Improving Automatic Music Transcription Through Key Detection

Emmanouil Benetos[1]*, Andreas Jansson[1], and Tillman Weyde[1]

[1]*Music Informatics Research Group, Department of Computer Science, City University London, UK*

Correspondence should be addressed to Emmanouil Benetos (emmanouil.benetos.1@city.ac.uk)

ABSTRACT

In this paper, a method for automatic transcription of polyphonic music is proposed that exploits key information. The proposed system performs key detection using a matching technique with distributions of pitch class pairs, called Zweiklang profiles. The automatic transcription system is based on probabilistic latent component analysis, supporting templates from multiple instruments, as well as tuning deviations and frequency modulations. Key information is incorporated to the transcription system using Dirichlet priors during the parameter update stage. Experiments are performed on a polyphonic, multiple-instrument dataset of Bach chorales, where it is shown that incorporating key information improves multi-pitch detection and instrument assignment performance.

1. INTRODUCTION

Automatic music transcription (AMT) refers to the process of converting an acoustic musical signal into some form of musical notation [11]. It has many applications in music information retrieval and computational musicology, but still remains an open problem especially for transcribing multiple-instrument polyphonic music. A possible reason for this limitation is that the vast majority of AMT algorithms does not exploit musicological information, but relies exclusively on acoustic models [4].

Key is a musicological cue that has been utilised in the past for informing music transcription systems. Ryynänen and Klapuri used key information for estimating possible note transitions for a melody and bass line transcription system [15]. In [14], Raczyński et al. exploit chord and key information for improving a spectrogram factorisation-based AMT systems in a post-processing step.

In this work, we propose a system for automatic music transcription which utilises key as a prior information. Using the Zweiklang algorithm for key detection [10], the proposed model incorporates key information as Dirichlet priors to the pitch activation matrix of a multiple-instrument polyphonic music transcription system [3]. Contrary to [15], the proposed system is not limited to transcribing melody and bass lines. Also, contrary to [14], the proposed system does not exploit musicological information in a post-processing step, but jointly combines the key information within an acoustic model, in order to derive a more musically meaningful transcription. The proposed system is evaluated on multiple-instrument recordings from the Bach10 database [6], both for multi-pitch detection and instrument assignment, the latter a considerably more challenging task. Results show that utilising key information can lead to a clear improvement when transcribing tonal music.

The outline of this paper is as follows. In Section 2, the key detection system is described. The proposed automatic transcription system using key information is presented in Section 3. The dataset used for experiments, evaluation metrics, and results, are shown in Section 4. Finally, conclusions are drawn and future directions are indicated in Section 5.

2. KEY DETECTION

To estimate keys from audio we use the Zweiklang algorithm which was presented and publicly evaluated in [10]. The algorithm uses template matching, in a fashion similar to the key detection algorithms of Pauws [13] and Izmirli [9], where the mean chromagram of an au-

*Emmanouil Benetos is supported by a City University London Research Fellowship.

dio piece is correlated with pre-defined profiles for the 24 Western major and minor keys.

The Zweiklang algorithm differs from traditional template matching based methods in two major ways. Firstly, instead of matching chromagrams, the algorithm matches Zweiklang profiles. A Zweiklang is an ordered pair of pitch classes. As in previous approaches, the audio is segmented into fixed-size time windows and the 12-bin chromagram is computed for each window. We apply quantile filtering to remove low-energy bins before folding into one single octave. The Zweiklang $z_{i,j}$ of a non-zero chromagram is defined as the indices i, j of the two strongest chroma bins, ordered by their strength. A Zweiklang profile Z consists of relative frequencies f of Zweiklangs in a piece or a corpus in a 12×12 matrix with $Z_{i,j} = f(z_{i,j})$.

The second difference is that instead of pre-defining profiles, profiles are learned from annotated audio. The training data consists of 960 synthesised Lilypond scores annotated with key information. The Lilypond scores were downloaded from the Mutopia project[1] and consist of mainly classical music. The audio was synthesized at a sample rate of 11025 Hz, and the window size was set to 4096 samples. A Zweiklang profile is learned from each piece and transposed to either C major or C minor, depending on the mode. The major and minor profiles are summed and normalised, and rotated to all the other keys. The C major and C minor profiles derived from the Zweiklang algorithm are shown in Figures 1 and 2, respectively.

