# Harmonising Melodies: Why Do We Add the Bass Line First?

**Raymond Whorley** and **Christophe Rhodes**
Department of Computing
Goldsmiths, University of London
New Cross, London, SE14 6NW, UK
{r.whorley, c.rhodes}@gold.ac.uk

**Geraint Wiggins** and **Marcus Pearce**
School of Electronic Engineering and Computer Science
Queen Mary, University of London
Mile End Road, London, E1 4NS, UK
{geraint.wiggins, marcus.pearce}@eecs.qmul.ac.uk

## Abstract

We are taking an information theoretic approach to the question of the best way to harmonise melodies. Is it best to add the bass first, as has been traditionally the case? We describe software which uses statistical machine learning techniques to learn how to harmonise from a corpus of existing music. The software is able to perform the harmonisation task in various different ways. A performance comparison using the information theoretic measure *cross-entropy* is able to show that, indeed, the bass first approach appears to be best. We then use this overall strategy to investigate the performance of specialist models for the prediction of different musical attributes (such as pitch and note length) compared with single models which predict all attributes. We find that the use of specialist models affords a definite performance advantage. Final comparisons with a simpler model show that each has its pros and cons. Some harmonisations are presented which have been generated by some of the better performing models.

## Introduction

In our ongoing research, we are developing computational models of four-part harmony such that alto, tenor and bass parts are added to a given soprano part in a stylistically suitable way. In this paper we compare different strategies for carrying out this creative task. In textbooks on four-part harmony, students are often encouraged to harmonise a melody in stages. In particular, it is usual for the bass line to be added first, with harmonic symbols such as Vb (dominant, first inversion) written underneath. The harmony is then completed by filling in the inner (alto and tenor) parts. This paper sets out to show what information theory has to say about the best way to approach harmonisation. Is adding the bass line first optimal, or is there a better approach?

In order to investigate questions such as this, we have written software based on *multiple viewpoint systems* (Conklin and Witten 1995) which enables the computer to learn for itself how to harmonise by building a statistical model from a corpus of existing music. The multiple viewpoint framework allows different attributes of music to be modelled. The predictions of these individual models are then combined to give an overall prediction. The multiple viewpoint systems are selected automatically, on the basis of minimising the information theoretic measure *cross-entropy*. We have developed and implemented three increasingly complex versions of the framework, which allow models to be constructed in different ways. The first two versions are particularly pertinent to the aims of this paper, since they facilitate precisely the comparisons we wish to make without the time complexity drawbacks of the more complex version 3. The latter is therefore not utilised in this part of our research.

The fact that the resulting models are statistical (and indeed self-learned from a corpus) means that harmonies are generated in a non-deterministic way. The harmonies are more or less probable, rather than right or wrong, with an astronomical number of ways for a melody to be harmonised from the probability distributions. Of course, there is little point in producing something novel if it is also deemed to be bad. Our aim is to hone the models in such a way that the subjective quality and style of the generated harmony is consistently similar to that of the corpus, whilst retaining almost infinite variety. In this way, the computational models can be thought of as creative in much the same way as a human composer (or at the very least that they imitate such creativity). Finding a good overall strategy for carrying out the harmonisation task is an important part of this improvement process.

## Multiple Viewpoint Systems

There follows a brief description of some essential elements of multiple viewpoint systems. In order to keep things simple we look at things from the point of view of melodic modelling (except for the subsection entitled Cross-entropy and Evaluation).

### Types of Viewpoint

*Basic* viewpoints are the fundamental musical attributes that are predicted, such as `Pitch` and `Duration`. The *domain* (or alphabet) of `Pitch` is the set of MIDI values of notes seen in the melodies comprising the corpus. A semibreve (or whole note) is divided into 96 `Duration` units; therefore the domain of `Duration` is the set of integer values representing note lengths seen in the corpus.

*Derived* viewpoints such as `Interval` (sequential pitch interval) and `DurRatio` (sequential duration ratio) are derived from, and can therefore predict, basic types (in this case `Pitch` and `Duration` respectively). A B4

(MIDI value 71) following a G4 (MIDI value 67) has an `Interval` value of 4. Descending intervals have negative values. Similarly, a minim (half note) following a crotchet (quarter note) has a `DurRatio` value of 2.

*Threaded* viewpoints are defined only at certain positions in a sequence, determined by Boolean test viewpoints such as `Tactus`; for example, `Pitch` $\ominus$ `Tactus` has a defined `Pitch` value only on `Tactus` beats (*i.e.*, the main beats in a bar).

A *linked* viewpoint is the conjunction of two or more simple (or *primitive*) viewpoints; for example, `DurRatio` $\otimes$ `Interval` is able to predict both `Duration` and `Pitch`. If any of the constituent viewpoints are undefined, then the linked viewpoint is also undefined. These are just a few of the viewpoints we have implemented. See Conklin and Witten (1995) for more information about viewpoints.

### N-gram Models

So far, *N-gram* models, which are Markov models employing subsequences of $N$ symbols, have been the modelling method of choice when using multiple viewpoint systems. The probability of the $N^{\text{th}}$ 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. Only defined viewpoint values are used in N-grams; sequence positions with an undefined viewpoint value are skipped. See Manning and Schütze (1999) for more details.

