Journal of New Music Research

ISSN: 0929-8215 (Print) 1744-5027 (Online) Journal homepage: http://www.tandfonline.com/loi/nnmr20

Predicting the Composer and Style of Jazz Chord Progressions

Thomas Hedges, Pierre Roy & François Pachet

To cite this article: Thomas Hedges, Pierre Roy & François Pachet (2014) Predicting the Composer and Style of Jazz Chord Progressions, Journal of New Music Research, 43:3, 276-290, DOI: 10.1080/09298215.2014.925477

To link to this article: http://dx.doi.org/10.1080/09298215.2014.925477

Journal of New Music Research, 2014
Vol. 43, No. 3, 276–290, http://dx.doi.org/10.1080/09298215.2014.925477

Predicting the Composer and Style of Jazz Chord Progressions

Thomas Hedges[1], Pierre Roy[1] and François Pachet[1,2]

[1]*Sony Computer Science Laboratory, France ;* [2]*University Pierre et Marie Curie, France*

(Received 5 July 2013; accepted 12 May 2014)

Abstract

Jazz music is a genre that consists mainly of improvising over known tunes, represented as a *lead sheet*. This study addresses the question 'to what extent does a lead sheet carry information about its composer?' Primarily, this study considers chord progressions alone, and secondarily melodic and temporal information combined with various multiple viewpoint models. Using these classifiers, a novel subsequence selection algorithm is presented to trace stylistic similarities within a lead sheet. We conclude that composers can, to a reasonable extent, be recognized from their chord progressions, and that the consideration of melodic and temporal information improves classification accuracy by a small but statistically significant amount.

Keywords: harmony, Markov models, prediction, multiple viewpoints, jazz, classification

1. Introduction

Like most artistic activities, music composition is an intimate process in which composers use their skills and talents to express their identity. However, it is well known that music evolves not only through individuals, but proceeds in larger-scale temporal epochs. In the case of jazz, this history is widely studied and composers and styles are relatively well defined from a musicological perspective. For instance, the jazz Wikipedia page (www.wikipedia.org/wiki/jazz) lists several subgenres (or styles) of jazz, for example swing, bebop, hard bop, and Latin. Each of these genres has specific features, well-known composers and representative jazz standards. So the question 'to what extent does a jazz standard carry information about its composer?' is natural. Musicology has addressed this issue in classical music for decades, for example, the seminal work of Rosen (1971) defines the Classical style precisely by the compositions of Haydn, Mozart and Beethoven. By contrast, musicological studies in jazz typically focus on sociological issues and improvisation, with some notable exceptions such as Larson (1998) who applies Schenkerian analysis to Bill Evans improvisations, Williams (1982) who presents a comprehensive analysis of themes in the bebop style, and an analysis of early bop harmony (Strunk, 1979).

A computational study of jazz music throws up some interesting ontological problems. To a greater extent than classical music, jazz performers aim to freely reinterpret pieces depending on their skills, musical taste, audience, etc. The information that remains invariant between different interpretations *is* precisely the lead sheet. Lead sheets contain all of the information that is common to all performances of a piece: the chord progressions, main melody, time signature and performance style (e.g. medium swing, even 8ths, etc.).

The core focus of this paper is chord progressions, which hold a central role in jazz (Williams, 1982). Improvisers usually play the main melody at the beginning and end of the performance, with improvisations in the central section, but use the same chord progressions throughout the piece, both to underpin the main melody and to develop their solos. As such, the chord progressions can be considered as the fundamental element of a jazz standard.

After a review of related works (Section 2), and the presentation of a comprehensive jazz corpus (Section 3), this paper addresses the issue of identifying a composer's style computationally in the context of jazz lead sheets with quantitative machine-learning techniques. A collection of Markovian classifiers are presented and tested in Section 4, making classifications based on the maximum likelihood of chord sequences. These are contrasted with a novel subsequence matching classifier, which classifies based on the number of matching subsequences between a chord sequence and a style-specific model. Multiple viewpoint classifiers are introduced in Section 5 as Markovian-based classifiers capable of combining information from several features of musical structure, namely duration and melodic information. Applying these

Correspondence: François Pachet, Sony Computer Science Laboratory, Paris, France, 75005. E-mail: pachetcsl@gmail.com

techniques, Section 6 explores the identification of styles within the chord sequences of a single jazz standard.

2. Related works

The current study draws from works in two fields of computational musicology: the modelling of jazz as a computational object (Section 2.1), and genre classification of symbolic sequences with machine-learning techniques (Section 2.2).

2.1 Computational approaches to jazz

As a specific case of tonal music, several grammar-based approaches to jazz and improvisation have been investigated. Ulrich (1977) provides an initial system for the task of fitting melodic improvisatory material to harmonic structure. Chords are analysed functionally having been defined by a chord grammar, with tonal centres identified by preferring a minimal number of modulations. Improvisations are built from a juxtaposition of motifs taking into account the identified chord functions. However, the system lacks hierarchical structure and the quality of the improvisations suffers as a result. More promisingly, Steedman (1984) shows that 12-bar blues can be represented quite faithfully by a simple generative grammar. The hierarchical nature of the model allows a small set of six transformation rules to generate a large number of variations for the 12-bar blues. Chemillier (2004) extends Steedman's grammar to the task of real-time improvisation by identifying and precompiling cadential sequences.

Probabilistic or Markovian-based computational studies of jazz harmony and melody have also proved fruitful. In particular, Johnson-Laird's (2002) work on jazz improvisation in the field of music perception has spawned several computational models for the improvisation of melodies. Keller and Morrison (2007) investigate the use of probabilistic grammar formalisms to capture essential aspects of melodic improvisation, building from the core labelling of notes as 'chord tones', 'colour tones' and 'approach tones'. Gillick, Tang & Keller (2009) extend this approach, adding melodic contour information to the grammar. The study generates melodies in certain styles by learning style-specific grammars, building a Markovian transition matrix of one-bar abstract melodies represented as 'slope expressions' from a vocabulary of clusters identified by k-means clustering. Melodies generated by grammars inferred from three composers were received favourably in a listening test with 20 subjects who were able to correctly identify the composer grammar 90% of the time, and 95% of whom considered the melodies as 'somewhat close' or 'quite close' to their target style. In the context of music cognition of jazz harmony, Rohrmeier and Graepel (2012) assess the predictive performance of multiple viewpoint n-gram models, Hidden Markov Models (HMM), autoregressive HMMs and Dynamic Bayesian Network (DBN) models. A trigram multiple viewpoint model (Pearce, 2005) combining the dimensions of mode, chord and duration into a single probabilistic model, marginally out-performed the best DBN model which combined just mode and chord. Interestingly, further increases in predictive performance were not found by adding duration features to the DBN model, however, they still outperformed the optimum HMM and auto-regressive HMMs.

2.2 Style and genre classification

In the field of machine learning, both supervised and unsupervised techniques have been used extensively to classify various corpora of symbolic music data. A trio of studies (Conklin, 2013a; Hillewaere, Manderick & Conklin, 2009, 2012) assess the performance of various machine-learning techniques applied to folk song and dance melodies. Conklin (2013a) applies multiple viewpoint statistical modelling methods (Pearce, 2005) to classifying two corpora (Basque dance and song melodies, and European folk tunes) with respect to genre and geographical region classes. Various multiple viewpoint models combine the posterior probabilities of a class given a sequence with the geometric mean of all viewpoints. For classifying geographical regions, the best model classified 58.8%/79.2% of the Basque/European corpora correctly. For the genre classification task, the best model classified 77.6%/88.7% of the Basque/European corpora correctly. These results compare favourably to Hillewaere et al., (2009), who achieve a European folk tune genre classification accuracy of 69.7% with a Support Vector Machine classifier operating on global features. Likewise, probabilistic event-based techniques were also found to outperform various string methods (edit distances, compression distance, and string subsequence kernel methods) when classifying a similar corpus represented as sequences of melodic and inter-onset intervals (Hillewaere et al., 2012).

String compression is further explored by Cilibrasi, Vitányi and Wolf (2004) with an unsupervised clustering of rock, jazz and classical genres. The Natural Compression Distance (NCD) captures the mutual information between two strings to construct a pairwise distance matrix. The clustering is performed by a stochastic hill-climbing search with random mutation, the 'Quartet method', which attempts to find the optimum configuration of a tree structure. Clustering by genre returns results that confirm musical intuitions, however, the performance of subsequent classifications of symphonies and piano works deteriorates when the number of items clustered increases over 60.

Two studies closely related to the current paper classify jazz composers and subgenres by chord sequences. Ogihara and Li (2008) cluster jazz chord progressions by composer with a cosine similarity measure from n-gram chunks weighted by duration. They show that composers cluster relatively convincingly by date in graph and hierarchical structures, suggesting that a composer's style can be found in the chord symbols. They also invite a deeper exploration of classification by chord sequences for a larger corpus, taking into account melodic information, as well as partitioning a corpus not only by composer, but also other attributes.