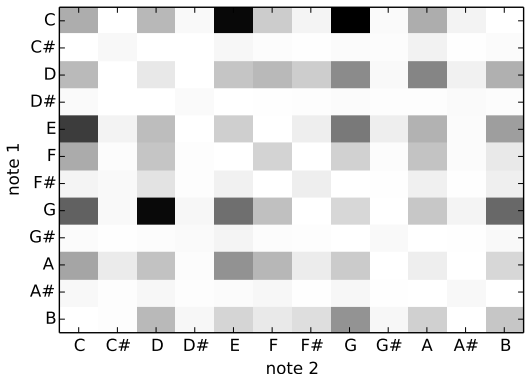

Fig. 1: Zweiklang C major profile

[1] http://www.mutopiaproject.org

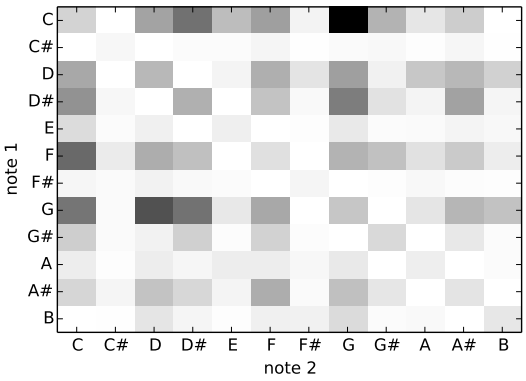

Fig. 2: Zweiklang C minor profile

To determine the key of an unknown piece of audio, the Zweiklang profile of the new piece is computed and compared to the 24 trained Zweiklang profiles. The key of the most similar profile (where the similarity measure is the dot product) is chosen as the key. The Zweiklang algorithm reached a weighted key score of 80% for the 2012 MIREX Audio Key Detection task [10].

3. AUTOMATIC MUSIC TRANSCRIPTION

The motivation behind the proposed model is to incorporate key information when automatically transcribing music. Essentially, when a (tonal) music piece is in a certain key, the probabilities of encountering notes that belong to that key are generally greater than for pitches that do not belong to that key [12]. To that end, we employ the model of [3], which supports the transcription of multiple-instrument polyphonic music and also supports pitch deviations or frequency modulations. The model of [3] is based on probabilistic latent component analysis (PLCA), which is a subspace analysis method which can be used for decomposing spectrograms [18]. As was shown in [20], PLCA can utilise priors for estimating unknown model parameters; for the present work, detected key can be used as prior for informing (multi-)pitch estimation.

The transcription model takes as input a normalised log-frequency spectrogram $V_{\omega,t}$ (ω is the log-frequency index and t is the time index) and approximates it as a bivariate probability distribution $P(\omega,t)$. $P(\omega,t)$ is decomposed into a series of log-frequency spectral templates per pitch, instrument, and log-frequency shifting

(which indicates deviation with respect to the ideal tuning), as well as matrices for pitch, instrument, and tuning distributions.

The model is formulated as:

$$P(\omega,t) = P(t) \sum_{p,f,s} P(\omega|s,p,f) P_t(f|p) P_t(s|p) P_t(p) \quad (1)$$

where p denotes pitch, s denotes instrument source, and f denotes log-frequency shifting. $P(t)$ is the energy of the log-spectrogram, which is a known quantity. $P(\omega|s,p,f)$ denote pre-extracted log-spectral templates per pitch p and instrument s, which are also pre-shifted across log-frequency. The pre-shifting operation is made in order to account for pitch deviations, without needing to formulate a convolutive model across log-frequency, as e.g. in [19]. $P_t(f|p)$ is the time-varying log-frequency shifting distribution per pitch, $P_t(s|p)$ is the time-varying source contribution per pitch, and finally, $P_t(p)$ is the pitch activation, which essentially is the resulting transcription. As a log-frequency representation we use the constant-Q transform (CQT) with a log-spectral resolution of 60 bins/octave [16].

