover, recent psychological investigations suggest that previously heard musical fragments are stored in memory (e.g., Saffran et al., 2000), and that fragments that are encountered more frequently are better represented in memory and consequently more easily activated than less frequently encountered fragments. The current availability of large annotated musical databases, such as the Essen Folksong Collection (Schaffrath, 1995; Huron, 1996), provides an excellent test domain for memory-based models of music analysis.

Although a purely memory-based model may not suffice as a theory of music analysis, it is important to study the merits of such a model so that its results may be used as a baseline against which other approaches can be compared. In the following we first describe the Essen Folksong Collection, after which we test three different memory-based parsing models on this collection. We will see that the best results are obtained by a model which combines two memory-based techniques: the Markov grammar technique of Collins (1999) and the Data-Oriented Parsing technique of Bod (1998). This combined model correctly predicts 85.9% of the phrases for a held-out test set of 1000 folksongs. A qualitative evaluation of our results reveals the existence of a class of patterns that are problematic for Gestalt-based/parallelism-based models, while these patterns are rather trivial for memory-based models. Our evaluation challenges two widely accepted grouping principles in music: the Gestalt principles of proximity/similarity (Wertheimer, 1923; Tenney & Polansky, 1980; Lerdahl & Jackendoff, 1983; Handel, 1989) and the higher-level principle of melodic parallelism (Lerdahl & Jackendoff, 1983; Cambouropoulos, 1998; Höthker et al., 2001). We argue that music perception may be much more memory-based than previously assumed.

2. The Essen Folksong Collection

The Essen Folksong Collection contains a large sample of (mostly) European folksongs that have been collected and encoded under the supervision of the late Dr. Helmut Schaffrath at the University of Essen (Schaffrath, 1993; 1995; Selfridge-Field, 1995). The ongoing development of the collection is now under the charge of Dr. Ewa Dahlig at the Helmut Schaffrath Laboratory of Computer Aided Research in Musicology, Warsaw. Currently, 6251 folksongs are publicly available at http://www.esac-data.org, although the total number of folksongs in the collection is reported to be over 20000. Each folksong is annotated with the Essen Associative Code (ESAC) which includes pitch and duration information, meter signatures and explicit phrase markers (the texts of the folksongs have not been entered; only their tonal representations are available). The presence of phrase markers makes the Essen Folksong Collection a unique test case for computational models of music segmentation.

The pitch encodings in the Essen Folksong Collection resemble "solfege": scale degree numbers are used to replace the movable syllables "do", "re", "mi", etc. Thus 1 corresponds to "do", 2 corresponds to "re", etc. Chromatic alterations are represented by adding either a "<#>" or a "b" after the number. The plus ("+") and minus ("–") signs are added before the number if a note falls resp. above or below the principle octave (thus –1, 1 and +1 refer al to "do", though on different octaves). Duration is represented by adding a period or an underscore after the number. A period (".") increases duration by 50% and an underscore ("_") increases duration by 100%; more than one underscore may be added after each number. If a number has no duration indicator, its duration corresponds to the smallest value (which is found in the KEY field preceding each folksong). A pause is represented by 0, possibly followed by duration indicators. No loudness or timbre indicators are used in the Essen Folksong Collection. Hard returns are used to indicate a phrase boundary (note that we use the terms "phrase" and "group" interchangeably). To make the Essen annotations readable for our memory-based parsers, we automatically converted its phrase boundary indications into bracket representations, where "('indicates the start of a phrase and')" the end of a phrase. For more information on the Essen Folksong Collection and the Essen Associative Code (ESAC), see Selfridge-Field (1995). The Essen Folksong Collection is also available in the Humdrum format (Huron, 1996).

Figure 1 gives an example of the encoding of folksong K0029 ("Schlaf Kindlein Feste") together with its phrase annotation (we leave out barlines and meter signature that will not be used by our parsers, but we will come back to metrical structure in Section 4):

Figure 1

(3_221_–5)(–533221_–5)(13335432)(13335432_)
(3_221_–5_)

Note that the Essen phrase annotations lack hierarchical structure: they neglect both phrase-internal structure, such as subphrases and motives, and phrase-external structure, such as periods and subsections (cf. Lerdahl & Jackendoff, 1983). Thus the first two phrases in folksong (1) could have been grouped together into a larger constituent, and the same holds for the two subsequent phrases. While there may in fact not be much internal structure in the phrases of folksong (1), the following annotation for folksong K0885 ("Schneckhaus Schneckhaus stecke deine Hoerner aus") shows that the lack of phrase-internal structure can lead to a rather impoverished annotation:

Figure 2

(5_3_5_3_)(1234553_)(1234553_)(12345_3_)(12345_3_)
(553_553_)(553_65432_1_)

A more fine-grained analysis of this folksong, we believe, would consist in subsegmentations of several of its phrases; for instance, the first phrase could be subsegmented into two equivalent subphrases (5_3_). Also a considerable amount of phrase-external structure could be added to this folksong, such as the addition of a larger group that includes the second and the third phrase. A more extreme case is provided by folksong Z0147 ("Besenbinders Tochter und kachelmachers Sohn"):

