AUTOMATIC TRANSCRIPTION OF NEY RECORDINGS IN
TRADITIONAL TURKISH ART MUSIC

Abstract
This paper presents an automatic music transcription (AMT) system which accepts monophonic instrumental ney (flute-like woodwind instrument) recordings of traditional Turkish art music (shortly Turkish music) as input and outputs transcriptions in the conventional staff notation format. Therefore, aim of the study is to obtain conventional Turkish staff notation for the use of analysis and performance. In order to obtain an output presented in staff notation from audio recordings the following steps are applied: f0 estimation, automatic makam and tonic detection, segmentation and quantization of the f0 curve, determination of pitch intervals, note labeling, quantization of duration and representation of transcription in a symbolic format (note names and durations). For testing purposes, both automatic and manual transcriptions are compared to the original scores of the pieces. Success rates of the automatic transcriptions and manual transcriptions are found to be comparatively close.

1. Introduction
Automatic music transcription (AMT) is defined as the conversion of acoustic music signals into symbolic music format (e.g. MIDI) and mainly applied for music information retrieval (MIR) (Klapuri 2006). However, the problem definition, in other words the meaning of transcription is not well-defined within the AMT literature. Automatic transcription is usually considered as the automatization of manual transcription. In other words, automatic transcription is considered within the context of conventional meanings of transcription. However, while music is visually represented by staff notation for performance or analysis in conventional transcription, AMT literature in general do not target an output for either performance or analysis, but for retrieval and thus do not require staff notation. On the other hand, our study aims to obtain staff notation for the use of analysis and performance. While the presented technology can be used in music access applications, such a use is not aimed or tested within the content of this particular manuscript.

There are only few AMT applications in the literature such as automatic music tutoring which is not designed for retrieval. Such studies seldom supply all relevant requirements of staff notation such as key detection, rhythm detection, pitch spelling etc. Few of the studies on automatic transcription of polyphonic/monophonic recordings also try to obtain western staff notation. The context of transcription in such studies are close to the conventional context of transcription in the sense that music is represented visually for the performance. Transcription applications for automatic music tutors aim to match the performance of the user with the original notation in order to help the music student align her/his performance visually. This process is also called as audio to score alignment (Mayor et al. 2009). Few of the automatic transcription of polyphonic/monophonic music applications also aim to help amateur musicians without proper music education to obtain scores of their musical compositions (Wang et al. 2003).
There is a difficulty in designing evaluation procedures for AMT studies due to the ambiguity in the definition of the automatic music transcription problem. Either manual transcription or original notation is used generally as a ground truth for the evaluation. However, the audio recording of a performance with manual transcription and original notation is an important subject of discussions in musicology.

Besides these drawbacks, there are certain challenging aspects of applying current AMT methods to non-western musics, since the techniques and methods of MIR for the automatic transcription is mainly developed for western music. Therefore, it is not possible to apply these techniques and methods directly to non-western musics due to the significant differences of acoustic and conceptual spaces between western and non-western musics.

A number of recent studies discuss the challenging aspects of applying current MIR methods to non-western musics. With a focus on musics of Central Africa, Moelants et al. (2007) mentions differences of African musics from western music. Similarly, application of MIR methods to Turkish music is also discussed by Gedik and Bozkurt (2010) in detail. Cornelis et al. (2010) and Lidy et al. (2010) discuss the challenges in a broader MIR spectrum considering the access and classification issues of non-western musics in turn.

More specifically, the problems of applying current MIR methods to Turkish music are considered in detail by Gedik and Bozkurt (2009; 2010). A challenging issue for applying AMT methods to Turkish music is about the notation system. Since notation system of Turkish music is a direct reflection of a theory that does not match practice, the relation of notation and performance is highly problematic even for the manual transcription of Turkish music. A final problem about applying current AMT methods to Turkish music is the lack of robust methods for the detection of ornamentations and performance styles in the MIR literature which are one of most the important characteristics of Turkish music.