The unknown model parameters ($P_t(f|p)$, $P_t(s|p)$, $P_t(p)$) can be iteratively estimated using the expectation-maximisation (EM) algorithm [5]. For the *Expectation* step, the following posterior is computed:

$$P_t(p,f,s|\omega) = \frac{P(\omega|s,p,f) P_t(f|p) P_t(s|p) P_t(p)}{\sum_{p,f,s} P(\omega|s,p,f) P_t(f|p) P_t(s|p) P_t(p)} \quad (2)$$

For the *Maximization* step, without using any priors, unknown model parameters are updated using the posterior from (2):

$$P_t(f|p) = \frac{\sum_{\omega,s} P_t(p,f,s|\omega) V_{\omega,t}}{\sum_{f,\omega,s} P_t(p,f,s|\omega) V_{\omega,t}} \quad (3)$$

$$P_t(s|p) = \frac{\sum_{\omega,f} P_t(p,f,s|\omega) V_{\omega,t}}{\sum_{s,\omega,f} P_t(p,f,s|\omega) V_{\omega,t}} \quad (4)$$

$$P_t(p) = \frac{\sum_{\omega,f,s} P_t(p,f,s|\omega) V_{\omega,t}}{\sum_{p,\omega,f,s} P_t(p,f,s|\omega) V_{\omega,t}} \quad (5)$$

Since we aim to incorporate key knowledge for informing the resulting transcription, we will impose a prior on the pitch activation $P_t(p)$. For enforcing a structure on pitch distributions, we employ the Krumhansl tone profiles for major and minor keys [12], which have been widely used in key recognition research (e.g. [21, 8]). By expanding the profiles on the considered note range (88 pitches, from A0 to C8) and by transposing them to all 12 keys, a $2\times12\times88$ pitch profile tensor is created, denoted $K_{m,d,p}$, where m denotes mode and d denotes root. The derived Krumhansl matrix for the 12 major keys can be seen in Fig. 3.

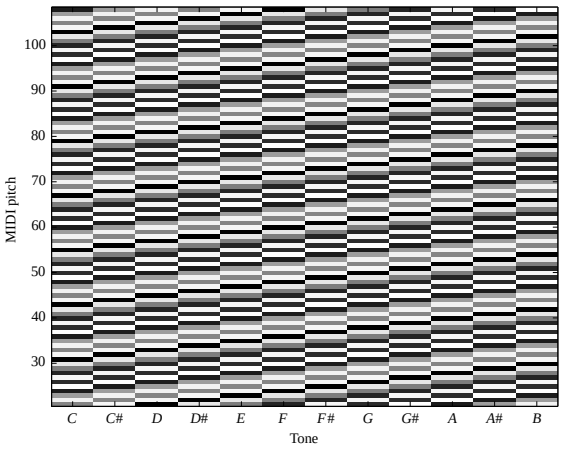

Fig. 3: The derived Krumhansl matrix for major keys.

As shown in [20], PLCA-based models use multinomial distributions. Since the Dirichlet distribution is conjugate to the multinomial, a Dirichlet prior can be used to enforce structure on the distribution of $P_t(p)$. Following the procedure of [20], we define the Dirichlet hyperparameter for the pitch structure (given a mode and root) as:

$$\alpha(p|t)_{m,d} = K_{m,d,p} P_t(p) \quad (6)$$

where $\alpha(p|t)_{m,d}$ essentially represents a modified transcription, giving higher probability to notes which are more frequently encountered in the key (e.g. tonic, dominant) than to notes which are not. Thus, the modified update for the pitch activation, given a detected mode and tone from the system of Section 2, is given by:

$$P_t(p) = \frac{\sum_{\omega,f,s} P_t(p,f,s|\omega) V_{\omega,t} + \kappa \alpha(p|t)_{m,d}}{\sum_{p,\omega,f,s} P_t(p,f,s|\omega) V_{\omega,t} + \kappa \alpha(p|t)_{m,d}} \quad (7)$$

where κ is a weight parameter expressing how much the prior should be imposed.