Figure 3

```
(5_4#_5_3_1__1_3_2_1#_2_-7_-5__.)
   (3_5_4#_5_3_1__1_3_221#_2_-7_-5__.)
(-5_-5_-5_-5-5-5_4__4_3_2_2_3_4_5__+1_)
   (3_5_4#_5_3_1_-7_1_332_1#_2_3_1__0__)
(-5_-5_-5_-5_444_4_3_2_2_3_4_5__+1_)
   (3_5_4#_5_3_1_1_1_3_2_1#_2_3_1__.)
(3_5_4#_5_3_1_1_1_3_2_1#_2_3_1__1_)
   (3_5_4#_5_3_1_-7_1_3_2_1#_2_3_1__1_0_)
(-5_-5_-5_-5_444_4_3_2_2_3_4_5__+1_)
   (3_5_4#_5_3_1_1_1_3_2_1#_2_3_1__)
```

We believe that every phrase in this folksong can be further subsegmented into subphrases. Yet, the annotation in Figure 3 is not wrong; it just represents the most basic phrase structure of the piece only. We want to emphasize that for our experiments in Section 3 we did not add (or modify) any structure in the Essen annotations. One might believe that the Essen Folksong Collection is therefore a relatively easy test case; yet it turned out to be surprisingly hard to predict the correct phrases for these folksongs.

This brings us to the problem of evaluation. To evaluate our memory-based parsing models for music, we employed the so-called *blind testing method* which has been widely used in evaluating natural language parsers (cf. Manning & Schütze, 1999). This method dictates that a collection of annotated strings is randomly divided into a training set and a test set, where the annotations in the training set are used to "train" the parser, while the *un*annotated strings in the test set are used as input to test the parser. The degree to which the predicted segmentations for the test set strings match with the correct segmentations in the test set is a measure for the accuracy of the parser. For our experiments in Section 3, we randomly divided the 6251 folksongs that are currently available into a training set of 5251 folksongs and a test set of 1000 folksongs.

There is an important question as to what kind of evaluation measure is most appropriate to compare the segmentations proposed by the parser with the correct segmentations in the test set. A widely used evaluation scheme in natural language parsing is the PARSEVAL scheme, which is based on the notions of *precision* and *recall* (see Black et al., 1991). PARSEVAL compares a proposed parse P with the corresponding test set parse T as follows:

$$\text{Precision} = \frac{\#\,\text{correct phrases in } P}{\#\,\text{phrases in } P}$$

$$\text{Recall} = \frac{\#\,\text{correct phrases in } P}{\#\,\text{phrases in } T}$$

A phrase is correct if both the start and the end of the phrase is correctly predicted. Note that these measures "punish" a parser which assigns too many phrases to a folksong: for example, an extremely overgenerating parser which assigns phrases to any combination of notes would trivially include all correct phrases, resulting in an excellent recall, but its precision would be very low. On the other hand, a very conservative parser which predicts extremely few, though correct phrases, will receive a high precision, but its recall will be low. A good parser will thus need to obtain both a high precision and a high recall. (It goes probably without saying that for computing the precision and recall for all test set strings, one needs to divide the total number of correctly predicted phrases in all proposed parses P by the total number of phrases in respectively all parses P and T.)

The precision and recall scores are often combined into a single measure of performance, known as the F-score (see Manning & Schütze, 1999):

$$\text{F-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

We will use these three measures of Precision, Recall and F-score to quantitatively evaluate our memory-based parsing models.

As a final pre-processing step, we (automatically) added to each phrase in the folksong the label "P" and to each whole song the label "S", so as to obtain conventional parse trees. Thus the structure in (1) becomes:

Figure 4

```
S(P(3_221_-5)P(-533221_-5)P(13335432)P(13335432_)
   P(3_221_-5_))
```

The advantage of this format is that we can now directly apply existing memory-based parsing models to the Essen Folksong Collection.

3. Experiments with the Essen Folksong Collection

In this section, we give a quantitative evaluation of three memory-based parsing models on the Essen Folksong Collection (we will go into a more qualitative evaluation of our results in Section 4). We consider the following memory-based parsing models from the literature: the Treebank grammar technique of Charniak (1996), the Markov grammar technique of Seneff (1992) and Collins (1999), and the Data-Oriented Parsing (DOP) technique of Bod (1993, 1998). Unless stated differently, we used the same random split of the Essen Folksong Collection

into a training set of 5,251 folksongs and a test set of 1000 folksongs.

3.1 The Treebank grammar technique

The Treebank grammar technique is an extremely simple learning technique: it reads *all* context-free rewrite rules from the training set structures, and assigns each rule a probability proportional to its frequency in the training set. For example, the following context-free rules can be extracted from the structure in Figure 4:

S –> PPPPP
P –> 3_221_–5
P –> –533221_–5
P –> 13335432
P –> 13335432_
P –> 3_221_–5_

Next, each rewrite rule is assigned a probability by dividing the number of occurrences of a particular rule in the training set by the total number of occurrences of rules that expand the same nonterminal as the particular rule. For instance, if we take folksong (4) as our only training data, then the probability of the rule P –> 3_221_–5 is equal to 1/5 since this rule occurs once among a total of 5 rules that expand the nonterminal P.

A Treebank grammar extracted in this way from the training set corresponds to a so-called Probabilistic Context-Free Grammar or PCFG (Booth, 1969). A crucial assumption underlying PCFGs is that the context-free rules are statistically independent. Thus, given the probabilities of the individual rules, we can calculate the probability of a parse tree by taking the product of the probabilities of each rule used therein. PCFGs have been extensively studied in the literature (cf. Wetherell, 1980; Charniak, 1993), and the efficient parsing algorithms that exist for Context-Free Grammars carry over to PCFGs (see Charniak, 1993 or Manning & Schütze, 1999 for the relevant algorithms).

