AUTOMATICALLY DETECTING KEY MODULATIONS IN J.S. BACH CHORALE RECORDINGS

Lesley Mearns, Emmanouil Benetos, and Simon Dixon
Centre for Digital Music, Queen Mary University of London, London E1 4NS, UK
{lesleym, emmanouilb, simond}@eecs.qmul.ac.uk

ABSTRACT

This paper describes experiments to automatically detect key and modulation in J.S. Bach chorale recordings. Transcribed audio is processed into vertical notegroups, and the groups are automatically assigned chord labels in accordance with Schönberg's definition of diatonic triads and sevenths for the 24 major and minor modes. For comparison, MIDI representations of the chorales are also processed. Hidden Markov Models (HMMs) are used to detect key and key change in the chord sequences, based upon two approaches to chord and key transition representations. Our initial hypothesis is that key and chord values which are derived from pre-eminent music theory will produce the most accurate models of key and modulation. The music theory models are therefore tested against models embodying Krumhansl's data resulting from perceptual experiments about chords and harmonic relations. We conclude that the music theory models produce better results than the perceptual data. The transcribed audio gives encouraging results, with the key detection outputs ranging from 79% to 97% of the MIDI ground truth results.

1. INTRODUCTION

Harmony, modulation and tonality are widely considered to be important indicators of individual composer and historical style [1]. However, harmony is not an exact science. A given chord sequence can imply more than one key, particularly in the absence of dominant harmony, and the precise moment of key change in diatonic modulation is difficult to demarcate precisely, due to the use of 'dual function' chords to smooth the transition between keys [1, 2]. Chords belonging to both the previous and new key may be reinterpreted to indicate the new key, a phenomenon referred to as 'revision' by Rorhmeier [3].

Thus there appears to be an incongruity in adopting a rigourous approach to harmony. However, a computational approach has advantages even for the experienced musicologist; the hidden or sub-conscious judgements of the analyst are rendered explicit, widely accepted facets of music theory or history may be systematically tested, and there is a pedadogical benefit, in that music analysis is made accessible to a broader community of people. In this paper, we aim both to test the possibility of obtaining musicological information directly from audio, which if successful, has the potential to open up new opportunities for musicological research based on musical recordings, and to ascertain whether perceptual or music theory data is more effective in the modelling of harmony.

To the authors' knowledge, this is the first study which utilizes polyphonic music transcription for systematic musicology research. Although key detection could also be achieved using an audio-based chord detection system, thus skipping the transcription step, we believe that fully transcribing audio is more appropriate, as it provides a framework for extracting information from a music piece that is not limited to a specific MIR task. We consider that such collaborative work has exciting potential, both for the improvement of automatic transcription, and for computational musicology.

The outline of the paper is as follows: Section 2 of the paper describes the data and the transcription methods. Section 3 outlines the automatic chord recognition method. Section 4 describes the different HMMs. Section 5 evaluates the results, and Section 6 presents conclusions and ideas for future work. In Fig. 1, a diagram for the proposed key modulation detection system can be seen.

2. MUSIC TRANSCRIPTION

Twelve J.S. Bach chorales were selected for experiments from www.jsbchorales.net, which provides organ-synthesized recordings along with aligned MIDI reference files. The size of the dataset is appropriate for transcription experiments [4, 5]. The list of the chorales employed for the key detection experiments can be seen in Table 1. Sample excerpts of original and transcribed chorales are available online [1] .

Firstly, the chorale recordings are transcribed into MIDI files using a modified version of the automatic transcription system that was proposed in [5]. The system is based on joint multiple-F0 estimation and note onset/offset detection. The constant-Q resonator time/frequency image (RTFI) [6] is employed due to its suitability for representing music signals. The number of bins per octave is set to 120, and the frequency range is set from 27.5 Hz (A0) to 12.5 kHz (the 3rd harmonic of C8). In order to suppress

[1] http://www.eecs.qmul.ac.uk/~emmanouilb/chorales.html

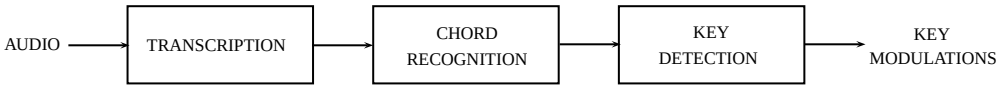

Figure 1. Key modulation detection diagram.