Pérez-Sancho, Rizo and Iñesta (2009) classify pieces from three different genres (academic, jazz and popular) with naive Bayes and *n*-gram (Markov) classifiers. A pre-processing procedure transposes all pieces into the same key (C major/A minor) and simplifies chord types. Promising classification accuracies of 85.3% were returned for classification over the three broad genres, but the more difficult task of classifying eight subgenres spread over the three genres returned a highest accuracy of 49.8% over a baseline of 12.5%. They note with the aid of a confusion matrix that it is more difficult to classify within broad genres than between them.

2.3 Positioning of the current study

Interestingly, there have been a limited number of attempts to differentiate between a large number of composers *of the same genre* (Ogihara and Li (2008) and Pérez-Sancho et al. (2009) excepted). As noted by Pérez-Sancho et al. (2009), the task of classifying subgenres within a single genre can be considered more challenging than simply classifying between broad genres, since the similarity between two pieces in the same genre is likely to be less than for two pieces in different genres.

The current study aims to make the following specific contributions to the field. Firstly, building on the works of Ogihara and Li (2008) and Pérez-Sancho et al. (2009), this paper presents the classification of a large number of classes from several different partitionings (composer, subgenre, etc.) of a complete, closed-world corpus (Pachet, Martín & Suzda, 2013) of jazz standards. Secondly, the study assesses the impact of various representations of chord sequences on classification performance, contrasting representations presented by Pérez-Sancho et al. (2009), multiple viewpoint representations (Conklin, 2010; Pearce, 2005) and representations presented below (Section 3.2). Thirdly, this paper aims to compare the classification performance of a novel subsequence matching classifier (Section 4.3) with other traditional probabilistic classifiers (Sections 4.1, 4.2 and 5.1). Finally, the current study presents a novel algorithm for identifying style specific subsequences within a piece of music (Section 6).

3. Methodology

Style identification is explored with a series of supervised learning tasks, which involve classifying four different partitionings of a corpus.

3.1 Corpus

The present study builds its corpus from an online database of lead sheets described in Pachet et al. (2013). The database presents over 5700 jazz standards collected from the 'Real Books' and various composer-specific songbooks ('The Michel Legrand Songbook', 'The Bill Evans Fake Book', etc.).

The machine learning tasks in Sections 4 and 5 partition the database corpus by *composer*, *subgenre*, *performance style* (or tempo indication) and *meter* (Table 1), resulting in four separate classification tasks. Intuitively, classification by *subgenre* should perform comparably to *composer* since the *subgenre* collection consists of groups of composers similar in style. Classification by *performance style* and *meter* should be less successful as chord sequences do not contain explicit information relating to how they should be performed or their meter. Indeed, metrical analysis, (Chew, Volk & Lee, 2005) or beat-tracking algorithms (Krebs & Widmer, 2012), would be better suited to this task. Their inclusion in the study is to check that classifiers do not simply find arbitrary patterns in any partitionings of a corpus. A minimum limit of around 30 standards for each class ensures sufficient data for reliable models to be built, and a maximum cap (60 for *subgenre* and *performance style*, 90 for *meter*) prevents large classes dominating the classification space. Where classes would exceed the maximum cap, jazz standards are selected randomly. *Composer*, *performance style* and *meter* collections can be compiled simply using the metadata tags available in the database. For the *subgenre* collection, standards were labelled by a human jazz expert using the Wikipedia (http://wikipedia.org/wiki/Jazz) definitions for jazz subgenres. In this case, Wikipedia is used to represent a general, universal understanding of subgenres of jazz, which are typically ill-defined.

Chords appear in typical jazz notation as chord symbols (e.g. GM7) corresponding as closely as possible to the original

Table 1. The four collections and their classes. Majority class percentages indicate the proportion of the largest class per collection.

Composer **(447)**	***Performance Style*** **(434)**	***Subgenre*** **(437)**	***Meter*** **(180)**
Majority Class: 14.8%	Majority Class: 13.7%	Majority Class: 13.9%	Majority Class: 50.0%
Thelonius Monk (66)	Latin (60)	Ballad (60)	Quadruple (90)
John Coltrane (64)	Vocal Standards (60)	Medium Up Swing (60)	Triple (90)
Bill Evans (56)	Bebop (60)	Medium Swing (60)	
Charlie Parker (54)	European Songwriters (60)	Up Tempo Swing (59)	
Richard Rodgers (47)	Swing (60)	Medium (49)	
Michel Legrand (45)	Blues (60)	Bossa Nova (47)	
Duke Ellington (43)	Hard Bop (51)	Jazz Waltz (39)	
Pepper Adams (40)	Post Bop (26)	Latin (31)	
Wayne Shorter (32)		Rock (29)	

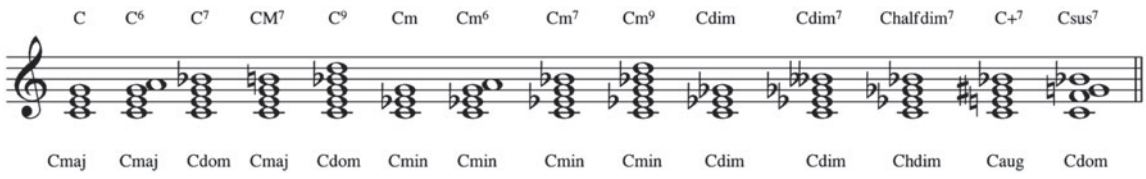


Fig. 1. Chord symbols as they appear in the database (above stave), in staff notation, and after applying chord simplification rules (below).

source. Melodies are represented as a sequence of notes, each consisting of a pitch class (e.g. C, D♯, E♭) and MIDI octave (e.g. 4). The duration in quarter notes of chords and melody notes is also available.

A notational problem arises from the variety of sources in the database, giving rise to a range of chord symbol representations. For example the first five chords of 'Giant Steps' are given as B, D^7, G, $B♭^7$, E♭, in 'The Real Book', but $Bmaj^7$, D^7, $Gmaj^7$, $B♭^7$, $E♭maj^7$, in 'The Music of John Coltrane'. In the vast majority of cases such discrepancies in notation do not change the fundamental harmonic function of chords, so can be normalized with a set of chord simplification rules (see Section 3.2).

3.2 Harmonic representation

The representation of musical structure can have a significant bearing on the quality of results for a computational analysis of a given corpus. In general, two approaches to representing harmonic information have emerged in computational musicology. The first represents harmony as the coincidence of polyphonic lines, which can be represented as a multiple viewpoint model (Whorley, Wiggins, Rhodes & Pearce, 2010). The second approach represents harmony more broadly, either by functional symbols (Tymoczko, 2003) or chord symbols, which is particularly appropriate in the case of jazz (Gillick et al., 2009; Ogihara & Li, 2008; Pérez-Sancho et al., 2009; Rohrmeier and Graepel, 2012). Conklin (2010) presents a multiple viewpoint representation for harmony, encoding information of root, type, root progression, duration and functional degree. The present study represents harmony by chord symbols as a musicologically rich representation able to provide sufficient information for analysis, whilst being general enough to incorporate notational discrepancies between sources (see Section 3.1).

A pre-processing procedure simplifies chord symbols found in the corpus (e.g. $E♭maj^7$) to their two essential attributes: fundamental root and chord type. Fundamental roots are always given by the prefix of the chord symbol (E♭) and are represented here as an integer from the set $\{-1, 0, 1, \ldots 11\}$ denoting pitch class assuming enharmonic equivalence, with -1 representing the case when no pitch class for the root is given. This case can arise when the 'No Chord' (N.C.) symbol appears, indicating no harmonic instruments should play. Bass notes (when given) are ignored, following a similar approach by Ogihara and Li (2008). Chord types are defined by applying a set of chord transformation rules to the rest of the chord symbol (e.g. maj^7) to normalize notation across sources, reduce sparsity of data and to group closely related or equivalent chords together. The transformation rules simplify any given chord symbol to a set of seven chord types {*dom, maj, min, dim, aug, hdim, NC*}. Dominant (*dom*) chords contain the major third of the triad and minor seventh (e.g. G^7, $D♯^9$, C^7alt). Major chords (*maj*) are any chords containing the major third of the triad that are not defined as dominant (e.g. G^6, D^{add9}, CM^7). Diminished chords are signified by 'dim' in the chord symbol. Minor chords are all chords with the minor third of the triad, but are not diminished (e.g. Gm, Dm^6, $Cm^{\#5}$). Augmented chords are signified by '+' or 'aug' in the chord symbol, and half-diminished chords by 'halfdim.' Chords with a suspended fourth are defined as *dom* if they also contain a minor seventh, otherwise are simplified to *maj*. Finally, N.C. signifies times of harmonic silence or where no specific chord is given. By way of example, Figure 1 shows 14 chords with their original chord symbols above the stave and simplified chord symbol below.