Any probabilistic grammar extracted from a training set faces the problem of data-sparseness: many of the rules in the training set are so infrequent that their observed probabilities are very bad estimates of their true population probabilities. A widely used method to cope with this problem is the Good-Turing method (Good, 1953). In general, Good-Turing estimates the expected population frequency f^* of a type by adjusting its observed sample frequency f. In order to estimate f^*, Good-Turing uses an additional notion, n_f, which is defined as the number of types which occur f times in an observed sample. Thus, n_f can be understood as the frequency of frequency f. The Good-Turing estimator uses this extra information for computing the adjusted frequency f^* as

$$f^* = (f+1)\frac{n_{f+1}}{n_f}$$

We thus compute the probabilities of our context-free rules in the Treebank grammar from their adjusted frequencies rather than from their raw observed frequences. Note that Good-Turing also adjusts the probabilities of unseen rules: if $f = 0$, then $f^* = n_1/n_0$. n_0 is the number of rules that have not been seen, and is usually estimated by $1 - n_1/N$ where N is the number of observed rules (see Good, 1953). However, Good-Turing does not differentiate among the rules that have not been seen: it assigns the same probability to all of them, which leads to rather inaccurate estimates for unseen rules. We will therefore introduce a more accurate way of assigning probabilities to unseen rules in Section 3.2. For an instructive paper on Good-Turing, together with a proof of the formula, see Church and Gale (1991).

The Treebank grammar that was obtained from the 5251 training folksongs was used to parse the 1000 folksongs in the test set. We computed for each test folksong the most probable parse using a standard best-first parsing algorithm based on Viterbi optimization (see Charniak, 1993; Manning & Schütze, 1999).

Using the evaluation measures given in Section 2, our Treebank grammar obtained a precision of 68.7%, a recall of 3.4%, and an F-score of 6.5%. Although the precision score may seem reasonable, the recall score is extremely low. This indicates that the Treebank grammar technique is a very conservative learner: it predicts very few phrases from the total number of phrases in the Essen Folksong Collection, resulting in a very low F-score. One of the problems with the Treebank grammar technique is that it learns only those context-free rules that literally occur in the training set (or otherwise assigns poor estimates to unseen rules), which is evidently not a very robust technique for musical parsing – while it has been shown to perform quite well in natural language parsing (cf. Charniak, 1996). We will see, however, that the results improve significantly if we slightly loosen the way of extracting rules from the training set.

3.2 The Markov grammar technique

A technique which overcomes the conservativity of Treebank grammars is the Markov grammar technique (Seneff, 1992; Collins, 1999). While a Treebank grammar can only accurately assign probabilities to context-free rules that have been seen in the training set, a Markov grammar can compute probabilities for any possible context-free rule, thus resulting in a more robust model. This is accomplished by decomposing a rule and its probability by a Markov process (see Collins, 1999: 44–48). For example, a third-order Markov process estimates the probability p of a rule P –> 12345 by:

$$p(\text{P} \rightarrow 12345) = p(1) \times p(2 \mid 1) \times p(3 \mid 1, 2) \times p(4 \mid 1, 2, 3) \times p(5 \mid 2, 3, 4) \times p(\text{END} \mid 3, 4, 5).$$

The conditional probability p(END | 3, 4, 5) encodes the probability that a rule ends after the notes 3, 4, 5. Thus even if the rule P –> 12345 does not literally occur in the training set, we can still estimate its probability by using a Markov history of three notes. The extension to larger Markov

histories follows from obvious generalization of the above example.

However, also a Markov grammar suffers from data-sparseness: we may get low counts, including zero counts, for some Markov histories. Zero counts are especially problematic: if one of the decomposed probabilities in the formula above has a zero occurrence in the training set, then the whole rule is assigned a zero probability. A widely used technique to solve the data-sparseness problem in Markov models is the *linear interpolation* technique (see Manning & Schütze, 1999: 218–219). This technique smooths a Markov history by taking into account its shorter histories. Let n_1, n_2 and n_3 denote three notes, then the conditional probability $p(n_1 \mid n_2, n_3)$ is smoothed ("interpolated") as

$$p(n_1 \mid n_2, n_3) = \lambda_1 p(n_1) + \lambda_2 p(n_1 \mid n_2) + \lambda_3 p(n_1 \mid n_2, n_3)$$

where $0 \leq \lambda_i \leq 1$ and $\lambda_1 + \lambda_2 + \lambda_3 = 1$. These λ-weights may be set by hand, but in general one wants to find the combination of weights λ_i which works best. A simple algorithm that finds the optimal weights is Powell's algorithm (see Press et al., 1988), which is also discussed in Manning & Schütze (1999: 218). We used this algorithm to assign weights to the lambdas in the linear interpolation technique, which in turn was used to estimate the conditional probabilities in the Markov grammar technique. Furthermore, each of the probabilities $p(n_1)$, $p(n_1 \mid n_2)$ and $p(n_1 \mid n_2, n_3)$ were not directly estimated from their observed relative frequencies in the training set, but were adjusted by the Good-Turing method, just as with Treebank grammars (Section 3.1). Note that the extension to any larger Markov history follows from simple generalization of the formulas above. The probability of a parse tree of a musical piece is computed by the product of the probabilities of the rules that partake in the parse tree, just as with Treebank grammars.