	BWV	Title
1	1.6	Wie schön leuchtet der Morgenstern
2	2.6	Ach Gott, vom Himmel sieh' darein
3	40.6	Schwing dich auf zu deinem Gott
4	57.8	Hast du denn, Liebster, dein Angesicht gänzlich verborgen
5	85.6	Ist Gott mein Schild und Helfersmann
6	140.7	Wachet auf, ruft uns die Stimme
7	253	Danket dem Herrn heut und allzeit
8	271	Herzlich tut mich verlangen
9	359	Werde munter, mein Gemüte
10	360	Werde munter, mein Gemüte
11	414	Danket dem Herrn, heuf und allzeit
12	436	Wie schön leuchtet der Morgenstern

Table 1. The list of organ-synthesized (top) and real (bottom) chorales used for key detection experiments.

timbral information, spectral whitening is applied [4], followed by a two-stage median filtering for noise reduction.

A log-frequency pitch salience function $s[n, p]$, is extracted, along with tuning and inharmonicity parameters. Here, $p = 1, \ldots, 88$ is the pitch index and n is the time frame. Onset detection is performed using a combination of a spectral flux-based and a salience function-based descriptor. For each segment defined by two consecutive onsets, multi-pitch estimation is applied in order to detect the pitches present. Pitch candidates are selected, and a pitch set score function combining several spectral and temporal features evaluates each possible pitch combination. Since the application of the transcription system concerns chorale recordings, the pitch range was limited to C2-A#6 and the maximum polyphony level was restricted to 4 voices. The pitch candidate set that maximizes the score function is selected as the pitch estimate for the current frame. Finally, note offset detection is also performed using HMMs trained on MIDI data from the RWC database [7]. Since the recordings are synthesized, tempo is constant and it can be computed using the onset detection functions from [5]. The estimated pitches in the time frames between two beats are averaged, resulting in a series of chords per beat. Transcription accuracy is 33.1% using the measure of [5], which however also takes into account note durations, hence the low value. An example of the transcription output of BWV 2.6 '*Ach Gott, vom Himmel sieh' darein*' is given in Fig. 2.

3. CHORD RECOGNITION

Transcribed audio, and for comparison, ground truth MIDI files, are segmented into a series of vertical notegroups according to onset times. Every new rhythmic value prompts the creation of a new vertical notegroup, so that notes which occur simultaneously or overlap in time are grouped. The pitch values within a group are converted to pitch classes 0

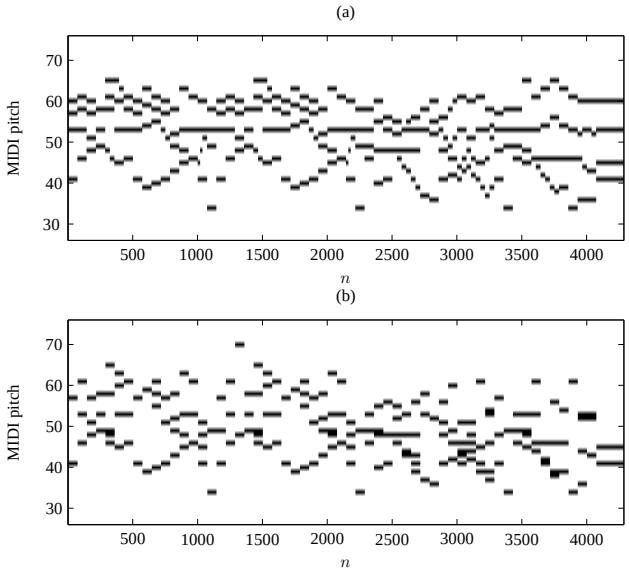

Figure 2. (a) The pitch ground-truth of BWV 2.6 '*Ach Gott, vom Himmel sieh' darein*'. (b) The transcription output of the same recording. The abscissa corresponds to 10 ms frames.

to 11, (0=C, 1=C# etc), and repeated tones are removed in order from the bass note to create a unique ordered set. For example, MIDI pitches {53,57,60,65}, (bass, tenor, alto, soprano) would become pitch classes {5,9,0,5} (modulo 12), which would become unique set {5,9,0}. The Bach chorales most commonly have a harmonic rhythm, (i.e. rate of harmonic change), of a crotchet beat, consequently for these experiments the vertical notegroups are organized into higher level groups which contain all of the notes present within this timing division. Thus, if the four notes of MIDI pitch {53,57,60,65} occurred at a metrical position of 1, but the MIDI note of pitch 65 (soprano voice) gave way to the seventh on the quaver offbeat, (metrical position 1.5), to MIDI pitch 63, the complete set of pitch classes within the crotchet beat would be {5,9,0,3}.

The notegroups are classified using a chord dictionary of templates expressed as ordered sets of pitch classes, (e.g. a C Major chord is {0,4,7}). No metrical, durational, or other type of weights are attached to the tones in the notegroup. All tones, including those occurring on offbeats, are equally operative as a possible part of the harmony. The approach is deliberate in order to capture elaborated seventh chords where the seventh note is introduced on the offbeat but is still an integral part of the harmony [8].

