THE SOUSTA CORPUS: BEAT-INFORMED AUTOMATIC TRANSCRIPTION OF TRADITIONAL DANCE TUNES

Andre Holzapfel
Austrian Research Institute
for Artificial Intelligence (OFAI)
andre@rhythmos.org

Emmanouil Benetos
Centre for Digital Music
Queen Mary University of London
emmanouil.benetos@qmul.ac.uk

ABSTRACT

In this paper, we present a new corpus for research in computational ethnomusicology and automatic music transcription, consisting of traditional dance tunes from Crete. This rich dataset includes audio recordings, scores transcribed by ethnomusicologists and aligned to the audio performances, and meter annotations. A second contribution of this paper is the creation of an automatic music transcription system able to support the detection of multiple pitches produced by lyra (a bowed string instrument). Furthermore, the transcription system is able to cope with deviations from standard tuning, and provides temporally quantized notes by combining the output of the multi-pitch detection stage with a state-of-the-art meter tracking algorithm. Experiments carried out for note tracking using 25ms onset tolerance reach 41.1% using information from the multi-pitch detection stage only, 54.6% when integrating beat information, and 57.9% when also supporting tuning estimation. The produced meter aligned transcriptions can be used to generate staff notation, a fact that increases the value of the system for studies in ethnomusicology.

1. INTRODUCTION

Automatic music transcription (AMT), the process of converting a music recording into notation, has largely focused on genres of eurogenetic [17] popular and classical music and especially on piano repertoire; see [3] for a recent overview. This is reflected in various AMT datasets, which consist of audio recordings along with a machine readable reference notation that specifies the time values of note onsets and offsets. Such datasets include the RWC database [14], the MAPS dataset [12], and the Bach10 dataset [9]. The reasons for the focus on certain styles seem manifold: Some aspects that might play a role are the cultural background of the AMT engineers, the relative ease of compiling reference notations for a piano using *MIDI*, and the predominant goal of transcription, i.e. the piano-roll, being closely related to the piano. As a point of fundamental importance, eurogenetic music lends itself nicely to the task of transcription, because in most cases a composition is first notated, and then performed using this notation. Hence, the notation can be interpreted as the ground-truth for an AMT system. The attempt to reconstruct this ground-truth, which is seen as a hidden generative concept for the performance [6], appears, at least at first glance, to be a well-defined task.

However, in the field of ethnomusicology, the process of transcribing a music performance mainly serves the means to analyse the structure of previously not notated music [11]. As a first contribution of this paper, we align a set of such recordings to transcriptions by ethnomusicologists, this way compiling an evaluation corpus for AMT that can enable us to monitor the performance of AMT systems on the music of a specific oral tradition. The music consists of Cretan dance tunes that were performed by Cretan musicians and recorded and transcribed by ethnomusicologists in the Crinnos project [2] that targeted the documentation of that specific music idiom. Only a small subset of the pieces recorded in the Crinnos project were transcribed, due to the large amount of effort that manual transcription takes. While it is clear that the building blocks of the tunes are small melodic phrases (see Section 2 for more detail), the way these phrases are strung together is largely improvised in the performance. These choices are not verbalized by the musicians, and an accurate transcription method will constitute an important tool to infer the grammar that underlies folk dance tunes in the area of the Eastern Mediterranean and beyond.

Therefore, as the second contribution of this paper, we extend an existing transcription algorithm [5] to be able to cope with tuning deviations and to take into account the metrical structure of the dance tunes. In Cretan music, as well as in many other music styles in the world, musicians tune their instruments according to personal preference. To the authors' knowledge, whilst several AMT systems support the extraction of multiple pitches in a high frequency resolution (e.g. [9, 13]), no AMT system has yet exploited that information for estimating the overall tuning level and to compensate for tuning deviations during the pitch quan-

AH is supported by the Austrian Science Fund (FWF: M1995-N31), and by the Vienna Science and Technology Fund (WWTF, project MA14-018).

EB is supported by a UK Royal Academy of Engineering Research Fellowship (grant no. RF/128).

