Abstract

We are investigating the utility of Markov models in relation to the learning of the task of four-part harmonisation, which is a creative musical activity. A program is described which uses statistical machine learning techniques to learn this task from a suitable corpus of homophonic music. The task is decomposed into a series of more manageable sub-tasks; these are each modelled by Markov models, which can use contexts drawn from symbols describing past, current and future chords. The results of a number of initial studies, for example comparing different types of model and the effect of corpus size, are given. There is also some discussion about harmonisations that have been generated by the program by random sampling of the probability distributions in the models. Following this, a procedure for the systematic evaluation and “optimisation” of the sub-task models, involving the application of an information-theoretic measure, is presented, along with some more results. An appraisal of the procedure’s shortcomings is made, and ideas for its improvement are put forward. Finally, an indication of the future direction of the work (which is currently in its early stages) is given.

Keywords: Machine learning, harmonisation, Markov models, evaluation.

1 Introduction

The current work contributes to the study of computational creativity in music by seeking to improve the computational simulation of four-part harmonisation, which is a creative human activity. We are investigating the utility of Markov models in relation to the learning of this task, in pursuit of which a program is being developed which seeks to learn a style of four-part harmonisation from a suitable corpus of homophonic music by means of statistical machine learning techniques. There are two primary motivations for this work. The first is that the models so produced should be of use in addressing stylistic and musicological issues, and the second is that the models could form the basis of a cognitive model of harmonic perception (Pearce et al., 2002). It is intended that the program will eventually handle music which is not completely homophonic, such as the set of chorale melodies harmonised by J. S. Bach.

The approach adopted breaks the overall harmonisation task into a series of more manageable sub-tasks (Hild et al., 1992; Allan, 2002). The sub-tasks implemented so far relate to the harmonic function of the chords; sub-tasks that for example assign particular notes to the alto, tenor and bass parts will follow in due course. An information-theoretic measure is used to evaluate the various sub-task models (Allan and Williams, 2005), and a systematic approach to evaluation is taken which guides the creation of the models (Pearce, 2005). Having learned a style of harmonisation in this way, the program is equipped to generate harmonisations to “unseen” melodies.

It should be noted that this is very much “work in progress”; indeed, work at a very early stage. The results obtained so far can only therefore be taken as indicative, rather than definitive.

2 Corpus, Test Data and Input

The training corpus currently comprises fifteen major key hymn tunes with harmonisations taken from Vaughan Williams (1933) and Nicholson et al. (1950). In addition, there are two sets of five major key test melodies with harmonisations from the same sources (although only three of these melodies are used for initial studies).

All of the training corpus melodies, along with a certain amount of annotation (Ponsford et al., 1999), are assigned to a single sequence (array) of symbols. Annotation symbols are placed at the beginning and end of each melody, and at phrase boundaries. All of the training corpus bass parts are assigned to another symbol sequence in a completely analogous way. Melody and bass notes are normalised to seven scale degree symbols which are not dependent upon key; for example a tonic note is represented by “1” and a dominant note by “5”. These scale degrees can be chromatically altered by post-fixing with “#” or “b” (sharp or flat). No account is currently taken of
note duration or metrical importance.

Symbols representing the harmonic function of chords in the training corpus (determined by the principal author) are assigned, along with annotation symbols, to yet another sequence. The chords are normalised to root position triads and seventh chords; for example a tonic chord is represented by “I” and a supertonic seventh chord by “ii7”. Chords containing chromatically altered notes are deemed to be triads or seventh chords in the nearest possible related key; for example if a chord containing the notes D, F♯ and A appears when the key is C major, it is represented by “V/V” (dominant of the dominant). Since this program does not keep track of modulations, the chord could, alternatively, be described as “II”, a major triad on the second degree of the original key. The latter approach requires less music-theoretic knowledge and gives an equally good description; therefore it might be adopted later. On the other hand, in their multiple viewpoint technique, Conklin and Cleary (1988) propose a viewpoint to detect modulations; something of this sort could be incorporated into the program. Note that since the hymns are homophonic (i.e., consisting of block chords only, with no extra-chord movement such as passing notes), it is simple to associate all symbols comprising a particular chord with the same array index for ease of retrieval.