For our experiments, we used a Markov grammar with a history of four notes. This grammar obtained a precision of 63.1%, a recall of 80.2%, and an F-score of 70.6%. These results are to some extent complementary to the Treebank grammar: although the precision is somewhat lower, the recall is (much) higher than for the Treebank grammar. Thus, while the Treebank grammar predicts too few phrases, the Markov grammar predicts (a bit) too many phrases. The combined F-score of 70.6% shows an immense improvement over the Treebank grammar technique. Experiments with higher or lower order Markov models diminished our results.

3.3 Extending the Markov grammar technique with the DOP technique

Although the Markov grammar technique obtained considerably better scores than the Treebank grammar technique, it does not take into account any global context in computing the probability of a parse tree. Knowledge of global context, such as the number of phrases that appear in a folksong, is likely to be important for predicting the correct segmentations for new folksongs. In order to include global context, we conditioned over the S-rule higher in the structure in computing the probability of a P-rule. This approach corresponds to the Data-Oriented Parsing (DOP) technique (Bod, 1998) which can condition over any higher or lower rule in a tree. In the original DOP technique, any fragment seen in the training set, regardless of size, is used as a productive unit. But in the Essen Folksong Collection we have only two levels of constituent structure in each tree, thus resulting in a much simpler probabilistic model. As an example take again the rule P –> 12345 and a higher S-rule such as S –> PPPP; then a DOP-Markov model based on a history of three notes computes the (conditional) probability of this rule as:

$$\begin{aligned} p(\text{P} \text{ --> } 12345 \mid \text{S} \text{ --> } \text{PPPP}) &= p(1 \mid \text{S} \text{ --> } \text{PPPP}) \\ &\times p(2 \mid \text{S} \text{ --> } \text{PPPP}, 1) \times p(3 \mid \text{S} \text{ --> } \text{PPPP}, 1, 2) \\ &\times p(4 \mid \text{S} \text{ --> } \text{PPPP}, 1, 2, 3) \times p(5 \mid \text{S} \text{ --> } \text{PPPP}, 2, 3, 4) \\ &\times p(\text{END} \mid \text{S} \text{ --> } \text{PPPP}, 3, 4, 5). \end{aligned}$$

The extension to larger histories follows from obvious generalization of the above example. For our experiments, we used a history of four notes, extended with the same smoothing techniques as in Section 3.2 (i.e., linear interpolation combined with Good-Turing). The most probable parse of a folksong is again computed by maximizing the product of the rule probabilities that generate the folksong.

Using the same training/test set division as before, this DOP-Markov parser obtained a precision of 76.6%, a recall of 85.9%, and an F-score of 81.0%. The F-score is an improvement of 10.4% over the Markov parser. Note that the DOP-Markov parser is relatively well-balanced: it is neither terribly conservative nor does it predict too many redundant phrases – keeping in mind the idiosyncracy of the Essen Folksong annotations. While there is no reason to expect a near to 100% accuracy for the shallowly annotated Essen Folksong Collection (especially in the absence of harmonic structure), our results show the importance of including global context in computing the probability of a parse. We also checked the statistical significance of our results, by testing on 9 additional random splits of the Essen Folksong Collection (into training sets of 5251 folksongs and a test sets of 1000 folksongs). On these splits, the DOP-Markov parser obtained an average F-score of 80.7% with a standard deviation of 1.9%, while the Markov parser obtained an average F-score of 70.8% with a standard deviation of 2.2%. These differences were statistically significant according to paired *t*-testing.

Before we go into a more qualitative evaluation of our results, we were interested in testing the impact of the training size on the F-score. As mentioned in the introduction, there is some psychological support for the hypothesis that previously heard musical fragments are stored in memory, and that more frequent fragments are more easily activated than less frequent fragments. Yet, it seems unlikely that humans store more than, 5,000 folksongs to analyze new folksongs. It is from this perspective that we were interested in investigating how our DOP-Markov parser performs with

Table 1. F-score as a function of training set size.

Training size	F-score
500	31.1%
1000	47.4%
1500	56.9%
2000	64.4%
2500	69.0%
3000	73.2%
3500	76.1%
4000	78.3%
4500	79.9%
5000	80.7%
5251	81.0%

smaller training sets. In the following experiments we started with an initial training set of only 500 folksongs (randomly chosen from the full training set of 5251 folksongs). We then increased the size of this initial training set with 500 folksongs each time (randomly chosen from the full training set). The test set was kept constant at 1000 folksongs. The results are shown in Table 1.

The table shows that the F-score rapidly increases when the size of the training set is enlarged from 500 to 2000 folksongs. The accuracy continues to increase at a lower rate if the training set is further enlarged. We note that at around 4000 folksongs, relatively good F-scores are obtained. We may question the cognitive reality of a memory of 4000 folksongs. But we must keep in mind that our parser has no knowledge of the Gestalt rules of proximity and similarity, or whatsoever. The inclusion of such knowledge might boost our results or reduce the size of the training set. On the other hand, we might also argue that we could just as well eliminate all memory-based knowledge if we have access to the Gestalt rules. We will discuss this issue in the following section.

