Modelling and Generating Music using Multiple Viewpoints

Darrell Conklin and John G. Cleary
Knowledge Sciences Institute, Department of Computer Science
The University of Calgary
2500 University Drive NW, Calgary, Canada T2N 1N4

Abstract

A technique for modelling tonal music using multiple viewpoints is described. Markov models are used with a different model for each of a number of different viewpoints: for example, the durations of notes or their relative intervals. The models are extracted from existing pieces, or can be constructed by hand. The technique is evaluated by using models to generate new pieces. Examples are given of pieces generated using models of monodic Gregorian chant and two voice polyphony.

Introduction

The composition of music with the computer requires the development of formal systems with which music can be modeled. The modelling of music presents an interesting challenge. Like many other phenomena such as written text or visual patterns it contains much regularity but also displays tremendous internal variability. Any attempt to model music must be able to capture as much regularity as possible and must also be able to accommodate the variability.

Recent research in the related problem of modelling natural (for example English) written text has shown that excellent results can be obtained with simple techniques (Bell *et al*, 1989), (Witten and Cleary, 1986), (Cleary and Witten,1984), (Cormack and Horspool, 1987). Quantitative results in these references indicate that most of the regularity of English can extracted by simple Markov models of the surface structure.

In text it is sufficient to consider a single sequence of symbols but the greater complexity of music requires that different aspects of it be modeled: for example, both the pitch and duration and in polyphony the separate voices and their relationships. This is dealt with below by the concept of multiple viewpoints, each viewpoint having its own variable order Markov model.

The rest of this paper describes how we extract and build up these models from existing pieces of music. The value and success of these models can be assessed in a number of ways. The first and most precise is to use an information theoretic measure of how well the models predict the music. This assessment was used with great success in the natural text work references above. Unfortunately, it requires large amounts of data which were not available to us. A second way is to examine the models and judge the extent to which they "explain" particular pieces of music. Unfortunately this is difficult (the models are quite opaque) and very subjective. The third technique, which we have chosen, is to use the models to compose or generate new pieces. The extent to which the new pieces are in the style of the original indicates how well the models have captured stylistic musical devices.

An experimental system for music modelling and generation called SONG/1 (Stochasticly Oriented Note Generator) has been developed in Prolog. The system consists of a control mechanism or rule interpreter and a modelling facility whereby rules are constructed from existing pieces of music. The next section describes the theory behind the design of the system and details of the construction of SONG/1. The following section shows results from modelling a Gregorian chant and a two voice piece. Finally we summarize our results and consider future work which needs to be done.

Markov Models

When describing a musical piece we assume the base symbols are discrete musical events. Each musical event is composed of two primitives: pitch and duration. Pitch is restricted to some fixed set (for example the standard 12 tone system) and duration is similarly restricted to whole, half, quarter notes *etc.* A piece of music can be represented as a single sequence of these events (monodic music) and as parallel sequences (polyphonic music). Any model of such a piece must be probabilistic to accommodate the many possible pieces. The basic model we use is the Markov model.

A Markov model predicts the probabilities of the next symbol (event) in a sequence using only some fixed number of preceding symbols to compute the probability. These symbols are referred to as the *context* and the number of them is referred to as the *order* of the model. For example, an order 2 model of English text would predict the probability of the next character in the sequence below using the last two letters "he" as the context:

```
He was rathe....
```

Such a model would probably predict a space with high probability (because of the frequency of the word "he" and "the") and "r" with a small probability. An order 5 model which had "rathe" available for consideration would predict "r" with a high probability. It seems obvious that as high an order as possible should be used for a model. However, when models are being extracted from an existing sequence a new problem arises for high order models. Most of the long contexts will never have been seen, and so there is no basis on which to make a prediction. The usual response to this is to use a default such as predicting all possible symbols with equal probability – a very poor prediction in general. A compromise which works very well for text is to use a range of models, for example, all models from order 0 to order 5 (Cleary and Witten, 1984). When making a prediction the order 5 model is first checked, if it has a prediction it is used, if not the order 4 model is checked and so on.

The technique just outlined has the property that it is adaptive and able to model any natural text, say French, rather than being restricted to a particular language or style. This is because the model is built up from the text itself as it is read. Thus the early part of text is used to predict the latter parts. The success of the approach rests on this use of adaptation, on the use of variable order models and on the acceptance that very large models need to be built up.