Test melodies, bass parts and harmonic symbols are assigned to sequences in the same way as for the training corpus, except that they are individually assigned. Only one test melody/harmony combination is used per program run.

A second version of the harmonic symbol sequence also exists, which contains chord inversion information in the symbols. This is used for purposes of comparison.

3 The Models and Their Construction

3.1 Markov Models

Markov chains, or n-gram models, widely used in natural language processing (Manning and Schütze, 1999), are sequences of n symbols (e.g., a sequence of musical note names), where it is assumed that the probability of the last symbol appearing is dependent upon the previous n − 1 symbols (the context). This is an approximation of the probability of the last symbol being dependent upon its entire history; it can therefore be expected that the longer the context, the more accurate the transition probabilities. The size of the context is known as the order of the Markov chain; therefore for example a 2-gram (or bigram) is first-order (n.b., a Markov “chain” with no context, a unigram, is zeroth-order).

Markov chains have been used in the prediction of both harmony (Ponsford et al., 1999) and melody (Pearce, 2005). Allan (2002) tries using Markov models with additional context to predict harmonic symbols, but decides that it would be better to use hidden Markov models. The contexts he uses, however, are restricted to past and current melodic symbols, and past harmonic symbols. Bearing in mind that a composer can look ahead while harmonising a melody, we extend the scope of possible contexts to include future symbols (i.e., those to the right of the melody note currently being harmonised) wherever they exist at the various stages of the harmonisation process; for example, future melodic symbols can be used from the beginning of the process.

3.2 Smoothing

It is often the case that contexts occurring while, for example, generating a harmony to an “unseen” melody, cannot be found in the model. In order to be able to make informed predictions in these circumstances, a number of models using different contexts are used in conjunction with a smoothing technique. Conklin and Cleary (1988) and Allan (2002) use a simple method, which has worked well with text, known as back-off smoothing (Katz, 1987). For Markov chains, a number of models of different order are created. When searching for a particular context, the highest order model is checked first. If the context is not found, the second-highest order model is checked, and so on, until the (progressively smaller) context is found. For the “mixed” contexts that we are using, however, back-off is not limited to successive contexts of decreasing size; different contexts of the same overall size can be involved in the back-off sequence.

It should be noted that an escape method is normally associated with back-off smoothing, whereby some of the probability mass is assigned to unseen Markov chains, the probabilities of which are estimated during back-off. No such escape method has so far been implemented here; if a context is not found in a particular model, then the model is discarded. The next model to be tried is assumed to contain all of the probability mass.

There are other more complicated smoothing methods that are also known to work well, but one of the purposes of this paper is to investigate how an optimum, or close to optimum, back-off sequence can be determined, and how well it performs.

3.3 Decomposition of the Harmonisation Task

Allan (2002) follows Hild et al. (1992) in breaking up the harmonisation task into three sub-tasks, in the following order: “harmonic skeleton”, “chord skeleton” and “ornamentation”. In the “harmonic skeleton” sub-task, harmonic symbols are assigned to each beat; the actual notes of each chord are filled out in the “chord skeleton” sub-task; and additional notes such as passing notes are supplied during “ornamentation”. Wiggins (1998) suggests that the problem should be broken up even further, by for example choosing cadences first.

Here, we break up the task of assigning harmonic symbols to melody notes as much as possible. The training corpus has been augmented with symbols representing root position triads and seventh chords (see Section 2). The program assigns these root position harmonic symbols on a first pass, and then bass note scale degree symbols during a second pass, thereby effectively assigning chord and inversion information. This is preferred to the use of harmonic symbols incorporating inversion information, in keeping with the dictum of using as little music-theoretic knowledge as possible in the training of such models. Knowledge about inversions is effectively induced from the training corpus as the models learn the
sequential structure of the given symbolic representations. Having said that, the corpus has also been augmented with harmonic symbols which incorporate chord inversion information (see Section 2), for purposes of comparison. It should be noted that in this case the appropriate bass note scale degrees still need to be induced from the corpus, since no logic has been encoded to interpret the harmonic symbols; however, they can be determined with a probability of 1.0 during the second pass for all harmonic symbols that have been “seen” in the corpus.

Each of the two passes is broken up into three sub-tasks, each requiring its own model. The first deals with cadence chords, which have been taken to be the last two chords in each phrase; the second is concerned with all chords except the final three in each phrase; and the third specialises in the antepenultimate chord of each phrase, with the objective of knitting together the bulk of the harmony in each phrase with its cadence.