Although there are a few MIR studies on AMT of non-western musics, they are also far from presenting a solution for the challenging aspects of applying current MIR methods to non-western musics. One of these few papers (Nesbit et al. 2004) presents transcription of Australian Aboriginal music, consists of two simple accompaniment instruments, while other 5 papers explore specific facets of transcription problem. Other studies on automatic transcription of non-western musics usually either reduce the pitch space to 12 tones equal temperament (Krishnaswamy 2003; Al-Taee et al. 2009) or simply do not mention the characteristics of pitch space of non-western music considered (Kapur et al. 2007).

As a result, current MIR literature seems to be insufficient for the development of AMT system for non-western musics. On the other hand, the discipline of ethnomusicology supplies some useful insights for the computational studies on non-western musics. For example, the problem of “transcription” of non-western musics, as well as western music, is as old as the ethnomusicology itself. The issue was subject to hot discussions for the founders and leading figures of the discipline such as Ellis (1814-90), Stumpf (1848-1936) and Hornbostel (1877-1935), and Seeger (1886-1977). The distinction between original notation and transcription has been defined fifty years ago by Charles Seeger in 1958 (Ellingson 1992a:111). While prescriptive notation (original notation) defines how a specific piece should be performed, the descriptive notation (transcription) defines how a specific performance actually sounds.

1-In this manuscript, Turkish music theory refers to the Arel theory (Arel, 1993) throughout the paper which is the most widely appreciated and used theory in Turkey.
Transcription, from the ethnomusicological point of view rather corresponds to the description of a musical piece. On the other hand, notation corresponds to visual representation of musical features for the purpose of prescription (Ellingson 1992b: 153). Therefore, transcription and notation are interrelated concepts since transcription is only possible for a definite notation system. As a result, naturally both are crucial concepts for the automatic transcription which take little attention within the literature of AMT. This problem especially reveals itself obviously by the use of either original notation or manual transcription for the evaluation of automatic transcriptions, as aforementioned.

While the use of original notation for evaluation disregards the difference between notation and performance, the use of manual transcription disregards the fact that the transcription of the same recording by different musicians does not result in unique transcriptions. Only very few of the studies within the MIR literature consider the significant differences between original notation and transcription of a performance (Dixon 2000; Orio 2010) and define automatic transcription as obtaining a human readable description of performance (Cemgil et al. 2004; Hainsworth and Macleod 2004) which is more reasonable. Hainsworth and Macleod (2004) figure out that manual transcription strategies can be quite different resulting various degrees of divergence from the original performance. Similarly, the study of Cemgil et al. (2004) shows that there is no unique ground truth for manual transcription even among well-trained musicians. Therefore, the dominant perspective of MIR clearly results with disappearance of the important distinction between a notation of a piece and a transcription of a performance even for western music. Given the said difficulties in finding a ground truth for testing, we follow an alternative method; both automatic transcription and manual transcription results are compared with the original scores of the pieces and then test results are compared.

To summarize, our system accepts monophonic instrumental ney (flute-like woodwind instrument) audio recordings of Turkish music and outputs conventional Turkish staff notation. Therefore, our system can be considered within the context of the conventional meaning of transcription in contrast to AMT studies in the literature.

Briefly the algorithm presented in this study consists of the following steps which also reflect the organization of the paper:

- f0 estimation
- Automatic makam classification and tonic pitch detection.
- Segmentation and quantization of the f0 curve, and determination of pitch intervals.
- Note labeling.
- Rhythmic analysis and quantization of duration.
- Representation of transcription in the symbolic format and staff notation.

Finally, we present the evaluation results of automatic transcription based on 5 monophonic instrumental ney recordings in comparison to manual transcriptions from 2 musicians. Automatic and manual transcriptions are evaluated with a reference to the original scores. This is a new proposal for the evaluation problem of automatic transcription. As a result, while automatic transcription outperforms manual transcriptions for 2 recordings, success rates of automatic transcription for the rest of 3 recordings are found close to the success rates of manual transcription. We also discuss the evaluation results qualitatively and present the future work.
2. Automatic transcription of Turkish music

Automatic transcription of Turkish music as a problem, demonstrates resemblance with automatic transcription of singing, humming or performance of fretless pitched instruments such as violin within MIR studies, due to the resulting continuous pitch space. As Ryynanen mentioned (2006) most of the singing transcription applications are designed as the front-end of query-by-humming (QBH) systems in contrast to our study. The most challenging task in singing transcription is converting a continuous f0 curve to note labels (Ryynanen 2006: 362). However, despite the resemblance of pitch-spaces in singing and Turkish music, it should be kept in mind that it is always a matter of quantization of the f0 curve to the nearest pitch-class in western music. Of course, a simple rounding operation gives poor results for quantization of f0 curve, depending on the following two important characteristics of singing:

i. The intonation of a singer while singing a piece can deviate in time.

ii. Performance of ornamentations such as vibrato, legato and glissando.