4. Discussion: Challenging the Gestalt principles

We have seen that a memory-based parsing model, known as the DOP-Markov parser, can quite accurately predict the preferred grouping structures for western folksongs. However, we have also seen that a large amount of hand-annotated training data is needed to achieve this result. In fact, to learn that a grouping boundary tends to occur at a large intervallic distance of pitch or time, our memory-based parser must encounter several specific instances of such intervals before it can assign a high probability to a boundary occurring at such an interval. This may seem a serious drawback since such intervallic boundaries may just as well be predicted by only a few rules that formalize the Gestalt notions of proximity and similarity (such as Lerdahl & Jackendoff, 1983: 39, or Cambouropoulos, 1997). However, there are many patterns in the Essen Folksong Collection that are problematic for Gestalt-based parsers, even when such parsers are extended with a parallelism-detection mechanism (as in Cambouropoulos, 1998), while they are rather trivial for memory-based models. These are patterns that contain a jump (a large pitch interval) at the beginning or end of a phrase (or both). As an example, consider the first 12 notes from folksong K0029, which was given in Figure 1, and which corresponds to the two groups in (5):

Figure 5

$$(3\_221\_-5)(-533221\_-5)$$

A Gestalt-based parser would probably assign one of the following grouping structures to these notes:

Figure 6

$$(3\_221\_)(-5-533221\_)(-5\ldots$$

or:

Figure 7

$$(3\_221\_-5-5)(33221\_-5)$$

While these grouping structures are possible, in that they *can* be perceived, they do not correspond to the structure that is actually perceived. The problem arises from the relatively large intervals of pitch (and time) between the notes 1_ and −5, and between the notes −5 and 3, from which a Gestalt-based parser would infer a grouping boundary at one of these intervals. What phenomenon overrules the perception of a boundary here? For this particular example, one could argue that the very strong melodic parallelism between the first five notes (i.e., 3_221_−5) and the last five notes (i.e., 3_221_−5_) of this folksong (See Figure 1) overrules the boundary at the local intervallic distance, thus resulting in the correct segmentation – provided that we have a mechanism which can discover these parallel patterns (cf. Cambouropoulos, 1998). However, there are also (many) folksongs where no such parallelism occurs and yet there is a group boundary between two equivalent notes that are preceded and followed by relatively large intervals. For example in folksong K0690 ("Ruru Rinneken"):

Figure 8

$$(3\_\_2\_\_1\_1\_-5\_)(-5\_3\_3\_2\_2\_1\_1\_-5\_)$$
$$(-5\_1\_2\_3\_1\_4\_\_2\_)(1\_-7\_1\_2\_-5\_3\_\_1\_)$$
$$(3\_1\_\_-5\_3\_1\_1\_-5\_3\_1\_\_-5\_)$$
$$(-5\_1\_2\_3\_1\_4\_3\_223\_1\_\_1\_0\_)$$

Here we have again two relatively large pitch intervals, or jumps, between the two notes at the end of the first group (1_ and −5_) and at the beginning of the second group (−5_ and 3_). Since there is no discontinuity in time here, one would expect a grouping boundary at the largest jump, i.e., between

−5_ and 3_, which would also be predicted by the Gestalt rules (see Lerdahl & Jackendoff, 1983: 39). Yet, the boundary occurs between the two equivalent notes −5_ and −5_! And now there is no higher-level parallelism that could enforce the correct grouping structure. On the contrary: a mechanism that would enforce musical parallelism would assign the same boundary between −5_ and 3_ as predicted by the Gestalt rules, since it would result in two very similar or parallel groups:

Figure 9

(3 __ 2 __ 1 _ 1 _ − 5 _ − 5 _)(3 _ 3 _ 2 _ 2 _ 1 _ 1 _ − 5 _ − 5 _)(. . .

What phenomenon overrules these phrase boundaries? Before trying to answer this question, we should be confident that the annotation of folksong K0690 is correct, i.e., that its annotation corresponds to the structure as perceived by a human listener. While we found the last two groups of the annotation of K0690 overly shallow, the group boundaries provided by the Essen Folksong Collection matched with our perception of group boundaries, to the best of our intuitions. Although we admit that the correctness of an annotation should preferably be established by an independent psychological experiment with more than one subject (which falls beyond the scope of this paper), we feel confident that the anomalous grouping boundaries of K0690 do not depend on some kind of annotation error.

A possible cause for the peculiar grouping structure of K0690 may be the lyrics, i.e., the text, of the folksong. It could be that the prosodic structure of the text enforces a certain melodic grouping structure which might explain the perceived "jump-phrases" in K0690. However, the texts of the folksongs have not been entered in the Essen Folksong Collection, and only very few texts are available at all (Dr. Ewa Dahling, personal communication). Moreover, we already established that our grouping intuitions for folksong K0690, *without* having access to its text, agreed with the segmentations in the Essen Folksong Collection. Thus we can rule out the influence of the text as a cause for the peculiar grouping of folksong K0690. (Note also that it is rather uncontroversial to study the melodies of vocal music without considering the texts – see for instance the many examples of songs, chorales and arias in Lerdahl & Jackendoff (1983) or Narmour (1990).)

So far, we have not considered the metrical structures of the Essen Folksongs. One might wonder whether meter can enforce the perceived grouping structure of K0690. It is widely acknowledged, however, that grouping structure is *independent* of metrical structure, which leads virtually all theories of music cognition to formulate separate models for grouping and meter. Lerdahl and Jackendoff convincingly show that "groups do not receive metrical accent, and beats do not possess any inherent grouping" (Lerdahl & Jackendoff, 1983: 26). But even if the metrical structure of K0690 did enforce and thus matched the grouping structure of this folksong, it would assign the same incorrect phrases as given in Figure 9, since the beats appear exactly on the first notes of these phrases. Thus, metrical structure would not help either to explain the anomalous grouping structure in (8).