3.4 Model Construction

Cadences are important structural components of music, and as such deserve particular attention. The idea here is to create models of symbols used at cadences, with the intention not only of generating acceptable cadences in isolation, but also generating appropriate sequences of cadences (e.g., avoiding too many cadences of the same type following each other). The sub-task which predicts root position harmonic symbols for melody notes in cadential positions will be used to exemplify the construction of models.

The first step is to create a sequence from pairs of melodic scale degree symbols at the end of each phrase, along with all of the annotation symbols. Harmonic symbols are treated in a completely analogous way; thus sequences containing only symbols at cadential positions (plus annotations) have been formed.

Markov models are then constructed using maximum likelihood estimation (Manning and Schütze, 1999), recording the probability of a particular harmonic symbol appearing given a particular context. The context is taken from the cadential symbol sequences created above. A series of models is created, each successive one having either a different context of the same size or a smaller context compared with the one before, going all the way down to no context at all (zeroth-order overall).

Models for the other five sub-tasks are constructed in a similar way. The main difference is that for the four models not directly concerned with cadences, all of the input symbols are used (even for the pre-cadence model, in spite of being used for only one chord per phrase).

To begin with, the back-off sequences were fairly arbitrary, consisting mainly of a succession of ever-smaller contexts. Some initial studies were performed using these back-off strategies, before using an information-theoretic measure to guide the construction of back-off sequences that are likely to be much closer to optimal.

4 Evaluation Study

4.1 Evaluation Method

Some initial studies were performed using arbitrary back-off strategies. The same information-theoretic measure used later to improve these strategies, cross-entropy, was used here both to evaluate models, using test data, and to compare generated harmonies. If we define $P_m(S_i|C_{i,m})$ as the probability of the $i^{th}$ musical symbol given its context for a particular model $m$, and assume that there are a total of $n$ sequential symbols, then cross-entropy is given by $-(1/n) \sum_{i=1}^{n} \log_2 P_m(S_i|C_{i,m})$. It is a useful measure, because it gives a “per symbol” value; it can therefore be used to compare sequences of any length. All else being equal, the model which assigns the lowest cross-entropy to a test sequence is the best descriptor of the data (Allan, 2002). This is equivalent to saying that the model which assigns the highest geometric mean of the conditional probabilities in a test sequence is the best descriptor of the data.

4.2 Description of Studies

The studies compared two broadly different model types; one using root position harmonic symbols, and the other employing harmonic symbols incorporating inversion information. (Note that in each case, two passes of a melody are required: one to predict harmonic symbols and the other to predict bass note scale degree symbols. See Section 3.3). Within each of these models, the effect of corpus size on both test data and generated harmony was investigated. Three hymns were used as test data: St. Anne, Winchester Old and Winchester New. The melodies of these same three hymns were also used in the random generation of harmony (i.e., these melodies were automatically harmonised by the program, sampling from the statistical models created from the training corpus). Each melody was harmonised ten times, and an overall cross-entropy for the three melodies was calculated for each pass (i.e., for allocation of harmonic symbols and bass note scale degree symbols). The results are summarised in Table 1.

4.3 Results

In order to achieve a degree of brevity in the following discussion, the model using root position harmonic symbols will be referred to as the root position model, and the one using harmonic symbols incorporating inversion information will be referred to as the inversions model. It can be seen from Table 1 that the root position model better describes the test data (i.e., it has lower overall cross-entropies), in spite of the inversions model having a cross-entropy of zero for the second pass. It should be borne in mind, however, that the arbitrary back-off strategies might have favoured one model over the other; therefore a new comparison should be made after optimising (as far as is practical) the back-off sequences for each model. A start will be made on this in the next section.

It is to be expected that models benefit from larger training corpora, since more previously “unseen” contexts
are likely to come to light with each addition to the corpus. The figures in the table bear this out; but it should be noted that although in general there is an improvement between corpus sizes of ten and fifteen hymns, it is very small compared with the marked improvement between five and ten hymns. One possible reason for this is that most of the “knowledge” concerning the basic harmonisation of homophonic hymns can be garnered from a relatively small corpus. Another possibility is that although there might be additional information in larger corpora, the system is, for one reason or another, unable to model it. A third possibility is that the additional five hymns introduce new structure that is mostly not found in the test data, resulting in only a very small improvement. The latter possibility can easily be tested by further increasing the corpus size, using a larger or different set of test data, or all of the above. One way of testing the other two possibilities is to repeat the studies after completion of the improved back-off strategies.