### Modelling Viewpoints

What we call a *viewpoint model* is a weighted combination of various orders of N-gram model of a particular viewpoint. The combination is achieved by *Prediction by Partial Match* (Cleary and Witten 1984). 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* (in this research, method C) determines prediction probabilities, which are generally high for predictions appearing in high-order models, and *vice versa*. If necessary, a probability distribution is completed by backing off to a uniform distribution.

### Combining Viewpoint Models

A multiple viewpoint system comprises more than one viewpoint; indeed, usually many more. The prediction probability distributions of the individual viewpoint models must be combined. The first step is to convert these distributions into distributions over the domain of whichever basic type is being predicted at the time. A weighted arithmetic or geometric (Pearce, Conklin, and Wiggins 2004) combination technique is then employed to create a single distribution. A run-time parameter called a *bias* affects the weighting. See Conklin (1990) for more information.

### Long-term and Short-term Models

Conklin (1990) 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 peculiar to that piece. Currently, the same multiple viewpoint system is used for each. The LTM and STM distributions are combined in the same way as the viewpoint distributions, for which purpose there is a separate bias (L-S bias).

### Cross-entropy and Evaluation

*Cross-entropy* is used to objectively compare the prediction performance of different models. 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 $-(1/n)\sum_{i=1}^{n}\log_2 P_m(S_i|C_{i,m})$. Jurafsky and Martin (2000) note that because the cross-entropy of a sequence of symbols (according to some model) is always higher than its true entropy, the most accurate model (*i.e.*, the one closest to the true entropy) must be the one with the lowest cross-entropy. In addition, because it is a "per symbol" measure, it is possible to similarly compare generated harmonisations of any length. Harmonisations with a low cross-entropy are likely to be simpler and more predictable to a listener, while those with a high cross-entropy are likely to be more complex, more surprising and in the extreme possibly unpleasant. See Manning and Schütze (1999) for more details on cross-entropy.

### Model Construction

Cross-entropy is also used to guide the automatic construction of multiple viewpoint systems. Viewpoints are added (and sometimes removed) from a system stage by stage. Each candidate system is used to calculate the average cross-entropy of a ten-fold cross-validation of the corpus. The system producing the lowest cross-entropy goes on to the next stage of the selection process. For example, starting with the basic system {`Duration`, `Pitch`}, of all the viewpoints tried let us assume that `ScaleDegree` lowers the cross-entropy most on its addition. Our system now becomes {`Duration`, `Pitch`, `ScaleDegree`}. `Duration` cannot be removed at this stage, as a `Duration`-predicting viewpoint must be present. Assuming that on removing `Pitch` the cross-entropy rises, `Pitch` is also retained. Let us now assume that after a second round of addition we have the system {`Duration`, `Pitch`, `ScaleDegree`, `Interval`}. Trying all possible deletions, we may now find that the cross-entropy decreases on the removal of `Pitch`, giving us the system {`Duration`, `ScaleDegree`, `Interval`}. The process continues until no addition can be found to lower the cross-entropy by a predetermined minimum amount. When selection is complete, the biases are optimised.

## Development of Multiple Viewpoints

The modelling of melody is relatively straightforward, in that a melody comprises a single sequence of non-overlapping notes. Such a sequence is ideal for creating N-grams. Harmony is much more complex, however. Not only does it consist (for our purposes) of four interrelated parts, but it usually contains overlapping notes. In other words, music is usually not homophonic; indeed, very few of the major key hymn tune harmonisations (Vaughan Williams 1933) in our corpora are completely homophonic. Some preprocessing of the music is necessary, therefore, to make it amenable to modelling by means of N-grams. We use *full expansion* on our corpora (corpus 'A' and corpus 'B' each contain fifty harmonisations), which splits notes where necessary to achieve a sequence of block chords (*i.e.*, without any overlapping notes). This technique has been used before in relation to viewpoint modelling (Conklin 2002). To model harmony correctly, however, we must know which notes have been split. Basic viewpoint `Cont` is therefore introduced to distinguish between notes which are freshly sounded and those which are a continuation of the preceding one. Currently, the basic viewpoints (or attributes) are predicted at each point in the sequence in the following order: `Duration`, `Cont` and then `Pitch`.

### Version 1

The starting point for the definition of the strictest possible application of viewpoints is the formation of vertical viewpoint elements (Conklin 2002). An example of such an element is $\langle 69, 64, 61, 57\rangle$, where all of the values are from the domain of the same viewpoint (*i.e.*, `Pitch`, as MIDI values), and all of the parts (SATB) 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

In this version, it is hypothesised that predicting all unknown symbols in a vertical viewpoint element at the same time is neither necessary nor desirable. It is anticipated that by dividing the overall harmonisation task into a number of subtasks (Allan and Williams 2005; Hild, Feulner, and Menzel 1992), each modelled by its own multiple viewpoint system, an increase in performance can be achieved. Here, a subtask is the prediction or generation of at least one part; for example, given a soprano line, the first subtask might be to predict the entire bass line. This version allows us to experiment with different arrangements of subtasks. 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.

## Comparison of Subtask Combinations

In this section we carry out the prediction of bass given soprano, alto/tenor given soprano/bass, tenor given soprano, alto/bass given soprano/tenor, alto given soprano, and tenor/bass given soprano/alto (*i.e.*, prediction in two stages), in order to ascertain the best performing combination for subsequent comparisons. Prediction in three stages is not considered here because of time limitations.