3.3 Classification procedure

The supervised classification procedure is implemented as a 10-fold cross-validation, dividing a corpus partition randomly into 10 approximately equal validation sets to estimate classification accuracies (the percentage of standards correctly classified). To counter any bias in the random allocation of songs into validation sets, each classification task is run 100 times, randomly re-allocating validation sets at the start of each run. A majority classifier acts as a baseline, classifying all songs into the largest class, returning a baseline accuracy (Equation 1). The F-measure (Equation 2) for each class, c, is calculated punishing both false negatives (an incorrectly classified item belonging to the given class) and false positives (an item not belonging to the given class, but is classified as such) by taking into account precision (Equation 3) and recall (Equation 4) for the given class.

$$baseline\,accuracy = \max_{c \in C}\left(\frac{|c|}{\sum_{c \in C}|c|}\right), \tag{1}$$

$$F_c = 2 \cdot \frac{precision_c \cdot recall_c}{precision_c + recall_c}, \tag{2}$$

$$precision_c = \frac{true\ positives_c}{true\ positives_c + false\ positives_c}, \tag{3}$$

$$recall_c = \frac{true\ positives_c}{true\ positives_c + false\ negatives_c}. \tag{4}$$

4. Supervised classification of chord sequences

Three supervised learning techniques address the extent to which composers can be identified purely by their chord

sequences. Further classification tasks on the *subgenre*, *performance style* and *meter* collections offer insights into the role of chord sequences as class predictors. A collection of probabilistic methods compare likelihoods of a chord sequence given a series of basic Markov models (Section 4.1) built from each class. For comparison, four n-gram methods for classification presented in Pérez-Sancho et al. (2009) are implemented (Section 4.2) to assess the impact of representation on the classification task. A novel subsequence matching method (Section 4.3) is proposed, classifying chord sequences with a fitting score based on the number and lengths of subsequences that occur in the chord sequence and a given class' model.

4.1 Markovian classifier

Probabilistic methods for classification compare the likelihoods of a set of data given various probabilistic models. Markov (n-gram) models (Norris 1997) are at the core of many probabilistic methods for modelling sequences of musical events (Collins, 2011; Cope, 2005; Pearce, 2005), making the assumption that musical sequences are generated from high-order Markovian sources. In the context of chord sequences, let e_i^j represent a sequence of chords from i to j, and $p(e_i|e_{i-n+1}^{i-1})$ the probability of a chord e_i given its predictive context e_{i-n+1}^{i-1}. The likelihood of a whole jazz standard of length T given a model order $n-1$ can therefore be estimated by Equation 5. At the start of the sequence (when $n > i$), $n-1$ padding symbols are inserted to provide the necessary predictive context.

$$p(e_1^T) = \prod_{i=1}^{T} p\left(e_i|e_{i-n+1}^{i-1}\right). \tag{5}$$

Witten–Bell method C smoothing (Witten & Bell, 1991) counters the zero-frequency problem, selected after a comprehensive review of smoothing methods on monophonic melodies (Pearce & Wiggins, 2004). The recursive interpolated smoothing algorithm terminates at the -1st order with a uniform distribution over the vocabulary size (Cleary & Witten, 1997), creating a bounded variable order Markov model (Begleiter, El-Yaniv and Yona, 2004). To determine the optimal global order bound for the present study, a 10-fold cross-validation of all collections (removing songs which appear in more than one collection so that each song appears only once) compared the average cross-entropies of various orders (Figure 2). Cross-entropy is a commonly used performance measure, calculating the divergence in entropies between an estimated probability distribution and its source (Manning & Schütze, 1999; Pearce & Wiggins, 2004). For a model m of order n and sequence e_1^{j-1}, the cross-entropy H_m is approximated by Equation 6 with the assumptions that j is sufficiently large, and that the sequence is generated by a stationary and ergodic stochastic process. Figure 2 shows the third global order bound to have the lowest cross-entropy (3.600), and is therefore selected for the Markovian Classifier.

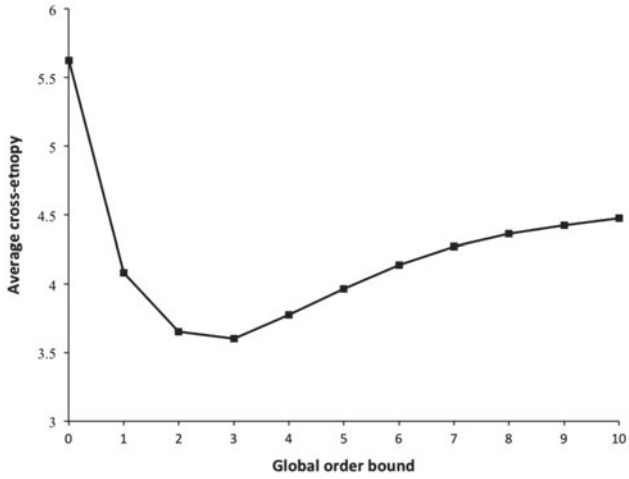


Fig. 2. Relative performances of bounded variable order Markov models measured by average cross-entropy per symbol of a 10-fold cross validation of all collections.

$$H_m(p_m, e_1^j) = -\frac{1}{j}\sum_{i=1}^{j} \log_2 p\left(e_i|e_{i-n+1}^{i-1}\right). \tag{6}$$

Each jazz standard is classified using Bayesian inference to select the most probable class, c^* (Equations 7 and 8), given the chord sequence e_1^T. The prior probability of the class, $p(c_s)$, is the class' proportion of the collection and the prior probability of the chord sequence, $p(e_1^T)$, is calculated with the total probability rule (Equation 9).

$$c^* = \underset{c_s \in C}{\operatorname{argmax}}\, p\left(c_s|e_1^T\right), \tag{7}$$

$$p\left(c_s|e_1^T\right) = \frac{p\left(e_1^T|c_s\right)\cdot p(c_s)}{p\left(e_1^T\right)}, \tag{8}$$

$$p\left(e_1^T\right) = \sum_{c_s \in C} p\left(e_1^T|c_s\right)\cdot p(c_s). \tag{9}$$

Before building models, all jazz standards are transposed 12 times, allowing identical chord sequences with different tonal centres to be considered as equivalent. The key and mode of a standard need not be determined since major mode standards will be transposed to all 12 major keys and those in minor modes to all 12 minor keys. Furthermore, any modulations within a standard will be accounted for without being identified explicitly. This is particularly important for a computational analysis of jazz music since key, mode and modulations are often ambiguous in jazz. For example, many standards by Bill Evans are strictly modal (Mawer, 2011).

Two variations of the Markovian classifier are presented, firstly (*Markovian1*) with chord type simplification (Section 3.2) and secondly (*Markovian2*) where chord types are left unedited. The state space for *Markovian1* can be conceptualized as the Cartesian product of chord roots and types, $root \times type$, where $root \in \{-1, 0, \ldots 11\}$ and $type \in \{dom, maj, min, dim, aug, hdim, NC\}$, producing a vocabulary of 93 including the $start$ and end padding symbols. The state space

for *Markovian2* is considerably larger, with the same set of *roots* but a set of 151 *types* creating a vocabulary of 1965.

4.2 Pérez-Sancho *n*-gram classifier

An alternative n-gram classifier (Pérez-Sancho et al., 2009) is presented, exploring the impact of contrasting chord sequence representations on classification performance. Each jazz standard is transposed to C major/A minor by considering its key signature and mode. Roots are represented either as note names so that enharmonic equivalent notes (e.g. C#/D♭) are distinct, or as scale degrees (e.g. I, V#) relative to the transposed key of the jazz standard. Chord types (extensions) are either left intact or are mapped to a set of five triad types: *major, minor, diminished, augmented, suspended 4th*. Four different representations, or feature sets, are possible with a combination of the two root and two chord type representations. Feature set 1 (FS1) comprises of scale degrees with chord type extensions, FS2: root names with extensions, FS3: scale degrees without extensions and FS4: root names without extensions. Table 2 shows a sample chord sequence from the opening of ''Round Midnight' by Thelonius Monk as it appears in its original key (E♭ minor) and transposed to A minor in the four feature sets. Note that since altered (*alt*) chords may sharpen or flatten the fifth of the triad (Levine, 1995, p. 70–71) they are simplified to *major* for FS3 and FS4.

The probability of a chord sequence is estimated with a smoothed (method C, Witten & Bell, 1991) n-gram model with $n \in \{2, 3, 4, 5\}$. Instead of classification by Bayesian inference (Section 4.1), the chord sequence is assigned to class by lowest perplexity, shown by Equations 10 and 11. As in Section 4.1, the classification task is undertaken as a 10-fold cross-validation.

$$c^* = \underset{c_s \in C}{\operatorname{argmin}}\, pp\left(e_1^T | c_s\right), \tag{10}$$

$$pp\left(e_1^T | c_s\right) = p\left(e_1^T | c_s\right)^{-1/T}. \tag{11}$$

4.3 Subsequence matching classifier

A novel supervised learning method is proposed for comparison with the Markovian methods described in Sections 4.1 and 4.2. The primary motivation behind the subsequence matching method is that for a chord sequence to be ubiquitous with a composer it is not necessarily the case that it must be repeated a large number of times in that composer's canon, as is assumed by a probabilistic model. Rather, it is possible for a unique chord sequence to appear only a handful of times in a few very popular jazz standards for it to be associated with that composers' style. A further motivation is to overcome the limitations of global order bounded Markov models and to consider longer chord sequences as complete entities, rather than segmented into n-gram chunks.