The chord matching process undergoes a series of iterations to find the template or templates that most closely match the presented notegroup in terms of edit distance. An exact match, (edit distance 0), would be for example a root position triad (e.g. {0,4,7}). An unordered exact

match, (edit distance 0.5), would be an inverted chord (e.g. {4,7,0}). The process continues, adding 1 for each insertion or deletion, up to a maximum edit distance of 2. If a match has not been found at this stage, the offbeat notes are removed from the group, and the match process is repeated with the set of notes which occurred on the beat. Due to the requirement of the HMM for a discrete sequence of chord symbols, groups of tones returning more than one possible chord classification are reduced to a single chord choice firstly by preferring root position chords, secondly by context matching with near neighbours, (two chords in either direction), and finally by random choice.

To measure the competence of the chord labelling process, the automatically generated chord sequences are compared to hand annotated sequences. Due to the laboriousness of hand annotation, half of the files in the set have been annotated with ground truth chord sequences. Each pair of chord index values in the sequences is compared, and a basic difference measure is calculated by counting the number of matches. The final counts are normalised, resulting in a proportional measure of matched or mismatched values between the two files (Table. 2). If two index values differ, the Levenshtein distance is calculated for the two pitch class sets represented as strings, to find out the degree of difference between the pitch class sets. Many of the index value mismatches found are in fact extremely close pitch class set matches, for example, {t, 2, 5} compared to {t, 2, 5, 9}, (t=10, e=11), generating a Levenshtein difference of 1. The Levenshtein distances calculated for each file are summed and normalised by the length of sequence to produce a combined measure of accuracy and distance.

	Transcribed Audio		**Ground Truth Midi**	
BWV	Match	Levenshtein	Match	Levenshtein
1.6:	0.45	1.20	0.86	0.30
2.6:	0.60	0.70	0.88	0.22
40.6:	0.55	0.95	0.83	0.28
57.8:	0.56	0.81	0.82	0.35
253:	0.55	0.75	0.83	0.35
436:	0.63	0.61	0.88	0.21
Totals Avg:	0.56	0.64	0.85	0.15

Table 2. Chord match results for transcribed audio and ground truth MIDI against hand annotations.

A greater quantity of label mismatches are found with the transcribed files than the ground truth MIDI files, depicting some of the pitch and timing errors resulting from the transcription process. Total chord mismatches between the transcribed data and the hand annotated data (i.e. where there are no pitches in common between the two pitch class sets), indicate an error in timing or quantisation. The greatest difficulty posed to the chord algorithm by the transcribed data however is the frequent presence of diads rather than triads in the groups. Resolving a diad correctly is not straightforward; if the diad is a third apart, this could imply either the upper or lower portion of a triad, equally, a diad a fifth apart could be either a major or a minor triad, a problem also encountered by Pardo [9]. The transcription algorithm has a low false alarm error rate and a high mis-detection rate, consequently the transcription process produces output which assists the chord method where the MIDI data poses problems; groups with suspended 9th and 13th notes, or other notegroups containing complex chord tones which are not defined in the chord dictionary, are captured from the transcribed data as simple triads whereas the MIDI data may result in a 'no chord' value. Complex chords such as 9ths and 13ths are less adaptable to the pitch class set match approach due to the fact that internal tones must be omitted from such chords to fit with four part harmony. Overall, the average accuracy levels for the ground truth files are in the upper range of accuracy results reported by Pardo [9]. The transcribed audio achieves an average of 65% correct of the ground truth result.

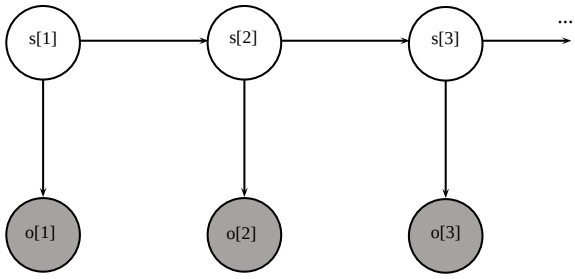

Figure 3. Graphical structure of the employed HMM for key modulation detection.

4. KEY MODULATION DETECTION

4.1 Hidden Markov Models

Key change detection is performed using a set of HMMs [10]. The observation sequence $O = \{o[n]\}, n = 1, \ldots, N$ is given by the output of the chord recognition algorithm in the previous section. The observation matrix (**B**) therefore defines the likelihood of a key given a chord. Likewise, the hidden state sequence which represents keys is given by $S = \{s[n]\}$. Each HMM has a key transition matrix **A** = $P(s[n]|s[n-1])$. There are two dimensions, of size 24×24, (representing the 12 major and 12 minor keys) which defines the probability of making a transition from one key to another. For a given chord sequence, the most likely key sequence is given by:

$$\hat{S} = \arg\max_{s[n]} \prod_{n} P(s[n]|s[n-1])P(o[n]|s[n]) \qquad (1)$$

which can be estimated using the Viterbi algorithm [10]. In Fig. 3, the graphical structure of the employed HMM model is shown.

