INCORPORATING PITCH CLASS PROFILES FOR IMPROVING AUTOMATIC TRANSCRIPTION OF TURKISH MAKAM MUSIC

Emmanouil Benetos
City University London
emmanouil.benetos.1@city.ac.uk

Andre Holzapfel
Boğaziçi University, Istanbul
andre@rhythmos.org

ABSTRACT

In this paper we evaluate the impact of including knowledge about scale material into a system for the transcription of Turkish makam music. To this end, we extend our previously presented appoach by a refinement iteration that gives preference to note values present in the scale of the mode (*i.e.* makam). The information about the scalar material is provided in form of pitch class profiles, and they are imposed in form of a Dirichlet prior to our expanded probabilistic latent component analysis (PLCA) transcription system. While the inclusion of such a prior was supposed to focus the transcription system on musically meaningful areas, the obtained results are significantly improved only for recordings of certain instruments. In our discussion we demonstrate the quality of the obtained transcriptions, and discuss the difficulties caused for evaluation in the context of microtonal music.

1. INTRODUCTION

The derivation of a suitable notation for a music performance is an topic that has been discussed with many different goals and forms of music in mind. In ethnomusicology, the process of this derivation is referred to as transcription. Throughout the decades of the twentieth century, it was attempted to find extensions to the normal staff notation in order to capture micro-tonal information as well as other aspects related to rhythm and timbre that are hard to capture in usual staff notation (Abraham & von Hornbostel, 1994). A prominent point in the discussion about transcription was a symposium on transcription and analysis in 1963 (England, 1964), in which several experts conducted transcriptions of a Bushman song, with the output giving an interesting example of how transcriptions differ depending on personal interpretation and focus of analysis.

While at least Anglo-American ethnomusicology since then has taken a turn towards anthropology, the importance of deriving an adequate graphical representation for music remains an important task in many contexts. For instance, in many cultures notations are used in order to support the memorization of music, such as for Turkish makam music or for Eurogenetic classical music. In these contexts, a notation is a culturally conventional tool in order to compose and teach music and to facilitate performances. Apart from the immediate practical value, notations can also enable for a more focused analysis of musical aspects. Typically, when using a form of notation that is similar to Eurogenetic staff notation, the aspect that is most adequately represented is melody. Therefore, a notation is a useful tool to arrive, for instance, at an understanding of melodic phrases present in a repertoire, of the scale material that they use, and typical modulations between tonal modes.

Automatic music transcription, then, is the process of automatically converting some music signal into a form of notation. It is one of the pivotal areas of research in the field of Music Information Retrieval (MIR), where the difficulty of the task prevails after decades of research especially for the transcription of ensemble recordings (Klapuri & Davy, 2006). The results of an automatic transcription system can be improved by including the knowledge of the tonal material present in the recording, as it was demonstrated by Benetos et al. (2014) for Eurogenetic music. In the present paper, we will transfer these results to the context of Turkish makam music. This is facilitated by formulating the transcription as a probabilistic model, in which pitch class profiles can be included to enforce tonal structure assumed to be present in the piece which is to be transcribed. We apply a large set of pitch class profiles derived by Bozkurt (2008) from a large set of instrumental recordings. This way, we will continue our research on transcription of Makam music, which will hopefully help to shed light on the challenges of the AMT task for musics not part of Eurogenetic classical or popular music, which were at the focus of most MIR research so far.

In our paper we will start with a summary of the notation conventionally used in Turkish makam music practice, and the specific challenges for an AMT system when targeting such a transcription. We will then give a detailed description of our transcription system in Section 3. In Section 4, we describe the music collection which we aim to transcribe using our system, and we will conduct a quantitative evaluation of our system. We will then give some qualitative examples for the transcriptions, clarifying the shortcomings and the potentials of the method in Section 5. Finally, we will summarize our findings and give some overview of potential future research directions in Section 6.