Thus, for estimating unknown model parameters, we iterate Eqs. (2)-(4) and (7); for the proposed system we set the number of iterations to 50. As in [20], weight

κ gradually decreases from 1 to 0 throughout iterations (initialising the model but letting it converge in the end). The resulting transcription is given by $P(p,t) = P(t)P_t(p)$. After performing 5-sample median filtering for note smoothing, thresholding is performed on $P(p,t)$ followed by minimum note duration pruning set to 40ms (corresponding to the length of one time frame) in order to convert $P(p,t)$ into a binary piano-roll representation.

4. EVALUATION

In this section, the dataset used for experiments will be presented, the evaluation metrics are defined, and finally the experimental results are presented and discussed.

4.1. Dataset

We employ the Bach10 dataset for testing the proposed transcription system [6]. The Bach10 dataset is a freely available multi-track collection of multiple-instrument polyphonic music, suitable for multi-pitch detection and instrument assignment experiments. It consists of ten recordings of J.S. Bach chorales, performed by violin, clarinet, saxophone, and bassoon. Pitch ground truth for each instrument is also provided. Due to the tonal content of the dataset, it is suitable for testing the incorporation of key information in a transcription system.

For training the system, pre-extracted and pre-shifted log-spectral templates are extracted for the instruments present in the dataset, using isolated note samples from the RWC database [7]. For comparative purposes, we also extract templates directly from the individual Bach10 tracks (the same recordings that are used for testing), which can indicate the upper performance limit of the proposed system.

4.2. Metrics

For evaluating the performance of the proposed system for multi-pitch detection and instrument assignment, we employ the precision, recall, and F-measure metrics, which are commonly used in transcription evaluations:

$$Pre = \frac{N_{tp}}{N_{sys}}, \quad Rec = \frac{N_{tp}}{N_{ref}}, \quad F = \frac{2 \cdot Rec \cdot Pre}{Rec + Pre} \qquad (8)$$

where N_{tp} is the number of correctly detected pitches, N_{sys} is the number of detected pitches, and N_{ref} is the number of ground-truth pitches. In all cases, due to the fact that the system might converge to different solutions, we perform 10 runs for each recording and present the average performance of the system.

As in the MIREX note tracking evaluations [1], a detected note is considered correct is if its pitch is the same as the ground truth pitch and its onset is within a 50ms tolerance interval of the ground-truth onset. For multi-pitch detection evaluations, we compare the transcription output of the system with the ground-truth, and denote the corresponding F-measure as F_{mp}.

For the instrument assignment evaluations we use the pitch ground-truth of each instrument separately. We derive the instrument-specific output as $P(s = i, p, t) = P_t(s = i|p)P_t(p)P(t)$, where i corresponds to a specific instrument. As was done for the multi-pitch evaluation metrics, we define the following instrument assignment metrics: F_v, F_c, F_s, F_b, which correspond to violin, clarinet, saxophone, and bassoon, respectively (each instrument metric is compared with the instrument pitch ground-truth). We also use a mean instrument assignment metric, averaging all the above instrument-specific metrics, denoted as F_{ins}.

4.3. Results

Multi-pitch detection and instrument assignment experiments are performed using the proposed system, compared with the system of [3], which uses the same PLCA model and post-processing procedure without any key priors. Multi-pitch detection results using training samples from the RWC database can be seen in Table 1. In terms of F-measure, there is an improvement of over 1% when using key information. As explained in [2], this improvement in the context of multi-pitch detection evaluation is statistically significant. The improvement is most evident in recordings that stay in the same key for their entire duration (about half of the pieces modulate to another key). For comparison with the method of [6] (where the Bach10 was first introduced), the proposed method using the frame-based accuracy metric defined in [6] reaches 71.4%, whereas the method of [6] reaches 69.7% (with unknown polyphony).

As an example, the transcription output using both the proposed model and that of [3], along with the pitch ground truth, is given for piece No.9 in Fig. 4. It can be seen that key information (G minor in this case) has resulted in an improved transcription (e.g. for MIDI pitch 50, i.e. B♭, or in the removal of certain higher octave errors).

System	F_{mp}	Pre_{mp}	Rec_{mp}
[3]	61.48%	59.76%	63.42%
Proposed	62.54%	61.53%	63.77%

Table 1: Multi-pitch detection results using the proposed model, compared with the model of [3] (training data from the RWC database).