4.2 Model Definitions

Five observation matrices (**B**) and four key transition (**A**) matrices are compared in total. Three of the observation matrices are derived from music theory, and are designed to represent and test Schönberg's theory with regard to the chord membership of the 24 major and minor modes [2]. Two further observation matrices use data from Krumhansl's

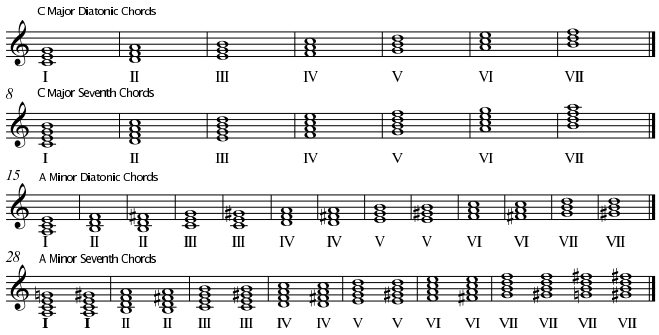

Figure 4. Diatonic chords for major and minor mode

perceptual experiments [11]. The four different versions of the key transition matrix (defined in section 4.4) are used in conjunction with all five of the observation matrices.

4.3 Observation Matrices

4.3.1 Music Theory Models

The extent to which a chord infers a key is modelled heuristically in the music theory observation matrices. The intention is to logically produce a set of musically plausible chord rankings per key across the full range of chords observed. The diatonic chords of a key are all indicative of the home key, however progressions containing chords II or IV with V or V7 are strongly indicative of the home key because they would have to be chromatically altered to imply a different key [1]. Similarly, the tonic triad, although it could be a member of several keys, tends to be prominent in the establishment of a tonal centre. Such chords may therefore be expected to rank highly compared to the lower values achieved by less characteristic chords. The relationship and interdependencies of individual tones, chords, and keys to human cognitive processing of tonality is not well understood. Consequently, to arrive at a score for a chord in relation to a key, points are given for both tone and chord properties. These include, points for each constituent tone per scale degree membership, partial points for ambiguous scale degree membership (i.e. 6th and 7th degrees in the minor key), for tonic chord status, and for being defined as a diatonic chord for the key by Schönberg. The points are then summed to give a total score for the chord in that key context.

Two of the Schönberg observation matrices symbolise the complete set of major, minor, diminished and augmented triads plus a 'no chord' value, resulting in a total of 49 possible chord symbols. The two matrices are weighted differently, inspired by Parncutt's psychoacoustical work suggesting that chords are heard as having singular identities which are prior to the constituent pitches [12]. Matrix *BSchCh* therefore assigns double points to the diatonic chord as whole and gives single points for individual tones, whereas *BSchP*, therefore gives double points to constituent tones and single points for diatonic chord status. The precise rules and values used are listed in Table 3.

For example, the chord rating for a 'C Major' triad in the key of 'C Major' for *BSchP* would be as follows:

Feature	**BSchP**	**BSchCh**
Diatonic chord	1	2
Scale degree	2	1
Dim/aug scale degree	1	0.5
Ambig scaledegree	1	0.5
Dim/aug ambigscaledegree	0.5	0.25
Tonicchord	1	1

Table 3. Rules for Schönberg observation matrix.

- C,E,G, three diatonic scale degrees = 2+2+2
- C,E,G, tonic triad = +1,
- C,E,G is listed by Schönberg as one of the diatonic triads = +1,
- Chord total = 8.

The third observation matrix *BSch7* symbolises the full set of triads and seventh chords elucidated by Schönberg [2] resulting in 22 chord definitions for the major key, and 30 chords for the minor key. (Please see Fig. 4.) The disparity in chord quantity is due to the optional raising of the 6th and 7th degree in the minor mode. A total number of 132 unique pitch class sets plus a 'no chord' value are therefore defined, bringing the total number of possible chord observations to 133.

The values assigned to each chord in the *BSch7* model are the same as those used for *BSchP*. In this model, the value for the dominant seventh of 'C Major' would be:

- G,B,D,F, four diatonic scale degrees = 2+2+2+2
- G,B,D,F, is listed by Schönberg as one of the diatonic sevenths for C Major = +1,
- Chord total = 9.

The dominant seventh chord thus is the highest signifier in the matrix for its key, satisfactorily articulating common practice in tonal harmony.