There is a large disparity between the cross-entropy figures for test data and generated harmony; the figures for generated harmony are much lower. We shall look at this issue in more detail in Section 5.

### 4.4 Discussion of Generated Harmony

Having harmonised a melody, the program outputs triples of melody note scale degree, harmonic symbol and bass note scale degree. In order to listen to realisations of these basic outputs, MIDI files are created “by hand”. Wherever a chord chosen by the program is in agreement with the hymnal harmony, the same arrangement of notes within the chord as appears in the hymnal is used in the MIDI file, provided that the usual voice-leading rules can be satisfied.

The harmonies are often quite good, especially the ones with lower cross-entropies. As an indication of this (using the root position model) the harmonies with the lowest cross-entropy for Winchester Old (five- and ten-hymn corpora) and Winchester New (fifteen-hymn corpus: see Figure 1) contain between 54 and 57 percent of chords that are exactly the same as in the hymnal. Other chords are good substitutions (e.g., the antepenultimate chord in Figure 1, a root position subdominant, is a perfectly good alternative to the first inversion supertonic seventh in the hymnal), while yet others sound distinctly out of place (e.g., the second inversion chord at the end of the first full bar of Figure 1). In contrast, the higher cross-entropy harmonies (see Figure 2) sound further removed from the style of the corpus, often containing chord progressions that are displeasing to the ear (e.g., the first two chords of the second full bar of Figure 2 inescapably contain parallel octaves, and the progression from the last chord of the third full bar of Figure 2 to the following chord sounds far from smooth).

### 4.5 Ten-fold Cross-validation

In addition to the above studies, a ten-fold cross-validation was performed for both the root position and inversions models. Each of the hymns in turn was removed from the ten-hymn corpus and used as test data for a model trained using the remaining nine hymns. The results are summarised in Table 2. The most obvious difference between the two models is the number of harmonic symbols encountered in the test data that had not been observed in the training corpus; there are far fewer such symbols associated with the root position model. This is unsurprising, bearing in mind that there are far fewer root position harmonic symbols to discover.

Overall, the figures confirm the earlier finding that the root position model gives a better description of the test data (of course, the same proviso about the arbitrary back-off strategies applies). In this case, it is due at least in part to the greater number of unseen harmonic symbols, which are assigned a very low probability. In addition, when an unseen harmonic symbol is encountered, the model has little or no idea which bass note scale degree to assign to that chord, resulting in another low probability. The final thing to note from this table is that overall, there is a narrower spread of cross-entropies for the root position model. For example, the difference between the highest and lowest cross-entropies for the root position model is 0.823, compared with 1.018 for the inversions model; this superior consistency might also be an indication that the root position model is better.

### 5 Systematic Improvement of Back-off Strategies

It should be noted from the beginning that the method described below is not guaranteed to produce an optimal back-off sequence; the method is less than rigorous in several respects. It is, however, expected to result in a close to optimal solution.
Figure 1: Low cross-entropy harmonisation of Winchester New, using the root position model trained on a fifteen-hymn corpus, with an arbitrary back-off strategy

Figure 2: High cross-entropy harmonisation of Winchester New, using the root position model trained on a fifteen-hymn corpus, with an arbitrary back-off strategy

Table 2: Cross-entropies from ten-fold cross-validation: inversions and root position models