Since we are interested in instrumental ney recordings in Turkish music, the first characteristic is out of our scope. The second characteristic is one of the most important characteristic of Turkish music as aforementioned.

Automatic transcription task roughly consists of three steps: f0 estimation, segmentation of f0 curve and labeling each segment with note names. There are various methods for the f0 estimation: methods based on time-domain, frequency domain or auditory model. Methods for segmentation and labeling of f0 curve mainly follow two approaches: cascade approach where f0 curve is first segmented and then labeled, and statistical method where segmentation and labeling are jointly performed (Ryynanen 2006: 363).

One of the most popular approaches to statistical methods for automatic transcription is the use of Hidden Markov Modeling (HMM). However, as mentioned by Orio (2010) the use of HMM for automatic transcription requires collection of scores for training HMM which are hardly available for non-western musics. Although there are plenty of scores of Turkish music, there is a great divergence between the notation and performance in Turkish music as stated by Ayangil (2008) and Kaçar (2005).

Similarly, for non-western musics, collecting manual transcriptions as training data for HMM, is also problematic. Manual transcription of non-western musics either requires existence of a notation system or a notation system in accordance with performance as in western music, at least.

Among the problems of using HMM, only lack of enough number of manual transcriptions is a serious problem for Turkish music which is planned as a future work. Therefore, we preferred cascade approach in our AMT system as shown in Figure 1. The system accepts monophonic audio recordings of instrumental Turkish music. After the f0 estimation, pitch-frequency histogram is calculated for the automatic makam and the tonic detection which are jointly applied. Both the knowledge of makam and tonic pitch are crucial for transcription, since without the determination of tonic pitch, it is not possible to find a reference pitch for the specific recording given that we lack a global reference as in western music where A4 approximately equals to 440 Hz. Then pitch intervals can be found with respect to a reference pitch. The makam defines the note name for that tonic pitch. The automatic makam recognition and tonic detection is jointly performed as described in Gedik and Bozkurt (2010) using template matching on pitch histograms and this
supplies both f0 value and the name of the tonic pitch. Knowledge of makam also provides the accidentalss to be used in the transcription.

As the f0 value of the tonic pitch is found, it is possible to express f0 curve with respect to tonic pitch and then to obtain f0 intervals. This conversion is made by subtracting the value of tonic pitch from the f0 curve (in the logarithmic scale). In order to label resulting f0 curve by note names, firstly it is necessary to segment the f0 curve. Segmentation corresponds to finding the onset of the notes. Secondly f0 curve within each segment is quantized which corresponds to eliminating ornamentations such as appoggiatura, acciaccatura, vibrato and glissandos. Rule-based approach is applied for segmentation and quantization where parameters are heuristically determined depending on the musicological knowledge specific to Turkish music. Parameters used for segmentation and quantization is also tested and tuned by a training process applied on 5 recordings which are not used for evaluation. This process is handled by a graphical user interface (GUI) which enables to observe the resulting f0 curve both auditory and visually and give feedback for the parameters.

![Block diagram of AMT system.](image)

After segmentation and quantization, representation of pitch intervals in terms of Hc2 gives a resolution of 53 Hc/octave for f0 curve which is much higher than the number of pitch classes defined in theory as 24 pitch-classes/octave. Since notation system of Turkish music is a direct reflection of theory and in order to obtain a readable notation, pitch intervals are converged to the nearest pitch-classes which have distinct names for 2 octaves in theory. As the last step before transcription, note durations corresponding to the segment lengths are quantized by using a duration histogram for deciding the length of an eighth note. Finally note names, onset time and note durations are used as an input to a notation software MUS23 which is specifically designed

---

2-It is common practice to use the Holdrian comma (Hc) (obtained by the division of an octave into 53 logarithmically equal partitions) as the smallest intervallic unit in Turkish music theoretical parlance.

for Turkish music and outputs conventional Turkish music staff notation. Since each block has a
definite success rate, GUI enables user to correct any faulty information such as makam name,
tonic pitch etc. in order to obtain a more robust transcription result.