4.3.2 Music Perception Models

An HMM has been used previously to infer the overall key of a piece using Krumhansl's perceptual data, specifically the correlations between harmonic hierarchies as a representation of key distance, and harmonic hierarchy chord ratings, which are used to populate the key transition matrix and the observation matrix respectively [13, 11]. However, many of Krumhansl's chord ratings appear to contradict music theory. For example, in the C Major context, all of the twelve major triads, irrespective of which tone is the root, are rated as inferring the key of C Major more highly than any of the diatonic chords belonging to the key of C Major which are minor or diminished in profile. The data seems to suggest that in human perception, any major chord is more indicative of any major key, than the diatonic chords which make up that key, because it sounds major. From the perspective of music theory and common

	Key Context			Key Context			Key Context	
Chord ↓	**C Major**	**C Minor**	Chord ↓	**C Major**	**C Minor**	Chord ↓	**C Major**	**C Minor**
C Maj	6.6 (I)	5.30	C min	3.75	5.90	C dim	3.27	3.93
C#/Db Maj	4.71	4.11	C#/Db min	2.59	3.08	C#/Db dim	2.70	2.84
D Maj	4.60	3.83	D min	3.12	3.25	D dim	2.59	3.43
D#/Eb Maj	4.31	4.14	D#/Eb min	2.18	3.50	D#/Eb dim	2.79	3.42
E Maj	4.64	3.99	E min	2.76	3.33	E dim	2.64	3.51
F Maj	5.59	4.41	F min	3.19	4.60	F dim	2.54	3.41
F#/Gb Maj	4.36	3.92	F#/Gb min	2.13	2.98	F#/Gb dim	3.25	3.91
G Maj	5.33	4.38	G min	2.68	3.48	G dim	2.58	3.16
G#/Ab Maj	5.01	4.45	G#/Ab min	2.61	3.53	G#/Ab dim	2.36	3.17
A Maj	4.64	3.69	A min	3.62	3.78	A dim	3.35	4.10
Bb Maj	4.73	4.22	Bb min	2.56	3.13	Bb dim	2.38	3.10
B Maj	4.67	3.85	B min	2.76	3.14	B dim	2.64	3.18

Table 4. Krumhansl ratings of chords in harmonic-hierarchy experiments.

compositional practice, the data is counterintuitive and one could expect inconsistent results when used with common practice musical works.

The perceptual observation matrices symbolise the same chord set as the previously described triad based Schönberg models. The four triad based models therefore process identical chord sequences, allowing a direct comparison of the models based on music theory against those based on perceptual data.

The first matrix *BKrumOrig* is formulated using Krumhansl's chord ratings (Table 4) as per previous work by Noland [13], with the slight difference that all of Krumhansl's chord data is used without modification. In the absence of data for augmented triads, these plus the 'no chord' value are given a uniform low value of 1.0. As an experiment, a second observation matrix *BKrumMod* is also created, in which the apparently contradictory values for minor chords in the major key context which are part of the key, are swapped with the major chord values which are not part of the key. For example, in the 'C Major' context, the values for the 'D Major' chord are swapped with the value for the 'D Minor' (chord II), 'E Major' with 'E Minor' (chord III), 'A Major' with 'A Minor' (chord VI), and 'B Major' with 'B Diminished' (chord VII). Performing this swap leads to disproportionately high values for the remaining major chords which also belie the home key without a parallel minor or diminished chord with which to exchange the rating. Such chords have 1 subtracted from their rating value to bring the data more in line with the swapped changes, for example the chord rating of 4.36 for 'F# Major' becomes 3.36. The values for minor chords in the minor key context in this model are left unmodified.

4.4 Key Transition Matrices

Four different versions of the key transition matrix are formalized and used for all five of the observation matrices. The first matrix *ANeutral* is neutral, so that a move to any key is equally likely. The second transition matrix *AKrum* features Krumhansl's correlations between key profiles summed with 1 [11], similar to previous work by Noland [13]. The third and fourth matrices, referred to as *ASchEq*, and *ASchNL* respectively, are implementations of Schönberg's table of key circles, in which seven circles of increasing key distance from a given tonic are delineated [2]. Using pitch class set representations there are six unique circles only, the seventh containing the enharmonically equivalent keys of previous circles. Therefore the *ASchEq* subtracts an equal value of 0.25 for each key circle, commencing with an upper boundary of 2.0, and moving through the relative minor and then each successive circle, ending on the 6th circle. The *ASchNL* implementation uses an exponentially decreasing value, halving the deducted value for each circle. In *ASchNL* therefore, the numeric distance between the first circle and the sixth circle is smaller than the distance between the same two circles in the *ASchEq* matrix. For all key transition matrices except the neutral matrix, the central diagonal is slightly weighted, to give a small preference to stay in the current key. These values were determined empirically.