2. CHALLENGES AND MOTIVATIONS

In our previous work we gave a detailed outline of the challenges related to the computational analysis of Turkish makam music (Bozkurt et al., 2014), and addressed challenges specific to automatic music transcription in Benetos & Holzapfel (2013). We will shortly summarize the most important musical aspects, that make a transcription system for Turkish makam music a challenging target for research.

First, Turkish music performance could be described as having a concept of relative pitch, which means that the frequency value of the tonic (*karar*) of a piece can vary depending on the choice of the performer(s). That means for a given performance, the pitch value of the tonic has to be determined either manually or using computational analysis in order to arrive at a valid notational representation of the piece.

Furthermore, once performers agree on a certain pitch for the tonic, certain instruments typically play the main melody in the distance of an octave due to their pitch range. This is the case for instance for the *tanbur* and *ney* instruments, which represent an often encountered combination in this music practice. Typically, melodies allow for a certain degree of freedom, resulting in a music practice that demands for the application of certain ornamentations that are supposed to give life and color to an interpretation. Especially in ensemble performances such ornamentations lead to deviations between the melodies played by the individual instruments, which results in an increased difficulty for a transcription.

Finally, the conventional notation system for Turkish makam music applies Eurogenetic staff notation with additional accidentals that signify certain micro-tonal intervals that deviate from well-tempered tuning. However, as discussed by Bozkurt (2008), these notated accidentals do often not match with the intervals encountered in performance practice. This further complicates AMT, since the set of note intervals to be expected in a piece strongly varies depending on tonal mode, performer, instrument, and possibly other parameters.

Our initial transcription system (Benetos & Holzapfel, 2013) for Turkish music targeted a transcription that included micro-tonal intervals. In the present publication we will evaluate, if the accuracy of our system can be further improved by including knowledge of the note intervals typically encountered in the performances of a certain makam. The inclusion of scale information was shown to improve AMT performance for Eurogenetic music Benetos et al. (2014), but it is an open question if this holds for Turkish makam music; There is a significant dissent between theory and practice, and the ongoing discussions among musicians indicate that the choice of certain intervals depends on personal choice, instrument, and historical period to some extent.

3. PROPOSED SYSTEM

The proposed transcription system takes as input a recording and information about the makam. Multi-pitch detection is performed using the efficient transcription system that was proposed in Benetos et al. (2013), modified for using ney and tanbur templates. Note tracking is performed as a post-processing step, followed by tonic detection. Given the tonic, the piece is then re-transcribed, using information from makam pitch profiles and the detected tonic. The final transcription output is a list of note events in cent scale centered around the tonic. A diagram of the proposed transcription system can be seen in Fig. 1.

3.1 Pitch Template Extraction

We use pitch templates extracted from 3 solo ney and 4 solo tanbur recordings, originally created in Benetos & Holzapfel (2013). The templates were extracted using probabilistic latent component analysis (PLCA) Smaragdis et al. (2006) with one component. The time/frequency representation used is the constant-Q transform (CQT) with a spectral resolution of 60 bins/octave, with 27.5Hz as the lowest bin (Schörkhuber & Klapuri, 2010). The range (in MIDI scale) for ney is 60-88 and the range for tanbur is 39-72.

3.2 Transcription Model

The proposed transcription model expands the probabilistic latent component analysis (PLCA) method, by supporting the use of multiple pre-extracted spectral templates per pitch and instrument, that are also pre-shifted across log-frequency for supporting frequency and tuning deviations; the latter is particularly useful for performing transcribing micro-tonal music. The model takes as input a log-frequency spectrogram $V_{\omega,t}$ (ω is the log-frequency index and t the time index) and approximates it as a bivariate distribution $P(\omega, t)$, which in turn is decomposed 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 $P(\omega|s, p, f)$ are the pre-extracted spectral templates for pitch p, instrument s, which are also pre-shifted across log-frequency according to parameter f. $P_t(f|p)$ is the time-varying log-frequency shift per pitch, $P_t(s|p)$ is the time-varying instrument contribution per pitch, and $P_t(p)$ the pitch activation over time (i.e. the transcription).