Following subtitles presents the details of blocks of the AMT system shown in Figure 1. The
same example of a hüzzam recording, “Alma Tenden Canımı” performed by ney player Salih
Bilgin, is used to explain each block.

2.1. f0 estimation, automatic makam recognition and tonic detection
The methods used for f0 estimation and tonic pitch detection for Turkish music recordings are
presented by Bozkurt (2008). f0 is estimated by the YIN algorithm (de Cheveigne & Kawahara
2002) with post-filters designed to correct octave errors and remove noise on the f0. Pitch fre-
cuity histograms are used for tonic detection. Once the makam of a given piece is known,
Bozkurt (2008) showed that its tonic can be found by aligning the histogram of the given piece
with makam histogram template. Therefore, in order to find the tonic pitch, it is necessary to find
the makam for a given piece.

Similar to key/tonality finding studies on western music where the tonality of a given piece
is found by comparing its pitch-class distribution with the pitch-class distributions of 24 tonalities
(12 major and 12 minor), Gedik and Bozkurt (2010) designed an automatic makam recognition
system. In this system, each makam histogram template is found by a data-driven approach; pitch
histograms of recordings of the same makam are aligned and summed up to construct a makam
template. Thus, pitch histogram of a given piece is compared by each makam template and the
makam histogram template which has the smallest distance gives the makam of the piece. His-
tograms are compared by shifting one over the other where City-Block (L1 norm) distance is used
as distance measure. Since the tonic of the template is known, the matching operation also gives
the tonic of the sample.

2.2. Segmentation
Cascade approach to segmentation and labeling of f0 curve usually apply a model similar to a
blackboard system proposed by Bello et al. (2000). Segmentation and labeling is usually done by
applying each step based on rules, thresholds or tuning parameters. However, onset detection as
one of the research domain in MIR and a robust method for segmentation takes little attention in
studies based on cascade approach. Onset detection is either applied by algorithms far from the
state-of-art (e.g. McNab and Smith 2000; Bruno and Nesi 2005; Antonelli and Rizzi 2008; Paiva
et al. 2008) or simply not applied (e.g. Haus and Pollastri 2001; Clarisse et al. 2002; De Mulder et
al. 2004).

Holzapfel, Stylianou, Gedik and Bozkurt (2010) studied problems of onset detection for
Turkish music for the first time. They proposed a fusion algorithm for pitched instruments of Turk-
ish music in comparison with western music instruments. Fusion algorithm consists of three on-
sset detection algorithms, spectral flux (SF) and phase slope (PS) and f0 change (F0). While 57
recordings corresponding to 1829 onsets are used for evaluation, 21 recordings corresponding to
674 onsets are used for training. As a result, following success rates in terms of F-measure are
obtained: 74.1% for F0, 73.9% for SF, 73.7 % for PS and %82.1 for fusion algorithm. Especially for
wind instruments F0 method considerably outperforms both SF and PS methods. Since we only
used a wind instrument, ney, due to the simplicity, computational costs and close success rates, we preferred onset detection based on f0 change in our AMT system for the segmentation block. If the difference between two successive f0 intervals is more than 2 Hc (threshold issues are discussed in the mentioned paper), then it is decided that there is an onset, since 4 Hc is defined as the smallest pitch interval in theory.