The subsequence matching method builds a model simply by concatenating all the chord sequences in a given class, transposed 12 times as in Section 4.1. To prevent false chord sequences which bridge songs being learnt, each standard is padded with starting and ending symbols. To assess how well a given jazz standard with a chord sequence length T matches a model, all possible subsequences from length T to 1 are selected and searched for in that model. The count c_t for all subsequences length t that occur both in the standard and the model is recorded. A score, s, is then returned, summing all counts multiplied by their length (Equation 12). The classification system favours long subsequences that, in contrast to Markov models, need only occur once in the training corpus to be counted.

$$s = \sum_{t=1}^{T} c_t \cdot t\,. \tag{12}$$

4.4 Results

Classification accuracies for the three classifiers are tabulated in Table 3, showing classification accuracy averaged over 100 runs with confidence intervals at the 95% confidence level. *Markovian2* (without chord type simplifications) achieves the highest classification accuracies for the *composer* (63.9%), *subgenre* (46.8%), *performance style* (31.3%), and *meter* (70.2%) collections. Classification accuracy will not give a full indication of performance when comparing collections containing a different number of classes, reflected in the baseline accuracies obtained from the majority classifier (see Section 3.3, Equation 1). Therefore, for each classifier, the t-statistic from a pairwise t-test over all 100 runs against the baseline accuracy is used as a performance measure.[1] These 19 t-statistics for each collection are then used to compare overall performance between collections with a further paired t-test. Across all 19 classifiers (two Markovian, 16 Pérez-Sancho n-gram and the subsequence matching classifier) a paired t-test at the 0.01 level shows classification by *composer* to be significantly easier compared to *subgenre* ($t(18) = 8.238$, $p < 0.001$, corrected[2]) and subsequently *subgenre* is significantly easier to classify compared to *performance style* ($t(18) = 18.877$, $p < 0.001$, corrected) and finally classification by *performance style* is significantly more successful ($t(18) = 3.854$, $p < 0.001$, corrected) compared to classification by *meter*.

Markovian2 (without chord type simplifications) outperforms the next most successful classifier significantly in the *composer* ($t(99) = 50.443$, $p < 0.001$), *subgenre* ($t(99) = 28.448$, $p < 0.001$), *performance style* ($t(99) = 36.932$, $p < 0.001$) and *meter* ($t(99) = 20.046$, $p < 0.001$) with significance judged by a paired t-test of classification accuracies across all 100 runs.

It is highly possible that classifiers not simplifying chord names (*Markovian2*, Pérez-Sancho FS1 and Pérez-Sancho

[1] $t = \sqrt{N}\frac{\bar{x}-\theta}{\sigma}$ where the average observed classification accuracy $\bar{x}$, standard deviation σ, is obtained over N repeated runs and compared to θ, the null hypothesis equating to the baseline accuracy.

[2] All corrected p-values are Bonferroni corrected by dividing the significance level, α, by the number of simultaneous hypotheses.

Table 2. Opening chord sequence of ''Round Midnight' by Thelonius Monk as it appears in the Real Book in the original key and encoded into the four feature sets.

Real Book:	E♭ *m,*	C *halfdim7,*	F *halfdim7,*	B♭ *alt7,*	E♭ *m7,*	A♭*7,*	B *m7,*	E *7,*
FS1:	I *m,*	VI# *halfdim7,*	II *halfdim7,*	V *alt7,*	I *m7,*	IV *7,*	V# *m7,*	I# *7,*
FS2:	A *m,*	F# *halfdim7,*	B *halfdim7,*	E *alt7,*	A *m7,*	D *7,*	E# *m7,*	A# *7,*
FS3:	I *min,*	VI# *dim,*	II *dim,*	V *maj,*	I *min,*	IV *maj,*	V# *min,*	I# *maj,*
FS3:	A *min,*	F# *dim,*	B *dim,*	E *maj,*	A *min,*	D *maj,*	E# *min,*	A# *maj,*

Table 3. Classification accuracies averaged over 100 10-fold classification tasks for classification over *composer, subgenre, performance style* and *meter* collections. Best performing classifiers judged by t-statistic are indicated in bold for each collection and classifier type. All t-statistics are significant at the 0.01 level after Bonferroni correction. Classifiers potentially biased by not simplifying chord types are indicated (*).

Classifier	Global order bound	*Composer* (9 classes) Baseline acc. 14.8%		*Subgenre* (8 classes) Baseline acc. 13.7%		*Performance Style* (9 classes) Baseline acc. 13.9%		*Meter* (2 classes) Baseline acc. 50.0%	
		Accuracy	$t(99)$	Accuracy	$t(99)$	Accuracy	$t(99)$	Accuracy	$t(99)$
Markovian1	3	59.0%±0.2	438.2	43.7%±0.2	248.8	27.4%±0.2	131.3	62.4%±0.3	79.0
*Markovian2**	3	**63.9%±0.2**	**526.9**	**46.8%±0.2**	**275.6**	**31.3%±0.2**	**151.1**	**70.2%±0.3**	**125.4**
Pérez-Sancho FS1*	1	53.9%±0.2	454.4	44.0%±0.2	270.1	25.7%±0.2	99.5	62.8%±0.4	71.0
Pérez-Sancho FS1*	2	55.0%±0.2	406.9	45.3%±0.2	270.0	24.7%±0.2	101.6	62.5%±0.4	61.6
Pérez-Sancho FS1*	3	55.0%±0.2	412.9	42.4%±0.2	313.9	26.4%±0.3	96.3	63.6%±0.4	74.5
Pérez-Sancho FS1*	4	55.4%±0.2	420.6	39.5%±0.2	246.7	23.8%±0.2	81.4	63.0%±0.4	66.9
Pérez-Sancho FS2*	1	58.7%±0.2	438.1	**43.1%±0.2**	**274.0**	**27.1%±0.2**	**117.2**	57.1%±0.4	38.2
Pérez-Sancho FS2*	2	59.7%±0.2	409.4	39.3%±0.2	235.8	26.8%±0.2	105.2	64.8%±0.3	85.2
Pérez-Sancho FS2*	3	**59.5%±0.2**	**479.6**	40.0%±0.2	261.4	23.1%±0.2	94.2	58.7%±0.3	50.4
Pérez-Sancho FS2*	4	58.8%±0.2	382.2	39.9%±0.2	224.9	25.0%±0.2	99.3	**67.0%±0.3**	**97.2**
Pérez-Sancho FS3	1	47.7%±0.2	276.6	30.9%±0.2	194.3	23.2%±0.2	83.1	62.2%±0.4	63.3
Pérez-Sancho FS3	2	50.6%±0.2	288.6	36.7%±0.2	257.5	25.4%±0.2	109.3	63.3%±0.4	62.3
Pérez-Sancho FS3	3	49.9%±0.2	347.1	40.4%±0.3	200.9	24.4%±0.2	100.3	61.0%±0.4	49.1
Pérez-Sancho FS3	4	50.2%±0.2	296.2	40.4%±0.2	238.0	24.4%±0.2	88.4	60.5%±0.4	50.5
Pérez-Sancho FS4	1	38.8%±0.2	270.8	30.9%±0.2	194.3	21.5%±0.2	76.1	55.7%±0.3	35.5
Pérez-Sancho FS4	2	40.2%±0.2	285.6	36.8%±0.2	257.5	20.3%±0.2	57.6	53.5%±0.4	17.7
Pérez-Sancho FS4	3	37.9%±0.2	236.7	30.3%±0.2	161.2	18.6%±0.2	48	53.4%±0.4	15.2
Pérez-Sancho FS4	4	36.5%±0.2	191.5	32.4%±0.2	166.1	16.5%±0.2	24.5	61.4%±0.3	69.4
Subsequence Matching	N/A	**55.6%±0.2**	**427.3**	**37.1%±0.2**	**227.0**	**23.8%±0.2**	**13.9**	**60.4%±0.3**	**58.2**
harmonicVP1	3	61.1%±0.2	419.7	45.4%±0.2	298.0	37.9%±0.2	204.7	**99.4%±0.0**	**2081.6**
harmonicVP2	3	58.8%±0.2	496.4	47.2%±0.2	273.1	26.7%±0.2	108.9	65.3%±0.4	76.9
melodicVP	3	50.2%±0.2	322.4	46.2%±0.2	288.1	31.1%±0.3	133.9	89.2%±0.2	393.9
allVP	3	**67.3%±0.2**	**533.5**	**57.6%±0.2**	**462.0**	**38.8%±0.2**	**206.5**	90.6%±0.2	396.1

FS2) gain a bias because of notational differences between sources (see Pachet et al. (2013) and arguments for chord simplification in Section 3.2). This is particularly problematic for the *composer* collection as composer classes are typically built from separate sources. For example, the Michel Legrand Songbook provides detailed chord symbols in comparison to the Real Books and many fakebooks, resulting in high recalls of 0.927 and 0.898 (Table 4) for *Markovian2* and Pérez-Sancho 4-gram FS2 respectively. These drop noticeably to 0.787 and 0.463 respectively for *Markovian1* and Pérez-Sancho 4-gram FS4, which simplify chord types but are otherwise identical. Therefore, removing the nine affected classifiers, the highest performing classifier is found to be the *Markovian1* (59.0%) which outperforms the subsequence matching classifier (55.6%) by a statistically significant ($t(99) = 34.778$, $p < 0.001$) amount.