MARKOV MODELS OF MUSIC

The use of simple Markov models in music generation was pioneered by Hiller and Isaacson (1959) and Xenakis (1970). Their basic idea was to use Markov models constructed by hand to generate random works based on the models. This work was difficult to evaluate because it sought to establish a new style of music at the same time as it introduced a new way of composing it. In contrast, our approach is to stay within well understood and established musical traditions so that we have some hope of evaluating the value of the modelling techniques *per se*.

Subsequent work on Markov models in music has tapered off, although interest is again developing (Jones, 1981), (Holtzman, 1981), (Zicarelli, 1987). It has been realized that music is too rich a communication form to be reduced to a one-dimensional Markov process. A deeper notion and representation of musical context is needed.

In order to accommodate the richness of musical structure we have used a framework in which music is viewed from many natural facets. Each facet, or viewpoint, is given a variable order Markov model and helps to predict possible events during music generation. The individual models' predictions are combined to produce a final overall prediction. Such multiple viewpoints have been used in adaptive and learning systems for robots (Andreae, 1977), (Andreae and Cleary, 1976), and have been shown formally to have more power than single viewpoint systems (Cleary, 1980). Multiple viewpoint systems have been used with success in the generation of music (Ebcioglu, 1986).

A viewpoint is a way of looking at music and musical events, for example, a piece can be viewed as a sequence of pitches, a succession of harmonic movements, or a succession of themes or motifs. This sequence is referred to as a viewpoint chain. Each item of the chain is called a viewpoint element. For the "pitch" viewpoint, a viewpoint element might be an integer pitch number, or symbolic pitch class name and octave. For the "harmonic movement" viewpoint, an element might be a chord and inversion symbol. For the "theme" viewpoint, each element of a chain might be a symbolic motif indicator.

Several viewpoints, from among many possibilities, have been used in the results reported below, five for pitch and one for duration:

- MELODY - the absolute pitch class and octave of an event;
- MELINT - the interval an event forms with its predecessor;
- VINT - the interval between an event and another simultaneous event in a different voice;
- PITCHCLASS - the class of pitches an event belongs to;
- VOICERANGE - all pitches between a highest and lowest prediction are included in the overall viewpoint prediction;
- DURATION - the duration of an event.

Each viewpoint is defined by the mapping from event primitives (pitch and duration) to the particular viewpoint elements. During prediction the current context is mapped to the particular viewpoint and after prediction each predicted element is translated back to one or more musical event primitives. Table I shows the viewpoint chains and models that would result from two bars of Palestrina's "Pope Marcellus Mass".

The models are written in the form of probabilistic transition rules. The following rule taken from the DURATION viewpoint says that in the cases when a quarter note was followed by a half note the next duration was a half twice and a quarter once:

[quarter, half] ⇒ {half:2, quarter:1}

GOOD MODELS

In a musical context we are looking for a modelling technique which, like the text models, is universal and applicable to any musical style or system. Indeed one of the major thrusts of our work is to find the viewpoints and ways of combining them which are best able to model musical structure. This presupposes some way of measuring what is "best".

One way to do this which worked very well in the case of textual prediction was an information theoretic measure of the complexity of particular piece with respect to a particular model. The result of this was an average entropy per symbol. Another way of assessing models is to compare them with standard rules of composition and harmony. The models will be doing well if they include the rules within their predictions. Again, this needs more data than we have available. It can be very difficult to work out whether the restrictions or recommendations of a standard rule are actually encompassed within a set of Markov models.

The third way to evaluate the models is to regenerate pieces from them. This is done by taking a model, priming it with the first few musical events (in order to establish a context) and then choosing the next event from among those the model predicts, updating the context with the chosen event, choosing the next event and so on. The choices are made randomly in proportion to the probability the model predicts for them. For example, consider a rule from Table I:

[quarter, half] ⇒ {half:2, quarter:1}

This says that after a quarter note and a half note have been generated the next generated note will be a half note 67% of the time and a quarter note 33% of the time. If a prediction ever comes out empty then SONG/1 is able to backtrack and alter the choice of a previous event then continue forward with the composition. Backtracking can conceivably extend to the first event in the piece. If all primitives have been unsuccessfully tried for the first event, the system would report a failure.

Care is needed when doing such a regeneration. If the order of the model is high enough then all the predictions will have a probability of unity and what will be regenerated is a fragment of the original piece. In the results given below the order of the generating models are varied to give a range from near regeneration of the original to very novel but random compositions. The evaluation of the model is then on the basis of how good or musical the generated pieces are. If all the modeled pieces were in a particular style, the criterion should be how closely the generated piece adheres to that style.

COMBINED PREDICTIONS

A complete SONG/1 system consists of many different models including the different orders of models within a particular viewpoint as well as the different viewpoints. The predictions of each of these models must be combined to arrive at a final overall prediction. The first step is to compute a prediction set for each viewpoint. This is done as with the text models by first considering the highest order model then the next and so on.