Earlier studies in the realm of melodic modelling revealed that the model which performed best was an LTM updated after every prediction in conjunction with an STM (a *BOTH+* model) using weighted geometric distribution combination. Time constraints dictate the assumption that such a model is likely to perform similarly well with respect to the modelling of harmony. In addition, only corpus 'A', a bias of 2 and an L-S bias of 14 are used for viewpoint selection (as for the best melodic BOTH+ runs using corpus 'A'). As usual, the biases are optimised after completion of selection. Here, we predict `Duration`, `Cont` and `Pitch` together (*i.e.*, using a single multiple viewpoint system at each prediction stage). We also use the *seen* `Pitch` domain at this juncture (*i.e.*, the domain of `Pitch` vertical viewpoint elements seen in the corpus, as opposed to all possible such elements).

It is appropriate at this point to make some general observations about the bar charts presented in this paper. Comparisons are made for a range of $\hbar$ (maximum N-gram order) from 0 to 5. Each value of $\hbar$ may have a different automatically selected multiple viewpoint system. Please note that all bar charts have a cross-entropy range of 2.5 bits/prediction, often not starting at zero. All bars have standard errors associated with them, calculated from the cross-entropies obtained during ten-fold cross-validation (using final multiple viewpoint systems and optimised biases).

Figure 1 compares the prediction of alto given soprano, tenor given soprano, and bass given soprano. The first thing to notice is that the error bars overlap. This could be taken to mean that we cannot (or should not) draw conclusions in such cases; however, the degree of overlap and the consistency of the changes across the range of $\hbar$ is highly suggestive of the differences being real. A clinching quantitative argument is reserved until consideration of Figure 3. Prediction of the alto part has the lowest cross-entropy and prediction of the bass has the highest across the board. This is very likely to be due to the relative number of elements in the `Pitch` domains for the individual parts (*i.e.*, 18, 20 and 23 for alto, tenor and bass respectively). The lowest cross-entropies occur at an $\hbar$ of 1 except for the bass, which has its minimum at an $\hbar$ of 2 (this cross-entropy is only very slightly lower than that for an $\hbar$ of 1, however).

There is a completely different picture for the final stage of prediction. Figure 2 shows that, having predicted the alto part with a low cross-entropy, the prediction of tenor/bass has the highest. Similarly, the high cross-entropy for the prediction of the bass is complemented by an exceptionally low cross-entropy for the prediction of alto/tenor (notice that the error bars do not overlap with those of the other prediction combinations). Once again, this can be explained by the number of elements in the part domains: the sizes of the cross-product domains are 460, 414 and 360 for tenor/bass, alto/bass and alto/tenor respectively. Although we are not using cross-product domains, it is likely that the seen domains are in similar proportion. The lowest cross-entropies occur at an $\hbar$ of 1.

Combining the two stages of prediction, we see in Fig-

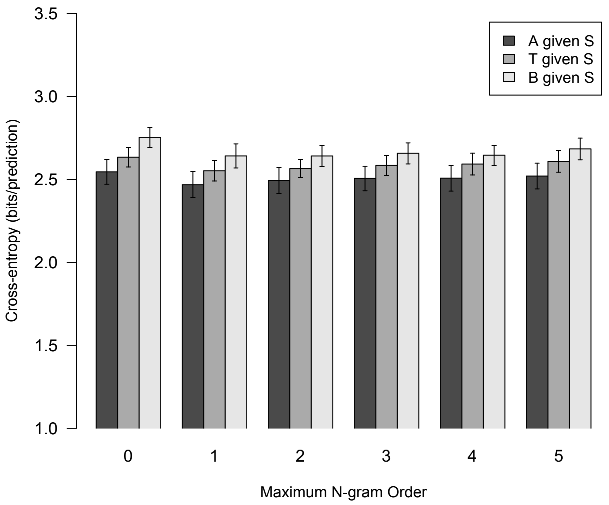

Figure 1: Bar chart showing how cross-entropy varies with $\hbar$ for the version 2 prediction of alto given soprano, tenor given soprano, and bass given soprano using the seen `Pitch` domain. `Duration`, `Cont` and `Pitch` are predicted using a single multiple viewpoint system at each prediction stage.

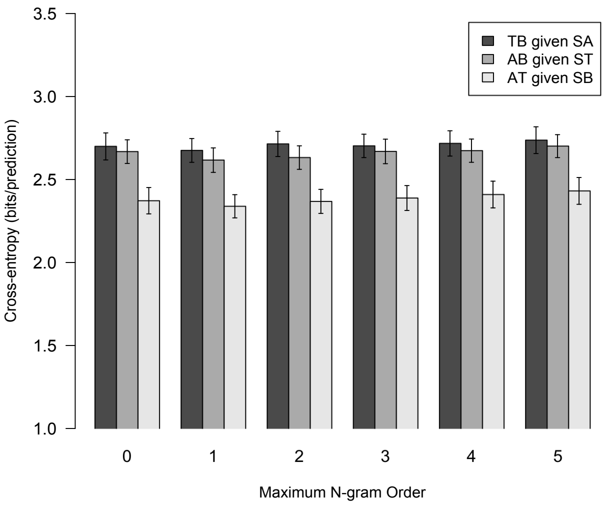

Figure 2: Bar chart showing how cross-entropy varies with $\hbar$ for the version 2 prediction of tenor/bass given soprano/alto, alto/bass given soprano/tenor and alto/tenor given soprano/bass using the seen `Pitch` domain. `Duration`, `Cont` and `Pitch` are predicted using a single multiple viewpoint system at each prediction stage.