As the easiest to classify collection, and the main focus of the current study, Table 4 provides further insight into the classification of the *composer* collection with the highest performing Markovian, Pérez-Sancho and subsequence matching classifiers. Certain patterns are maintained across all three classifiers, in particular that Michel Legrand, Bill Evans and Charlie Parker return high recalls for all classifiers. Additionally, Bill Evans returns a relatively low precision in comparison with recall, implying this part of the model contains high probabilities for universally common 4-grams. Finally, it is noticeable that Duke Ellington, John Coltrane and Wayne Shorter are difficult to classify, returning low recalls,

Table 4. Performance measures averaged over 100 10-fold classification tasks for the *composer* collection classified by three classifiers.

Classifier	Class	Recall	Precision	F-measure
Markovian2 Accuracy: 63.9%±0.2	Thelonius Monk (66)	.540	.710	.613
	John Coltrane (64)	.315	.433	.364
	Bill Evans (56)	.866	.596	.706
	Charlie Parker (54)	.739	.807	**.772**
	Richard Rodgers (47)	.783	.658	.714
	Michel Legrand (45)	**.927**	**.816**	.867
	Duke Ellington (43)	.330	.363	.345
	Pepper Adams (40)	.771	.762	.766
	Wayne Shorter (32)	.563	.554	.559
Pérez-Sancho 4-gram classifier, FS2 Accuracy: 59.5%±0.2	Thelonius Monk (66)	.456	.570	.506
	John Coltrane (64)	.345	.455	.392
	Bill Evans (56)	.857	.525	.651
	Charlie Parker (54)	.715	**.882**	**.789**
	Richard Rodgers (47)	.696	.744	.718
	Michel Legrand (45)	**.898**	.690	.780
	Duke Ellington (43)	.319	.278	.297
	Pepper Adams (40)	.680	.789	.730
	Wayne Shorter (32)	.405	.545	.464
Subsequence Matching Accuracy: 55.6%±0.2	Thelonius Monk (66)	.549	.625	.584
	John Coltrane (64)	.395	.506	.444
	Bill Evans (56)	**.768**	.638	697
	Charlie Parker (54)	.736	**.667**	**.699**
	Richard Rodgers (47)	.707	.468	.563
	Michel Legrand (45)	.765	.572	.654
	Duke Ellington (43)	.182	.346	.238
	Pepper Adams (40)	.578	.484	.526
	Wayne Shorter (32)	.179	.587	.273

although in the case of Wayne Shorter this may be because the small class size creates a sparse model.

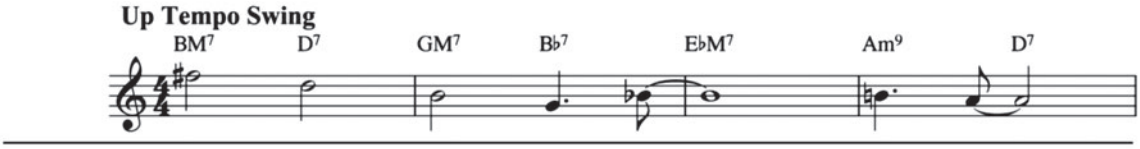
5. Supervised classification with multiple viewpoint classifiers

Musical structure is a complex multi-dimensional landscape, a property that has been modelled by multiple viewpoint Markov models, applied to melodic structure by Pearce (2005) and extended for classification tasks by Conklin (2013a). Intuitively, it seems beneficial to model the interaction between melody and harmony as it captures the composer's choice of chords to support melodies and vice versa. Likewise, since music is perceived as a temporal sequence, information of duration should also improve model performances.

Different structural features of music (such as *root*, *duration* and *pitch*) are modelled as primitive viewpoints and their inter-relations as linked viewpoints (such as *pitch*⊗*duration*). All selected primitive and linked viewpoints are modelled as separate Markov models and the likelihood of a sequence as the geometric mean across all selected viewpoints (Conklin, 2013b). Multiple viewpoint models are able to combine the performance of individual expert models to outperform a single model with the same information (Pearce, Conklin & Wiggins 2005), reducing the sparsity of complex representations allowing for better generalization of training data. This increase in model performance seems likely to extend to classification. Conklin (2013a) reports that a multiple viewpoint model of melodic attributes consistently outperforms a model that represents the same information as a single linked viewpoint.

5.1 Multiple viewpoint representation

Five primitive viewpoints represent the harmonic, melodic and temporal structure of a jazz standard (Figure 3). The work of Conklin (2010) is drawn on for the representation of chords, with *root* and *type* viewpoints representing the chord attributes exactly as described in Section 3.2 with chord type simplification. $RootInterval \in \{-1, 0, 1, \ldots 11\}$ represents the interval in semitones between successive roots. The *pitch* viewpoint represents melodic pitch as an integer from the set $pitch \in \{-1, 0, \ldots 11\}$ where -1 represents a rest. *Duration* is represented as a positive integer $\in \{0, 1, \ldots 15120\}$ where 2520 represents one quarter note. A 'timebase' (Pearce, 2005, p. 63) for the database of 2520 is calculated from the lowest common multiple of 5, 7, 8, 9, 12 representing the number of division in a quarter note; for quintuplet 16th notes, septuplet 16th notes, 32nd notes, nontuplet 32nd notes and triplet 64th notes (all of which are present in the database). The vocabulary size is given by multiplying the timebase by the longest possible duration in quarter notes (six). Lead sheets



root	11	2	7	10	10	3	9	9	2
rootInterval	⊥	3	5	3	0	5	6	0	5
type	maj.	dom.	maj.	dom.	dom.	maj.	min.	min.	dom.
pitch	6	2	11	7	10	10	11	9	9
duration	5,040	5,040	5,040	3,780	1,260	10,080	3,780	1,260	5,040

Fig. 3. The opening four bars of John Coltrane's 'Giant Steps' represented by the five primitive viewpoints of (chord) *root*, (chord) *type*, (chord) *rootInterval*, (melodic) *pitch* and *duration*.

Table 5. Four multiple viewpoint models with primitive and linked viewpoints.

harmonicVP1	*harmonicVP2*	*melodicVP*	*allVP*
root	*root*	*pitch*	*root*
type	*type*	*duration*	*type*
rootInterval	*rootInterval*	*ptich⊗duration*	*rootInterval*
duration	*root⊗type*		*duration*
root⊗type	*rootInterval⊗type*		*pitch*
rootInterval⊗type			*root⊗type*
root⊗type⊗duration			*rootInterval⊗type*
			root⊗type⊗duration
			ptich⊗duration
			root⊗type⊗pitch

are segmented at every chord change if a harmonic viewpoint is present and at every note onset if a melodic viewpoint is present. Four viewpoint models (Table 5) are constructed comparing viewpoint models with (*harmonicVP1*) and without (*harmonicVP2*) temporal information, with melodic and temporal information (*melodicVP*), and with harmonic, melodic and temporal information combined (*allVP*).

The global order bound of the multiple viewpoint Markov model was determined with a 10-fold cross-validation entropy test of all the collections for all five primitive viewpoints (Table 6). The third order is retained as the global order bound for the multiple viewpoint models as the optimal order for the *root* and *type* viewpoints. Although the average cross-entropy for the *rootInterval*, *duration* and *melody* viewpoints is lower for the second order, as the difference is small and the interpolated smoothing will incorporate lower order models, a third-order model is retained.

5.2 Results

The supervised classification by multiple viewpoint Markov models is implemented with a 10-fold cross-validation procedure, transposing all jazz standards 12 times, with results for the four collections tabulated in Table 3. The *allVP* classifier (67.3%) significantly outperforms its nearest rival (including classifiers from Section 4) in the *composer* ($t(99) = 58.953$, $p < 0.001$), *subgenre* ($t(99) = 90.991$, $p < 0.001$) and *performance style* ($t(99) = 7.415$, $p < 0.001$) collections, whilst *harmonicVP1* significantly outperforms ($t(99) = 118.977$, $p < 0.001$) *allVP* in classification by *meter*. Taking all 23 classifiers into account, a comparison of t-statistics by paired t-test shows classification by *composer* to still be significantly more successful than by *subgenre* ($t(22) = 8.761$, $p < 0.001$, corrected) and *performance style* ($t(22) = 18.110$, $p < 0.001$, corrected), but it is no longer significantly easier to classify compared to *meter* ($t(22) = -0.932$, $p = 0.819$, corrected).