It is not obvious how to combine the different viewpoints. Some of the desirable properties of a combination technique are:

- any primitive predicted by all viewpoints should have a very high probability (near unity) in the final prediction;
- primitives which are predicted by few viewpoints and with low probability, should have a probability of zero or near zero in the final prediction;

The simple combining technique we use is to first expand the predictions of all the viewpoints into the common form of (pitch, duration) pairs. If a particular viewpoint prediction is expanded into more than one pair then the probability is spread equally between these pairs. The probabilities from the different viewpoints are then added to give a single predicted probability for every possible (pitch, duration) pair. It was necessary to modify this simple process for the PITCHCLASS and VOICERANGE viewpoints.

Results

We turned to early music for initial modelling and generation; monodic Gregorian chant and sixteenth century polyphony. These forms were chosen for a number of reasons:

- there is a "regularity" which characterizes this music;
- there is a large body of knowledge formalizing the rules of this style, making it easy to see how well generations "followed the rules";
- the music of this period has a simple yet flexible vocabulary, unaffected by complex harmonic structures;
- after experimentation with music of these periods the framework could be enhanced and extended to handle more complex musical forms.

The intention was never to duplicate "correct" sixteenth century polyphony. The project was taken on as a creative experiment, not as a formalization and coding of strict stylistic rules.

MONODIC GENERATIONS

The model for the monodic generations presented here is a fragment of Gregorian chant. The source for the chant is (Soderlund and Scott, 1971). The chant, all of which was modeled, is reproduced in Fig. 4a. It is in the D Dorian ecclesiastical mode. A two bar motif is presented in bars 1 and 2. The motif reappears throughout the chant, separated by other material. The chant has a smooth, flowing texture and a melancholy tone.

All the monodic pieces generated from this model were primed with the initial sequence:

[(d-4,eighth), (f-4,eighth), (a-4,eighth)].

Every generation is restricted to be approximately five bars in length.

The pieces in Figures 1a, 1b and 1c used the MELODY viewpoint for pitch prediction. This constrains all notes in the generation to be within the absolute pitch domain of the chant. This pitch domain is expressed in the order 0 MELODY rule:

[] ⇒ {d-4:32, a-4:25, g-4:20, f-4:18, e-4:11, rest:9, c-4:8, c-5:6, d-5:6, a-3:1, b-3:1]}

As the highest order of match is decreased, the generations become more original. Fig. 1a, at an order of 8 for MELODY very nearly reproduces the original chant (there is always a highest order at which a modeled piece will be exactly reproduced). Note how the generation presents the original phrase, but then goes on to present a motif that does not appear until much later in the original.

At an order of 2, Fig. 1b presents some interesting original melodic movement. The characteristic smoothness of the chant is preserved. In Fig. 1c this smoothness is broken, replaced by wild, random melodic leaps because at an order of 0 any note seen in the model is a valid possibility.

Composers of chant thought in terms of smooth, vocal melodic intervallic motion, constrained within a modal system. The MELODY viewpoint will never present individual pitches that are not in the model, and will thus display limited originality. The melodic interval viewpoint is more general, and generality leads to originality, flexibility, and variation.

At high orders, the MELINT viewpoint will also reproduce the modeled piece. As we decrease the order, however, the generation quickly wanders out of the modal system. Fig. 1d demonstrates this unpleasant result. The cure for these meandering modal anomalies is to combine the MELINT viewpoint with the PITCHCLASS viewpoint in a different fashion. The predictions from MELINT were intersected with the set of predictions by PITCHCLASS so that only those pitches predicted by both viewpoints were included in the final prediction. Fig. 1e is the result. Even a generation done at order 0 displays a smooth melodic contour without notes straying out of the given PITCHCLASS. The order 0 PITCHCLASS model used for Fig. 1e is:

$[] \Rightarrow \{d,e,f,g,a,b,c\}$.

Contrast Fig. 1e with Fig. 1c, also an order 0 generation, and the power of combining viewpoints in this way will be evident.

We have a sampling problem because of the small amount of data available: roughly 60 melodic intervals, 120 notes, and 120 durations in the chant. The model for the chant was strengthened by modelling another chant, also in the D Dorian mode, and combining the models. The result of this (not reported here) show that the melodic line became more original, displaying fewer similarities with either modeled piece.