Since the log-frequency representation has a resolution of 60 bins/octave, and f is constrained to one semitone range, f has a length of 5. The unknown parameters of the model, $P_t(f|p)$, $P_t(s|p)$, and $P_t(p)$, can be iteratively estimated using the Expectation-Maximization algorithm of Dempster et al. (1977), with 30 iterations being sufficient for convergence. The spectral templates $P(\omega|s, p, f)$ are kept fixed using the pre-extracted pitch templates from Section 3.1 and are not updated.

The output of the transcription model is a MIDI-scale pitch activation matrix given by:

$$P(p, t) = P(t) P_t(p) \quad (2)$$

as well as a high pitch resolution time-pitch representation, given by:

$$P(f', t) = [P(f, p_{low}, t) \cdots P(f, p_{high}, t)] \quad (3)$$

where $P(f, p, t) = P(t) P_t(p) P_t(f|p)$. In (3), f' denotes pitch in 20 cent resolution. As an example of a time-pitch representation, Fig. 2 displays $P(f', t)$ for a ney recording.

3.3 Post-processing

The transcription output is a non-binary representation that needs to be converted into a list of of note events, listing

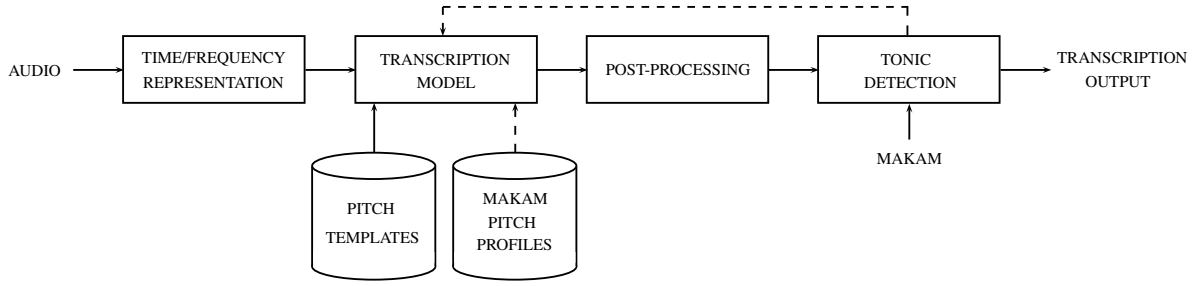


Figure 1: Proposed transcription system diagram. Dashed lines indicate operations taking place at re-transcription.

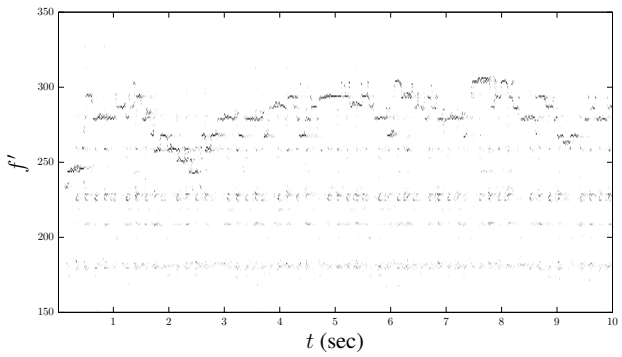


Figure 2: The time-pitch representation $P(f', t)$ for the 'Huseyni Peşrev' piece performed by ney. Note the percussive elements in the lower pitch range.

onset, offset, and pitch. Thus, we perform median filtering and thresholding on $P(p, t)$ (for converting it into a binary representation), followed by minimum duration pruning (with a minimum note event duration of 130ms).

As in Benetos & Holzapfel (2013), a simple 'ensemble detector' is used in order to detect heterophonic recordings. Subsequently, if a piece is detected as such, each octave interval is processed by merging the note event of the higher note with that of the lowest one. Then, using information from $P(f', t)$, each detected note is converted into the cent scale.