ure 3 that predicting bass first and then alto/tenor has the lowest cross-entropy. Notice, however, that the error bars of this model overlap with those of the other models. This is a critical comparison, requiring a high degree of confidence in the conclusions we are drawing. Let us look at the $\hbar = 1$ and $\hbar = 2$ comparisons in more detail, as they are particularly pertinent. In both cases, all ten cross-entropies produced by ten-fold cross-validation are lower for B then AT than for A then TB; and nine out of ten are lower for B then AT than for T then AB. The single increase is 0.11 bits/chord for an $\hbar$ of 1 and 0.09 bits/chord for an $\hbar$ of 2 compared with a mean decrease of 0.22 bits/chord for the other nine values in each case. This demonstrates that we can have far greater confidence in the comparisons than the error bars might suggest. A likely reason for this is that there is a range of harmonic complexity across the pieces in the corpus which is reflected as a range of cross-entropies (ultimately due to compositional choices). This inherent cross-entropy variation seems to be greater than the true statistical variation applicable to these comparisons.

We can be confident, then, that predicting bass first and then alto/tenor is best, reflecting the usual human approach to harmonisation. The lowest cross-entropy is 4.98 bits/chord, occurring at an $\hbar$ of 1. Although having the same cross-entropy to two decimal places, the very best model combines the bass-predicting model using an $\hbar$ of 2 (optimised bias and L-S bias are 1.9 and 53.2 respectively) with the alto/tenor-predicting model using an $\hbar$ of 1 (optimised bias and L-S bias are 1.3 and 99.6 respectively).

Table 1 gives some idea of the complexity of the multiple viewpoint systems involved, listing as it does the first six viewpoints automatically selected for the prediction of bass given soprano ($\hbar = 2$) and alto/tenor given soprano/bass ($\hbar = 1$). Many of the primitive viewpoints involved have already been defined or are intuitively obvious. `LastInPhrase` and `FirstInPiece` are either true of false, and `Piece` has three values: first in piece, last in piece or otherwise. `Metre` is more complicated, being an attempt to define metrical equivalence within and between bars of various time signatures. Notice that only two of the viewpoints are common to both systems. In fact, of the twenty-four viewpoints in the B given S system and twelve in the AT given SB system, only five are common. This demonstrates the degree to which the systems have specialised in order to carry out these rather different tasks. The difference in the size of the systems suggests that the prediction of the bass part is more complicated than that of the inner parts, as reflected in the difference in cross-entropy.

## The Effect of Model Order

Figure 1 indicates that, for example, there is only a small reduction in cross-entropy from $\hbar = 0$ to $\hbar = 1$. The degree of error bar overlap means that even this small reduction is questionable. Is it possible that there is no real difference in performance between a model using unconditional probabilities and one using the shortest of contexts? Let us, in the first place, examine the individual ten-fold cross-validation cross-entropy values. All ten of these values are lower for an $\hbar$ of 1, giving us confidence that there is indeed a small improvement. Having established that, however, it would be useful to explain why the improvement is perhaps smaller than we might have expected.

One important reason for the less than impressive improvement is that although the $\hbar = 0$ model is nominally unconditional, the viewpoints `Interval`, `DurRatio` and `Interval` $\ominus$ `Tactus` appear in the $\hbar = 0$ multiple view-

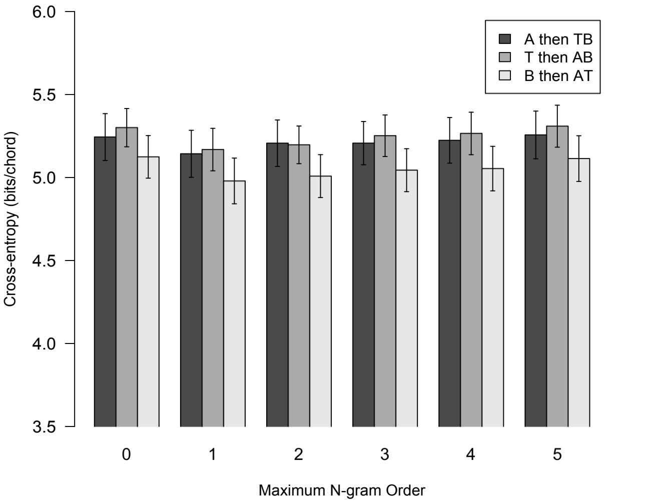

Figure 3: Bar chart showing how cross-entropy varies with $\hbar$ for the version 2 prediction of alto then tenor/bass, tenor then alto/bass and bass then alto/tenor given soprano using the seen Pitch domain. Duration, Cont and Pitch are predicted using a single multiple viewpoint system at each prediction stage.

point system (linked with other viewpoints). These three viewpoints make use of attributes of the preceding chord; therefore with respect to predicted attributes Duration and Pitch, this model is partially $\hbar = 1$. This hidden conditionality is certainly enough to substantially improve performance compared with a completely unconditional model.