POLYPHONIC GENERATIONS

From a musician's point of view, monodic generations can only interest us to a limited degree. Polyphony poses new, unique, and difficult challenges. Experiments in polyphony were first performed using no viewpoint to monitor intervoice relationships. The unfortunate results can be seen in Fig. 2a which used the MELINT viewpoint from the chant for the pitch primitives in each voice. As each voice is completely independent any vertical interval could occur and there is much crossover between voices.

A partial cure for voice crossover is to combine the MELINT viewpoint with the VOICERANGE viewpoint. VOICERANGE was combined by intersecting prediction lists in the same way the PITCHCLASS viewpoint was. Fig. 2b is a three voice generation demonstrating this combination. Although voice crossover is now under control, the incorrect use of vertical intervals makes it clear that some viewpoint is needed to constrain the inter-voice relationships.

The next examples (Figs. 2c and 2d) were generated using the VINT viewpoint at the quarter note gradation. Part of "Oculus Non Vidit" by Orlando Lasso was used as a model of two voice counterpoint. The part modeled is given in Fig. 4b.

Fig. 2c is a generation resulting from using the VINT, PITCHCLASS, and VOICERANGE viewpoints as the pitch predictors. VINT was set to trigger at every quarter note gradation: if one looks at any vertical interval at this gradation in Fig. 2c, one finds it in the piece modeled at the same gradation. This demonstrates reasonable two voice counterpoint. The parallel octaves at the beginning were actually part of the primer.

Composers in the the sixteenth century did not use vertical intervals as the sole criterion for note selection. Smooth, flowing melodies were just as important. Fig. 2d tries to simulate this by including MELINT as well (the MELINT model was constructed by combining both voices into one model). The predictions were combined by first computing the individual viewpoint prediction sets. The VINT and MELINT prediction probabilities were then added and the result of this was finally intersected with each of the PITCHCLASS and VOICERANGE prediction sets..

Although the melodic contour of Fig. 2d is better than that of 2c., disallowed vertical intervals have crept in. Note, for example, the presence of parallel fifths, tritones, and unresolved dissonances. This makes it evident that the prediction sets of the VINT and MELINT viewpoints should also be intersected. Only those events predicted by both viewpoints should be considered valid choices.

Two voice generations are interesting, but three and four voice counterpoint allows for a richer, more complete harmonic texture. Fig. 3 demonstrates the result of allowing only consonant intervals between pairs of voices. It can be seen that this interesting harmonic texture is at the expense of melodically interesting material. The rules used in the for the generation of Fig. 3 were:

VINT: [] ⇒ {minor3:1,major3:1,perfect5:1,minor6:1,major6:1,octave:1}.
DURATION: [] ⇒ {quarter:1,eighth:1}.
PITCHCLASS: [] ⇒ {d,e,fs,g,a,b,cs}.

This example demonstrates that in a three or four voice texture, the VINT viewpoint is not powerful enough to capture the many intervoice relationships. It is evident that more complex viewpoints must be constructed. A HARMONICMOTION viewpoint would use some algorithm (Hindemith 1937, Winograd 1968) to parse the root and inversion of a chordal structure. Modulations of tonality would have to be detected, perhaps using another viewpoint. A VOICING viewpoint would take care of placement of chord tones. A NONCHORDTONE viewpoint could parse ornamenting tones, and reapply them during generation.

Discussion

Whereas the musical generations resulting from the application of this framework are interesting, they lack purpose, structure, and direction. Most notably lacking is any thematic material treatment, phrase structure, and harmonic motion. It seems that the basic Markov models need to be extended to account for effects which range over larger parts of the piece than just a few notes. The framework has trouble, for example, in capturing a piece which presents an initial motif, then reworks the motif with structural cadential points or phrase endings. An idea for tackling this problem is to use a grammar (Holtzmann, 1981), (Lerdahl and Jackendoff, 1983), (Winograd, 1968) to overlay and direct the generation process.

Other adaptive modelling techniques should also be explored, one possibility is neural networks. Multiple viewpoints lend themselves to neural network techniques of parallel, cooperative computation and weight adaptation. Also it will be interesting to see if they can model the longer range effects discussed above.

The work reported here is very preliminary although it does show that it is not implausible that a system like SONG/1 can model much of the surface structure of polyphony. The major difficulty in extending this work is getting enough musical pieces of a particular style in machine readable form. We have obtained Bach's Forty Eight, but feel the system is not yet able to handle these complex forms. Once the system is extended, we will do more extensive experiments and apply information theoretic techniques to the evaluation of models.