2.3. Quantization of f0 segments

Once the f0 curve is segmented it is necessary first to quantize and then to label each segment with note names. Median is the most frequent operation used for quantization of f0 segments (Haus and Pollastri 2001; Clarisse et al. 2002; Adams et al. 2006; Paiva et al. 2008; Typke 2011). Before labeling, ornamentations such as vibrato and glissando or articulations such as legato are detected in contrast to the statistical approach. Haus and Pollastri (2001) detect both vibrato and legato. Similarly, Pollastri (2002) applies 0.8 semitone threshold for the detection of vibrato and legato. De Mulder et al. (2004) apply legato detection for segments longer than 300 ms by looking for multiple stable intervals having gaps in between. Paiva et al. (2008) present detection of vibrato and glissando in their automatic transcription study where threshold of 1 semitone is used for vibrato and constant increase or decrease of successive short notes for glissando.

We again applied a rule-based algorithm for quantization in our system. Each segment is searched for the existence of vibrato and glissando. According to the type of ornamentation, two different methods of quantization is applied, one for vibrato and the other for glissando. Firstly, segments are classified as vibrato or segments including glissandos according to the following rules: If the difference between maximum and minimum f0 values of a segment is less than or equal to 3 Hc which corresponds to almost a semitone, it is classified as a vibrato segment. Otherwise the segments are classified as segments including glissando if their durations are also more than 150 ms. Therefore, possible appoggiatura and acciacatura segments with a duration less than 150ms left unclassified either as vibrato or glissando. Similarly, such short ornamentations are left almost unchanged if classified as a vibrato segment. Such segments are useful for onset detection in the proceeding blocks but they are cancelled at the transcription block by a duration filter for the sake of obtaining a simple notation.

2.3.1. Quantization of vibrato segments

Median of the f0 values of each vibrato segment is calculated and set as the quantized value. Figure 2 shows some of the quantized vibrato segments marked with ellipses.
2.3.2 Quantization of segments including glissando

A segment including glissando can also be a combination of long ornamentations such as glissando and vibrato and as well as short ornamentations such as appoggiaturas and acciaccaturas. As a result, quantization of a segment including glissando corresponds to the quantization of long ornamentations and cancellation of short ornamentations. Therefore, it is possible to quantize each segment including glissando separately. Firstly, f0 values of a segment including glissando is filtered by a median filter as follows:

\[ f_{0_{\text{med}}}(n) = \text{median}\{ f_0(n-M) \ldots f_0(n+M) \} \], \quad M= 150 \text{ ms} 

Then heuristically determined rule-based algorithms are applied for quantization of segments including glissandos. Final quantization result is shown in Figure 3. As can be seen from the figure there are short ornamentations, since they are either classified as vibrato segment or remained unclassified due to their durations less than 150 ms. Finally, frequency values less than 30 Hz are also marked as silence, before the note labelling step.

![Figure 3: Quantization of segments including glissando.](image)

2.4 Note labelling

In order to label quantized f0 curve with note names, a list of note names for 8 octaves is used. 53 note names for each octave is listed which corresponds to the resolution of quantized f0 curve (53 Hz/octave). Therefore, successive notes are listed in the list in order to have 1 Hz difference in between, as follows: C1…C4#1 C4#2 C4#3 C4#4 C4#5 C4#6 C4#7 C4#8 D4...C8. The distance between each pitch in the f0 curve and the tonic pitch is calculated from the note list and the corresponding note name is found.

In order to get rid of short ornamentations in the quantized f0 curve for the sake of an easy readable notation, a final filter is applied which cancels the notes shorter than 150 ms by sharing durations of canceled notes equally between neighboring notes. Finally, Figure 4 shows the resulting onsets and note names above the f0 curve depending on pitch interval values of quantized f0 curve and the tonic note name found for the same excerpt of hüzzam recording.
The pitch space of performance in Turkish music is much richer than the pitch space defined in theory. Therefore, pitches of the performance obtained according to the resolution of 53 Hz/octave is converted to the closest pitches defined in theory, since we aim to obtain conventional staff notation which is used by musicians. Following rules are applied for this conversion:

- #1 is converted to natural if the note is not F, since #1 is only defined for F in theory.
- #2 is converted to #1
- #3 is converted to #4
- #6 is converted to #5
- #7 is converted to #8

Again, for the sake of easy readability of notation, sharps more than 4 are expressed in terms of flats; eg. A4#8 is written as B4b1 which is the actual representation of note segah. Similarly, e.g. D5#5 is written as E5b4 which is the actual representation of note hisar.