The improved classification by *meter* is exemplified by an average classification accuracy of 99.4% for the *harmonicVP1* classifier. It is clear that the improved performance is gained from the *duration* viewpoint as the *harmonicVP2* achieves an average accuracy of only 65.3% and is significantly ($t(99) = 26.714$, $p < 0.001$) outperformed by the *Markovian2* classifier (70.2%).

Further insight into classification by *composer* (as the primary classification task of the current study) is shown in Table 7, with associated recall, precision and F-measures for the four multiple viewpoint classifiers. As observed in Section 4.4 high recalls are returned for Bill Evans, although Charlie Parker and Michel Legrand are less consistent. Again, Bill Evans returns a low precision despite high recalls, especially for the *melodicVP* classifier (0.323). Duke Ellington and Wayne

Table 6. Relative performance of bounded variable-order Markov models that use primitive viewpoints, measured by average cross-entropy per symbol of a 10-fold cross-validation over all collections.

Global order bound	*root*	*type*	*rootInterval*	*duration*	*melody*
0	3.742	2.116	1.836	2.865	3.739
1	1.629	1.129	1.585	2.262	3.443
2	1.613	1.122	**1.555**	**2.198**	**3.283**
3	**1.582**	**1.115**	1.580	2.254	3.300
4	1.583	1.126	1.645	2.441	3.571
5	1.593	1.151	1.787	2.768	3.934
6	1.646	1.214	1.988	3.198	4.153
7	1.720	1.313	2.247	3.685	4.247
8	1.817	1.449	2.546	4.171	4.287
9	1.924	1.631	2.883	4.607	4.307
10	2.050	1.852	3.232	4.977	4.322

Shorter are consistently the lowest ranked composers by recall and F-measure. The fact that these patterns are consistent across a wide variety of classification methods strongly suggests that they are not merely coincidental, but an intrinsic property of a composer's style. It is interesting to note that the *harmonicVP1* (61.1%) outperform the *harmonicVP2* (58.8%) model by only a small amount, although this is found to be statistically significant ($t(99) = 22.233$, $p < 0.001$). This implies the addition of temporal information does not improve the classification of chord sequences by composer.

6. Classifying subsequences within compositions

The classification methods presented in Sections 4 and 5 can be used as the basis for an analysis of chord subsequences within a composition. Arguably, this is more interesting than simply classifying a piece with a label, since jazz musicians in particular are adept at borrowing and manipulating subsequences of chords from the œuvre of other musicians. For the examples presented in this section the *harmonicVP1* classifier is chosen as the best performing classifier on chord sequences only.

Such an analysis may be able to shed some light on the certainty of classifications, as shown in two extracts from 'Boo Boo's Birthday' by Thelonious Monk (Tables 8 and 9). The transition probabilities, $p(e_i^j|c)$, are calculated with the third-order *harmonicVP1* classifier and the posterior class probabilities, $p(c|e_i^j)$, used to find the most likely class (indicated in bold). The opening four bars (Table 8) show considerable uncertainty within the classifications, with four different composers returned and all posterior probabilities below 0.4. On the other hand, the chromatic descent over the following four bars (Table 9) shows more certainty, with four of six transitions classified correctly. The whole standard is classified correctly as Thelonious Monk at probability 0.962, giving some indication of the uncertainty in the opening bars, but no clue as to where the uncertainty might lie, or what precisely is stylistically typical. Such feedback is particularly useful for style specific generation in identifying idiomatic sequences and patterns (Collins, 2011).

6.1 Subsequence selection algorithm

With these points in mind, this section presents an algorithm to identify and label subsequences of chords within a jazz standard to find all of the maximal length subsequences (see Figure 4) classified for a given set of composers. Maximal length subsequences are defined as subsequences labelled by class that cannot be extended forwards or backwards without re-classification.

A subsequence selection algorithm is applied to find the maximal length subsequences classified for a given set of composers. First, the classifications of all possible subsequences for all possible lengths down to a minimum threshold of 8 are calculated. The subsequences are arranged in a directed acyclic graph (Figure 4) with the longest subsequence spanning the whole piece at the root and the shortest subsequences at the leaves. Each vertex representing a subsequence e_i^j has two parents: e_i^{j+1} and e_{i-1}^j respectively. To select all subsequences that cannot be extended any further without being reclassified, a vertex is selected for return if it is classified in a different class to both its parents (or its only parent if it is at the start or finish). To reduce the number of subsequences returned, pieces are divided into sections (defined on the original lead sheet) preventing subsequences from bridging sections.

6.2 'Giant Steps' by John Coltrane

Figure 5 displays a global map of the selected subsequences for 'Giant Steps' by John Coltrane with chord sequences of the jazz standard in Table 10. For the classification process 'Giant Steps' was removed from the training corpus to prevent a trivial classification of its subsequences. Subsequences classified as John Coltrane (green) and Bill Evans (red) are identified, with the subsequence spanning the whole song correctly classified to John Coltrane. In particular bars 1–4 outline an idiomatic *Coltrane Changes* progression (see

Table 7. Performance measures averaged over 100 runs for the *composer* collection classified by four multiple viewpoint classifiers.

Classifier	Class	Recall	Precision	F-measure
	Thelonius Monk (66)	.527	.625	.571
	John Coltrane (64)	.454	.538	.492
	Bill Evans (56)	**.897**	.579	.704
harmonicVP1	Charlie Parker (54)	.656	**.741**	.695
Accuracy:	Richard Rodgers (47)	.733	.619	.670
61.1%±0.2	Michel Legrand (45)	.832	.696	**.757**
	Duke Ellington (43)	.249	.380	.300
	Pepper Adams (40)	.675	.729	.700
	Wayne Shorter (32)	.442	.510	.473
	Thelonius Monk (66)	.483	.594	.532
	John Coltrane (64)	.459	.508	.482
	Bill Evans (56)	**.886**	.579	.700
harmonicVP2	Charlie Parker (54)	.666	.749	.705
Accuracy:	Richard Rodgers (47)	.655	.598	.625
58.8%±0.2	Michel Legrand (45)	.815	.661	**.729**
	Duke Ellington (43)	.203	.287	.238
	Pepper Adams (40)	.689	**.736**	.711
	Wayne Shorter (32)	.378	.452	.412
	Thelonius Monk (66)	.374	**.875**	.523
	John Coltrane (64)	.638	.380	.476
	Bill Evans (56)	**.763**	.323	.454
melodicVP	Charlie Parker (54)	.679	.767	**.720**
Accuracy:	Richard Rodgers (47)	.467	.804	.590
50.2%±0.2	Michel Legrand (45)	.470	.769	.582
	Duke Ellington (43)	.164	.356	.224
	Pepper Adams (40)	.610	.512	.556
	Wayne Shorter (32)	.148	.539	.231
	Thelonius Monk (66)	.577	.795	.668
	John Coltrane (64)	.667	.541	.597
	Bill Evans (56)	**.877**	.481	.621
allVP	Charlie Parker (54)	.743	.838	.787
Accuracy:	Richard Rodgers (47)	.711	.873	.783
67.3%±0.2	Michel Legrand (45)	.777	**.910**	**.838**
	Duke Ellington (43)	.368	.508	.426
	Pepper Adams (40)	.860	.794	.825
	Wayne Shorter (32)	.377	.650	.476

Table 8. Chord sequence and associated transition and posterior class probabilities for bars 1–4 of Thelonious Monk's 'Boo Boo's Birthday'.

Class (c)	CM7		B7		E7		E7	
	$p(e^1_{-3}\|c)$	$p(c\|e^1_{-3})$	$p(e^2_{-2}\|c)$	$p(c\|e^2_{-2})$	$p(e^3_{-1}\|c)$	$p(c\|e^3_{-1})$	$p(e^4_0\|c)$	$p(c\|e^4_0)$
Thelonious Monk:	.455	(.133)	.093	(.166)	.047	(.181)	.349	(.261)
John Coltrane:	**.533**	**(.156)**	.078	(.138)	.002	(.010)	.178	(.133)
Bill Evans:	.217	(.064)	.003	(.006)	.046	(.179)	.089	(.067)
Charlie Parker:	.466	(.136)	.019	(.033)	.005	(.021)	**.414**	**(.309)**
Richard Rodgers:	.328	(.096)	.172	(.306)	.001	(.004)	.069	(.052)
Michel Legrand:	.000	(.000)	.004	(.007)	**.088**	**(.340)**	.068	(.051)
Duke Ellington:	.371	(.109)	**.186**	**(.331)**	.006	(.024)	.045	(.034)
Pepper Adams:	.655	(.192)	.002	(.003)	.033	(.129)	.027	(.020)
Wayne Shorter:	.391	(.114)	.005	(.010)	.029	(.113)	.100	(.074)

Figure 3) reflected in the fact that no other composers are returned until the start of bar 4. Bars 4–15 suggest stylistic similarity with Bill Evans, which is plausible given they shared part of their careers in the Miles Davis Sextet.

Table 9. Chord sequence and associated transition and posterior class probabilities for bars 4–8 of Thelonious Monk's 'Boo Boo's Birthday.'