In order to detect the tonic frequency of the recording we apply the procedure described in Bozkurt (2008), which computes a histogram of the detected pitch values, and aligns it with a template histogram for a given makam using the cross-correlation function. A final post-processing step is made after centering the detected note events by the tonic, where note events that occur more than 1700 cents or less than -500 cents apart from the tonic are eliminated.

3.4 Pitch class profiles

For incorporating information on the pitch structure of the recording given a makam, we re-transcribe the recording having as additional information its detected tonic, and we impose a prior on the pitch activation $P_t(p)$. For enforcing a structure on pitch distributions, we employ the 44 makam pitch class profiles that were computed from large sets of instrumental recordings in Bozkurt (2008). An example of a pitch class profile is given in Fig. 3, for the Beyati makam.

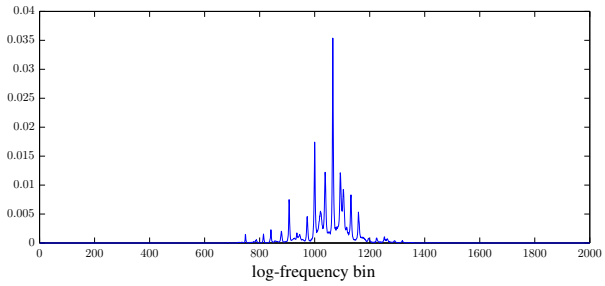


Figure 3: The pitch class profile for the Beyati makam, computed in Bozkurt (2008). The tonic is aligned with the bin 1000, and the resolution is 159 bins/octave (i.e. 1/3Hc).

As was shown in Smaragdis & Mysore (2009), PLCA-based models can use Dirichlet priors for enforcing structure on their distributions. We thus define the Dirichlet hyper-parameter for the pitch structure given a makam m and tonic τ as:

$$\alpha(p|t)_{m,\tau} = K_{m,\tau} P_t(p) \tag{4}$$

where $K_{m,\tau}$ is a pitch profile for makam m centered around tonic τ detected from 3.3. Essentially, $\alpha(p|t)_{m,\tau}$ represents a modified transcription, giving higher probability to pitches which are more frequently encountered in the specific makam than to pitches which are not.

Thus, a modified update rule is created for $P_t(p)$, which is as follows:

$$P_t(p) = \frac{\sum_{\omega,f,s} P_t(p, f, s|\omega) V_{\omega,t} + \kappa \alpha(p|t)_{m,\tau}}{\sum_{p,\omega,f,s} P_t(p, f, s|\omega) V_{\omega,t} + \kappa \alpha(p|t)_{m,\tau}} \tag{5}$$

where $P_t(p, f, s|\omega)$ is the posterior of the model and κ is a weight parameter expressing how much the prior should be imposed, which as in Smaragdis & Mysore (2009) gradually decreases from 1 to 0 throughout iterations (initialising the model but letting it converge in the end). After the modified transcription step, the output is then post-processed using the steps from Section 3.3.

4. EVALUATION

4.1 Music Collection

The music collection which is used is identical with the collection described in Benetos & Holzapfel (2013). It consists of 16 recordings of metered instrumental pieces of Turkish makam music, 10 of them solo performances, and

the remaining 6 recordings ensemble performances. The performances were note-to-note aligned with the microtonal notation available from the collection published in Karaosmanoğlu (2012). This results in a ground-truth notation that consists of a list of time-instances and note-values in cent assigned to each time instance, with the tonic of the piece at 0 cent. It is important to point out that the available ground-truth does not represent a descriptive transcription of the performance, but rather a summary of the basic melody played in the piece without ornamentations, as it is typical for notations in Turkish makam music.