<table>
<thead>
<tr>
<th>Hymn</th>
<th>Inversions model</th>
<th>Root position model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pass 1</td>
<td>Pass 2</td>
</tr>
<tr>
<td>1</td>
<td>2.708</td>
<td>0.326</td>
</tr>
<tr>
<td>2</td>
<td>1.141</td>
<td>0.099</td>
</tr>
<tr>
<td>3</td>
<td>2.610</td>
<td>0.515</td>
</tr>
<tr>
<td>4</td>
<td>2.260</td>
<td>0.372</td>
</tr>
<tr>
<td>5</td>
<td>1.753</td>
<td>0.099</td>
</tr>
<tr>
<td>6</td>
<td>2.281</td>
<td>0.127</td>
</tr>
<tr>
<td>7</td>
<td>1.740</td>
<td>0.080</td>
</tr>
<tr>
<td>8</td>
<td>1.460</td>
<td>0.078</td>
</tr>
<tr>
<td>9</td>
<td>1.089</td>
<td>0.000</td>
</tr>
<tr>
<td>10</td>
<td>1.861</td>
<td>0.302</td>
</tr>
<tr>
<td>Overall</td>
<td>1.977</td>
<td>0.227</td>
</tr>
</tbody>
</table>
5.1 Procedure Description

First of all, two new hymns are added to the test data, making five in all. Having more test data will tend to smooth out any statistical anomalies. In the first instance, the back-off strategies to be improved are those of the root position model, using a fifteen-hymn corpus. In order to set a baseline standard, cross-entropies for the two passes are initially calculated using zeroth-order models for all of the sub-tasks. Following this, for each of the sub-task models in turn, and with all of the other sub-task models set to zeroth-order, all relevant first-order contexts are evaluated (in each case backing off to zeroth-order), and then a back-off sequence is assembled for a complete first-order sub-task model. The sequence starts with the context achieving the lowest cross-entropy and ends with no context at all, which has the highest cross-entropy.

For example, the single-symbol contexts relevant to the sub-task model assigning harmonic symbols to the majority of melody notes in each phrase are the previous harmonic, previous melodic, current melodic and future melodic symbols. It turns out that the current melodic symbol achieves by far the lowest cross-entropy (which is no great surprise), followed by the past harmonic, past melodic and future melodic symbols. The back-off sequence is therefore assembled in this order, finishing off with no context. Generally speaking, there will be no back-off from the current melodic context; but on rare occasions a previously “unseen” melodic symbol might be encountered, necessitating back-off (e.g., the melodic symbol “2b” does not exist in the current corpus, but could theoretically occur).

5.2 Assumptions and Restrictions

Once all of the sub-task models have been through this process, work can begin on marshalling the second-order contexts. In order to avoid searching a huge number of contexts (the number of possible different contexts increases dramatically with overall context size) we have assumed that the best combinations (i.e., those that achieve the lowest cross-entropies) will contain the best context from the next-lowest order of model. In the case of our example sub-task model, therefore, the relevant dual-symbol contexts all contain the current melodic symbol in addition to the past harmonic, past melodic and future melodic symbols respectively. Once again, all of the other sub-task models are set to zeroth-order, and the sub-task model under consideration backs off directly from a dual-symbol context to zeroth-order. We have assumed that a larger context which subsumes a smaller one will always be better than that smaller context, even if its cross-entropy is higher, so that the new contexts need only be added to the head of the back-off chain. There is some logic to this. It is likely that smaller contexts will be matched more often than larger ones; therefore even though the resulting individual probabilities might be lower, they can easily contribute to a greater improvement in the overall cross-entropy by virtue of the number of symbols assigned. The assignment of fewer, but higher-probability symbols earlier in the back-off sequence will improve things further.

The process continues for an arbitrary number of over-all context sizes. It should be pointed out at this stage that another simplifying restriction that we have imposed is that symbols can only be added to the context such that constituent parts of the context are chains, radiating outwards from the chord which is the focus of attention at the time. For example, if the context contains harmonic symbol $i - 1$ (immediately preceding current chord $i$), and another past harmonic symbol is to be added, it must be harmonic symbol $i - 2$ in the sequence from which the context is taken.

The above assumptions, and issues around them, mean that the method is not guaranteed to produce an optimum solution. It is conceivable that larger contexts not containing the best context from the previous stage of the process could achieve a lower cross-entropy; this can be tested. Even if this were not the case, however, the poorer contexts could still contribute by being inserted somewhere further down the back-off sequence. It is also conceivable that some larger contexts that do contain the best previous context might actually contribute to a deterioration in performance. This can be guarded against by testing contexts in conjunction with the full previously assembled back-off sequence for a sub-task model.

Probably the best way of improving this “optimisation” method, however, is to calculate cross-entropies only for the assignments that a particular context makes. It will then be perfectly clear, for example, that a model assigning only one harmonic symbol to a melody note with a probability of 0.9 should be higher up the back-off sequence than a model assigning harmonic symbols to ten melody notes, each with a probability of 0.8.

