RECOGNITION OF VARIATIONS USING AUTOMATIC SCHENKERIAN REDUCTION

Alan Marsden
Lancaster Institute for the Contemporary Arts, Lancaster University, UK
A.Marsden@lancaster.ac.uk

ABSTRACT

Experiments on techniques to automatically recognise whether or not an extract of music is a variation of a given theme are reported, using a test corpus derived from ten of Mozart's sets of variations for piano. Methods which examine the notes of the 'surface' are compared with methods which make use of an automatically derived quasi-Schenkerian reduction of the theme and the extract in question. The maximum average F-measure achieved was 0.87. Unexpectedly, this was for a method of matching based on the surface alone, and in general the results for matches based on the surface were marginally better than those based on reduction, though the small number of possible test queries means that this result cannot be regarded as conclusive. Other inferences on which factors seem to be important in recognising variations are discussed. Possibilities for improved recognition of matching using reduction are outlined.

1. SCHENKERIAN REDUCTION

Earlier work [6] has shown that Schenkerian analysis by computer is possible, though not easy. (Currently only short segments of music can be analysed, and confidence in the analyses produced cannot be high.) The aim of the research reported here is a first attempt at testing whether these automatic analyses produce information which is useful for information retrieval.

Schenkerian analysis is a technique, with a long pedigree in music theory, which aims to discover the structural 'framework' which is believed to underlie the 'surface' of a piece of music (see [1], for example). Reduction according to the theory of Lerdahl & Jackendoff, which has also been subject to computational implementation [2], is broadly similar. Figure 1 shows the first four bars of the theme of a set of variations for piano by Mozart, and its reduction as derived by the software used here. (This is by far the simplest of the themes used here; to show other themes and their reductions would take more space than is available.) Schenker's reductions were notated in a different fashion, and also included information not given here, but the basic information of which pitches are regarded as more 'structural', and so included in the higher levels, is similar.

The research reported here fits into that body of MIR research which aims to improve MIR procedures through the application of ideas from music theory.

2. VARIATIONS

A common type of composition in classical music is 'theme and variations'. In this kind of piece, a theme is presented, followed by a number of variations of that theme. There is no single and established definition of what constitutes a variation of a theme, but in the Classical period (the period of Haydn, Mozart and Beethoven) it is clear that a variation is not simply the presentation of the same melody in different arrangements (as it was for some later composers) but rather a composition which has the same structural features as the theme. This is particularly clear in Mozart's variations: they are almost always the same length as the theme, have the same number of phrases, and have matching cadences for those phrases (at least in their harmony; often in other features also). The

G5 C5 C3 C4							
G5 C5 C3 C4				G5 C4			
C5 C3		G5 C4		A5 C4		G5 C4	
C5 C3	C5 C4	G5 E4	G5 C4	A5 F4	A5 C4	G5 E4	G5 C4

Figure 1. The first four bars of the theme of Mozart's variations K.265 (using a tune known in English as 'Twinkle, twinkle little star'), and the highest-scoring reduction derived from these bars by the software.

Figure 2. The beginning of two variations of the theme shown in Figure 1.

internal structure of those phrases can also show common features: the harmony is often similar; there can be common notes, especially in important positions like beginnings and endings, and the variation sometimes clearly gives a decorated version of the melody and/or bass of the original. Figure 2 shows the first four bars of two variations of the theme shown in Figure 1.

If Schenkerian analysis validly reveals musical structure, then the analysis of each variation should, to some degree, match the analysis of the theme. To test this requires analyses of variations and themes which are unbiased in the sense that the analyses of each variation should be made with no knowledge of the theme. To achieve unbiased analyses with human analysts would be very difficult: expert analysts are required, and one would have to recruit as many analysts as there are variations in a set. Furthermore, it is well known that different analysts produce different analyses, and it would be difficult to neutralise these personal differences. The computer software described below gives a means for generating unbiased analyses, and so allows this kind of empirical test of the validity of Schenkerian analysis.

