Probabilistic Grammars for Music

Rens Bod

ILLC, University of Amsterdam
Nieuwe Achtergracht 166, 1018 WV Amsterdam
rens@science.uva.nl

Abstract

We investigate whether probabilistic parsing techniques from Natural Language Processing (NLP) can be used for musical parsing. As in NLP, the main problem in music is ambiguity: several different structures may be compatible with a musical sequence while a listener typically hears only one structure. Our best probabilistic parser can correctly predict 85.9% of the phrases for a test set of 1,000 folksongs from the Essen Folksong Collection.

1. Introduction

We investigate whether probabilistic parsing techniques from Natural Language Processing (NLP) can be used for musical parsing. As in natural language, a listener segments a sequence of notes into groups or phrases that form a grouping structure for the whole piece (Longuet-Higgins 1976; Tenney & Polansky 1980; Lerdahl & Jackendoff 1983). For example, according to Lerdahl & Jackendoff (1983: 37) a listener hears the following grouping structure for the first few bars of melody in the Mozart G Minor Symphony, K. 550.

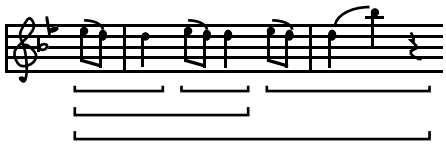

Figure 1. Grouping structure for the opening theme of Mozart's G Minor Symphony

Each group is represented by a slur beneath the musical notation. A slur enclosed within a slur means that a group is heard as part of a larger group. This hierarhical structure of melody can, without loss of generality, also be represented by a phrase structure tree, as in figure 2.

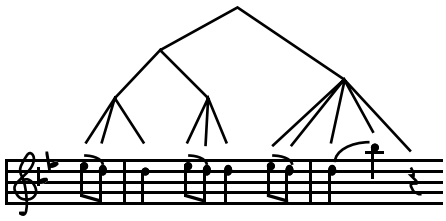

Figure 2. Tree structure for the grouping structure in figure 1

Although visually quite different, the two representations in figures 1 and 2 are mathematically equivalent. Note the analogy with phrase structure trees in linguistics: a tree describes how parts of the input combine into constituents and how these constituents combine into larger constituents and into a representation for the whole input. Apart from this analogy, there is also an important difference: while the nodes in a linguistic tree structure are typically labeled with syntactic categories such as S, NP, VP etc., musical tree structures are unlabeled. This is because in language there are syntactic constraints on how words can be combined into larger constituents (e.g. in English a determiner can be combined with a noun only if it precedes that noun, which is expressed by the rule NP -> Det N), while in music there are no such restrictions: in principle any note may be combined with any other note. This makes the problem of ambiguity in music much harder than in language. Longuet-Higgins and Lee (1987) note that "Any given sequence of note values is in principle infinitely ambiguous, but this ambiguity is seldom apparent to the listener.".

For example, the first few bars of Mozart's G Minor Symphony could also be assigned the following, alternative grouping structure (among the many other possible structures):

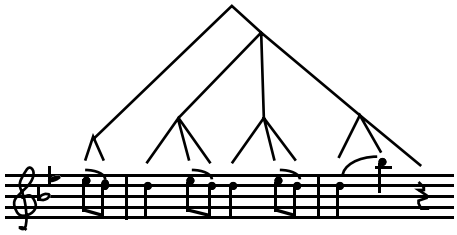

Figure 3. Alternative grouping structure for Mozart's opening theme

While this alternative structure is possible in that it *can* be perceived, it does not correspond to the structure that is actually perceived by a human listener. There is thus an important research question as to how to select the perceived tree structure from the total, possibly infinite set of possible tree structures of a musical input.

In the field of natural language processing (NLP), the use of probabilistic corpus-based parsing techniques has become increasingly influential for solving ambiguity (see Manning & Schütze 1999 for an overview). Instead of using a predefined set of rules, a probabilistic corpus-based parser learns how to parse new input by generalizing from examples of previously annotated data, and in case of ambiguity, such a parser computes the most probable phrase structure for a given input. State-of-the-art probabilistic parsers which use the Wall Street Journal portion in the Penn Treebank (Marcus et al. 1993) as a test domain, obtain around 90% correctly predicted phrases (e.g. Charniak 2000; Bod 2001a). With the current availability of large annotated musical corpora, such as the Essen Folksong Collection (Schaffrath 1995), we may wonder whether such probabilistic parsing techniques carry over to musical parsing.