5.3 Results

The results are summarised in Table 3. Evaluation is carried out using a set of test data comprising harmonisations, found in hymnals, of five different hymn tunes. These hymn tunes do not appear in the training corpus or in the test data used in the “optimisation” process; they are completely “unseen” from the model’s point of view. Cross-entropies for the test data used during the “optimisation” procedure are also shown for comparison. The generated harmony results are derived from harmonising the set of melodies in the evaluation test data, since these melodies are still “unseen” on any given run of the program.

As expected, there is a huge improvement in the model between zeroth-order and first-order. There is a much smaller, but still very significant improvement between first-order and second-order. Unfortunately, due to the fact that results have only so far been produced for three context sizes, it is difficult to judge how the trend in improvement will continue. The cross-entropy of 0.880 ("unseen" test data) for the second order model is significantly higher than the 0.529 for the model with arbitrary back-off. This latter model, however, has much longer back-off sequences in terms of overall context size (although it has far fewer instances of different contexts of the same size).

Another thing to notice is that the models predicting bass note scale degrees (pass 2) perform significantly better than those predicting harmonic symbols (pass 1). This is also true for the model with arbitrary back-off. Finally,
the cross-entropies of the generated harmony are similar to those of the test data. This is in stark contrast with the results of the model with arbitrary back-off, which shows much lower cross-entropies for the generated harmony. It seems likely that the similarity in this case, and dissimilarity in the former case, directly result from the very different back-off strategies employed.

6 Conclusions and Future Work

We have discussed ways in which Markov models used for learning the creative human task of four-part harmonisation may be improved. Some preliminary studies have been carried out, and the results presented.

The root position model describes the test data better than the inversions model when the arbitrary back-off strategies are employed; since it is possible that these strategies are biased towards the root position model, another comparison will be made using back-off strategies thought to be close to optimal for each of the models.

Models using the arbitrary back-off strategies improve significantly between corpus sizes of five and ten hymns, but only very slightly between ten and fifteen hymns. The corpus size will be increased further to test whether this is a statistical aberration or a real effect. The studies will be repeated using close to optimal back-off strategies to find out if more "knowledge" can be extracted. This time, for the generation side of the studies, hymn tunes not in the test data (nor, of course, in the training corpus) will be used, since the test data will have been employed for "optimisation" purposes.

Models using close to optimal back-off strategies improve greatly between zeroth-order and first-order. There is a much smaller, but still very significant improvement between first-order and second-order. Further work will be done on the improvement of back-off strategies, including: extending to higher-order contexts; calculating cross-entropies only for the assignments that a particular back-off strategies employ; since it is possible that these assignments are biased towards the root position model, another comparison will be made using back-off strategies thought to be close to optimal for each of the models.

Models using the arbitrary back-off strategies improve significantly between corpus sizes of five and ten hymns, but only very slightly between ten and fifteen hymns. The corpus size will be increased further to test whether this is a statistical aberration or a real effect. The studies will be repeated using close to optimal back-off strategies to find out if more "knowledge" can be extracted. This time, for the generation side of the studies, hymn tunes not in the test data (nor, of course, in the training corpus) will be used, since the test data will have been employed for "optimisation" purposes.

Models using close to optimal back-off strategies improve greatly between zeroth-order and first-order. There is a much smaller, but still very significant improvement between first-order and second-order. Further work will be done on the improvement of back-off strategies, including: extending to higher-order contexts; calculating cross-entropies only for the assignments that a particular context makes; and investigating the effectiveness of contexts not containing the previous best context.

Future work also includes comparing the performance of different arrangements of the root position model; for example reversing the passes such that bass note scale degree symbols are assigned before harmonic symbols, and assigning both symbols on a chord by chord basis rather than in two complete passes of the melody. An investigation into the effect of increasing the number of chords covered by the cadential models is also proposed, as well as ascertaining whether or not the pre-cadential models really make a useful contribution.

Consideration of the effect of metre and note duration is also planned for the not-too-distant future. Clearly, extending the system to predict actual bass notes (rather than just the scale degree), inner parts and, for example, passing notes, is a little further in the future.

References