5. EVALUATION

5.1 Metrics

To provide a rigorous measure of accuracy of the outputs of the HMMs, each key value in the output sequences is compared to the corresponding hand-annotated key, and an error rate (Err), distance measure ($Dist$), measure of modulation concurrency ($Conc$), and modulation percentage ($Mods$) are calculated. Given N_{diff} the number of differences between output key and hand annotated key, N_{len} the length of the sequence, N_{cmod} the number of concurrent modulations, N_{hmod} the number of hand annotated modulations, and N_{omod} the number of modulations in the output, Err, $Conc$ and $Mods$ are defined as:

$$Err = \frac{N_{diff}}{N_{len}}, \quad Conc = \frac{N_{cmod}}{N_{hmod}}, \quad Mods = \frac{N_{omod}}{N_{hmod}} \quad (2)$$

The distance value $Dist$ captures both the number of differences and the extent of each difference relative to the circle of fifths when two key values are found to conflict. For example, the distance value for a key with another key on the same circle, i.e. its dominant, subdominant, or relative minor, is 1 whereas a key difference two fifths apart on the circle of fifths (in either direction) would result in a difference value of 2, and so on. $Conc$ refers to whether the HMM sequence changes key at precisely the same moment as the hand annotated sequence, regardless of whether the actual key change matches or not. Finally, $Mods$ shows

the percentage of the number of modulations in the HMM sequences compared to the number of modulations in the hand annotated key sequences. The results tables show the mean of all of the normalised data.

5.2 Results of Triadic Models

The results for all combinations of key transition matrices and observation matrices for the triadic models are shown in Table 5.2.

Error rates range from 0.26 to 0.35 for the transcribed data and 0.20 to 0.33 for the ground truth MIDI data sets. When the results are ordered by error, key distance measure, or the number of modulations relative to the number of modulations in the hand annotated data, the Schönberg observation matrices expose a pattern of consistently higher accuracy levels than the perceptual data matrices. The key transition matrices, for both the music theory models and the Krumhansl model, are less easily distinguished, however the *ANeutral* matrix gives the poorest performance overall.

Matching the exact moment of key change between the HMM and the hand annotated sequences is a predicament because the hand annotated sequences take into account phrasing; the key designations of preceding chords may be revised depending upon subsequent harmonic movement. The HMM has no phrase information, hence will change key solely on the basis of chord and key transition data. The models often display a key change timing lag of approximately one beat behind the annotated data. The modulation concurrence results are therefore quite low overall, but are significantly higher for the Schönberg observation matrices, with the combination of *AKrum* and *BSchCh* showing the best results. Fig. 5, which includes harmony annotations by Piston [1] demonstrates the issue. The *BSchCh* observation matrix changes to g# minor on precisely the same chord as Piston and holds the key for four beats. *BSchP* also changes to the correct key, but a beat later. Although Piston annotates the g# minor triad of bar 20 in the excerpt as III of E Major, it could equally be classed as chord I of g# minor, as per the HMM outputs. The music theory data also appears to illustrate greater sensitivity to short digressions through other keys than the perceptual data. In terms of recognising global key, the perceptual models, which tend to stay in the home key when harmonic divergence is only for the length of a couple of beats, could be a preferred choice. If closer recognition of secondary dominants is desired, the music theory based models appear to be the more suitable option.

The key output accuracy of the transcribed audio for all models is encouragingly high when compared to the ground truth MIDI, achieving an average of 79% of the the accuracy of the ground truth accuracy, despite the higher quantity of chord recognition errors for the transcribed data. The implication is that the transcribed audio is of sufficient quality for musicological work based on predominantly homophonic textures.

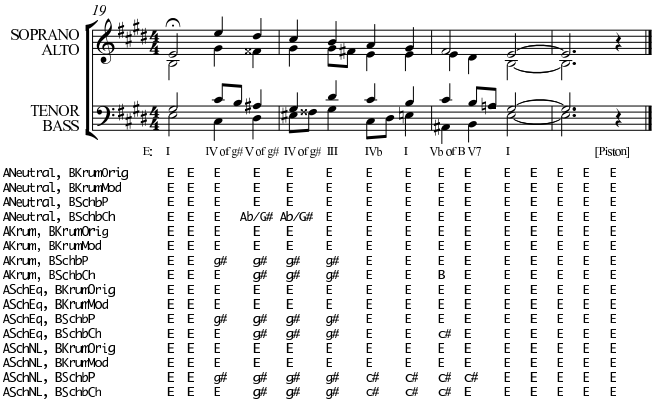

Figure 5. Key outputs of final bars of BWV 436 for all triad model combinations with harmony annotations by Piston [1]

5.3 Results of Sevenths Model

The results for the *BSch7* model in combination with all four key transition matrices are shown in Table 6. This more complex HMM containing 132 chords demonstrates a greater level of disparity from the hand annotated key sequences than the triad based models. The MIDI data marks an increase of 'no chord' values resulting from unclassified complex notegroups (especially suspended 9ths, 11ths and 13ths), however further research is required to understand precisely why the representation of complex chords produces more equivocal results. It is possible that the model results substantiate the notion that triads are more indicative of key than complex chords, excepting the dominant 7th. For this model, the error rates for the transcribed data are very close to the MIDI data achieving a relative best accuracy of 97%.