Finally, we should consider the role of harmony. It is well-known that the internal harmony of a piece *does* often influence its melodic grouping structure. So one might hope that by taking into account the implied or internal harmony of folksong K0690, we can explain and predict its grouping into jump-phrases. However, the two alternative groupings, expressed by the first two phrases in Figures 8 and 9, display the same internal harmony: both are melodic elaborations of the basic triad 1, 3, 5. Thus, harmonic grouping preferences, as proposed in e.g., Lerdahl and Jackendoff (1983) or Narmour (1990, 1992), are not of any help in predicting the peculiar grouping structure of K0690.

So there seems to be no musical factor that can overrule the incorrect predictions made by the Gestalt principles for this folksong: neither melodic parallelism, nor metrical structure, and not even internal harmony. One might put forward that grouping structures with jump-phrases are highly exceptional and limited to only a few folksongs that are not representative for the Essen Folksong Collection. Yet, a detailed analysis of the test data (1000 folksongs) shows that more than 32% of the folksongs contained at least one jump-phrase and that the total percentage of phrases that start or end with a jump (or both – as in the second phrase in (8)) is at least 15%. Thus folksongs with jump-phrases are not epiphenomenal.

It is noteworthy that our DOP-Markov parser predicted to a very high degree (98.0%) the correct grouping boundaries for these 15% jump-phrases (although it often assigned additional subphrases within these phrases). A Gestalt-based/parallelism-based parser, on the other hand, would definitely predict the wrong grouping boundaries for all these jump-phrases – except if there are parallel phrases in the piece that may enforce the correct grouping boundaries, as we discussed for figure 1, but such parallel phrases occurred less than 1% in the test set. Other things being equal, our parser would improve with about 12% over a Gestalt-based/parallelism-based parser – given the 15% jump-phrases, the 98.0% performance on these phrases by our parser, and the less than 1% of these phrases for which parallelism may override the Gestalt principles. Moreover, we could not find any test folksong for which a Gestalt-based/parallelism-based parser might possibly improve over our memory-based parser, though we fully admit that this needs to be checked by an actual experiment with an implementation of such a parser. The patterns that were problematic for our DOP-Markov parser seem to be entirely due to the shallowness of the annotations in the Essen Folksong Collection (i.e., our parser still predicts too many phrases); this shallowness is equally problematic for a Gestalt-based/parallelism-based parser, we trust. (We should perhaps mention that jumps in the *middle* of phrases are also problematic for Gestalt-based models, but such jumps would only lead to additional subphrases which are not annotated in the Essen Folksong

Collection and can therefore not be tested here. Only jumps at the beginning or at the end of phrases lead to wrong predictions by Gestalt-based/parallelism-based models.)

We may thus conclude that jump-phrases provide serious evidence against the Gestalt principles of proximity and similarity, and that a model which is solely based on musical factors, such as intervallic distances, parallelism, meter and harmony, can never learn jump-phrases like in Figure 8. The following figure gives some other folksongs from the Essen Folksong Collection that involve jumps from or to note –5 (there are also jumps from other notes, such as –4 and –6, which are not present in this example).

Figure 10

Folksong K0641

(11 – 7 – 511 – 5)(–511 – 721 _ – 50)(11 – 7 – 5222 _)(11 – 721 _ – 5)
(–511 – 5 – 511 – 5 _)(11 – 7 – 5222)(211 – 721 _ – 50)

Folksong A0214

(1 _ 1 _ 1 _ 1 _ 1 _ – 7b _ – 7b _ – 5 _)(–5 _ 3b _.3b4 _3b _2 __1 _)
(1 _ 1 _ 1 _ 1 _ 1 _ – 7b _ – 7b _ – 5 _)
(–5 _ 3b _.3b4 _3b _)(2 __1 _ 1 _ 3b _.45 _3b _4 __1 __)
(4 _.43b _21 – 7b _.12 _)(3b _4 _1 _3b _2 _1 __.)

Folksong B0752

(1 _ 5 __5 _ 5 _ 4 _ 3 _ 5 _ 4 _ 3 _ 2 _ 1 _)
(–5 _ 4 _ 3 _ 2 _ 3 __3 _ 5 _ 4 _ 3 _ 2 __)
(1 _ 5 __5 _ 5 _ 4 _ 3 _ 5 _ 4 _ 3 _ 2 _ 1 _)
(–5 _ 4 _ 3 _ 2 _ 3 __3 _ 5 _ 4 _ 3 _ 2 __)
(2 _ 2 __2 _ 3 __3 _ 4 _ 3 _ 4 _ 5 __)
(5 _ +1 _ 7 _ 6 _ 5 _.6543 __2 _ 1 __)

Folksong B0179

(–5 _ 5 _.43 _2 _1 _ – 6 _ – 5)(–5 _ 2 __0 _ – 5 _ 3 __0 _)
(3 _ 6 _.54# _3 _2 _1 _ – 7 _)(3 _ 2 __0 _ 1 _ – 7b __0 _)
(234 _ 432 _ 234 _ 3 _ 2 _)(45654 _ 4321 – 7 _ – 712 _)
(–5 _ 3 _.32 _5 _1 __0 _)(–5 _ 3 _.32 _5 _1 __0 _)

One can of course argue that there may still be a more fundamental principle or rule, which we do not (yet) know of, and which *does* predict the correct grouping boundaries for jump-phrases. The search for such a principle or rule, which seems to go beyond the harmonic, metric, and melodic nature of music, will be part of future research. But we should neither rule out the possibility that this particular grouping phenomenon is inherently memory-based. This possibility may be supported by Huron (1996) who observed that phrases in western folksongs tend to exhibit an "arch" shape, where the pitch contour rises and then falls over the course of a phrase. Thus the group (–5_3_3_2_2_1_1_–5_) in folksong K0690 displays such an arch contour, while the group (3_3_2_2_1_1_–5_–5_) does not. Assuming that Huron's observation is correct, arch-like patterns may either express a universal tendency in music, in which case they ought to be formalized by a rule or principle (but there is no evidence for this universality), or arch-like patterns may be strictly idiom-dependent, in which case they can be best captured by a memory-based model that tries to mimic the musical experience of a listener from a certain culture. Thus, music perception may be much more memory-based as previously assumed.