4.2 Metrics

For the proposed evaluations, we use the note-based onset-only metrics that were defined in Benetos & Holzapfel (2013), and are based on the metrics used for the MIREX Note Tracking tasks for Music Information Retrieval: SymbTr (MIREX). Specifically, we consider a note to be correct if its F0 is within a +/-20 cent tolerance around the ground-truth pitch and its onset is within a 100ms tolerance, and use the proposed Precision, Recall, and F-measure metrics from Benetos & Holzapfel (2013), which are defined as follows:

$$\mathcal{P} = \frac{N_{tp}}{N_{sys}}, \;\; \mathcal{R} = \frac{N_{tp}}{N_{ref}}, \;\; \mathcal{F} = \frac{2\mathcal{R}\mathcal{P}}{\mathcal{R}+\mathcal{P}} \qquad (6)$$

where N_{tp} is the number of correctly detected notes, N_{sys} the number of notes detected by the transcription system, and N_{ref} the number of reference notes. Duplicate notes are considered as false alarms.

4.3 Results

We perform two types of evaluations, as in Benetos & Holzapfel (2013); The first uses the tonic automatically detected by the system of Bozkurt (2008), which attempts to automatically determine the frequency value in Hz of the tonic in a recording by comparing the pitch profile of the recording with a reference pitch profile for the makam. In the second evaluation, we use a manually annotated tonic. The proposed method is able to transcribe the 75min dataset in less than one hour, i.e. less than real time. Results comparing the proposed system with the system of Benetos & Holzapfel (2013) using the manually and automatically detected tonic can be seen in Tables 1 and 2, respectively. It can be seen that the F-measure increases by about 1.2% in both cases, indicating that the incorporation of prior information on pitch structure can improve transcription performance. This improvement is consistent for almost all 16 recordings, and is mostly evident for the subset of tanbur recordings (about 3% improvement). As in Benetos & Holzapfel (2013), there is a significant performance drop when comparing results with automatically detected tonic over the manually supplied one. This is attributed to the fact that the F0 tolerance for the evaluation is 20 cent, so even a slight tonic miscalculation might lead to a performance decrease.

In Tables 3 and 4, detailed results for ney, tanbur, and ensemble recordings can be seen, using manually aligned and automatically detected tonic, respectively. As in Benetos & Holzapfel (2013), the performance drops when the automatic tonic detection method is used. Likewise, the best results are reported for the subset of tanbur recordings, followed by the subset of ney recordings. The ensemble recordings, which additionally contain percussion along with heterophonic music, provide a greater challenge, with the F-measure reaching 47% for the case of manually annotated tonic.

	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
Benetos & Holzapfel (2013)	51.58%	52.85%	51.24%
Proposed system	52.72%	54.10%	52.42%

Table 1: Transcription results using manually annotated tonic.

	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
Benetos & Holzapfel (2013)	41.23%	42.07%	40.89%
Proposed system	44.44%	41.79%	42.09%

Table 2: Transcription onset-based results using automatically detected tonic.

5. DISCUSSION

The results we obtained when including pitch class profile information into our transcription system draw a slightly ambiguous picture. While we can observe a significant increase over the previous system for tanbur examples, for both ney and ensemble recordings no such conclusion can be drawn. This is astonishing since the pitch class profiles were derived by (Bozkurt, 2008) using a wide variety of different solo instrument recordings. Apparently, the fretted tanbur seems to be characterized by a more stable interval structure that fits well to the used profiles, while the pitch class profiles seem not to match with the ney recordings. It is an open question if this is due to the playing style or the variation between instruments. We will depict two examples in this section: One tanbur example, in which the inclusion of pitch class profiles proved to be of clear advantage (sample 8 in Table 1 of Benetos & Holzapfel (2013)), and a negative outlier from the set of ney recordings, which resulted in F-measures below 30%, independent from the usage of pitch class profiles (piece 10 on Table 1 of Benetos & Holzapfel (2013)). Audio of the depicted examples along with reference scores can be obtained from the second author's web-page [1] .