Both authors contributed equally to this paper.

tisation step. In addition, for many music styles, especially when related to dance, a clear metrical organization and a predictable tempo development enable for synchronisation between dancers and musicians in performances. For that reason, we apply a state-of-the-art meter tracking algorithm [15] for tracking beats and measures, and apply this information in order to achieve a temporal quantisation of note positions obtained from our AMT system. This way, we can obtain a transcription with temporal precision that is clearly increased to that of previously presented systems. In addition, this step enables us to obtain a visualisation of the transcription in a staff notation including bar positions, a perspective that marks an important step beyond the piano-roll as AMT output.

Our paper is structured as follows; Section 2 provides some detail about the musical idiom and corpus, and describes the process that was followed to align transcriptions with performances on the note level. Section 3 summarizes the chosen AMT system, and describes the extensions proposed in this paper. We then evaluate the performance of our systems, and provide illustrative examples in Section 4. Section 5 concludes the paper.

2. THE SOUSTA CORPUS

2.1 Background and motivation

The recordings that constitute the Sousta corpus presented in this paper were conducted in 2004 within the Crinnos project [2] in Rethymnon, Crete, Greece. Within the Crinnos project 444 pieces of Cretan music were recorded, and 40 of these performances were transcribed by ethnomusicologists. The transcriptions contain the melody played by the main melody instrument, as well as the vocal melody if vocals are present in a piece, and ignore the rhythmic accompaniment. Half of the 40 transcriptions regard a specific dance called *Sousta*. These transcriptions were chosen for a note-to-note alignment for several reasons.

First, this way we obtain a music corpus that is highly consistent in terms of musical style, which made a unified alignment strategy applicable to the recordings. The Sousta dance is usually notated in $2/4$ meter, and is characterized by a relatively stable tempo that lies between 110-130 beats per minute (bpm). The instrumental timbres are highly consistent, with usually two Cretan lutes playing the accompaniment, and one Cretan lyra (a pear-shaped fiddle) playing the main melody. All recordings were performed in the same studio, but with differing musicians. Apart from supporting our alignment procedure, the consistency of the recordings will enable a style comparison between individual musicians as part of our future work.

The **second** reason to choose the Sousta tunes lies with their value for music segmentation approaches. Like many tunes in the Eastern Mediterranean, the Sousta dance follows an underlying syntax that has been termed as *parataxis* [18]. In parataxis tunes, the building elements are short melodic phrases that are strung together in apparently arbitrary order without clear conjunctive elements. These phrases have a length of typically two measures for the Sousta dance. The 20 transcriptions were analysed within the Crinnos project and its elementary melodic phrases were identified by the experts. This way, a catalogue of 337 phrases was compiled that describes the melodic content of the tunes. Each measure of the corpus is assigned to a particular phrase. The note-to-note alignment that is made available in this paper enables to identify the phrase boundaries within the recordings, and this way the corpus can serve for music segmentation experiments. Such a corpus can form a basis for the development of an accurate system for syntactic analysis of music styles in the Eastern Mediterranean and elsewhere.

Such an analysis system, however, needs to be built on an AMT system that works as accurately as possible, in order to be able to analyse performances for which no manual transcription is available. We take this as a motivation to use for the first time, to the best of our knowledge, a set of performance transcriptions as the source for what is usually called ground-truth in MIR. This way, as our **third** motivation for choosing this specific style, we contribute to a larger diversity in available AMT datasets, by providing access to the aligned data for research purposes. The challenging aspects for AMT systems are the high density of notes, the tuning deviations, and the necessary focus on a bowed string instrument (lyra) within a pitched percussive accompaniment (lutes).

2.2 Alignment procedure

The first step to obtain a note-to-note alignment is to correct for transpositions between transcription and performance. Four out of the 20 pieces were played either one or two semitones higher than the transcription implied. Apparently, transcribers preferred to notate the upper empty string of the lyra as the note A, even if the player tuned the instrument one or several semitones higher.

As a second step, we conduct a meter tracking to obtain estimations for beat and measure positions, using the algorithm presented in [15]. The meter tracker was trained on the meter-annotated Cretan music used in [15], and then applied to track the meter in the 20 Sousta performances (for more details on the tracking algorithm see Section 3.4).