Another reason is quite simply that the corpus has failed to provide sufficient conditional statistics; in other words, the corpus is too small. This is the fundamental reason for the performance dropping off above an $\hbar$ of 1 or 2. We would expect peak performance to shift to higher values of $\hbar$ as the quantity of statistics substantially increases. Supporting evidence for this is provided by our modelling of melody. Much better melodic statistics can be gathered from the same corpus because the Pitch domain is very much smaller than it is for harmony. A BOTH+ model shows a large fall in cross-entropy from $\hbar = 0$ to $\hbar = 1$ (with error bars not overlapping), while peak performance occurs at an $\hbar$ of 3.

| Viewpoint | B | AT |
|---|---|---|
| Pitch | × | |
| Interval ⊗ InScale | × | |
| Cont ⊗ TactusPositionInBar | × | × |
| Duration ⊗ (ScaleDegree ⊖ LastInPhrase) | × | × |
| Interval ⊗ (ScaleDegree ⊖ Tactus) | × | |
| ScaleDegree ⊗ Piece | × | |
| Cont ⊗ Interval | | × |
| DurRatio ⊗ TactusPositionInBar | | × |
| ScaleDegree ⊗ FirstInPiece | | × |
| Cont ⊗ Metre | | × |

Table 1: List of the first six viewpoints automatically selected for the prediction of bass given soprano ($B, \hbar = 2$) and alto/tenor given soprano/bass ($AT, \hbar = 1$).

Figure 2 reveals an even worse situation with respect to performance differences across the range of $\hbar$. For TB given SA, for example, it is not clear that there is a real improvement from $\hbar = 0$ to $\hbar = 1$. In this case, there is a reduction in five of the ten-fold cross-validation cross-entropy values, but an increase in the other five. This is almost certainly due to the fact that, having fixed the soprano and alto notes, the number of tenor/bass options are severely limited; so much so, that conditional probabilities can rarely be found. This situation should also improve with increasing corpus size.

## Separate Prediction of Attributes

We now investigate the use of separately selected and optimised multiple viewpoint systems for the prediction of Duration, Cont and Pitch. Firstly, however, let us consider the utility of creating an *augmented* Pitch domain. Approximately 400 vertical Pitch elements appear in corpus 'B' which are not present in corpus 'A', and there are undoubtedly many more perfectly good chords which are absent from both corpora. Such chords are unavailable for use when the models generate harmony, and their absence must surely skew probability distributions when predicting existing data. One solution is to use a full Cartesian product; but this is known to result in excessively long run times. Our preferred solution is to transpose chords seen in the corpus up and down, a semitone at a time, until one of the parts goes out of the range seen in the data. Such elements not previously seen are added to the augmented Pitch domain. Derived viewpoints such as ScaleDegree are able to make use of the extra elements. We shall see shortly that this change increases cross-entropies dramatically; but since this is not a like-for-like comparison, it is not an indication of an inferior model.

Figure 4 shows that better models can be created by selecting separate multiple viewpoint systems to predict individual attributes, rather than a single system to predict all of them. The difference in cross-entropy is quite marked, although there is a substantial error bar overlap. An $\hbar$ of 1 is optimal in both cases. All ten cross-entropies produced by ten-fold cross-validation are lower for the separate system case, providing confidence that the improvement is real. The lowest cross-entropy for separate prediction at $\hbar = 1$ is 5.44 bits/chord, compared with 5.62 bits/chord for prediction together. The very best model for separate prediction, with a cross-entropy of 5.35 bits/chord, comprises the best performing systems of whatever the value of $\hbar$.

## Comparison of Version 1 with Version 2

A comparison involving Duration, Cont and Pitch would show that version 2 has a substantially higher cross-entropy than version 1. This is due to the fact that whereas the duration of an entire chord is predicted only once in version 1, it is effectively predicted twice (or even three times)

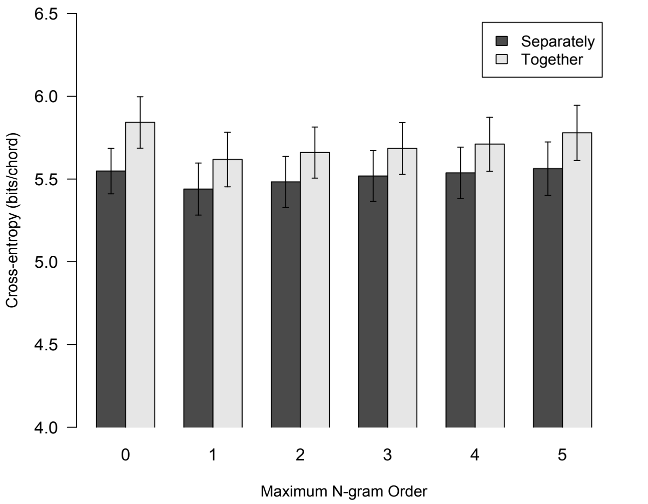

Figure 4: Bar chart showing how cross-entropy varies with $\hbar$ for the version 2 prediction of bass given soprano followed by alto/tenor given soprano/bass using the augmented `Pitch` domain. The prediction of `Duration`, `Cont` and `Pitch` separately (*i.e.*, using separately selected multiple viewpoint systems) and together (*i.e.*, using a single multiple viewpoint system) are compared.

in version 2. Prediction of `Duration` is set up such that, for example, a minim may be generated in the bass given soprano generation stage, followed by a crotchet in the final generation stage, whereby the whole of the chord becomes a crotchet. This is different from the prediction and generation of `Cont` and `Pitch`, where elements generated in the first stage are not subject to change in the second. The way in which the prediction of `Duration` is treated, then, means that versions 1 and 2 are not directly comparable with respect to that attribute.