In Figures 4a and 4b the tonic (and its octave), and the dominant of the Rast makam are marked with dashed, and dotted lines, respectively. Black rectangles indicate the onsets of annotated notes, with the size of the rectangles determined by the given tolerances for pitch and timing inaccuracy. The red crosses designate the obtained annotations. It can be seen that the inclusion of pitch class profiles reduces the number of spurious notes and leads to a slightly

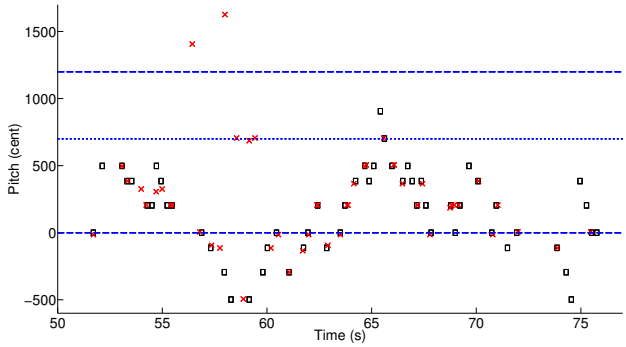
[1] www.rhythmos.org/shareddata/FMA2014

	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
Ney recordings	52.01%	49.25%	50.41%
Tanbur recordings	64.67%	51.67%	57.30%
Ensemble recordings	41.07%	58.37%	47.90%

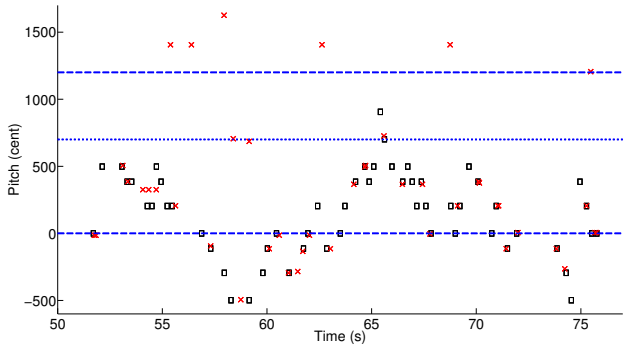
Table 3: Transcription onset-based results for each group of recordings, using manually annotated tonic.

	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
Ney recordings	47.16%	48.81%	46.62%
Tanbur recordings	48.42%	33.02%	39.14%
Ensemble recordings	35.20%	44.78%	38.95%

Table 4: Transcription onset-based results for each group of recordings, using automatically detected tonic.



(a) Transcription using pitch class profile



(b) Transcription without pitch class profile

Figure 4: Example for the transcription of the Teslim section of a Rast Peşrev.

more focused transcription. It is apparent, however, that the obtained result can only be considered a rough approximation of a precise transcription.

In Figure 5 we depict the allegedly negative outlier in our data, a Segah Peşrev played by a single ney. The astonishing insight provided by the figure is that the much lower F-measure for the Segah example (25.6% compared to 70.0% for the Tanbur example) is not related to a clearly worse automatic transcription. The detected notes are strongly related to the annotations, with one important exception. The fourth note of the Segah makam would be at an interval of approximately 500 cent above the tonic, which is a perfect fourth, and this interval was applied in the ground truth. However, the detected notes imply that this interval on the applied ney is sharper, with a size of about 550 cent, a fact not unusual for this interval in this specific makam in practice. Hence, it is apparent that in this case the dissent between theory and practice lead to an artificial underestimation of the actual quality of the transcription. In addition, the values of the metrics might have been influenced by the recording quality, since this recording is the only recording in the collection with a large reverberation, which further impedes a precise detection of note onsets in time.

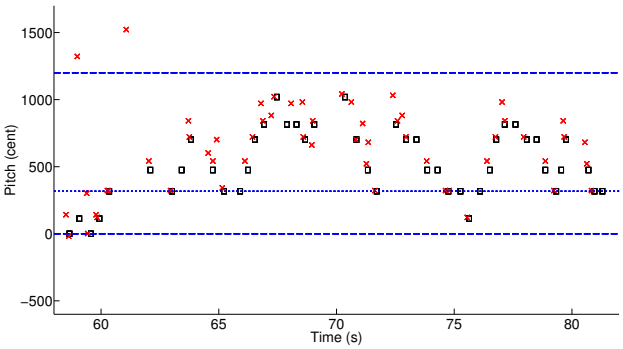


Figure 5: Transcription using pitch class profiles of the teslim section in a Segah Peşrev.