3. REDUCTION SOFTWARE

The method of reduction used here is described more fully in [6]. There is space here only to give a brief outline. An analysis of a piece is a binary tree whose leaves are the 'segments' of the surface of the music (the notes of the score). A segment is a span of music, containing all the notes sounding at that time. At least one note begins at the start of the segment and at least one note finishes at its end. No notes begin or finish at other points within the span. A note is defined by its pitch and by whether or not it is tied to a note in the preceding segment. A single note in the score can be split into a series of tied notes across several segments.

Segments above the surface are related to a pair of 'child' segments through a set of 'atomic elaborations'. These define how a note in a higher-level segment can be elaborated to become two shorter notes (or a note and a rest) in the child segments. The set of atomic elaborations is derived from Schenkerian theory and consists of such things as repetitions, neighbour notes, anticipations, consonant skips, etc. Atomic elaborations can imply that certain pitches are consonant, and the implications of the set of atomic elaborations relating a higher-level segment to its children must be consistent (i.e., the consonant pitches must form an acceptable harmony).

An analysis is therefore a kind of parse tree employing a grammar defined by the atomic elaborations. The software used here effectively employs a chart parser [4] as a step towards generating such a tree, but the computational complexity of the algorithm is of order $O(n^4)$ time. With typical computing resources, it is therefore possible to derive a parse chart from extracts of simple piano music up to only four to eight bars in length.

The parse chart is a triangular matrix whose cells contain the possible reductions at each stage of reduction. The bottom (longest) row contains the segments of the surface. The first row above contains segments which result from reduction of each of the pairs of consecutive segments below. Rows further above contain segments which result from reductions of those with other segments, etc., until the top row, with just one cell, contains the segments which derive from reduction of the entire extract. The top part of Figure 1 shows a reduction chart in which the best-scoring analysis has been selected (see below). Most of the cells of this chart are empty and those that are not contain just one segment, each containing two to four notes. Before an analysis is selected from a chart, its cells are generally fuller, and each contains a number of segments corresponding to the different ways in which a group of surface segments may be reduced. Each derived segment has an associated score, intended to suggest how likely that segment is to be a part of a complete 'good' analysis of the entire extract.

An analysis can be derived from the chart by selecting a high-scoring segment in the top cell, and then recursively selecting its highest-scoring children until a complete tree to all the segments of the surface has been derived. However, complications of context-sensitivity mean that selecting the locally highest-scoring children at each stage does not guarantee the highest-scoring complete analysis. The current procedure to ensure derivation of the highest-scoring analysis from the chart is of exponential complexity, so in some cases a chart containing information on possible analyses can be derived, but it is not practical, by current means, to derive a single best analysis from this chart.

The research reported in [6] derived some scoring mechanisms by comparing the output of the analysis-derivation software with pre-existing analyses of the same pieces. One can therefore have some confidence in the scores the software derives, but because of a lack of readily available test material, the research so far has been based on a very small quantity of music (just five short themes by Mozart). At this stage, therefore, results from research in this general area can only really be regarded as provisional.

That earlier research also showed that low-scoring possible reductions can be omitted from the chart, vastly reducing the computation time required for its derivation, without jeopardising the derivation of a good analysis. This project has used the same limits as outlined in [6]. In deriving the reduction chart, no more than 25 segments were recorded in each cell, discarding lower-scoring possibilities if necessary. In [6] scores were computed from comparison of good analyses with random analyses containing an *Ursatz* (a structure Schenker regarded as indicating a complete musical statement). In this project, the extracts of music do not constitute complete statements (most importantly they often do not end on the tonic), so new scores were computed from the same raw data from comparisons of good analyses with random analyses, regardless of the presence of an *Ursatz*. The new scores were similar to the old ones.

Small changes were also made to the set of possible atomic reductions because certain configurations not found in the five themes used in [6] were found in the material used here. An 'échappée' (a following incomplete neighbour note) elaboration was added, with tight harmonic constraints. New harmonic constraints, looser in some respects but tighter in others, were defined for some elaborations to allow situations where a dissonant note can be elaborated by 'repetition', 'delay' or 'shortening' (i.e., being preceded or followed by a rest).