After that, the MIDI file obtained from the transcription is synthesized, and the algorithm from [16] is used to obtain an initial alignment of the MIDI file to the recorded performance. The timing of the measures is extracted from the aligned MIDI using the Matlab MIDI Toolbox [10]. Each of the estimated measures in the MIDI is then corrected to take the time value of the closest beat as obtained from the meter tracker from the recording. This step was included to compensate for timing inaccuracies of the automatic alignment. The obtained downbeats were manually corrected using Sonic Visualizer [1] . The output of this process is the exact timing of all measures that are notated in the transcription.

These manually corrected measure positions are then used as a source for the exact timing of the pre-aligned

[1] `http://sonicvisualiser.org/`

MIDI, by determining an alignment curve that corrects all note onsets accordingly. After that, also the note durations are edited to fit the notated length in seconds (*e.g.* a quarter note at 120 bpm should last 0.5 s). The result was again manually checked for inaccuracies. In addition, vocal sections were manually annotated. During vocal sections, the main instrument usually stops, and the transcription of musical phrases is of interest only for the instrumental sections in a recording. To the authors' knowledge, this measure-informed process is a novel and promising way to generate note-level transcriptions, as opposed to performing manual note corrections on an automatically aligned MIDI file, or by relying on an expert musician to follow and perform the recorded music in real-time [20].

The obtained corpus contains 35357 aligned notes in 4455 measures, distributed along the 20 recordings with a total length of 71m16s [2], 84% being instrumental. The average polyphony (on voiced frames only) is 1.08, and the average note duration is 108ms.

3. BEAT-INFORMED TRANSCRIPTION

This section describes the AMT system developed to transcribe the traditional dance corpus of Section 2. The main contributions of the proposed system are: (i) Supporting the transcription of lyra, a bowed string instrument that is present in all recordings of the corpus, by supplying the system with lyra templates; (ii) Estimating the overall tuning level and compensating for deviations from 440Hz tuning (cf. Section 1 for a discussion on related work for tuning estimation in AMT systems); (iii) Incorporating meter and beat information (from either manual meter annotations or estimations from a state-of-the-art meter tracking system [15]), resulting in a temporally quantised music transcription.

As a basis for the proposed work, the AMT system of [4] is adapted, which was originally aimed for transcribing 12-tone equal tempered music and supported Eurogenetic orchestral instruments. The system is based on probabilistic latent component analysis (PLCA), a *spectrogram factorization* method that decomposes an input time-frequency representation into a series of note templates and note activations. The system of [4] also supports the extraction of tuning information per transcribed note, which is used in this paper to estimate the overall tuning level. A diagram for the proposed system can be seen in Fig. 1, with all system components being presented in the following subsections.

3.1 Time-Frequency Representation

As input time-frequency representation for the transcription system, the variable-Q transform (VQT) spectrogram is used [19], denoted $V_{\omega,t}$ (ω is the log-frequency index and t is the time index). Here, the interpolated VQT spectrogram has a frequency resolution of 60 bins/octave (i.e. 20 cent resolution), using a variable-Q parameter $\gamma = 30$, with a minimum frequency of 36.7 Hz (i.e. at D1). As with the constant-Q transform (CQT), this VQT representation allows for pitch changes to be represented by shifts across the log-frequency axis, whilst offering an increased temporal resolution in lower frequencies compared to the CQT.

[2] For a list of recordings see www.rhythmos.org/ISMIR2016Sousta.html

3.2 Multi-pitch Detection

The multi-pitch detection model takes as input the VQT spectrogram of an audio recording and returns an initial estimate of note events. Here, we adapt the PLCA-based spectrogram factorization model of [4] for transcribing music produced by lyra. The model approximates $V_{\omega,t}$ as a bivariate probability distribution $P(\omega, t)$, which is in turn decomposed into a series of probability distributions, denoting note templates, pitch activations, tuning deviations, and instrument/source contributions.