6. CONCLUSIONS

In this paper, we evaluated if the inclusion of pitch class profiles can improve the performance of a transcription system for Turkish makam music. The results imply that for solo instrument recordings of tanbur a consistent improvement can be achieved, when including this information about the interval structure of the makam. For the ney and for ensemble recordings, no increase in performance can be observed. Due to the limited number of included makams, it cannot be determined if this decrease is caused by the different type of instrument, or if makam with more variable interval structure have lead to this result. However, our discussion demonstrates that the values of the evaluation metrics might stand in no direct relation to the quality of the actual transcription, mainly due to the variable interval structure of this music. While the applied metrics represent a convention in the evaluation of AMT, the examples clarify that apparently objective measures should not be trusted blindly.

The quality of the achieved automatic transcriptions seems adequate as a starting point for a more precise manual transcription, or for a qualitative summary of the melodic content of the pieces, and can therefore serve as a method of practical value for the analysis of Turkish makam music. The achieved quality of the transcriptions is comparable to values achieved for Eurogenetic music. However, the demand of detecting notes apart from well-tempered tuning necessarily increases the search space for the notes to be detected, and the need to detect the frequency of the tonic represents another aspect that adds to the complexity. We

plan to improve our system by including pitch class profiles that are adapted to the piece at hand, and to adjust the ground truth annotations to the intervals encountered in a performance, in order to avoid the distortion of evaluation results that was observed in this paper.

7. ACKNOWLEDGEMENTS

Emmanouil Benetos is supported by a City University London Research Fellowship. Andre Holzapfel is supported by a Marie Curie Intra European Fellowship (PIEF-GA-2012-328379).

8. REFERENCES

Abraham, O. & von Hornbostel, E. M. (1994). Suggested methods for the transcription of exotic music. *Ethnomusicology*, *38*(3), 425–456. Originally published in German in 1909: "Vorschläge für die Transkription exotischer Melodien".

Benetos, E., Cherla, S., & Weyde, T. (2013). An efficient shift-invariant model for polyphonic music transcription. In *6th International Workshop on Machine Learning and Music*, Prague, Czech Republic.

Benetos, E. & Holzapfel, A. (2013). Automatic transcription of Turkish makam music. In *14th International Society for Music Information Retrieval Conference (ISMIR)*, (pp. 355–360)., Curitiba, Brazil.

Benetos, E., Jansson, A., & Weyde, T. (2014). Improving automatic music transcription through key detection. In *AES 53rd Conference on Semantic Audio*, London, UK.

Bozkurt, B. (2008). An automatic pitch analysis method for turkish maqam music. *Journal of New Music Research*, *37*(1), 1–13.

Bozkurt, B., Ayangil, R., & Holzapfel, A. (2014). Computational analysis of makam music in Turkey: review of state-of-the-art and challenges. *Journal for New Music Research*, *43*(1), 3–23.

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

England, N. M. (1964). Symposium on transcription and analysis: A Hukwe song with musical bow. *Ethnomusicology*, *8*(3), 223–277.

Karaosmanoğlu, K. (2012). A turkish makam music symbolic database for music information retrieval: Symbtr. In *ISMIR*.

Klapuri, A. & Davy, M. (Eds.). (2006). *Signal Processing Methods for Music Transcription*. New York.

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

Schörkhuber, C. & Klapuri, A. (2010). Constant-Q transform toolbox for music processing. In *7th Sound and Music Computing Conference*, Barcelona, Spain.

Smaragdis, P. & Mysore, G. (2009). Separation by "humming": user-guided sound extraction from monophonic mixtures. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, (pp. 69–72)., New Paltz, USA.

Smaragdis, P., Raj, B., & Shashanka, M. (2006). A probabilistic latent variable model for acoustic modeling. In *Neural Information Processing Systems Workshop*, Whistler, Canada.