4. RECOGNISING VARIATIONS

The objective of the research reported here is to explore mechanisms for recognising whether or not a passage of music is a variation of a given theme, and in particular to test whether or not a procedure using reduction yields better recognition than one relying only on the 'surface' of the music. To be precise, if a procedure which uses reductions of the theme and variations produces better results than a similar procedure which does not use reductions, then we can conclude with some confidence that the reduction software does produce useful information concerning musical structure.

4.1 Materials

The materials used in this project are encodings made by myself of four bars from the theme and most of the variations of 10 sets of variations for piano by Mozart: K. 179, 180, 264, 265, 352, 354, 398, 455, 573 and 613. These are all the sets of variations in section 26 of the Neue Mozart Ausgabe—the source used—with the exception of two sets written when Mozart was nine years old, and which cannot therefore safely be regarded as mature compositions, one set in the metre 6/8, and one which has a theme beginning and ending half-way through a bar. In all but one case it is the first four bars which are used. In K. 613 the first four bars are taken from the theme proper, which begins after an introduction. In each case the four bars form a coherent phrase. Variations in a minor key, or in a different metre from the theme, were omitted. Some small changes to the music were required in order to facilitate successful reduction by the software: all anacruses (pickups) were omitted as the reduction software cannot cope with these; all grace notes, and trills plus any terminating turn, were omitted; in a very few cases notes from some middle voices were omitted because the software operated with a limit of 4 notes in a segment; notes at the end of the last bar which clearly led into the following bar rather than belonging to the first phrase were omitted.

The encoding gave the pitch of each note (the pitch spelling of the score is used in the encoding, but pitches are converted to MIDI values in the software) and its duration. Voices are indicated, and were determined by hand when the encoding was made. This information is used only when matching surfaces as the reduction procedure changes the composition of voices.

To neutralise differences of key, each theme and variation was transposed to the key of F major, a key selected because it allowed each entire set of theme and variations to be transposed in the same direction and still remain in range for the software. It is not so simple to neutralise differences of metre, so themes in a triple metre were only compared with variations in a triple metre, and similarly for themes in a duple metre. This made a corpus in two parts, for duple and triple metres, of 5+5 themes and 41+36 variations. This is not a sufficiently large corpus for definitive results, but further materials are not readily available.

4.2 Procedure

A reduction 'chart' (i.e., a matrix of the possible reductions) was derived from each of the extracts of themes and variations, using the software as described above. (This took about 24 hours of computing time.) The best-scoring analyses were derived for each of the themes. (This was not possible for the variations because of the excessive demand of computing time in some cases.)

There has been considerable research on techniques of measuring melodic similarity (see, for example [3]), but to ask if some extract of music is a variation of another, at least in the case of 'Classical' variations as described above, is not the same as to ask if two extracts are similar. Some work in measuring melodic similarity has attempted to make use of concepts of structure from music theory [5, 7], with encouraging results. Unlike that work, the research reported here is concerned with full textures rather than just melodies, and unlike [7], which shares some of the underlying concepts of this work, the comparison method requires no manual intervention (though it does make use of an encoding which gives the key and metre). Instead a large number of methods specialised to comparing extracts to determine if one is a variation of the other, both at the surface and comparing a best analysis to a reduction chart, were implemented in software. Each method resulted in a single match value for each pair compared. If a comparison method is successful, it will consistently yield higher values for comparisons between a theme and a variation of that theme than between a theme and a variation of a different theme.

4.2.1 Comparison Methods

Similar principles were used in the design of the methods for comparing both surfaces and reductions, as follows.