We are also looking at possible applications of this work. These include a real time SONG/1 system which adapts its model to what is happening around it and generates its output using this continually updated model. Another possibility is an interactive composer which is primed with pieces from a particular style. It then suggests possible compositions. It could learn from the composer by watching the changes and improvements made to its initial suggestions. The system might in this way become an "mind mirror" or apprentice to a particular composer, absorbing his particular style.

Acknowledgments

We would like to thank Tim Bell and Ian Witten for discussions about this work. Thanks also to Michael Mattson, who helped record the generations. Support was received from the Canadian Natural Sciences and Engineering Research Council.

References

Andreae, J.H. (1977) *Thinking with the Teachable Machine.* Academic Press, London, England

Andreae, J.H. and Cleary, J.G. (1976) "A New Mechanism for a Brain" *Int. J. Man Machine Studies* 8, 89-119.

Bell, T.C., Cleary, J.G, Witten, I.H. (1989) *Text Compression.* Prentice Hall (in press).

Cleary, J.G. (1980) *An Associative and Impressible Computer* Ph.D. Thesis, University of Canterbury, New Zealand.

Cleary, J.G. and Witten, I.H. (1984) "Data Compression using Adaptive Coding and Partial String Matching" *IEEE Trans Information Theory*, IT-30 (2) 306-315.

Cormack, G.V. and Horspool, R.N. (1987) "Data Compression Using Dynamic Markov Modelling" *Computer J.*

Ebcioglu, K. (1986) "An Expert System for Chorale Harmonization," *Proceedings AAAI-86 : Fifth National Conference on Artificial Intelligence*, 2:784-8.

Hiller, L. and Isaacson, L.M. (1959) *Experimental Music.* McGraw-Hill.

Hindemith, P. (1937) *The Craft of Musical Composition* : Book 1 : Theoretical Part. Schott and Co. Ltd., London.

Holtzman, S.R. (1981) "Using Generative Grammars for Music Composition," *Computer Music Journal* 5 (1):51-64.

Jones, K. (1981) "Compositional Applications of Stochastic Processes," *Computer Music Journal* 5 (2):45-61.

Lerdahl, F. and Jackendoff, R.(1983) *A Generative Theory of Tonal Music.* The M.I.T. Press.

Soderlund, G.F. and Scott, S.H. (1971) *Examples of Gregorian Chant and Other Sacred Music of the Sixteenth Century.* Prentice-Hall, Inc.

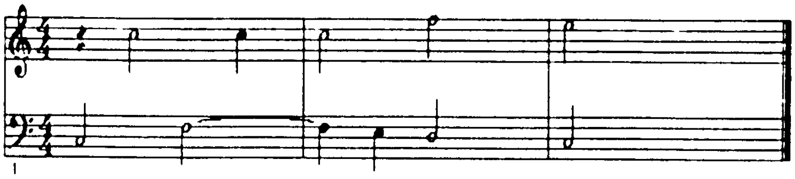
Winograd, T. (1968) "Linguistics and Computer Analysis of Tonal Harmony," *Journal of Music Theory* 12, Spring:2-49.

Witten, I.H. and Cleary, J.G. (1986) "Foretelling the Future by Adaptive Modeling," *ABACUS* 3 (3):16-36.

Xenakis, I. (1971) *Formalized Music.* Indiana University Press, 1970.

Zicarelli, D. (1987) "M and Jam Factory," Computer Music Journal 4 (11) 13-29.

Table I - viewpoint chains and models constructed from two bars of two voices in Palestrina's "Pope Marcellus Mass".



viewpoint	voice	viewpoint chain
MELODY	1	[rest, c-5, c-5, c-5, f-5, e-5]
MELODY	2	[c-3, f-3, e-3, d-3, c-3]
MELINT	1	[rest, 0, 0, +perfect4, -minor2]
MELINT	2	[+perfect4, -minor2, -major2, -major2]
VINT*	1, 2	[rest, fifth, fifth, minor3, major3]
PITCHCLASS	1	[rest, c, c, c, f, e]
PITCHCLASS	2	[c, f, e, d, c]
DURATION	1	[quarter, half, quarter, half, half, half]
DURATION	2	[half, dottedhalf, quarter, half, half]

* firing at every half note gradation, intervals scaled to be within one octave for this example

Some of the models extracted from this fragment:

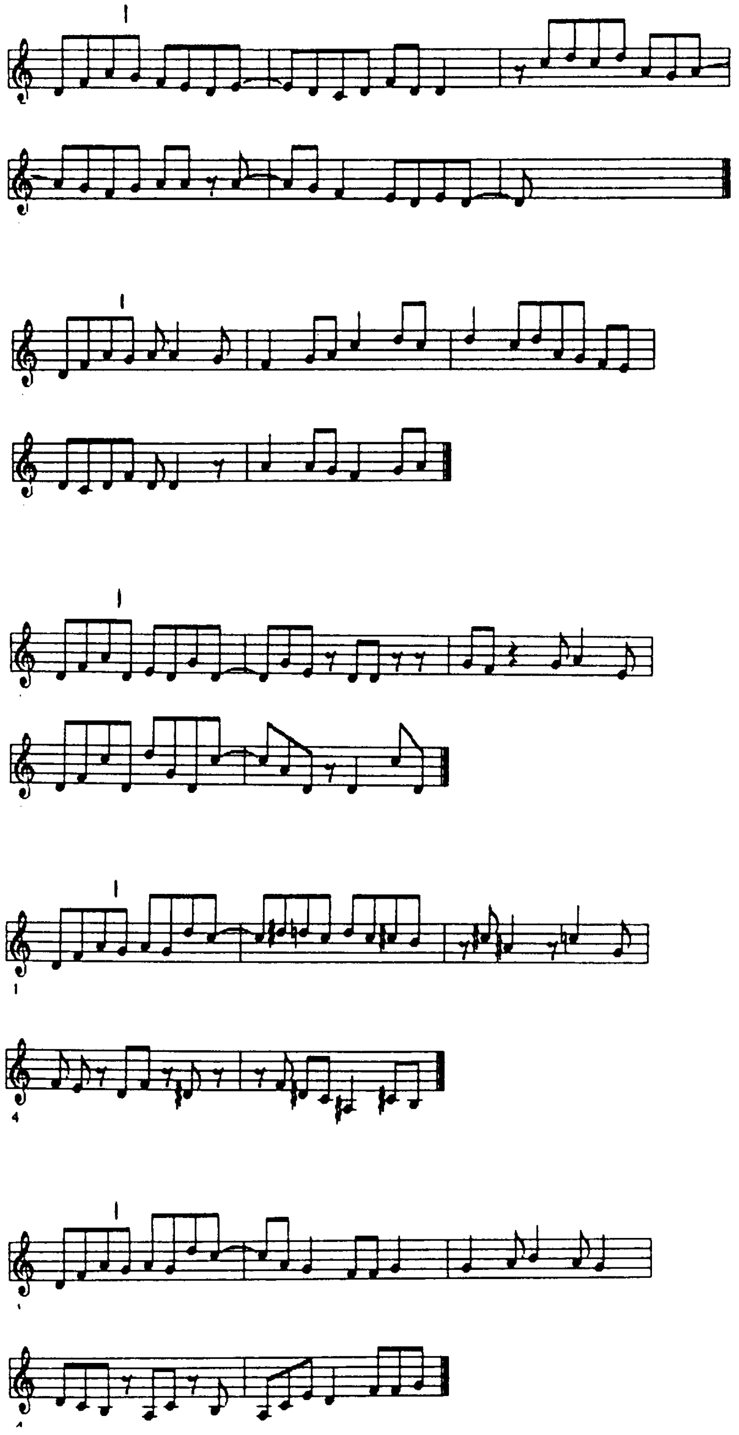
MELINT (combining both voices)

order 0:	[]	⇒ {rest:1, 0:2, +perfect4:2, -minor2:2, -major2:2}
order 1:	[rest]	⇒ {0:1}
	[0]	⇒ {0:1, +perfect4:1}
	[+perfect4]	⇒ {-minor2:2}
	[-minor2]	⇒ {-major2:1}
	[-major2]	⇒ {-major2:1}

DURATION (combining both voices)

order 0:	[]	⇒ {quarter:3, half:7, dottedhalf:1}
order 1:	[quarter]	⇒ {half:3}
	[half]	⇒ {quarter:1, half:3, dottedhalf:1}
	[dottedhalf]	⇒ {quarter:1}
order 2:	[quarter, half]	⇒ {half:2, quarter:1}
	[half, quarter]	⇒ {half:1}
	[half, half]	⇒ {half:1}
	[half, dottedhalf]	⇒ {quarter:1}
	[dottedhalf,quarter]	⇒ {half:1}

primed to |



A. melody viewpoint, order 8
duration viewpoint, order 8

B. melody viewpoint, order 2
duration viewpoint, order 2

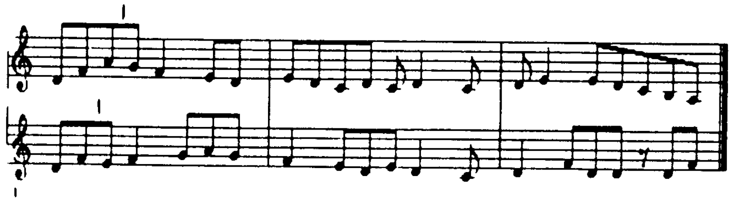
C. melody viewpoint, order 0
duration viewpoint, order 0