By ignoring `Duration` prediction, and combining only the directly comparable `Cont` and `Pitch` cross-entropies, we can make a judgement on the overall relative performance of these two versions. Figure 5 is strongly indicative of version 2 performing better than version 1. Again, there is an error bar overlap; but for an $\hbar$ of 1, nine out of ten cross-entropies produced by ten-fold cross-validation are lower for version 2; and for an $\hbar$ of 2, eight out of ten are lower for version 2. The single increase for an $\hbar$ of 1 is 0.07 bits/chord, compared with a mean decrease of 0.22 bits/chord for the other nine values. The mean of the two increased values for an $\hbar$ of 2 is 0.03 bits/chord, compared with a mean decrease of 0.20 bits/chord for the other eight values.

As one might expect from experience of harmonisation, predicting the bass first followed by the alto and tenor is better than predicting all of the lower parts at the same time. It would appear that the selection of specialist multiple viewpoint systems for the prediction of different parts is beneficial in rather the same way as specialist systems for the prediction of the various attributes. The optimal version 2 cross-entropy, using the best subtask models irrespective of the value of $\hbar$, is 0.19 bits/prediction lower than that of version 1.

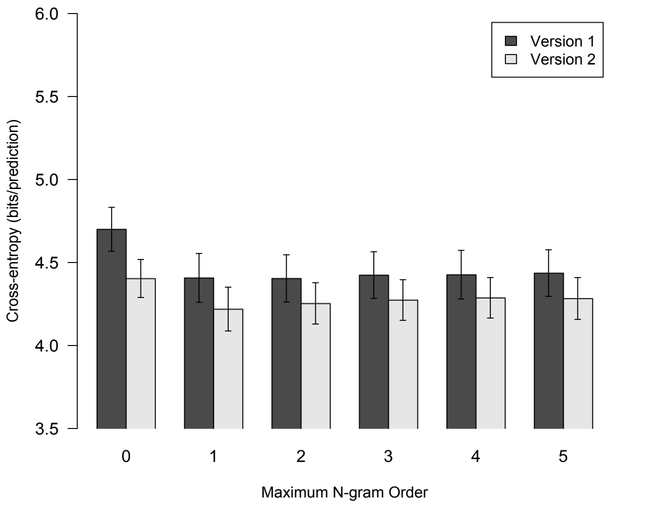

Figure 5: Bar chart showing how cross-entropy varies with $\hbar$ for the separate prediction of `Cont` and `Pitch` in the alto, tenor and bass given soprano using the augmented `Pitch` domain, comparing version 1 with version 2.

Finally, the systems selected using corpus 'A' are used in conjunction with corpus 'A+B'. Compared with Figure 5, Figure 6 shows a much larger drop in cross-entropy for version 1 than for version 2: indeed, the bar chart shows the minimum cross-entropies to be exactly the same. Allowing for a true variation smaller than that suggested by the error bars, as before, we can certainly say that the minimum cross-entropies are approximately the same. The only saving grace for version 2 is that the error bars are slightly smaller. We can infer from this that version 1 creates more general models, better able to scale up to larger corpora which may deviate somewhat from the characteristics of the original corpus. Conversely, version 2 is capable of constructing models which are more specific to the corpus for which they are selected. This hypothesis can easily be tested by carrying out viewpoint selection in conjunction with corpus 'A+B' (although this would be a very time-consuming process).

Notice that there are larger reductions in cross-entropy from $\hbar = 0$ to $\hbar = 1$ in Figure 6 than in Figure 5. The only difference between the two sets of runs is the corpus used; therefore this performance change must be due to the increased quantity of statistics gathered from a larger corpus, as predicted earlier in the paper.

## Generated Harmony

Generation is achieved simply by random sampling of overall prediction probability distributions. Each prediction probability has its place in the total probability mass; for example, attribute value X having a probability of 0.4 could be positioned in the range 0.5 to 0.9. A random number from 0 to 1 is generated, and if this number happens to fall between 0.5 and 0.9 then X is generated.