A final correction for the pitch intervals and note names is applied as follows: If the difference between two successive notes is 1 Hz, which is not a musical interval, then the most frequent one used in the whole piece is attended as the pitch interval of the less frequent one. As a result, Figure 8 shows the note names found below the f0 curve. As can be seen from the figure the second note G5#1 and third note A5#1 are converted to natural G5 and A5, etc. Successive notes F5#5 is expressed as F5b4 and similarly F5#8 is expressed as G5b1. List of transcribed notes are written to a table to be used for conventional staff notation.

2.5. Quantization of note durations
Duration of notes also should be quantized in order to be represented as conventional note durations such as 1/16, 1/8, 1/4 etc. There are mainly three approaches for duration quantization; statistical approach, ratio approach and an approach based on tempo set by the user. Viitaniemi et al. (2003) use distribution of durations obtained from EsAC-database for the quantization of note durations of a given piece. Adams et al. (2006) applies a uniform quantization where duration levels are assumed to be uniformly distributed.

Ratio approach is mainly applied in QBH systems where note durations need not to be represented conventionally. Both Haus and Pollastri (2001) and Unal et al. (2008) apply ratio of durations of consecutive notes. McNab et al. (1996) for a tempo of 120 beat/minute (bpm) set 125
ms as semiquaver (1/16) and use the resolution of semiquaver. Duggan et al. (2008) use a duration histogram where the highest peak is determined as the quaver note.

We apply the method based on duration histogram proposed by Duggan et al. (2008) for the quantization of durations which assumes that the most frequent note duration is the quarter note. In order to obtain a robust histogram, a rounding operation is applied to the first digit of millisecond for each note duration (e.g. 1317 ms is rounded to 1320 ms). Finally, the highest peak of the duration histogram is set as the eighth note (1/8). For the duration histogram of the same sample of hüzünzam recording, eighth note is found as 0.4 sec.

As a result, each note duration is divided to 0.4 sec. and expressed in terms of eighth note. In order to use these duration values for conventional staff notation, each value is expressed as simplest integer ratios such as 3/16, 5/8 etc., and numerator and denominator for each duration is written to a list beside note names. Due to the complexity of rhythmic structure in Turkish music involving compound rhythmic patterns such as 5/4, 7/8, 9/8, we leave this topic to future work.

2.6. Viewing the results

As aforementioned, a GUI is designed as a tool to enable users to correct any faulty information occurred at blocks of the automatic transcription such as correction of makam and tonic information. Automatic transcription produces a symbolic format and this format can be read by the software MUS2 which produces conventional Turkish staff notation and its corresponding MIDI file. Pitch bend function is used to specify the microtonal deviations of pitches from the standard pitches defined by MIDI numbers. The main motivation in using a MIDI output is to help the user listen to the output without the need of a specific synthesizer dedicated to Turkish music notation. Therefore, the user can check the produced notation also by listening to it. Figure 5.a shows the GUI. Figure 5.b shows the conventional staff notation produced by MUS2. Finally Figure 5.c shows the original notation. In overall Figure 5 both demonstrates the result of automatic transcription and the divergence between the notation and performance in terms of pitch intervals and duration. Although notation dictates F5#4, the performer plays F5#5 (G5b4) as can be seen from the f0 curve.
3. Evaluation

There are mainly two approaches for the evaluation of automatic transcription in the literature: comparison of transcription and reference notation (original notation or manual transcription) and simply measuring the success of retrieval operation. Since the latter approach to evaluation is for retrieval applications, it is out of our scope.