D. melint viewpoint, order 0
duration viewpoint, order 0

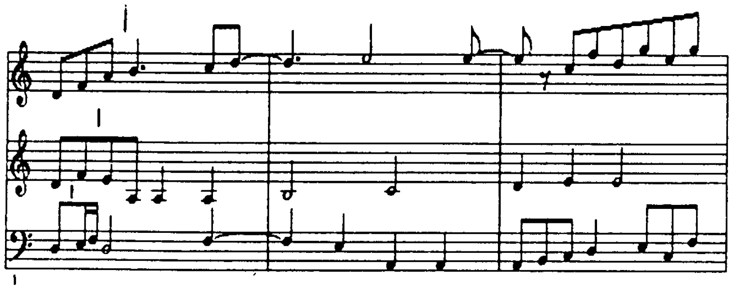
E. melint viewpoint, order 0
duration viewpoint, order 0
pitchClass viewpoint, order 0

Figure 1 : Single Voice Pieces Generated from Gregorian Chant (see Fig. 4a)

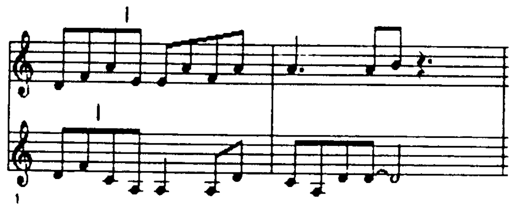
primed to |



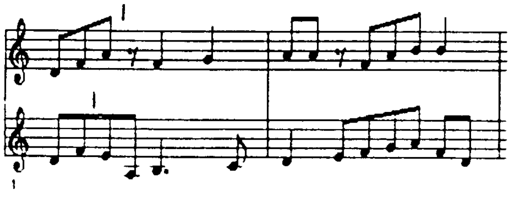
A. melint viewpoint, order 2
duration viewpoint, order 2
pitchClass viewpoint, order 0
(taken from Chant - Fig. 4a)



B. melint viewpoint, order 2
duration viewpoint, order 2
pitchClass viewpoint, order 0
pitchRange viewpoint, order 0*
(taken from Oculus non Vidit -
Fig. 4b)



C. vint viewpoint, order 2
duration viewpoint, order 2
pitchClass viewpoint, order 0
pitchRange viewpoint, order 0*
(taken from Oculus non Vidit -
Fig. 4b)

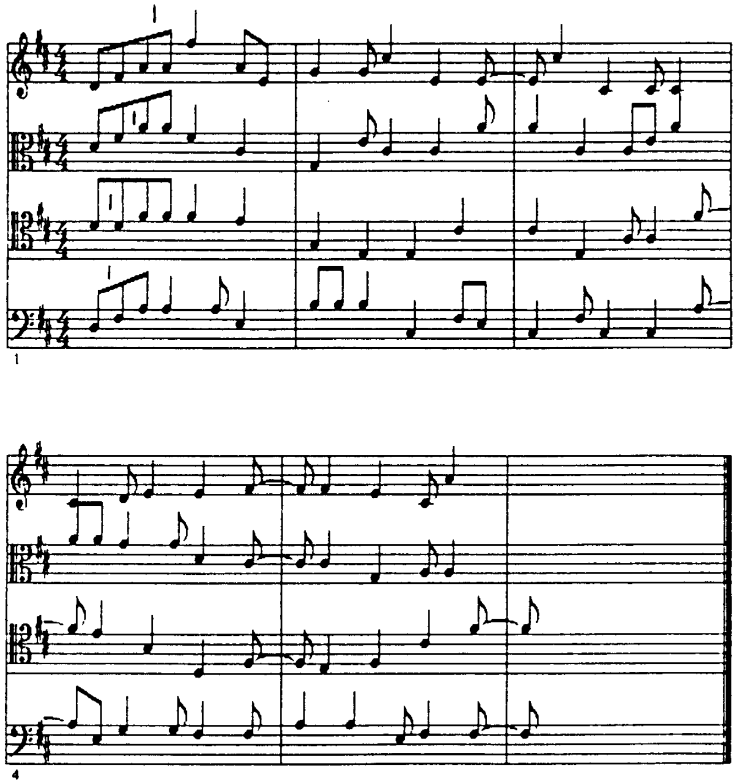


D. vint viewpoint, order 2
duration viewpoint, order 2
melint viewpoint, order 2
pitchClass viewpoint, order 0
pitchRange viewpoint, order 0*
(taken from Oculus non Vidit -
Fig. 4b)

* different model for each voice

Figure 2 : Two and Three Voice Generations.