The model is formulated as:

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

where q denotes the sound state (e.g. attack, sustain parts of a note), p denotes pitch, s denotes instrument source, and f denotes log-frequency shifting with respect to 12-tone equal temperament (12-TET) at a tuning of 440 Hz for note A4. In (1), $P(t)$ is the energy of the VQT spectrogram, which is known. $P(\omega|q,p,f,s)$ is a 5-dimensional tensor that represents the pre-extracted spectral templates of lyra notes, per sound state q, pitch p and instrument model s, which are also pre-shifted across log-frequency f (cf. Section 4.1 on the extraction of lyra templates). $P_t(f|p)$ is the time-varying log-frequency shifting distribution per pitch (used to estimate tuning deviations per produced note), $P_t(s|p)$ is the source contribution per pitch over time, $P_t(q|p)$ is the time-varying sound state activation per pitch, and finally $P_t(p)$ is the pitch activation, i.e. the resulting multi-pitch detection output. In the proposed model, $p \in \{1\ldots,88\}$, with $p = 1$ denoting A0 and $f \in \{1,\ldots,5\}$, which respectively denote $\{-40,-20,0,20,40\}$ cent deviation from ideal tuning using 12-TET.

The unknown model parameters ($P_t(f|p)$, $P_t(s|p)$, $P_t(p)$, $P_t(q|p)$) are iteratively estimated using the expectation-maximization (EM) algorithm [8], with the update rules described in [4]. With 30 iterations set in the system, the runtime for multi-pitch detection is approximately 3×real-time using a Sony VAIO S15 laptop. The output of the model is $P(p,t) = P(t)P_t(p)$, which represents pitch activation probability in semitone scale.

3.3 Tuning Estimation - Postprocessing

The output of the multi-pitch detection model, $P(p,t)$, is non-binary and needs to be converted into a list of note events or a MIDI file. Firstly, in order to compensate for any tuning deviations from A4=440 Hz, a tuning estimation step is proposed, utilising information from the pitch

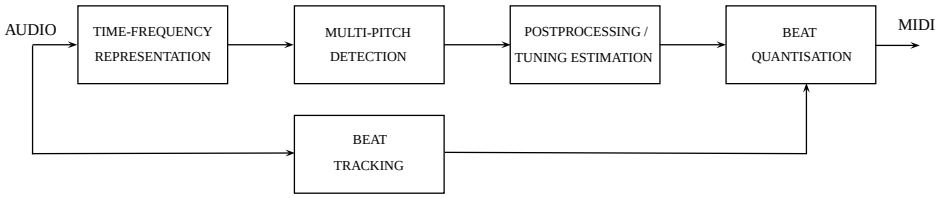


Figure 1: Diagram for the proposed system.

shifting parameter $P_t(f|p)$. The tuning probability vector is computed as:

$$P(f) = \sum_{p,t} P_t(f|p)P_t(p)P(t). \quad (2)$$

$P(f)$ provides an estimate on tuning deviations from 440 Hz tuning, with the various f values corresponding to $\{-40, -20, 0, 20, 40\}$ cent deviation. The final tuning estimate is given by $\mathrm{argmax}_f\, P(f)$. Then, a 20 cent resolution time-pitch representation is computed:

$$P(f', t) = [P(f, 1, t) \cdots P(f, 88, t)] \quad (3)$$

where $P(f, p, t) = P_t(f|p)P_t(p)P(t)$, and $f' = \{1, \ldots, 88 * 5\}$ denotes pitch values between 1 and 88 with 20 cent resolution. The time-pitch representation is subsequently shifted towards 440 Hz tuning by reassigning the index $f' = f' + \mathrm{argmax}_f\, P(f) - 3$ (since $f = 3$ represents 0 cent tuning deviation). Then, a tuning-compensated pitch activation $P(p, t)$ is re-computed from $P(f', t)$:

$$P(p, t) = \sum_{f'=5p-4}^{5p} P(f', t), \;\; \forall p \in \{1, \ldots, 88\}. \quad (4)$$

Following tuning compensation, thresholding is performed on $P(p, t)$, followed by a process for removing note events with a duration less than 40 ms. This results in a list of note events, denoted by onset, offset, and pitch, denoted $\mathrm{nmat}_m(on, \mathit{off}, p)$, with $m \in \{1, \ldots, M\}$ denoting the note index, with on and *off* being the onset and offset times, respectively.

3.4 Meter Tracking & Quantisation

Since most dance tunes have an underlying stable meter and a relatively predictable tempo that enables dancers to synchronize, a quantisation of the estimated notes onto a tight metrical grid is likely to improve transcription performance. In addition, the notes in the obtained transcription are assigned positions within the meter, and obtain quantised note durations, which enables for an immediate visualisation as staff notation including a time signature.