Class	F7		E7		E♭7		D7		DM7♯11		D♭7	
(c)	$p(e_1^5\|c)$	$p(c\|e_1^5)$	$p(e_2^6\|c)$	$p(c\|e_2^6)$	$p(e_3^7\|c)$	$p(c\|e_3^7)$	$p(e_4^8\|c)$	$p(c\|e_4^8)$	$p(e_5^9\|c)$	$p(c\|e_5^9)$	$p(e_6^{10}\|c)$	$p(c\|e_6^{10})$
TM:	**.536**	**(.245)**	**.029**	**(.481)**	**.438**	**(.199)**	.267	(.204)	.634	(.299)	**.009**	**(.289)**
JC:	.144	(.066)	.003	(.050)	.396	(.180)	**.442**	**(.337)**	.043	(.020)	.001	(.033)
BE:	.258	(.118)	.005	(.085)	.372	(.169)	.099	(.076)	**.636**	**(.300)**	.003	(.089)
CP:	.195	(.089)	.000	(.007)	.228	(.104)	.112	(.086)	.043	(.020)	.001	(.023)
RR:	.143	(.065)	.001	(.019)	.233	(.106)	.169	(.129)	.247	(.116)	.004	(.125)
ML:	.156	(.071)	.002	(.027)	.064	(.029)	.024	(.018)	.028	(.013)	.006	(.203)
DE:	.383	(.175)	.004	(.075)	.257	(.117)	.079	(.060)	.088	(.042)	.001	(.020)
PA:	.131	(.060)	.007	(.115)	.018	(.008)	.046	(.035)	.069	(.033)	.001	(.045)
WS:	.244	(.111)	.008	(.141)	.194	(.088)	.072	(.055)	.331	(.156)	.005	(.174)

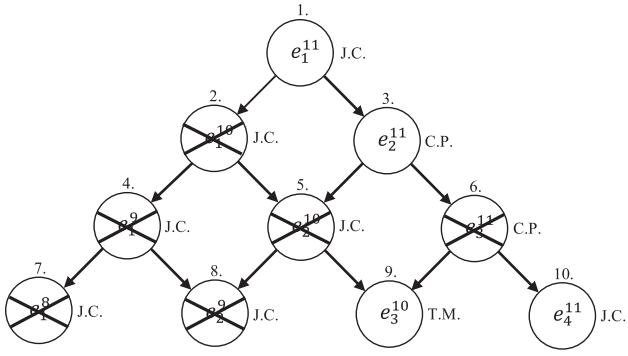


Fig. 4. The directed acyclic graph of all subsequences of all lengths (from 8) classified as J.C. (John Coltrane), C.P. (Charlie Parker) or T.M. (Thelonious Monk). A vertex is selected for return only if it is classified in a different class to both its parents. The above example would return subsequences e_1^{11}, e_2^{11}, e_3^{10} and e_4^{11}.

6.3 'Pretty Late' by Pachet and d'Inverno

'Pretty Late' by Pachet and d'Inverno (Table 11) provides an interesting case for the subsequence classifier. The piece is based on 'Very Early' by Bill Evans but without making direct quotations of substantial length. Interestingly, the classifier is sensitive to this influence, identifying the three subsequences spanning the three main sections as Bill Evans (Figure 6), strengthening the credibility of the classifier. The coda section closes with a Coltrane-esque chain of thirds in bars 58–61: BM^7, $A♭M^7$, $EM^{7♯11}$, $E♭M^7$, prompting the subsequence spanning the whole coda to be classified as John Coltrane.

7. Discussion and conclusion

The machine learning techniques presented in the current study have shown that to a large extent, composers can be identified computationally by their chord sequences alone. Markovian and novel subsequence matching classifiers (Section 4) returned similar results (accuracies of 59.0% and 55.6% respectively, compared to a baseline accuracy of 14.8%), reinforcing trends found in chord sequence classification of the *composer* collection. Multiple viewpoint representations for classifiers were implemented in Section 5 incorporating harmonic, melodic and temporal information improving classification accuracy to 67.3%. Finally, an algorithm for selecting stylistically prominent subsequences within a jazz standard found plausible interpretations of two lead sheets (Section 6).

Classification across different partitionings of the corpus provides useful information on what partitionings are relevant to the style of a chord sequence. Notably, classifying by *composer* (67.3%) was significantly more successful than by *subgenre* (57.6%), and classifications by *performance style* (38.8%). The poorer classification accuracies for the more arbitrary partitionings of the corpus by *performance style* imply that the classification models do not simply find patterns by chance in any given partitioning of a training set. These results suggest that individual composers have a distinctive harmonic style, which does not hold so well for subgenres. Another possible explanation is that while composers are unambiguous, subgenre is not. Therefore, the poor performance of style

Fig. 5. The subsequence selection algorithm applied to 'Giant Steps' by John Coltrane.

Table 10. Chord sequence for ‘Giant Steps’ by John Coltrane. Bars are represented by cells which are divided equally by a vertical bar where appropriate.

1	B | D^7	G | B♭ ^{7}	E♭	Am^7 | D^7
5	G | B♭ ^{7}	E♭ | F♯^{7}	B	Fm^7 | B♭ ^{7}
9	E♭	Am^7|D^7	G	C♯m^7|F♯^{7}
13	B	Fm^7 | B♭ ^{7}	E♭	C♯m^7|F♯^{7}

Table 11. Chord sequence for ‘Pretty Late’ by Pachet and d’Inverno. Bars are represented by cells which are divided equally by a vertical bar where appropriate.

Intro	1	CM^7	G^{7sus}	CM^7	G^{7sus}
	5	CM^7	G^{7sus}	CM^7	G^{7sus}
A	9	CM^7	Am^7	B♭ m^7	E♭ ^{7}
	13	A♭ M^7	Fm^7	F♯m^7	B^7
	17	EM^7	B♭ ^{7}	E♭ M^7	D^7
	21	GM^7	Dm^7	GM^7	$G^{7\sharp5}$
B	25	CM^7	$E^{7\sharp9}$	FM^7	B♭ ^{7}
	29	E♭ M^7	A^7	DM^7	A♭ ^{7}
	33	Gm^7	C^7	A/B	A/B
	37	EM^7	C♯m^7	Dm^7	$G^{7\flat9}$
A’	41	CM^7	Am^7	B♭ m^7	E♭ ^{7}
	45	A♭ M^7	Fm^7	F♯m^7	B^7
	49	EM^7	B♭ ^{7}	E♭ M^7	D^7
Coda	53	Am^7	B^7	EM^7	C♯m^7
	57	Cm^7	BM^7	A♭ M^7	$EM^{7\sharp11}$
	61	E♭ M^7			

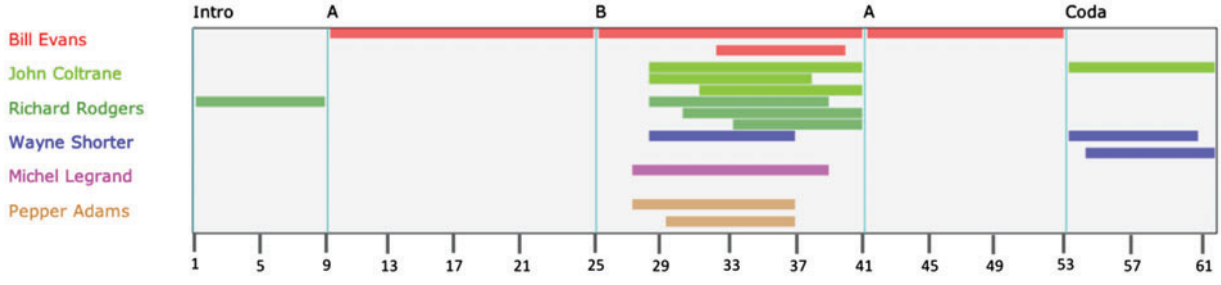


Fig. 6. The subsequence selection algorithm applied to ‘Pretty Late’ by Pachet and d’Inverno.

prediction may be explained in part by spurious labelling and in part by an inconsistent effect of style on chord progressions.

Markovian classifiers presented in Section 4.1 significantly outperformed similar *n*-gram classifiers presented by Pérez-Sancho et al. (2009), both with chord simplifications (63.9% to 59.5%) and without (59.0% to 50.6%). It is expected that these differences in performances are due to variations in representation, particularly how pieces in different keys are made equivalent. Whilst the current study transposes to all 12 tonal centres (regardless of mode), Pérez-Sancho et al. (2009) transpose all pieces to the same key. This is likely to be problematic since jazz standards are often ambiguous in key, modal, without key, or modulate.

It is particularly interesting that the subsequence matching classifier in Section 4.3, which is entirely independent of frequency of occurrences, finds similar results to the Markovian classifiers, which are probabilistic and therefore reliant on events occurring often in a training corpus. Additionally, the subsequence matching classifier considers subsequences of variable lengths, whilst the third-order Markovian classifiers only observe 4-gram chunks. Finally, the subsequence matching classifier considers subsequences as whole entities, whilst a Markovian classifier assigns high probabilities to chunks for which it can easily predict the suffix given a prefix. Despite the fundamental differences in these two approaches to classification, they return similar findings. This implies that identifiable stylistic patterns can be labelled as stylistically typical with fairly high confidence.