System	F_{mp}	Pre_{mp}	Rec_{mp}
[3]	66.27%	64.75%	67.97%
Proposed	67.56%	67.60%	67.58%

Table 2: Comparative multi-pitch detection results using training data from the Bach10 recordings.

Table 2 presents multi-pitch detection results using training data from the same source as in the test recordings. This demonstrates a potential upper limit of the performance of the proposed model, and it can be seen that the reported improvement is over 5% compared to using training data from a different source. Apart from that fact, a similar improvement is reported for the system incorporating key prior information over the system of [3].

Finally, instrument assignment results are displayed in Table 3. As can be seen, the performance of the proposed system regarding instrument assignment is much lower compared to multi-pitch detection; this can be explained by the fact that instrument assignment is a much more challenging problem, since it requires a correct identification of a note, which should also be assigned to the correct instrument. However, a clear improvement can be seen when incorporating key information. Interestingly, an improvement of about 2% is reported for violin and bassoon, whereas the improvement is minimal for clarinet and saxophone. For the saxophone, this can be attributed to the low score reported for the instrument, as the spectral shape of the RWC samples does not match the spectral shape of the Bach10 source (an alto sax is used in the recordings, whereas a tenor sax is used for training the system).

5. CONCLUSIONS

In this paper, we proposed a system for automatic music transcription which incorporated key detection information. Key detection was performed using the Zweiklang

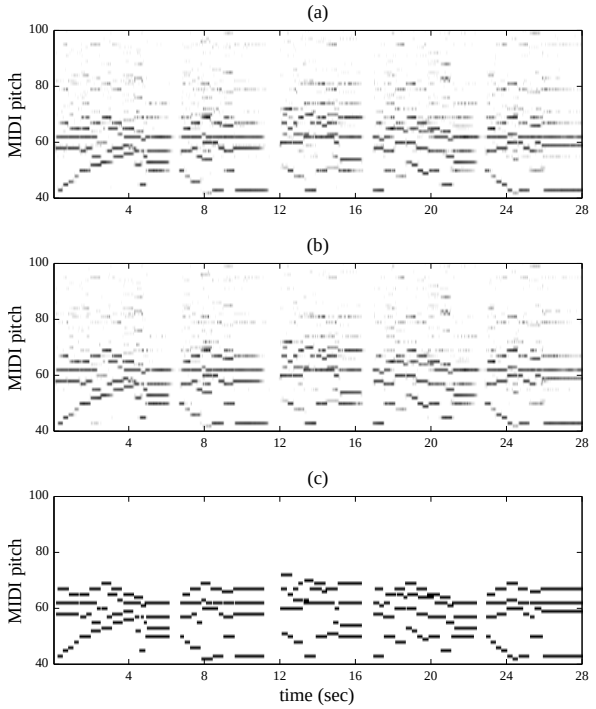

Fig. 4: Transcription results for piece No.9 (“Jesus”). (a) The pitch activation $P(p,t)$ using the model of [3]. (b) The pitch activation $P(p,t)$ using the proposed model. (c) The pitch ground truth.

algorithm [10], and the key information was incorporated as Dirichlet prior to a PLCA-based multiple-instrument transcription system [3]. Multi-pitch detection and instrument assignment experiments on the Bach10 dataset [6] showed that by imposing a prior structure on transcribed pitches based on detected key information can lead to a clear performance improvement. In a larger MIR context, this is a clear case when incorporating musicological knowledge can improve music content analysis [17].

The proposed method, however, has made two strong assumptions: that the input recordings are tonal and that the key stays the same throughout the recording. In reality, the vast majority of Western music is tonal, although modulations are frequently encountered. Even though there is a vast literature on key detection methods, a relatively small part of research has been carried out on key modulation detection. This can also be attributed by current public evaluation initiatives such as the MIREX Audio Key Detection task, which do not consider key mod-

System	F_v	F_c	F_s	F_b	F_{ins}
[3]	14.38%	35.59%	11.74%	39.78%	25.37%
Proposed	16.46%	35.79%	11.83%	41.25%	26.33%

Table 3: Instrument assignment results using the proposed model, compared with the model of [3].

ulations [1].