In this paper, beat and measure positions for a recording are computed using the Bayesian meter tracker presented in [15]. Given a series of observations/features $\mathbf{y}_k$, with $k \in \{1, ..., K\}$, computed from a music signal, a set of hidden variables x_k is estimated. The hidden variables describe at each analysis frame k the position Φ_k within a measure, and the tempo in positions per frame ($\dot{\Phi}_k$). The goal is to estimate the hidden state sequence that maximizes the posterior (MAP) probability $P(\mathbf{x}_{1:K}|\mathbf{y}_{1:K})$. If we express the temporal dynamics as a Hidden Markov Model (HMM), the posterior is proportional to

$$P(\mathbf{x}_{1:K}|\mathbf{y}_{1:K}) \propto P(\mathbf{x}_1) \prod_{k=2}^{K} P(\mathbf{x}_k|\mathbf{x}_{k-1})P(\mathbf{y}_k|\mathbf{x}_k) \quad (5)$$

In (5), $P(\mathbf{x}_1)$ is the *initial state distribution*, $P(\mathbf{x}_k|\mathbf{x}_{k-1})$ is the *transition model*, and $P(\mathbf{y}_k|\mathbf{x}_k)$ is the *observation model*. When discretising the hidden variable $\mathbf{x}_k = [\Phi_k, \dot{\Phi}_k]$, the inference in this model can be performed using the Viterbi algorithm. As in [15], a uniform initial state distribution $P(\mathbf{x}_1)$ was chosen. The transition model factorizes into two components according to

$$P(\mathbf{x}_k|\mathbf{x}_{k-1}) = P(\Phi_k|\Phi_{k-1}, \dot{\Phi}_{k-1})P(\dot{\Phi}_k|\dot{\Phi}_{k-1}) \quad (6)$$

with the two components describing the transitions of position and tempo states, respectively. The position transition model increments from Φ_{k-1} to Φ_k deterministically using the tempo $\dot{\Phi}_{k-1}$, starting from a value of 1 (at the beginning of a metrical cycle) to a value of 800. The tempo transition model allows for tempo transitions to the adjacent tempo states, allowing for gradual tempo changes. The observation model $P(\mathbf{y}_k|\mathbf{x}_k)$ divides the $2/4$-bars of meter-annotated Sousta tunes used in [15] into 32 discrete bins. Spectral-flux features are assigned to one of these metrical bins, and parameters of a Gaussian Mixture Model (GMM) are determined. The computation follows exactly the procedure described in [15], which lead to an almost perfect beat tracking for the Cretan tunes.

In order to quantise the detected note events nmat_m with respect to the estimated beat positions, firstly a metrical grid is created from the beat positions (beat_n, $n \in \{1, \ldots, N\}$). The metrical grid times are:

$$\mathrm{grid}_{D(n-2)+d+1} = \mathrm{beat}_{n-1} + (d/D)(\mathrm{beat}_n - \mathrm{beat}_{n-1}) \quad (7)$$

which are computed for $n = 2, \ldots, N$. In (7), D is the beat subdivision factor ($D = 4, 8$ corresponds to 16th and 32nd note subdivisions, respectively) and $d = \{0, \ldots, D-1\}$. Then, the beat-quantised transcription is produced by changing the onset time for each detected note nmat_m to the closest time instant computed from (7).

4. EXPERIMENTS

4.1 Training

Spectral templates for lyra are extracted from 20 short segments of solo lyra recordings, taken from the Crinnos

project [2] (disjoint from the recordings in the corpus described in Section 2). These are used as $P(\omega|q,p,f,s)$ in the model of (1). The recordings are partially annotated, identifying non-overlapping pitches. Then, for each recording the VQT spectrogram is computed as in Section 3.1 and spectral templates for each note are extracted using standard PLCA, whilst keeping the pitch activation matrix fixed to the reference pitch annotations. The templates are pre-shifted across log-frequency to account for tuning deviations, and templates for missing notes are created by shifting the extracted templates across the log-frequency axis. The resulting note range for the training templates is B3-F5.