1. **Pitch-matching: pitches/pitch classes.** Some methods count exactly matching pitches; some methods accept matching pitch classes (i.e., the matched pitch can be transposed up or down any number of octaves).
2. **Voices to test: all/melody+bass/melody/bass.** There are four different kinds of match under this heading: those which seek to match all notes of each segment from the theme, those which match only the melody and bass, those which match only the melody, and those which match only the bass. For reduction matches, the lowest note of a segment is taken to belong to the bass and the highest to the melody.
3. **Voice-matching: yes/no.** Some methods only accept matches of pitches in the same voice; some accept matches no matter in which voice the note occurs in the variation. The concept of voice used here is only 'melody', 'middle' and 'bass'. The middle contains all the notes which are not in the melody and bass.
4. **Match tied notes: yes/no.** Some methods seek to match only notes which are not tied to a preceding note, while others seek to match all notes.
5. **Weighting by duration: yes/no.** Some methods weight matches in proportion to the duration of the segment in the theme to be matched.
6. **Weighting by metre/level: yes/no.** In surface-matching methods, matches can or cannot be weighted by the metrical level of the beginning of the note, giving notes at the beginning of the bar the greatest weight. (The metre of a piece is specified in the encoding.) In reduction-matching methods, the corresponding weight is determined by the level of the segment in the analysis tree. Weight steadily decreases from the root to the leaves.
7. **Limiting by parent match: yes/no (reduction only).** Some matching methods for reductions limit the level of match found for child segments to be no greater than the level of match found for their parents, on the grounds that matches of children when the parent is not matched are accidental.
8. **Values: present/proportion/bar; maximum/average/score-weighted/score-weighted*2.** Different values can be recorded for any individual segment. In the case of surface matches, some methods only look for a matching pitch to be present within the time span occupied by the original pitch. In other cases, the proportion of the original time span during which a matching pitch is sounding in the variation is used. In yet others, it is sufficient merely for a matching pitch to be present somewhere with the same bar, since variations clearly sometimes involve changes in rhythm. For reduction-matching measures, a segment of the theme can be matched with up to 25 possible segments in the reduction chart for the variation. In different methods, four different values are recorded: the maximum match; the average match; the average match weighted by the score of the matching segment; and the average match weighted by the square of the score of the matching segment. Score weights are computed in relation to the maximum weight in a reduction chart so as to always fall in the range 0 to 1 and decrease exponentially in relation to decreases in score.

The combination of all these parameters results in 384 comparison methods for surfaces and 1024 for reductions. In each case, the match value for a segment is based on the number of notes from the segment of the theme which are matched in the corresponding segments of the variation, divided by the number of notes to be matched, weighted as appropriate by proportion for surfaces or score for reductions with parent-match limiting applied if appropriate. The overall result of a comparison between a theme and a variation is the average of the results from matching each segment of the theme (and its reduction, if appropriate) with the corresponding segments of the variation (and its reduction), weighted by duration and/or metre/level as appropriate.

4.2.2 Testing Methods

Every theme was compared with every variation in the same class of metre—those which were variations of this theme and those which were variations of another theme—using each of the comparison methods outlined above. Each test can be thought of as retrieval from a database using a theme as the query. A perfect response would retrieve all the variations of that theme, and none of the variations of other themes. An appropriate measure of success is therefore the F-measure, the harmonic mean of 'precision' (the proportion of correctly retrieved variations to the total retrieved) and 'recall' (the proportion of correctly retrieved variations to the total number of variations for that theme).

A simple query mechanism would retrieve all variations whose comparison with the theme yields a value above a certain threshold. Possible values for this threshold lie between the lowest value for any comparison between a theme and one of its variations, and the highest value for any comparison between a theme and a variation of a different theme. For each comparison method, the average F-measure, using each theme as a query, was computed, at each candidate value of the threshold. The best possible F-measure (on this corpus) using each comparison method was thus be computed.

An alternative test is to ask, for each variation, of which of the five candidate themes is it a variation. The simple answer would be the theme which yields the highest comparison value. This test will be called the 'recognition measure, and for each comparison method the value recorded is the percentage of variations whose theme is correctly recognised.

5. RESULTS

The main hypothesis of this study, that reduction will lead to better recognition of variations, is not confirmed by the results, as shown in Table 1. In fact twelve of the 384 methods comparing surfaces produced a better average F-measures than the best reduction-comparing method, and two produced better recognition measures. The difference is small, however. It is impossible to know without further research whether this is because the fundamental idea that variations share common reductions is mistaken, or whether it is because the reductions produced by this reduction software are incorrect. Currently there is no simple way of determining the correctness of an analysis.