In this paper we will test the usefulness of three probabilistic parsing techniques for music: the Treebank grammar technique of Charniak (1996) and Bod (1993), the Markov grammar technique of Collins (1999), and the Data-Oriented Parsing (DOP) technique of Bod (1998). We develop a new parser which combines two of these techniques, and which correctly predicts up to 85.9% of the phrases for a held-out test set of 1,000 folksongs from the Essen Folksong Collection

(Schaffrath 1995). To the best of our knowledge, this paper contains the first parsing experiments on the Essen Folksong Collection; moreover, it also contains the first experiments on a musical test set of non-trivial size.

2. The Essen Folksong Collection

The Essen Folksong Collection provides a large sample of (mostly) European folksongs that have been collected and encoded under the supervision of Helmut Schaffrath at the University of Essen (see Schaffrath 1993, 1995; Selfridge-Field 1995). Each of the 6,251 folksongs in the Essen Folksong Collection is annotated with the Essen Associative Code (ESAC) which includes pitch and duration information, meter signatures and explicit phrase markers. 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", but 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. A pause is represented by 0, possibly followed by duration indicators. No loudness or timbre indicators are used in ESAC.

Thus, the opening theme of Mozart's G Minor Symphony in figure 1 can be encoded in ESAC as follows.

+3b+2+2_+3b+2+2_+3b+2+2_+7b_0_

Figure 4. ESAC encoding for the opening theme of Mozart's G Minor Symphony

ESAC uses hard returns to indicate a phrase boundary. To make the Essen annotations readable for our probabilistic parsers, we automatically convert ESAC's phrase boundary indications into bracket representations, where "(" indicates the start of a phrase and ")" the end of a phrase. The phrase structures in figures 1 and 2 thus correspond to the following bracket representation.

(((+3b+2+2_) (+3b+2+2_)) (+3b+2+2_+7b_0_))

Figure 5. Bracket representation for the phrase structures in figures 1 and 2

Figure 6 gives an example of an encoding of an actual folksong from the Essen Folksong Collection (converted to our bracket representation):

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

Figure 6. Bracket representation for folksong K0029, "Schlaf Kindlein feste"

Note that the Essen annotations are very shallow; yet, we will see that it is surprisingly difficult to predict the correct phrases for the Essen folksongs. To evaluate our probabilistic parsers for music, we employed the *blind testing method* (see Manning & Schütze 1999), by randomly dividing the Essen Folksong Collection into a training set of 5,251 folksongs and a test set of 1,000 folksongs. As evaluation metrics we used the notions of *precision* and *recall* (see Black et al.

1991) that compare a proposed parse *P* with the corresponding test set parse *T* as follows:

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

A phrase is correct if both the start and the end of the phrase is correctly predicted. The precision and recall scores are often combined into a single measure of performance, known as the F-score (see Manning & Schütze 1999): F-score = 2 * Precision * Recall / (Precision + Recall). We will use these three measures to quantitatively evaluate our probabilistic parsing models for music.

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 annotation in figure 6 becomes:

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

Figure 7. Labeled-bracketing annotation for the structure in figure 6

3. Parsing the Essen Folksong Collection

3.1 The Treebank Grammar Technique

The Treebank grammar technique, coined by Charniak (1993) but used earlier in Bod (1993), 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 7:

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 in figure 7 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 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, and the efficient parsing algorithms that exist for Context-Free Grammars carry over to PCFGs (see Manning & Schütze 1999 for the relevant algorithms). The Treebank grammar obtained in this way from the 5,251 training folksongs was used to parse the 1,000 folksongs in the test set. We computed for each test folksong the most probable parse using a standard best-first parsing algorithm (Charniak 1993).

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, which 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 only learns those context-free rules that literally occur in the training set, which is evidently not a very robust technique for musical parsing (while it has been shown to perform quite well in natural language parsing -- see 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 assign probabilities to context-free rules that have been seen in the training set, a Markov grammar can in principle assign a probability to 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 $p(\text{END} \mid 3, 4, 5)$ 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.

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 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 occur 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 in the tree 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, which results in a much simpler probabilistic model. As an example take again the rule P -> 12345 and an S-rule such as S -> PPPP; a DOP-Markov model based on a history of three notes computes the (conditional) probability of this rule as:

p(P -> 12345 | S -> PPPP) =
p(1 | S -> PPPP) × p(2 | S -> PPPP, 1) × p(3 | S -> PPPP, 1, 2) ×
p(4 | S -> PPPP, 1, 2, 3) × p(5 | S -> PPPP, 2, 3, 4) ×
p(END | S -> PPPP, 3, 4, 5).

The extension to larger histories follows from obvious generalization of the above example. For our experiments, we used a history of four notes. 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. We also checked the statistical signifance of our results, by testing on 9 additional random splits of the Essen Folksong Collection (into training sets of 5,251 folksongs and a test sets of 1,000 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.