4.2 Metrics

For assessing the performance of the proposed system in terms of multi-pitch detection, we utilise the onset-based metric used in the MIREX note tracking evaluations [1]. Here, a note event is assumed to be correct if its pitch corresponds to the ground truth pitch and its onset is within a $\pm$25 ms range of the ground truth onset. This is in contrast with the $\pm$50 ms onset tolerance setting used in MIREX, since the current corpus has fast tempo with short note durations. Using the above rule, the precision ($\mathcal{P}$), recall ($\mathcal{R}$), and F-measure ($\mathcal{F}$) metrics are defined:

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

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. The above metrics are computed only for the recording regions that do not contain any vocal parts (a comparative experiment is done in Section 4.3).

4.3 Results

Using the evaluation metrics of Section 4.2, average results on the corpus described in Section 2 are presented in Table 1. Various configurations for the proposed system are used to evaluate the performance of each system component. Configuration 1 refers to simply using the output of the multi-pitch detection method from Section 3.2. Configuration 2 involves multi-pitch detection plus the proposed tuning estimation method from Section 3.3. Configuration 3 refers to multi-pitch detection combined with meter tracking from Section 3.4, thus producing a beat-aligned note output. Configuration 4 combines multi-pitch detection, tuning estimation, and meter tracking. Finally, Configuration 5 is an oracle version of Configuration 4, with the automatically estimated beats being replaced by the manually annotated measure positions, obtained as described in Section 2.2. In all configurations that utilise beat information the beat subdivision factor used is $D = 4$ (corresponding to 16th notes).

As can be seen from Table 1, when integrating tuning estimation the system performance improves by +2.2% in terms of F-measure. Likewise, by incorporating meter tracking, system performance improves by +13.5%, whereas when integrating both tuning and meter information the overall improvement is at +16.8%. Finally, using the reference measure annotations (Configuration 5) leads to an improvement of only +0.3% over the automatic beat extraction, indicating the reliability of meter tracking. Indeed, comparing the manually corrected measure annotations with those obtained from the automatic tracking, we obtain an F-measure [7] of 94.5%. This is an even higher meter tracking performance than observed on the Cretan recordings in [15], possibly caused by the fact that the recordings used in this paper were all conducted in the studio, and all tunes relate to the same dance. A discrepancy is also observed between average precision and average recall; the lower recall is mostly attributed to repeated notes in the ground truth, which are merged into single note events in the output transcription. The aforementioned results are approximately at the level of the state-of-the-art for AMT, when using other datasets [1]; results for individual recordings range from $\mathcal{F} = 70.9\%$ to 34.2% (the latter for a particularly idiosyncratic recording).

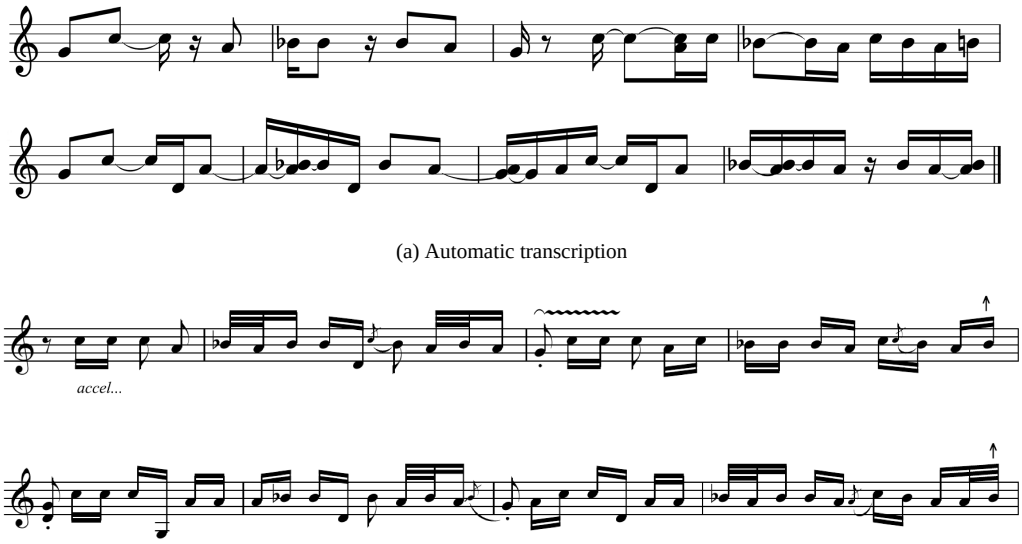
System	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$
Configuration 1	41.12%	45.33%	37.79%
Configuration 2	43.37%	48.12%	39.64%
Configuration 3	54.61%	66.38%	46.53%
Configuration 4	57.92%	70.71%	49.21%
Configuration 5	58.25%	71.14%	49.47%

Table 1: Average multi-pitch detection results using the corpus of Section 2, using various system configurations explained in Section 4.3.