The values of match between the analysis of a theme and the reductions of its variations are generally high, but they can also be high for reductions of variations of other themes. This is illustrated in Figure 3, which shows a graph of the match values for K. 265, using the best reduction-matching method (matching pitch classes from the melody and bass in the appropriate voice in the variation, but not matching tied notes; weighted by duration but not level and not limited; taking the maximum match among alternative segments). The best threshold value for this comparison method is 0.78, which causes one variation of this theme not to be recognised, and a number of false positives from variations of other themes. According to Schenkerian theory, pieces of tonal music become more alike each other the higher up the structural tree one looks, until all (proper) pieces share one of only three possible *Ursätze*. Perhaps the reduction-matching methods have been confounded by this underlying universal similarity.

The match values for surface matches are typically lower and more spread out, as illustrated in Figure 4, which shows the results for the same theme using the best surface-matching method (matching all pitch classes in the appropriate voice in the variation, including tied notes; weighted by duration but not metre; taking the proportion a pitch class is present in a segment's span). The best threshold for this method is 0.36, causing all variations of this theme to be correctly recognised but also a false positive.

	Surface methods		**Reduction methods**	
	Average F-meas.	Recog. measure	Average F-meas.	Recog. measure
Best	0.867	94.8%	0.842	90.9%
Average	0.776	74.8%	0.748	70.3%
Worst	0.540	42.9%	0.671	35.1%

Table 1. Summary results.

5.1 Factors leading to better recognition

Analysis of the results indicates that many of the factors listed above make little difference to the quality of a recognition method. One notable exception is that weighting by level in the case of reduction matches generally leads to worse results. This is consistent with the general conclusion above that reduction does not lead to better recognition of variations. Also consistent with this is a weaker result that weighting by duration does not improve recognition in the case of reduction matches, probably because higher-level segments are likely to have longer durations. In the case of surface matches, however, weighting by duration, but not by metre, leads to a slight improvement.

On average, counting a surface match simply by the presence of the required pitch or pitch class within the span of a segment gives slightly better results than measuring the proportion of the span in which it is present, and both give better results than counting matches anywhere within the bar. However, there are interdependencies among the various parameters. For example, when pitch classes are matched within voices, measuring the proportion gives consistently better results.

In the case of reduction-based methods, taking the maximum match among alternative segments yields the best results, on average. This is consistent with the idea that variations should have reductions which match the reductions of the theme. The listener hears the theme first, and so ambiguities in the structure of variations can be resolved by reference to the structure of the theme. It is therefore sufficient that there be some *possible* reduction of the variation which matches the theme.

In both surface- and reduction-based methods, the worst results come from matching only the bass, followed by matching only the melody. The difference between matching all notes and just the melody and bass is small. In every case, if pitch classes are matched, the best results come from matching them in the appropriate voices, whereas if pitches are matched, the best results come from ignoring the voice in which they occur in the variation. This might be because sometimes Mozart writes a new part *above* the melody, and in such cases the melody often occurs at its original register.

5.2 Possible Improvements

A half-way house has been tested, which looked for matches of segments at higher levels only if there was no match at a level below. However, this produced no better results than those given above. Better results might come from matching melody and bass voices separately, possibly at different levels, but this has not yet been tested.

In examination of some of the false negatives and false positives, similarities and dissimilarities are revealed in the reductions which are not present at the surface, but as yet no consistent pattern has been discerned which would lead to a consistently better variation-recognition method. It is possible that harmony should be taken into account. (Harmonic analysis is a bi-product of the reduction procedure.) Matching on harmony alone, however, would not produce good results because many of the themes have similar harmonic structures; it would have to be combined with other factors.

Overall, variation has been found to be more complicated than first thought. The quantitative results do not show reduction to reveal the relationship between theme and variations, but examination of false results suggests that further research might yet show this to be the case.

6. REFERENCES

[1] Forte, A. & Gilbert, S.E. *Introduction to Schenkerian Analysis*, Norton, New York, 1982.