4. Discussion: other approaches to musical parsing

There exists an extensive literature in the field of computational models of music analysis (see Cambouropoulos 1998 for an overview). Most if not all approaches to musical parsing are non-probabilistic and are based on the assumption that the perceived phrase structure of a musical piece can be predicted on the basis of a combination of low-level phenomena, such as Gestalt phenomena of proximity and similarity, and higher-level phenomena, such as melodic parallelism and internal harmony.

For example, Tenney & Polansky (1980), Lerdahl & Jackendoff (1983) and Cambouropoulos (1998) use the Gestalt principles (Wertheimer 1923) to predict the low-level grouping structure of a piece: phrase boundaries preferably fall on larger time intervals, larger pitch intervals, etc. While most models also incorporate higher-level 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 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 test sets of non-trivial size, such as the Essen Collection. Only very few, hand-selected passages are typically used to evaluate these models, which questions the objectivity of the results.

More importantly perhaps, is the fact that the Gestalt principles, which were originally proposed for *visual* perception (Wertheimer 1923), do not straightforwardly carry over to music perception. Elsewhere (Bod 2001b), we have shown that more than 15% of the phrase boundaries in the Essen Folksong Collection fall *before* or *after* large pitch or time intervals (which we called "jump-phrases"), rather than *at* such intervals, and that phrase boundaries can even appear

between *identical* notes, as in the folksong of figure 6. This goes against the predictions of any Gestalt-based parser, which would assign phrase boundaries exactly *at* large intervals rather than before or after them. We have shown in Bod (2001b) that higher-level phenomena such as melodic parallelism and internal harmony are not of any help for predicting the correct phrase boundaries for these 15% jump-phrases. On the contrary, for virtually all these phrases, melodic parallelism and harmony reinforces the incorrect predictions of the Gestalt principles. While our best parser is still far from perfect (it obtained a 79.4% F-score for jump-phrases and a 81.0% F-score for all phrases from the test set), a Gestalt-based parser would assign incorrect phrase boundaries to *all* of the jump-phrases. A probabilistic, corpus-based model seems more apt to deal with these phrases since it considers counts of any sequence of notes that has been observed with a certain structure rather than trying to capture these by a few formal rules.

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 (-533221_-5) in figure 6 displays such an arch contour, while its alternatively possible grouping (33221_-5) would not (see Bod 2001b for more details). 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 than often assumed. We surmise that a listener's melodic structuring depends partly on regularities in the input and partly on previous musical experiences. An adequate model of music analysis should do justice to both aspects of music.

5. Conclusion

We have shown that probabilistic parsing models from Natural Language Processing can be successfully applied to musical parsing. Our best parser can correctly predict up to 85.9% of the phrases for a test set of 1,000 folksongs from the Essen Folksong Collection. We hope that our results may serve as a baseline for other models of music analysis. Our parser may also be used to speed up the time-consuming annotation of newly collected folksongs, thereby contributing to the creation of larger musical databases in computer-assisted musicology.

A detailed evaluation of our results shows that there is a class of musical patterns, so-called jump-phrases, that challenge the Gestalt principles of proximity and similarity. 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). Elsewhere we have shown 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.

References

E. Black et al., 1991. A Procedure for Quantitatively Comparing the Syntactic Coverage of English, *Proceedings DARPA Speech and Natural Language Workshop*, Pacific Grove, Morgan Kaufmann.

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

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

R. Bod, 2001a. What is the Minimal Set of Fragments that Achieves Maximal Parse Accuracy? *Proceedings ACL'2001*, Toulouse, France.

R. Bod, 2001b. Memory-Based Models of Music Analysis: Evidence against the Gestalt Principles in Music. *Proceedings International Computer Music Conference 2001 (ICMC'2001)*, Havana, Cuba.

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

E. Cambouropoulos, 1998. *Towards a General Computational Theory of Musical Structure*, Ph.D. thesis, University of Edinburgh, UK.

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

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

E. Charniak, 2000. A Maximum-Entropy-Inspired Parser. *Proceedings ANLP-NAACL'2000*, Seattle, Washington.

K. Church and W. Gale, 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.

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

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

F. Lerdahl and R. Jackendoff, 1983. *A Generative Theory of Tonal Music*. Cambridge, The MIT Press.

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

H. Longuet-Higgins and C. Lee, 1987. The Rhythmic Interpretation of Monophonic Music. In: *Mental Processes: Studies in Cognitive Science*, Cambridge, The MIT Press.

C. Manning and H. Schütze, 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, The MIT Press.

M. Marcus, B. Santorini and M. Marcinkiewicz, 1993. Building a Large Annotated Corpus of English: the Penn Treebank, *Computational Linguistics* 19(2).

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

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

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

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

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

J. Tenney and L. Polansky, 1980. Temporal Gestalt Perception in Music, *Journal of Music Theory*, 24, 205-241.

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