In Figure 2, an example of transferring a beat-quantised transcription obtained with Configuration 4 ($\mathcal{F} = 56.17\%$ for this piece) to staff notation is depicted, along with the manually transcribed reference notation. Spurious differences occur (*e.g.* added note G in the first measure) and the style of notation seems artificial. However, the resemblance between melodic contour in reference and automatic transcription is apparent. According to the analysis in the Crinnos project, the phrase depicted in Figure 3 is repeated with slight variations four times in the eight bars of this example, and comparing the phrase with each two consecutive bars in the transcriptions, this structure can be recognized. Figure 4 depicts the VQT and the pitch activations for the same eight bars. Further examples, all transcriptions obtained with Configuration 4 (MIDI and audio), and the reference annotations will be available on the paper's website [3] .

A comparison with a state-of-the-art AMT method is also made, employing the system of [21], which is based on non-negative matrix factorization. The aforementioned system decomposes a pitched sound as the sum of narrowband spectra. Results using multi-pitch detection only reach $\mathcal{F} = 26.08\%$ (in contrast with 41.12% for the proposed system). By integrating multi-pitch detection with beat information, the performance of [21] reaches

[3] www.rhythmos.org/ISMIR2016Sousta.html



(a) Automatic transcription

(b) Manual transcription, source [2]

Figure 2: Four repetitions of a two-bar phrase.

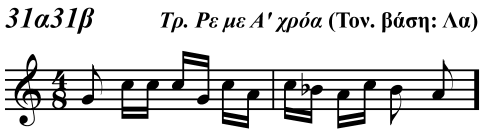


Figure 3: Sousta phrase that is repeated (in slight variations) in Figure 2 four times, source [2].

$\mathcal{F} = 35.94\%$ (as compared with 54.61% for the proposed method). It should be noted that tuning estimation cannot be achieved using the aforementioned system, as the output is quantised on the MIDI scale.

Experiments are also carried out using a larger onset tolerance for the metrics of Section 4.2, set to 50 ms (as in the MIREX evaluations [1]). When evaluating Configuration 4, $\mathcal{F} = 60.74\%$, while using the method of [21] $\mathcal{F} = 38.17\%$. The relatively small difference between using 25 ms or 50 ms tolerance is attributed to the fact that the employed corpus contains several short repeated notes; since the utilised evaluation metrics consider duplicate notes in the same temporal region as false alarms, a larger tolerance window penalises the systems' performance.

As mentioned in Section 4.2, the results presented in Table 1 are computed only for instrumental regions of the corpus, thus excluding any vocal parts. When also transcribing vocal parts, performance using Configuration 4 drops by 1.9% ($\mathcal{F} = 55.84\%$), due to the fact that the training data do not contain vocal templates; however, the transcription of vocal music is not in the scope of this work. Finally, experiments were carried out using a beat subdivision factor $D = 8$, which corresponds to 32nd notes. This results in $\mathcal{F} = 48.2\%$, which indicates that the onsets for some of the detected notes were placed in incorrect temporal positions on the metrical grid.

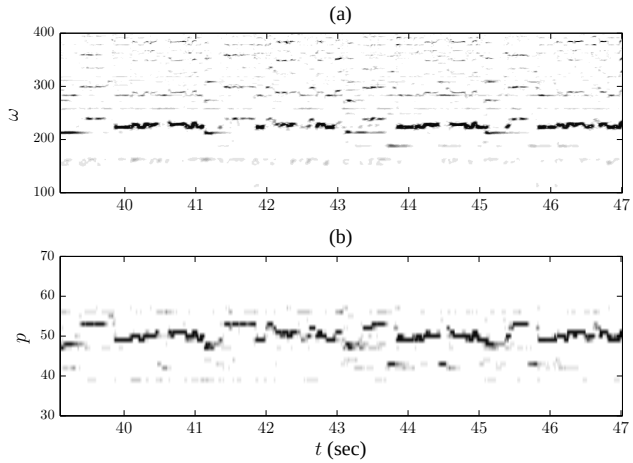


Figure 4: (a) The VQT spectrogram for the section transcribed in Figure 2. (b) The corresponding pitch activation $P(p, t)$.