It was quickly very obvious, judging by the subjective quality of generated harmonisations, that a modification

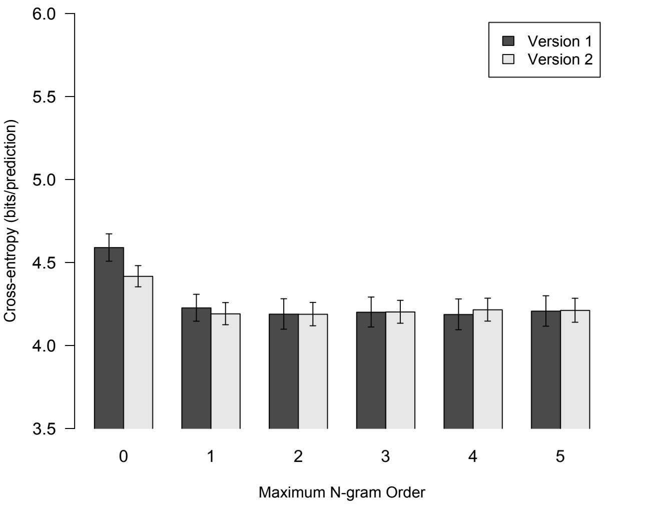

Figure 6: Bar chart showing how cross-entropy varies with $\hbar$ for the separate prediction of `Cont` and `Pitch` in the alto, tenor and bass given soprano using the augmented `Pitch` domain and corpus 'A+B' with systems selected using corpus 'A', comparing versions 1 and 2.

to the generation procedure would be required to produce something coherent and amenable to comparison. The problem was that random sampling sometimes generated a chord of very low probability, which was bad in itself because it was likely to be inappropriate in its context; but also bad because it then formed part of the next chord's context, which had probably rarely or never been seen in the corpus. This led to the generation of more low probability chords, resulting in harmonisations of much higher cross-entropy than those typically found in the corpus (quantitative evidence supporting the subjective assessment). The solution was to disallow the use of predictions below a chosen value, the *probability threshold*, defined as a fraction of the highest prediction probability in a given distribution. This definition ensures that there is always at least one usable prediction in the distribution, however high the fraction (*probability threshold parameter*). Bearing in mind that an expert musician faced with the task of harmonising a melody would consider only a limited number of the more likely options for each chord position, the removal of low probability predictions was considered to be a reasonable solution to the problem. Separate thresholds have been implemented for `Duration`, `Cont` and `Pitch`, and these thresholds may be different for different stages of generation. It is hoped that as the models improve, the thresholds can be reduced.

The probability thresholds of models used for generating harmony are optimised such that the cross-entropy of each subtask, averaged across twenty harmony generation runs using the ten melodies from test dataset 'A+B', approximately matches the corresponding prediction cross-entropy obtained by ten-fold cross-validation of corpus 'A+B'.

One of the more successful harmonisations of hymn tune *Das walt' Gott Vater* (Vaughan Williams 1933, hymn no. 36), automatically generated by the best version 1 model with optimised probability threshold parameters, is shown in Figure 7. It is far from perfect, with the second phrase being particularly uncharacteristic of the corpus. There are two parallel fifths in the second bar and another at the beginning of the fourth bar. The bass line is not very smooth, due to the many large ascending and descending leaps.

One of the more successful harmonisations of the same hymn tune, automatically generated by the best version 2 model with optimised probability threshold parameters, is shown in Figure 8. The first thing to notice is that the bass line is more characteristic of the corpus than that of the version 1 harmonisation. This could well be due to the fact that this version employs specialist systems for the prediction of bass given soprano. It is rather jumpy in the last phrase, however, and in the final bar there is a parallel unison with the tenor. The second chord of the second bar does not fit in with its neighbouring chords, and there should be a root position tonic chord on the third beat of the fourth bar. On the positive side, there is a fine example of a passing note at the beginning of the fifth bar; and the harmony at the end of the third phrase, with the chromatic tenor movement, is rather splendid.

## Conclusion

The first set of version 2 viewpoint selection runs, for attribute prediction together using the seen `Pitch` domain, compare different combinations of two-stage prediction. By far the best performance is obtained by predicting the bass part first followed by the inner parts together, reflecting the usual human approach to harmonisation. It is interesting to note that this heuristic, almost universally followed during harmonisation, therefore has an information theoretic explanation for its success.

Having demonstrated the extent to which multiple viewpoint systems have specialised in order to carry out these two rather different prediction tasks, we use an even greater number of specialist systems in a second set of runs. These show that better models can be created by selecting separate multiple viewpoint systems to predict individual musical attributes, rather than a single system to predict them all.

In comparing version 1 with version 2, only `Cont` and `Pitch` are taken into consideration, since the prediction of `Duration` is not directly comparable. On this basis, version 2 is better than version 1 when using corpus 'A', which again tallies with human experience of harmonisation; but when corpus 'A+B' is used, their performance is identical. We can infer from this that version 1 creates more general models, better able to scale up to larger corpora which may deviate somewhat from the characteristics of the original corpus. Conversely, version 2 is capable of constructing models which are more specific to the corpus for which they are selected.

## Acknowledgements

We wish to thank the three anonymous reviewers for their constructive and insightful comments, which greatly improved this paper.

