Development of Techniques for the Computational Modelling of Harmony

Raymond Whorley, Geraint Wiggins, Christophe Rhodes, and Marcus Pearce*

Centre for Cognition, Computation and Culture
Goldsmiths, University of London
New Cross, London SE14 6NW, UK.
*Wellcome Laboratory of Neurobiology
University College London
London WC1E 6BT, UK.
{r.whorley,g.wiggins,c.rhodes}@gold.ac.uk
marcus.pearce@ucl.ac.uk

Abstract. This research is concerned with the development of representational and modelling techniques employed in the construction of statistical models of four-part harmony. *Multiple viewpoint systems* have been chosen to represent both surface and underlying musical structure, and it is this framework, along with *Prediction by Partial Match* (PPM), which will be developed during this work. Two versions of the framework are described, starting with the strictest possible application of multiple viewpoints and PPM, and then extending and generalising a little. Some implementation details are reported, as are some preliminary results.

Key words: Harmony, statistical models, machine learning, evaluation

1 Introduction

The problem we are attempting to solve by computational means is this: given a soprano part, add alto, tenor and bass such that the whole is pleasing to the ear. This is not as easy as it might initially appear, as there are many rules of harmony to be followed, which have arisen out of composers' common practice. Rather than providing the computer with rules [1], however, we wish to investigate the process of learning such rules. The idea is to write a program which allows the computer to learn for itself how to harmonise in a particular style, by creating a model of harmony from a corpus of existing music in that style. In our view, however, present techniques are not sufficiently well developed for models to generate stylistically convincing harmonisations (or even consistently competent harmony) from both a subjective and an analytical point of view; although Allan and Williams [2] have demonstrated the potential of this sort of approach.

A means of representing music which, when combined with machine learning and modelling techniques, shows particular promise, is *multiple viewpoint systems* [3]. This framework allows us to model different aspects of the music, and then combine the individual predictions of these models to give an overall prediction. Our research aims to make a theoretical contribution to the field of computational creativity in the domain of music by extending the multiple viewpoint framework in order to cope with the complexities of harmony, such that improved computational models of four-part harmonisation can be created. This is not merely an application to harmony of the framework as it stands. This paper is concerned with two versions of the framework, beginning with a very strict application, and then extending and generalising a little.

2 Brief Description of Multiple Viewpoint Systems and Their Evaluation

See Table 1 for a list of basic and derived viewpoints (not exhaustive) and their meanings. *Basic types* are the fundamental attributes that are predicted, such as `cpitch` and `dur`. *Derived types* such as `cpint` and `dur-ratio` are derived from, and can therefore predict, basic types (in this case `cpitch` and `dur` respectively). *Threaded types* are defined only at certain positions in a sequence, determined by Boolean test viewpoints such as `tactus`; for example, (`cpitch` $\ominus$ `tactus`) has a defined `cpitch` value only on tactus beats (*i.e.*, the main beats in a bar). A *linked type*, or *product type*, is the conjunction of two or more viewpoints; for example, `dur-ratio` $\otimes$ `cpint` is able to predict both `dur` and `cpitch`. See also [3] for more details.

Table 1. Basic and derived viewpoint types (not exhaustive).

Viewpoint	Meaning	Viewpoint	Meaning
`dur`	duration of event	`barlength`	number of time units in a bar
`cont`	event continuation, or not	`phrase`	event at start or end of phrase
`cpitch`	chromatic pitch	`piece`	event at start or end of piece
`ioi`	difference in start-time	`contour`	descending, level, ascending
`posinbar`	position of event in the bar	`cpintfref`	pitch interval from tonic
`metre`	metrical importance of event	`inscale`	event in major scale, or not
`cpint`	sequential pitch interval	`dur-ratio`	sequential duration ratio
`fib`	on first beat of bar, or not	`liph`	last event in phrase, or not
`tactus`	event on tactus pulse, or not	`fip`	first event in piece, or not
`fiph`	first event in phrase, or not		

N-gram Models are Markov models employing sub-sequences of n symbols. The probability of the n^{th} symbol, the *prediction*, depends only upon the previous $n - 1$ symbols, the *context*. The number of symbols in the context is the *order* of the model. See [5] for more details.