The use of multiple viewpoint classifiers was motivated by a recent study (Conklin, 2013a) in folk melody classification. Accuracies for all four classification tasks improved by a small but statistically significant amount, with a classifier incorporating harmonic, melodic and temporal information performing best (67.3%). For chord sequences alone, it was found that temporal information increased the classification accuracy only from 58.8% (*harmonicVP2*) to 61.1% (*harmonicVP1*).

For classification by *meter*, the discrepancy between the performance of the *harmonicVP1* (99.4%) and *harmonicVP2* (65.5%) strongly suggests that chord duration alone is sufficient to classify between the two *meter* classes. This is perhaps unsurprising considering that the chord durations in quadruple meter are mainly four quarter notes long (occasionally two) and chord durations in triple meter are mainly three quarter notes long (occasionally one or two, but importantly never four). This intuition is confirmed, as a zeroth-order classifier comprising of the *duration* viewpoint segmenting only at chord changes, returns an average classification accuracy of 99.8%±0.0.

A subsequence selection algorithm returned plausible readings of two lead sheets in Section 6. This novel application of machine learning techniques could provide a useful feedback tool for composers and analysts, allowing them to discover how exact subsequences of chords relate to other composers. Additionally, such an application could provide the basis for style specific generation (Collins, 2011). It is important to note that it is very difficult to draw conclusions from the classifier on whether a piece was influenced by a certain composer in a historical sense. For example, the fact that 'Giant Steps' by John Coltrane contains long subsequences classified as Bill Evans does not necessarily imply that John Coltrane was influenced by Bill Evans or vice versa. It could also be possible that they were both separately influenced by an external composer and did not influence one another directly despite sharing stylistic qualities.

Acknowledgements

The authors would like to thank Daniel Martín, Jeff Suzda and Marcus Pearce for their contributions to the study.

Funding

This research was conducted within the Flow Machines project which received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement no. 291156.

References

Begleiter, R., El-Yaniv, R., & Yona, G. (2004). On prediction using variable order Markov models. *Journal of Artificial Intelligence Research, 22*, 385–421.

Chemillier, M. (2004). Toward a formal study of jazz chord sequences generated by Steedman's grammar. *Soft Computing, 8*(9), 617–622.

Chew, E., Volk, A., & Lee, C.-Y. (2005). Dance music classification using inner metric analysis. In B. Golden, S. Raghavan, & E. Wasil (Eds.), *The Next Wave in Computing, Optimization, and Decision Technologies* (Operations Research/Computer Science Interfaces Series, Vol. 29 pp. 355–370). Berlin: Springer.

Cleary, J.G., & Witten, W.J. (1997). Unbounded length contexts for PPM. *The Computer Journal, 40*(2/3), 67–75.

Collins, T. (2011). *Improved methods for pattern discovery in music, with applications in automated stylistic composition* (PhD thesis), The Open University, Milton Keynes, UK.

Conklin, D. (2010). Discovery of distinctive patterns in music. *Intelligent Data Analysis, 14*(5), 547–554.

Conklin, D. (2013a). Multiple viewpoint systems for music classification. *Journal of New Music Research, 42*(1), 19–26.

Conklin, D. (2013b). Fusion functions for multiple viewpoints. In Ramirez, R., Conklin, D. & Iñesta, J.M.. (Eds.) *Proceedings MML 2013: 6th International Workshop on Machine Learning and Music Prague, Czech Republic* (Retrieved from: https://docs.google.com/file/d/0B7a519JYo78Nelp3QjVENUsxSnM/edit?pli=1).

Cope, D. (2005). *Computer models of musical creativity*. Cambridge, MA: MIT Press.

Gillick, J., Tang, K. & Keller, R. (2009). Learning jazz grammars. F. Gouyon , Á. Barbosa & Serra, X.. Eds. *SMC 2009: 6th Sound and Music Computing Conference, Porto, Portugal* (pp. 23–25, http://smc2009.smcnetwork.org/proceedings/proceedings.pdf).

Hillewaere, R., Manderick, B., & Conklin, D. (2009). Global feature versus event models for folk song classification. In *ISMIR 2009: 10th International Society for Music Information Retrieval Conference, Kobe, Japan* (pp. 729–733). Canada: International Society for Music Information Retrieval.

Hillewaere, R., Manderick, B., & Conklin, D. (2012). String methods for folk tune genre classification In *ISMIR 2012: 13th International Society for Music Information Retrieval Conference, Porto, Portugal* (pp. 217–222). Canada: International Society for Music Information Retrieval.

Johnson-Laird, P. (2002). *How jazz musicians improvise. Music Perception, 19*(3), 415–442.

Keller, R.M. & Morrison, D.R. (2007). A grammatical approach to automatic improvisation. In *SMC 2007: 4th Sound and Computing Music Conference, Lefkada, Greece*. (pp. 330–337), http://smc07.uoa.gr/SMC07%20Proceedings/SMC07%20Paper%2055.pdf

Krebs, F. & Widmer, G. (2012). MIREX 2012 audio beat tracking evaluation: Beat.E. In *Music Information Retrieval eXchange (MIREX), Porto*. Canada: International Society for Music Information Retrieval.

Larson, S. (1998). Schenkerian analysis of modern jazz: questions about method. *Music Theory Spectrum, 20*(2), 209–241.

Levine, M. (1995). *The Jazz theory book*. Petaluma, CA: Sher Music Co.

Manning, C.D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Mawer, D. (2011). French music reconfigured in the modal jazz of Bill Evans. In J. Mäkelä (ed.), *9th Nordic Jazz Conference Proceedings, Helsinki, Finland* (pp. 77–89). Helsinki: The Finnish Jazz & Pop Archive.

Norris, J.R. (1997). *Markov chains*. Cambridge: Cambridge University Press.

Ogihara, M., & Li, T. (2008). N-gram chord profiles for composer style representation. In *ISMIR 2008: 9th International Society for Music Information Retrieval Conference, Philadelphia, USA* (pp. 671–676). Canada: International Society for Music Information Retrieval.

Pachet, F., Martín, D., & Suzda, J. (2013) A comprehensive online database of machine-readable lead sheets for jazz standards. In *ISMIR 2013: 14th International Society for Music Information Retrieval Conference, Curitiba, Brazil* (pp. 275–280). Canada: International Society for Music Information Retrieval.

Pearce, M., & Wiggins, G. (2004). Improved methods for statistical modelling of monophonic music. *Journal of New Music Research, 33*(4), 367–385.

Pearce, M., Conklin, D. & Wiggins, G. (2005). Methods for combining statistical models of music. In *Proceedings of the Second international conference on Computer Music Modeling and Retrieval* (pp. 295–312). Berlin: Springer-Verlag.

Pearce, M. (2005). *The construction and evaluation of statistical models of melodic structure in music perception and composition* (PhD thesis). City University, London, UK.

Pérez-Sancho, C., Rizo, D., & Iñesta, J.M. (2009). Genre classification using chords and stochastic language models. *Connection Science, 21*(2), 145–159.

Rosen, C. (1971). *The classical style: Haydn, Mozart, Beethoven*. New York: Norton.

Rohrmeier, M. & Graepel, T. (2012). Comparing feature-based models of harmony. In *CMMR 2012: 9th International Symposium on Computer Music Modeling and Retrieval, London, UK* (pp. 315–370). http://cmmr2012.eecs.qmul.ac.uk/sites/cmmr2012.eecs.qmul.ac.uk/files/pdf/papers/cmmr2012submission95.pdf

Steedman, M. J. (1984). A generative grammar for jazz chord sequences. *Music Perception, 2*(1), 52–77.

Strunk, S. (1979). The harmony of early bop: A layered approach. *Journal of Jazz Studies, 6*, 4–53.

Tymoczko, D. (2003). Function theories: A statistical approach. *Musurgia, 10*(3–4), 35–64.

Ulrich, W. (1977). The analysis and synthesis of jazz by computer. *Fifth International Joint Conference on Artificial Intelligence, Cambridge, MA* (Vol. 2, pp. 865–872). San Francisco, CA: Morgan Kaufmann.

Whorley, R., Wiggins, G., Rhodes, C. & Pearce, M. (2010). Development of techniques for the computational modelling of harmony. In D. Ventura, A. Pease, R. Pérez y Pérez, G. Ritchie & T. Veale , *ICCC 2010: 1st International Conference on Computational Creativity, Lisbon, Portugal* (pp. 11–15). Coimbra, Portugal: University of Coimbra.

Williams, J.K. (1982). *Themes composed by jazz musicians of the bebop era: A study of harmony, rhythm, and melody* (PhD thesis). Indiana University, Bloomington, IN, USA.

Witten, I.H., & Bell, T.C. (1991). The zero-frequency problem: estimating the probability of novel events in adaptive text compression. *IEEE Transactions on Information Theory, 37*(4), 1085–1094.