5. DISCUSSION

In this paper, we presented a corpus for evaluation of AMT systems that is based on performance transcriptions manually compiled by experts in ethnomusicology. We then proposed an AMT system that can cope with tuning deviations, and we improve the performance of the AMT system by quantising its output on a metrical grid that was estimated using a state of the art meter tracker. Apart from the performance improvement, this quantisation enables for a straightforward generation of staff notation. For future work, we intend to improve the transcription by including instrument templates for the accompaniment instruments, which will enable for a better estimation of the main melody. Furthermore, we plan to conduct a user study with ethnomusicologists, who will evaluate the performance of our AMT system.

6. REFERENCES

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

[2] Website of the Crinnos project. http://crinnos.ims.forth.gr. Accessed: 2016-03-16.

[3] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri. Automatic music transcription: challenges and future directions. *J. Intelligent Information Systems*, 41(3):407–434, December 2013.

[4] E. Benetos and T. Weyde. An efficient temporally-constrained probabilistic model for multiple-instrument music transcription. In *16th International Society for Music Information Retrieval Conference (ISMIR)*, pages 701–707, Malaga, Spain, October 2015.

[5] Emmanouil Benetos and Andre Holzapfel. Automatic transcription of Turkish microtonal music. *Journal of the Acoustical Society of America*, 138(4):2118–2130, 2015.

[6] A. T. Cemgil, H. J. Kappen, and D. Barber. A generative model for music transcription. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):679–694, 2006.

[7] M. E. P. Davies, N. Degara, and M. D. Plumbley. Evaluation methods for musical audio beat tracking algorithms. Technical Report C4DM-TR-09-06, Queen Mary University of London, Centre for Digital Music, 2009.

[8] 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.

[9] 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.

[10] Tuomas Eerola and Petri Toiviainen. *MIDI Toolbox: MATLAB Tools for Music Research*. University of Jyväskylä, Jyväskylä, Finland, 2004.

[11] Ter Ellingson. Transcription. In Helen Myers, editor, *Ethnomusicology: An Introduction*, pages pp. 110–152. MacMillan, London, 1992.

[12] V. Emiya, R. Badeau, and B. David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1643–1654, August 2010.

[13] B. Fuentes, R. Badeau, and G. Richard. Harmonic adaptive latent component analysis of audio and application to music transcription. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(9):1854–1866, September 2013.

[14] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: music genre database and musical instrument sound database. In *International Symposium for Music Information Retrieval*, October 2003.

[15] Andre Holzapfel, Florian Krebs, and Ajay Srinivasamurthy. Tracking the "odd": Meter inference in a culturally diverse music corpus. In *Proceedings of ISMIR - International Conference on Music Information Retrieval*, pages 425–430, Taipei, Taiwan, 2014.

[16] R. Macrae and S. Dixon. Accurate real-time windowed time warping. In *International Society for Music Information Retrieval Conference*, pages 423–428, Utrecht, Netherlands, 2010.

[17] Robert Reigle. Reconsidering the idea of timbre: A brief history and new proposals. In *MusiCult '14: Music and Cultural Studies Conference*, pages 233–243, 2014.

[18] Haris Sarris, Tassos Kolydas, and Panagiotis Tzevelekos. Parataxis: A framework of structure analysis for instrumental folk music. *Journal of interdisciplinary music studies*, 4(1):71–90, 2010.

[19] C. Schörkhuber, A. Klapuri, N. Holighaus, and M. Dörfler. A Matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution. In *AES 53rd Conference on Semantic Audio*, page 8 pages, London, UK, January 2014.

[20] L. Su and Y.-H. Yang. Escaping from the abyss of manual annotation: New methodology of building polyphonic datasets for automatic music transcription. In *Int. Symp. Computer Music Multidisciplinary Research (CMMR)*, June 2015.

[21] E. Vincent, N. Bertin, and R. Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):528–537, March 2010.