If we wish to propose a memory-based approach to music as a serious alternative to a Gestalt-based approach, we should address the question of how any structure can be acquired if we do not have any structured pieces in our corpus to start with. With an already analyzed corpus, we can at best simulate *adult* music perception – analogous to a corpus of analyzed natural language (see Bod, 1998). We conjecture that the acquisition of a structured corpus may be the result of a bootstrapping process where the discovery of similar recurrent patterns and distributional regularities plays an important role. As soon as a pattern appears more than once, it may be hypothesized as a group, and may be used as a productive unit to analyze new pieces. The frequency with which a pattern occurs is used to decide between conflicting groups. Much research in unsupervised language learning is concerned with bootstrapping syntactic structure on the basis of pattern similarity and statistics from large language corpora (e.g., Finch & Chater, 1994; Brent & Cartwright, 1996; van Zaanen, 2000). One of our future goals is to investigate whether such unsupervised learning techniques carry over to bootstrapping musical structure, and whether the learned structure corresponds to the structure as perceived by human listeners. On the other hand, there is already a considerable amount of work on unsupervised musical pattern induction (e.g., Cope, 1990; Mattusch, 1997; Crawford et al., 1998; Rolland & Ganascia, 2000). We hope to assess these models, along with unsupervised models of natural language learning, for the task of bootstrapping structure in a large musical corpus. Once an initial corpus of musical patterns has been learned, these patterns can be used by our supervised model to efficiently segment new pieces. Only for completely new sequences of notes that have never appeared before, unsupervised methods need still to be invoked. The exact interplay between unsupervised and supervised aspects of music perception needs to await further investigation.

But also if we limit ourselves to supervised music segmentation, this work triggers much new research. One of our projects is to convert the absolute pitch encodings in the Essen annotations into relative pitch encodings, such that our memory-based parser can more easily generalize over intervals that occur between different notes but that involve the same pitch or time distances; this may also reduce the size of the training set, which would increase the cognitive plausibility of our model. Another project is to manually enrich the Essen annotations with more fine-grained constituents, such as subphrases and subsections, and assign these constituents with labels that summarize regularities of the under-

lying patterns, as proposed by musical coding languages such as Collard et al. (1981) and Deutsch and Feroe (1981). We surmise that a listener's melodic structuring depends partly on regularities in the input patterns (as described by musical coding languages) and partly on previous musical experiences (as described by our memory-based approach). An adequate model for music perception should do justice to both aspects of music.

5. Conclusion

We have presented a memory-based approach to music which analyzes new pieces by combining fragments from structures of previously encountered pieces. In case of ambiguity, this approach computes the analysis that can be considered the most probable one on the basis of the occurrence-frequencies of the fragments. We successfully tested some instances of this approach on a set of 1000 folksongs from the Essen Folksong Collection, obtaining an F-score of up to 81.0%. To the best of our knowledge, this paper contains the first parsing experiment with the Essen Folksong Collection, which we hope may serve as a baseline for other computational models of music analysis.

A qualitative analysis of our results showed that there is a class of musical patterns, so-called jump-phrases, that challenge both the Gestalt principles of proximity and similarity and the principle of melodic parallelism. Jump-phrases provide evidence that grouping boundaries can appear *after* or *before* large pitch intervals, rather than *at* such intervals, and that grouping boundaries can even appear between *identical* notes (that are preceded and followed by relatively large intervals). We have seen that Gestalt-based, parallelism-based and/or harmony-based models are inadequate to deal with these patterns. Probabilistic, memory-based models seem more apt to deal with these gradient phenomena of music analysis since they can capture the entire continuum between jump-phrases and non-jump-phrases.

Acknowledgments

We are grateful to Emilios Cambouropoulos and two anonymous reviewers for useful comments on a previous version of this paper. We also thank Ewa Dahlig for providing information about the Essen Folksong Collection.

References

Black, E., Abney, S., Flickinger, D., Gnadiec, C., Grishman, R., Harrison, P., Hindle, D., Ingria, R., Jelinek, F., Klavans, J., Liberman, M., Marcus, M., Roukos, S., Santorini, B., & Strzalkowski, T. (1991). A procedure for quantitatively comparing the syntactic coverage of English, *Proceedings DARPA Speech and Natural Language Workshop*, Pacific Grove, Morgan Kaufmann.

Bod, R. (1993). Using an Annotated Language Corpus as a Virtual Stochastic Grammar. *Proceedings AAAI'93*, Morgan Kaufmann, Menlo Park.

Bod, R. (1998). *Beyond Grammar: An Experience-Based Theory of Language*, Stanford, CSLI Publications (distributed by Cambridge University Press).

Booth, T. (1969). Probabilistic Representation of Formal Languages, *Tenth Annual IEEE Symposium on Switching and Automata Theory*.