What we call a *viewpoint model* is a weighted combination of various orders of n-gram model of a particular viewpoint type. The n-gram models can be combined by, for example, *Prediction by Partial Match* (PPM) [6]. PPM makes use of a sequence of models, which we call a *back-off sequence*, for context matching and the construction of complete prediction probability distributions. The back-off sequence begins with the highest order model, proceeds to the second-highest order, and so on. An *escape method* determines prediction probabilities at each stage in the sequence.

A multiple viewpoint system comprises more than one viewpoint. The prediction probability distributions of the individual viewpoint models are combined by employing a weighted arithmetic or geometric [10] combination technique. See [7] for more information.

Conklin [7] introduced the idea of using a combination of a *long-term model* (LTM), which is a general model of a style derived from a corpus, and a *short-term model* (STM), which is constructed as a piece of music is being predicted or generated. The latter aims to capture musical structure particular to that piece.

An information-theoretic measure, *cross-entropy*, is used to guide the construction of models, evaluate them, and compare generated harmonisations. The model assigning the lowest cross-entropy to a set of test data is likely to be the most accurate model of the data. See [5] for more details.

3 Development of the Multiple Viewpoint and PPM Frameworks

Version 1: Strict Application of Multiple Viewpoints and PPM The starting point for the definition of the strictest possible application of viewpoints is the formation of vertical viewpoint elements [8]. An example of such an element is $\{69, 64, 61, 57\}$, where all of the values are from the domain of the same viewpoint, and all of the parts (soprano, alto, tenor and bass) are represented. This method reduces the entire set of parallel sequences to a single sequence, thus allowing an unchanged application of the multiple viewpoint framework, including its use of PPM. Only those elements containing the given soprano note are allowed in the prediction probability distribution, however. This is the base-level model, to be developed with the aim of substantially improving performance.

Version 2: Dividing the Harmonisation Task into Sub-tasks In this version, it is hypothesised that predicting all unknown symbols in a vertical viewpoint element (as in version 1) at the same time is neither necessary nor desirable. It is anticipated that by dividing the overall harmonisation task into a number of sub-tasks [2] [9], each modelled by its own multiple viewpoint system, an increase in performance can be achieved. For example, given a soprano line, the first sub-task might be to generate the entire bass line. This version allows us to experiment with different arrangements of sub-tasks. For example, having generated the bass line, is it better to generate the alto and tenor lines together, or one before the other? As in version 1, vertical viewpoint elements are restricted to using the same viewpoint for each part. The difference is that not all of the parts are now necessarily represented in a vertical viewpoint element.

4 Implementation

At present, the corpus comprises fifty major key hymn tunes, and the test data five, harmonised as in [4].

The Lisp implementation of version 1 is capable of predicting or generating the attributes `dur` (note duration), `cont` (note continuation, which is the part of an already sounding note which continues to be heard when a new note is sounded) and `cpitch` (chromatic pitch) for the alto, tenor and bass parts, given the soprano. More than forty viewpoints have been implemented, and any link between two viewpoints which is capable of predicting `dur`, `cont` or `cpitch` is allowed. A modification of the feature selection algorithm described in [10], which involves ten-fold cross-validation of the corpus, is used to optimise multiple viewpoint systems for the long-term model alone, the short-term model alone, or for both together (in which case the same system is used for both). The maximum order of the n-gram models can be varied, as can the method of combining prediction probability distributions, which are initially created using PPM with escape method C. Parameters (*biases*) affecting the weighting of distributions during combination can also be varied.

Version 2 extends version 1, and is implemented as described in Section 3.

5 Preliminary Results

Table 2 shows the lowest cross-entropy version 1 multiple viewpoint systems found so far for prediction of `dur`, `cont` and `cpitch`. These are for a combination of long-term and short-term models (LTM and STM, with a cross-entropy of 4.46 bits per event), LTM only (with a cross-entropy of 4.54 bits per event), and STM only (with a cross-entropy of 6.20 bits per event), using weighted geometric combination. This confirms the findings of previous research, for example that of Pearce [10], that using both LTM and STM results in a lower cross-entropy than the use of either of them alone. What is particularly interesting, however, is the fact that the STM system does not share a single viewpoint with the LTM + STM system, and has only one viewpoint in common with the LTM system; this is in stark contrast with the substantial overlap between the LTM + STM system and the LTM system. This prompted us to try using two different multiple viewpoint systems together, one optimised for the LTM and the other separately optimised for the STM; but with a cross-entropy of 4.51 bits per event, this turned out to be not as good a model as *LS* in Table 2.