A Matrix ↓	**Transcribed Midi**				**Ground Truth Midi**			
	Err	*Dist*	*Conc*	*Mods*	*Err*	*Dist*	*Conc*	*Mods*
ANeutral	0.36	0.57	21.65	153.22	0.34	0.47	18.93	70.13
AKrum	0.35	0.50	34.66	205.22	0.35	0.47	23.24	110.36
ASchEq	0.36	0.51	37.02	238.45	0.34	0.47	27.12	113.27
ASchNonLin	0.37	0.49	29.14	217.29	0.36	0.47	21.44	109.61

Table 6. Key data for BSch7 with all four A matrices: error average, key distance of differences average, modulation concurrence average. Ground truth MIDI and transcribed file sets.

The results data intimates minimal differences between the four key transition matrices with the *BSch7* observation data, however closer inspection of the outputs of the different versions can be interpreted as indicating the harmonic complexity of the individual chorales. The outputs for all file sets for all matrix combinations were ordered per file error rate and distance value, resulting in a highly consistent ordering of the chorales across the data sets, an example of which is shown in Table 7. The chorales of less complex harmony, i.e. those which are in a major key and which hardly deviate from this key, appear at or near

B Matrix →	BSchP				BSchCh				BKrumOrig				BKrumMod			
A Matrix ↓	*Err*	*Dist*	*Conc*	*Mods*	*Err*	*Dist*	*Conc*	*Mods*	*Err*	*Dist*	*Conc*	*Mods*	*Err*	*Dist*	*Conc*	*Mods*
Transcribed																
ANeutral	0.35	0.74	11.55	51.09	0.27	0.45	25.23	109.42	0.28	0.42	4.76	18.89	0.32	0.51	2.68	9.68
AKrum	**0.26**	0.42	22.31	78.63	0.30	0.54	**37.58**	132.59	0.30	0.47	7.82	52.98	0.31	0.47	2.68	33.63
ASchEq	**0.26**	0.41	23.54	87.38	0.31	0.56	36.07	124.84	0.31	0.53	7.82	53.67	0.30	0.52	6.50	34.26
ASchNonLin	**0.26**	0.39	28.40	81.72	0.30	0.47	33.86	118.57	0.31	0.53	7.82	56.31	0.31	0.54	5.80	33.00
Ground Truth Midi																
ANeutral	0.31	0.45	9.25	38.26	0.27	0.45	24.59	85.26	0.22	0.37	8.45	31.72	0.33	0.53	5.01	22.87
AKrum	0.23	0.33	33.01	87.25	**0.20**	0.34	**46.03**	120.84	0.28	0.40	15.66	87.24	0.26	0.35	13.31	59.34
ASchEq	0.21	0.32	32.81	85.66	0.21	0.31	43.05	109.18	0.27	0.35	15.66	109.18	0.25	0.33	16.72	52.68
ASchNonLin	0.21	0.30	29.38	72.47	**0.20**	0.30	38.54	113.79	0.26	0.36	15.66	83.70	0.28	0.36	17.06	55.52

Table 5. Key data: error average, distance value for key differences average, percentage of modulation timing match, number of modulations as a percentage of hand annotated number of modulations.

	ASchbEq / BSch7			AKrum / BSch7			ANeutral / BSch7		
	BWV	*Err*	*Dist*	**BWV**	*Err*	*Dist*	**BWV**	*Err*	*Dist*
1	**1.6**	0.18	0.20	**1.6**	0.09	0.09	**1.6**	0.11	0.11
2	**414**	0.20	0.25	**414**	0.20	0.28	**414**	0.23	0.32
3	**253**	0.23	0.70	**140.7**	0.21	0.23	**359**	0.27	0.50
4	**436**	0.25	0.27	**253**	0.23	0.70	**360**	0.28	0.38
5	**140.7**	0.27	0.29	**360**	0.23	0.30	**140.7**	0.29	0.31
6	**360**	0.33	0.44	**436**	0.27	0.30	**436**	0.33	0.36
7	**359**	0.34	0.39	**359**	0.36	0.50	**253**	0.38	0.78
8	**57.8**	0.35	0.46	**57.8**	0.39	0.50	**271**	0.41	0.80
9	**271**	0.42	0.88	**271**	0.42	0.77	**57.8**	0.44	0.87
10	**85.6**	0.45	0.46	**85.6**	0.45	0.48	**2.6**	0.45	0.60
11	**2.6**	0.55	0.67	**2.6**	0.60	0.67	**85.6**	0.48	0.66
12	**40.6**	0.78	1.08	**40.6**	0.80	1.14	**40.6**	0.69	1.16

Table 7. Chorales ordered by error rate and distance using transcribed audio and Sch7 models.