Brent, M. & Cartwright, T. (1996). Distributional Regularity and Phonotactic Contraints are Useful for Segmentation, *Cognition*, *61*, 93–125.

Cambouropoulos, E. (1996). A Formal Theory for the Discovery of Local Boundaries in a Melodic Surface. *Proceedings of the Troisièmes Journées d'Informatique Musicale (JIM-96)*, Caen, France.

Cambouropoulos, E. (1997). Musical Rhythm: A Formal Model for Determining Local Boundaries, Accents and Meter in a Melodic Surface. In: M. Leman (Ed.), *Music, Gestalt and Computing – Studies in Systematic and Cognitive Musicology*, Berlin, Springer-Verlag.

Cambouropoulos, E. (1998). Musical Parallelism and Melodic Segmentation, *Proceedings XII Colloquium on Musical Informatics*, Gorizia, Italy.

Charniak, E. (1993). *Statistical Language Learning*, Cambridge: The MIT Press.

Charniak, E. (1996). Tree-bank Grammars, *Proceedings AAAI-96*, Menlo Park, Ca.

Church, K. & Gale, W. (1991). A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams, *Computer Speech and Language*, *5*, 19–54.

Collard, R., Vos, P., & Leeuwenberg, E. (1981). What Melody Tells about Metre in Music. *Zeitschrift für Psychologie*, *189*, 25–33.

Collins, M. (1999). *Head-Driven Statistical Models for Natural Language Parsing*, PhD-thesis, University of Pennsylvania, PA.

Cope, D. (1990). Pattern-Matching as an Engine for the Computer Simulation of Musical Style, *Proceedings ICMC'1990*, Glasgow, UK.

Crawford, R., Iliopoulos, C., & Raman, R. (1998). String Matching Techniques for Musical Similarity and Melodic Recognition, *Computing in Musicology*, *11*, 71–100.

Deutsch, D. & Feroe, J. (1981). The Internal Representation of Pitch Sequences in Tonal Music, *Psychological Review*, *88*, 503–522.

Finch, S. & Chater, N. (1994). Distributional Bootstrapping: From Word Class to Proto-Sentence, *Proceedings 16th Annual Cognitive Science Society*, 301–306, Hillsdale, Lawrence Erlbaum.

Good, I. (1953). The Population Frequencies of Species and the Estimation of Population Parameters, *Biometrika*, *40*, 237–264.

Handel, S. (1989). *Listening. An Introduction to the Perception of Auditory Events*. Cambridge: The MIT Press.

Höthker, K., Hörnel, D., & Anagnostopoulou, C. (2001). Inves-

tigating the Influence of Representations and Algorithms in Music Classification. *Computers and the Humanities*, *35*, 65–79.

Huron, D. (1996). The Melodic Arch in Western Folksongs. *Computing in Musicology*, *10*, 2–23.

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

Longuet-Higgins, H. (1976). Perception of Melodies. *Nature* 263, October *21*, 646–653.

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

Mattusch, U. (1997). Emulating Gestalt Mechanisms by Combining Symbolic and Subsymbolic Information Processing Procedures. In: M. Leman (Ed.), *Music, Gestalt and Computing – Studies in Systematic and Cognitive Musicology*, Berlin, Springer-Verlag.

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill Companies.

Narmour, E. (1990). *The Analysis and Cognition of Basic Melodic Structures: The Implication-Realization Model*, The University of Chicago Press, Chicago.

Narmour, E. (1992). *The Analysis and Cognition of Melodic Complexity*, The University of Chicago Press, Chicago.

Press, W., Flannery, B., Teukolsky, S., & Vetterling, W. (1988). *Numerical Recipes in C*. Cambridge University Press.

Rolland, P. & Ganascia, J. (2000). Musical Pattern Extraction and Similarity Assessment. In E. Miranda (Ed.) *Readings in Music and Artificial Intelligence*, Harwood Academic Publishers.

Saffran, J., Loman, M., & Robertson, R. (2000). Infant Memory for Musical Experiences. *Cognition* 77, B16–23.

Schaffrath, H. (1993). Repräsentation einstimmiger Melodien: computerunterstützte Analyse und Musikdatenbanken. In: B. Enders & S. Hanheide (Eds.), *Neue Musiktechnologie*, 277–300, Mainz, B. Schott's Söhne.

Schaffrath, H. (1995). The Essen Folksong Collection in the Humdrum Kern Format. In: D. Huron (Ed.), Menlo Park, CA: Center for Computer Assisted Research in the Humanities.

Selfridge-Field, E. (1995). The Essen Musical Data Package. Menlo Park, California: Center for Computer Assisted Research in the Humanities (CCARH).

Seneff, S. (1992). TINA: A Natural Language System for Spoken Language Applications. *Computational Linguistics*, *18*(1), 61–86.

Stoffer, T. (1985). Representation of Phrase Structure in the Perception of Music. *Music Perception*, *3*(2), 191–220.

Tenney, J. & Polansky, L. (1980). Temporal Gestalt Perception in Music, *Journal of Music Theory*, *24*, 205–241.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt. *Psychologische Forschung*, *4*, 301–350.

Wetherell, C. (1980). Probabilistic Languages: A Review and Some Open Questions, *Computing Surveys*, 12(4).

van Zaanen, M. (2000). Bootstrapping Structure and Recursion Using Alignment-Based Learning, *Proceedings International Conference on Machine Learning (ICML'2000)*, Stanford, California.