In the future, we will carry out research in key modulation detection, and integrate a key modulation detection system with a system for transcribing tonal music. We will also incorporate additional musicological cues for improving automatic music transcription performance, such as combining polyphonic music language models with acoustic models [22].

6. REFERENCES

[1] Music Information Retrieval Evaluation eXchange (MIREX). `http://music-ir.org/mirexwiki/`.

[2] E. Benetos. *Automatic transcription of polyphonic music exploiting temporal evolution*. PhD thesis, Queen Mary University of London, December 2012.

[3] E. Benetos, S. Cherla, and T. Weyde. An effcient shift-invariant model for polyphonic music transcription. In *6th International Workshop on Machine Learning and Music*, Prague, Czech Republic, September 2013.

[4] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri. Automatic music transcription: breaking the glass ceiling. In *13th International Society for Music Information Retrieval Conference*, pages 379–384, Porto, Portugal, October 2012.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[6] Z. Duan, B. Pardo, and C. Zhang. Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):2121–2133, November 2010.

[7] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: music genre database and musical instrument sound database. In *International Conference on Music Information Retrieval*, Baltimore, USA, October 2003.

[8] D. Hu and L. K. Saul. A probabilistic topic model for unsupervised learning of musical key-profiles. In *10th International Society for Music Information Retrieval Conference*, pages 441–446, 2009.

[9] O. Izmirli. Template based key finding from audio. In *International Computer Music Conference*, pages 211–214, Barcelona, Spain, 2005.

[10] A. Jansson and T. Weyde. MIREX 2012: key recognition with zweiklang profiles. In *Music Information Retrieval Evaluation eXchange (MIREX)*, October 2012.

[11] A. Klapuri and M. Davy, editors. *Signal Processing Methods for Music Transcription*. New York, 2006.

[12] C. Krumhansl. *Cognitive Foundations of Musical Pitch*. Oxford University Press, Oxford, UK, 1990.

[13] S. Pauws. Musical key extraction from audio. In *International Symposium on Music Information Retrieval*, Barcelona, Spain, 2004.

[14] S. Raczynski, E. Vincent, and S. Sagayama. Dynamic Bayesian networks for symbolic polyphonic pitch modeling. *IEEE Transactions on Audio, Speech and Language Processing*, 21(9):1830–1840, April 2013.

[15] M. Ryynänen and A. Klapuri. Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal*, 32(3):72–86, fall 2008.

[16] C. Schörkhuber and A. Klapuri. Constant-Q transform toolbox for music processing. In *7th Sound and Music Computing Conference*, Barcelona, Spain, July 2010.

[17] X. Serra, M. Magas, E. Benetos, M. Chudy, S. Dixon, A. Flexer, E. Gómez, F. Gouyon, P. Herrera, S. Jorda, O. Paytuvi, G. Peeters, J. Schlüter, H. Vinet, and G. Widmer. *Roadmap for Music Information ReSearch*. 2013.

[18] M. Shashanka, B. Raj, and P. Smaragdis. Probabilistic latent variable models as nonnegative factorizations. *Computational Intelligence and Neuroscience*, 2008. Article ID 947438.

[19] P. Smaragdis. Relative-pitch tracking of multiple arbitary sounds. *Journal of the Acoustical Society of America*, 125(5):3406–3413, May 2009.

[20] P. Smaragdis and G. Mysore. Separation by "humming": user-guided sound extraction from monophonic mixtures. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 69–72, October 2009.

[21] D. Temperley. A Bayesian approach to key-finding. In C. Anagnostopoulou, M. Ferrand, and A. Smaill, editors, *Music and Artificial Intelligence*, volume 2445 of *Lecture Notes in Computer Science*, pages 195–206. Springer Berlin Heidelberg, 2002.

[22] F. Weninger, B. Schuller, C. Liem, F. Kurth, and A. Hanjalic. Music information retrieval: An inspirational guide to transfer from related disciplines. In M. Müller, Ma. Goto, and M. Schedl, editors, *Multimodal Music Processing*, volume 3 of *Dagstuhl Follow-Ups*, pages 195–216. 2012.