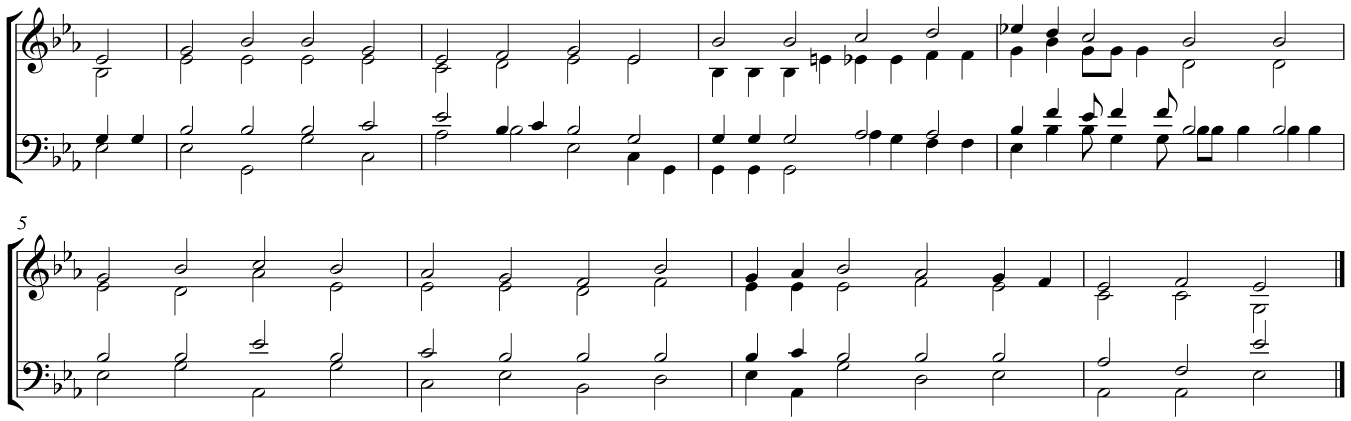

Figure 7: Relatively successful harmonisation of hymn tune *Das walt' Gott Vater* (Vaughan Williams 1933, hymn no. 36) automatically generated by the best version 1 model with optimised probability threshold parameters, using corpus 'A+B'.

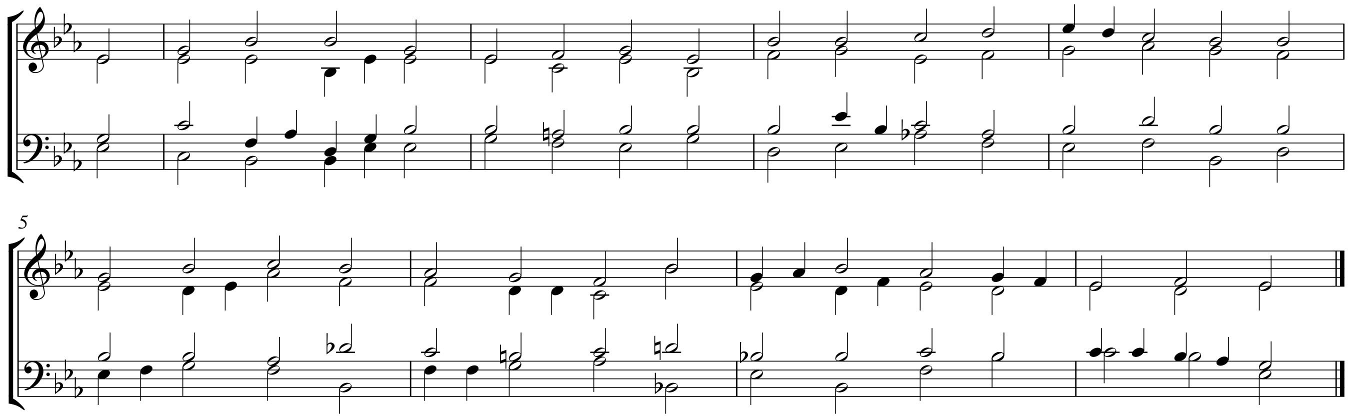

Figure 8: Relatively successful harmonisation of hymn tune *Das walt' Gott Vater* (Vaughan Williams 1933, hymn no. 36) automatically generated by the best version 2 model with optimised probability threshold parameters, using corpus 'A+B'.

## References

Allan, M., and Williams, C. K. I. 2005. Harmonising chorales by probabilistic inference. In L. K. Saul; Y. Weiss; and L. Bottou., eds., *Advances in Neural Information Processing Systems*, volume 17. MIT Press.

Cleary, J. G., and Witten, I. H. 1984. Data compression using adaptive coding and partial string matching. *IEEE Trans Communications* COM-32(4):396–402.

Conklin, D., and Witten, I. H. 1995. Multiple viewpoint systems for music prediction. *Journal of New Music Research* 24(1):51–73.

Conklin, D. 1990. Prediction and entropy of music. Master's thesis, Department of Computer Science, University of Calgary, Canada.

Conklin, D. 2002. Representation and discovery of vertical patterns in music. In C. Anagnostopoulou; M. Ferrand; and A. Smaill., eds., *Music and Artificial Intelligence: Proc. ICMAI 2002, LNAI 2445*, 32–42. Springer-Verlag.

Hild, H.; Feulner, J.; and Menzel, W. 1992. Harmonet: A neural net for harmonizing chorales in the style of J. S. Bach. In R. P. Lippmann; J. E. Moody; and D. S. Touretzky., eds., *Advances in Neural Information Processing Systems*, volume 4, 267–274. Morgan Kaufmann.

Jurafsky, D., and Martin, J. H. 2000. *Speech and Language Processing*. New Jersey: Prentice-Hall.

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

Pearce, M. T.; Conklin, D.; and Wiggins, G. A. 2004. Methods for combining statistical models of music. In *Proceedings of the Second International Symposium on Computer Music Modelling and Retrieval*.

Vaughan Williams, R., ed. 1933. *The English Hymnal*. Oxford University Press.