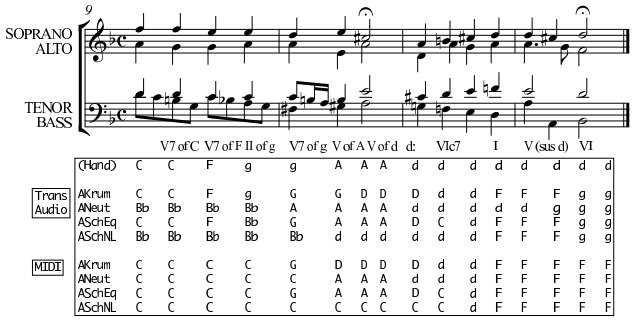

Figure 6. Mid bars of BWV 40.6 'Schwing dich auf zu deinem Gott' with HMM key outputs per transition matrix for BSch7 with hand annotated key and harmony labels.

the top of the list, with BWV 1.6 (in the key of F Major throughout), disclosing the least errors for almost every model. The three minor key chorales in the file set, BWV 85.6, 2.6, and 40.6, consistently show the greatest number of errors for all of the data sets.

The fragmentation of key sequence outputs identifies areas of harmonic complexity within the chorales. The mid section of BWV 40.6, (Fig. 6), exemplifies the difficulty of identifying a single key or exact point of key change in transitory sections; bar 9 implies several keys, and the end of bar 10 cadences in A, but is a secondary dominant of the home key of d minor.

6. CONCLUSIONS

This paper has presented an approach to key detection and key modulation using automatic chord classification of transcribed audio and ground truth MIDI data. A set of HMMs were explored using perceptual data and values calculated to represent formal music theory. Although the transcription error rate is quite high, key error rates for the audio recordings are only slightly higher compared to the key error rates for the ground-truth MIDI. Also, the key error rates are slightly higher for transcribed data using the triadic models, but the complex chord HMM exhibits remarkable alignment of results for both transcribed audio and MIDI data, suggesting that the quality of the transcribed chorales is of sufficiently high quality for the task. The music theory models were shown to outperform the perceptual data, with much of the variation between the models evincing the subtle and often ambiguous nature of musical harmony. Alignment of key boundaries is low overall with the HMM, due to the absence of phrase information, however the music theory observation matrix *BSchCh* showed a consistently better result for key change concurrence. Results are considered promising for the use of automatic transcription research in computational musicology. By combing key outputs with chord sequences, functional harmony could be obtained for the chorales measures of modulatory frequency and complexity could be derived.

Future work aims to improve the automatic chord recognition method to be able to classify complex chords and tone groups containing non-chord tones by identifying structural tones. Prior knowledge of key and harmony could also be used to improve the output of a transcription process; for example, initially transcribing the data, obtaining harmony information, and subsequently re-transcribing the data utilising this knowledge. For music research the combination of transcription and a musicology system could facilitate the analysis of large corpuses of audio data with the potential for some exciting discoveries about music.

Acknowledgments

Lesley Mearns is supported by an EPSRC DTA studentship. Emmanouil Benetos is supported by a Westfield Trust PhD Studentship (Queen Mary, University of London).

7. REFERENCES

[1] W. Piston and I. Ekeland, *Harmony*. W. W. Norton & Company, 1983.

[2] A. Schönberg, *Theory of Harmony*. University of California Press, 1911.

[3] M. Rohrmeier, "Modelling dynamics of key induction in harmony progressions," in *SMC 2007*, 2007.

[4] A. Klapuri and M. Davy, Eds., *Signal Processing Methods for Music Transcription*, 2nd ed. New York: Springer-Verlag, 2006.

[5] E. Benetos and S. Dixon, "Polyphonic music transcription using note onset and offset detection," in *IEEE International Conference on Audio, Speech and Signal Processing*, Prague, Czech Republic, May 2011.

[6] R. Zhou, "Feature extraction of musical content for automatic music transcription," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, Oct. 2006.

[7] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: music genre database and musical instrument sound database," in *Int. Conf. Music Information Retrieval*, Oct. 2003.

[8] C. H. Kitson, *Elementary Harmony*. Oxford University Press, 1920.

[9] B. Pardo and W. Birmingham, "Algorithms for chordal analysis," *Computer Music Journal*, vol. 26, no. 2, pp. 22–49, Summer 2002.

[10] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[11] C. L. Krumhansl, Ed., *Cognitive Foundations of Musical Pitch*, 1st ed. Oxford University Press, 1990.

[12] R. Parncutt, *Harmony: a psychoacoustical approach.* Springer-Verlag, 1989.

[13] K. Noland, "Computational tonality estimation: signal processing and hidden Markov models," Ph.D. dissertation, Queen Mary University of London, UK, Mar. 2009.