For prediction of `cpitch` only, the best version 1 LTM system found so far results in a cross-entropy of 3.29 bits per event. By comparison, the best version 2 LTM system found so far predicts the bass first (1.70 bits per prediction), followed by the alto and tenor together (1.55 bits per prediction), giving a total cross-entropy of 3.25 bits per event. For prediction of `cpitch` only, then, version 2 appears to be very slightly better than version 1. It is worth noting that the best version 2 system reflects the usual human approach to harmonisation: bass first, followed by alto and tenor together.

6 Conclusions and Future Work

We have described two versions of the multiple viewpoint framework and PPM, motivated by our aim to take account of the complexities of four-part harmony. The preliminary results weakly indicate that version 2 is better than version 1 for the prediction of `cpitch` only. They also suggest the perhaps counter-intuitive conclusion that optimising the LTM and STM together leads to a better model than optimising them separately. This latter result opens interesting routes for

Table 2. Best version 1 multiple viewpoint systems (predicting dur, cont and cpitch) for LTM + STM *(LS)*, LTM only *(L)* and STM only *(S)*.

Viewpoint	LS	L	S	Viewpoint	LS	L	S
cont ⊗ cpint	×	×		(cpintfref ⊖ fiph) ⊗ piece	×		
cont ⊗ (cpintfref ⊖ tactus)	×	×		cpitch		×	×
dur ⊗ (cpintfref ⊖ liph)	×	×		dur-ratio ⊗ (ioi ⊖ fib)		×	
cont ⊗ metre	×	×		dur-ratio ⊗ phrase		×	
dur ⊗ posinbar	×	×		dur ⊗ cont			×
cpintfref	×	×		cont ⊗ (cpitch ⊖ tactus)			×
dur ⊗ liph	×	×		inscale			×
(cpintfref ⊖ liph)	×	×		contour			×
(cpintfref ⊖ fiph) ⊗ fip	×	×		cpitch ⊗ tactus			×
cpint ⊗ cpintfref	×			cpitch ⊗ (cpintfref ⊖ liph)			×
(cpintfref ⊖ fib)	×			inscale ⊗ barlength			×
cont ⊗ (cpintfref ⊖ liph)	×			cpitch ⊗ (cpintfref ⊖ fiph)			×

further work. Finally, using the LTM alone is less good still; and the STM alone is, as expected, by far the least good model.

In the immediate future, we intend to implement other versions which push the development of the multiple viewpoint/PPM framework further.

References

1. Ebcioğlu, K.: An Expert System for Harmonizing Four-Part Chorales. Computer Music Journal, 12(3), 43–51 (1988)
2. Allan, M., Williams, C.K.I.: Harmonising Chorales by Probabilistic Inference. In: L.K. Saul, Y. Weiss, L. Bottou, editors, Advances in Neural Information Processing Systems, vol. 17. MIT Press (2005)
3. Conklin, D., Witten, I.H.: Multiple Viewpoint Systems for Music Prediction. Journal of New Music Research, 24(1), 51–73 (1995)
4. Vaughan Williams, R., editor. The English Hymnal. Oxford University Press (1933)
5. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press (1999)
6. Cleary, J.G., Witten, I.H.: Data Compression Using Adaptive Coding and Partial String Matching. IEEE Trans Communications, COM-32(4), 396–402 (1984)
7. Conklin, D.: Prediction and Entropy of Music. Master's Thesis, Department of Computer Science, University of Calgary, Canada (1990).
8. Conklin, D.: Representation and Discovery of Vertical Patterns in Music. In: C. Anagnostopoulou, M. Ferrand, A. Smaill, editors, Music and Artificial Intelligence: Proc. ICMAI 2002, LNAI, vol. 2445, pp. 32–42. Springer-Verlag (2002)
9. Hild, H., Feulner, J., Menzel, W.: Harmonet: A Neural Net for Harmonizing Chorales in the Style of J.S. Bach. In: R.P. Lippmann, J.E. Moody, D.S. Touretzky, editors, Advances in Neural Information Processing Systems, vol. 4, pp. 267–274. Morgan Kaufmann (1992)
10. Pearce, M.T.: The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition. Ph.D. Thesis, Department of Computing, City University, London (2005)