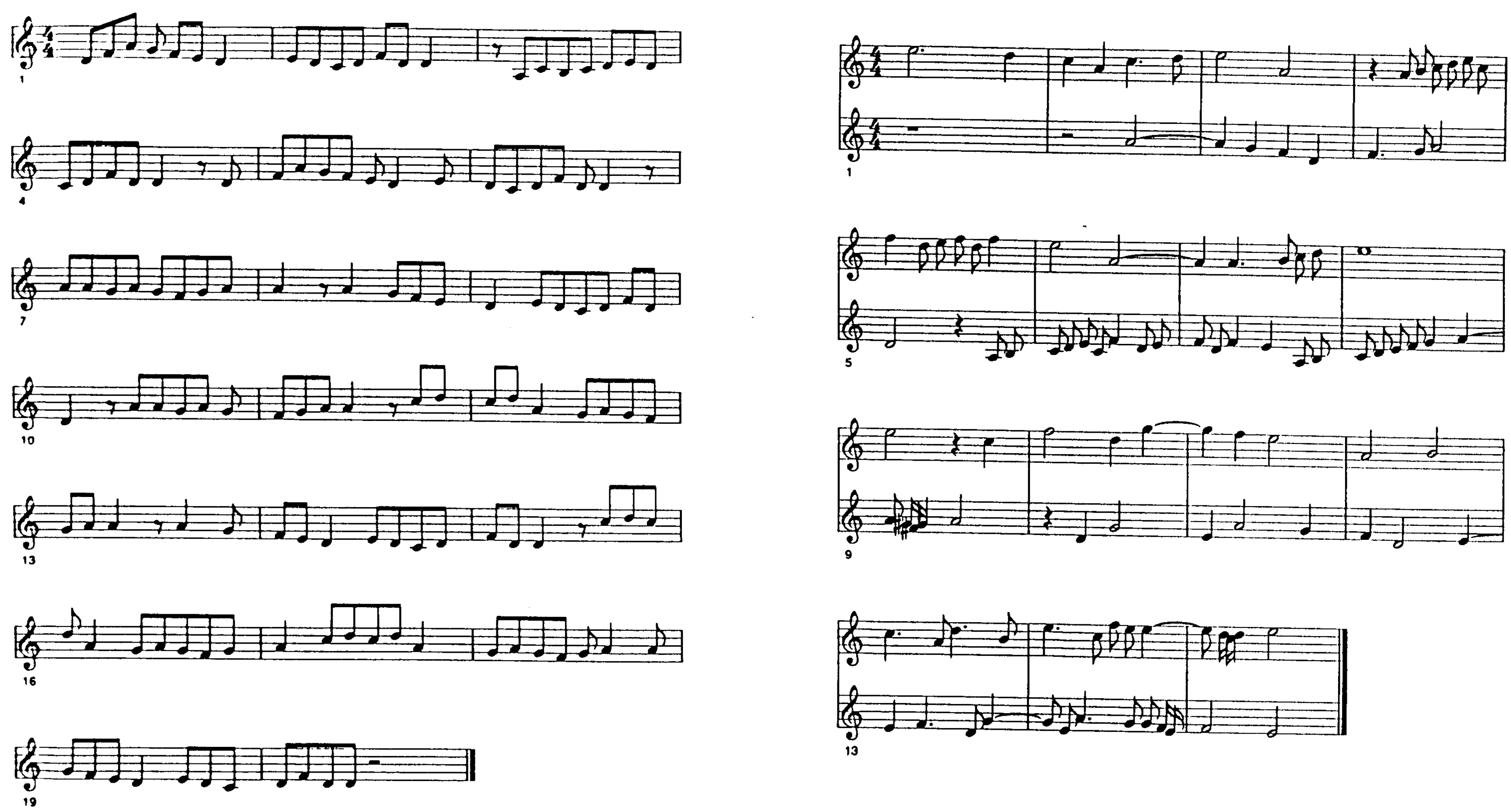
primed to |



vint viewpoint, order 0
pitchClass viewpoint, order 0
pitchRange viewpoint, order 0*
(model constructed by hand)

* different model for each voice

Figure 3 : a Four Voice generation



A. Gregorian Chant

B. Oculus Non Vidit, Lasso

Figure 4 : Modeled Pieces.