[2] M. Hamanaka, K. Hirata, and S. Tojo: "Implementing 'A Generative Theory of Tonal Music'", *Journal of New Music Research*, Vol. 35 No. 4, pp. 249–277, 2007.

[3] W. Hewlett and E. Selfridge-Field (eds.): *Melodic Similarity* (Vol. 11 of *Computing in Musicology*), MIT Press, Cambridge MA, 1998.

[4] D. Jurafsky and J.H. Martin: *Speech and Natural Language Processing* (2nd edition), Pearson, Upper Saddle River NJ, 2009.

[5] P. Kranenburg, A. Volk, F. Wiering and R.C. Veltkamp: "Musical Models for Folk-Song Melody Alignment", *Proceedings of the International Symposium on Music Information Retrieval*, pp. 507–512, 2009.

[6] A. Marsden: 'Schenkerian Analysis by Computer: A Proof of Concept', *Journal of New Music Research*, forthcoming.

[7] N. Orio and A Rodà: "A Measure of Melodic Similarity Based on a Graph Representation of the Music Structure", *Proceedings of the International Symposium on Music Information Retrieval*, pp. 543–548, 2009.

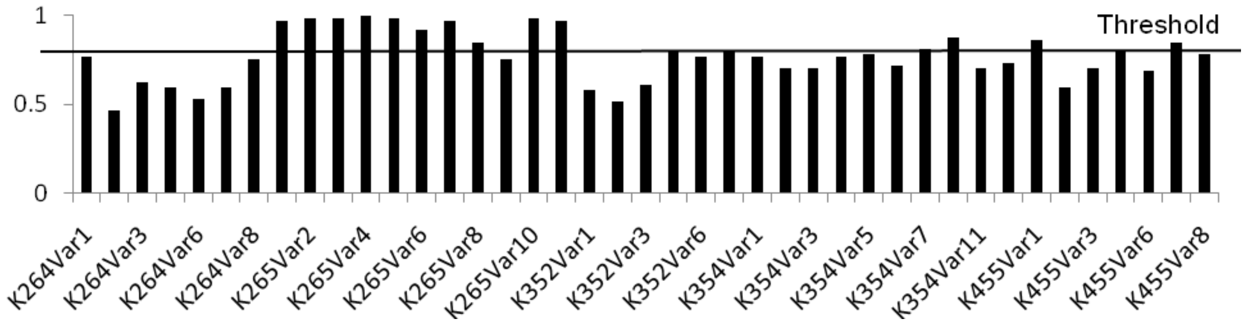

Figure 3. Match values for the theme of K. 265 using a reduction-based comparison method.

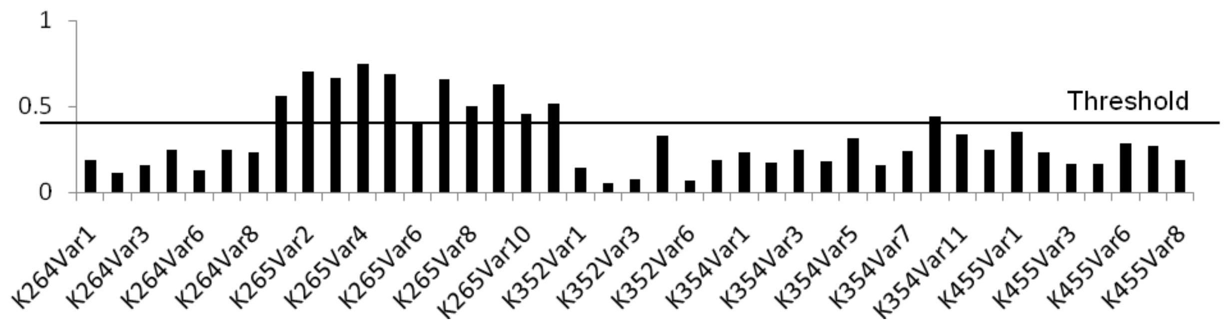

Figure 4. Match values for the theme of K. 265 using a surface-based comparison method.