There are various metrics for the evaluations based on comparison of transcription and reference notation. One of them is edit distance (ED) where the transcription is compared with reference notation on the basis of number of correct, inserted and deleted notes (e.g. Krige and Niesler 2006; Jiang et al. 2007; Unal et al. 2008). However, the effect of duration or onset/offset times on the success rate is not clear in these studies.
Fonseca and Ferreira (2009) classify evaluation metrics as frame-based and note-based approaches. Frame-based approach is based on the comparison of two notations for every 10 ms (Dixon 2000). Note-based approach is based on classification metrics such as false negatives, false positives, recall, precision and f-measure. Transcribed notes are classified as correct if they satisfy following two conditions: 1) onset of the transcribed note should be within a certain neighborhood of the onset of the reference note and 2) difference of transcribed and reference pitch values should be below a certain threshold value, usually half of a semi-tone. The neighborhood for onset is defined in the literature as a threshold between +25 - +150 ms (Fonseca and Ferreira 2009). Some of these studies also take duration into account by defining an overlap ratio for the definition of correctly detected notes (e.g. Rynanen and Klapuri 2004, 2006, 2008; Antonelli and Rizzi 2008). Overlap ratio determines the tolerance for the original and transcribed notes in terms of their overlapping onset and offset times.

As a result, evaluation of AMT is mainly based on quantitative measures leaving out the questions about false transcribed notes. This quantitative approach toward evaluation makes the details of the process inaccessible. Especially when manual transcriptions are used as reference data, the procedure applied is not clear in the literature. In this sense Daniel et al. (2008) focus on the perceptual evaluations of listeners for the transcription errors and use these data for developing a perceptual-based evaluation.

However, the use of original notation or manual transcription as reference data for evaluation is also problematic even for western music as our discussion on the concepts of descriptive notation and prescriptive notation showed. No doubt this approach is much more problematic for Turkish music due to the divergence of theory and performance. Two studies clearly focus on the problems of notation system in Turkish music. Ayangil (2008:445) especially underlines that although the performance styles such as melodic and rhythmic variations and ornamentations constitutes one of the most important characteristic of Turkish music, they are not represented in notation. Similarly, Kaçar (2005) empirically shows this problem by comparing the notation of pieces and the performances of pieces by master musicians.

Therefore, an objective evaluation of automatic transcription system is also problematic. In order to handle this challenging problem we applied a cross-evaluation method for the first time in the literature. In short, we asked 2 locally well-known performers having a formal education on Turkish music, to manually transcribe the pieces which are selected for the evaluation of our AMT system. Recordings are performed by a well-known neyzen (ney player) Salih Bilgin. Sound files were presented as mp3. Since our system accepts wav files we converted mp3 files to wav with a sampling frequency 44.1kHz. Names of the recordings and relevant information are presented in Table1.

<table>
<thead>
<tr>
<th>Piece #</th>
<th>Piece names</th>
<th>Makam</th>
<th>Composer</th>
<th>Duration (sec.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Piece #1</td>
<td>Alma Tenden Canımı</td>
<td>Hüz zam</td>
<td>Sadettin Kaynak</td>
<td>56</td>
</tr>
<tr>
<td>Piece #2</td>
<td>Aşkınla Çak Olsa Bu Ten</td>
<td>Uşşak</td>
<td>Nafiz Bey</td>
<td>72</td>
</tr>
<tr>
<td>Piece #3</td>
<td>Ben Bu Yolu Bilmem İdim</td>
<td>Hicaz</td>
<td>Unknown</td>
<td>147</td>
</tr>
<tr>
<td>Piece #4</td>
<td>Can-u Dilde Hane Kildin</td>
<td>Hüseyni</td>
<td>Ahmet Hatiboğlu</td>
<td>88</td>
</tr>
<tr>
<td>Piece #5</td>
<td>Kil Kudumunla Müderris</td>
<td>Saba</td>
<td>Unknown</td>
<td>101</td>
</tr>
</tbody>
</table>

4-Recordings and corresponding original notations are obtained from http://www.neyzen.com/.
They have 2 alternative approaches for transcription: one is simple transcription without ornamentations in accordance with the tradition of notation in Turkish music and the other is more complicated transcription including both ornamentations and performance styles. Although both musicians were familiar with both approaches, we asked them to transcribe as simple as possible, similar to the original notations in Turkish music.

As a result, outputs of automatic transcription and manual transcriptions are compared with original notations as a more objective measure for the success of our AMT system. Success rate of AMT system is measured by both note-based and frame-based evaluation methods.

Following measures are used in note-based evaluation by applying a 150ms tolerance for onset and threshold of 3 Hc (approximately half of a semitone) for pitch difference for a correctly transcribed note:

\[
\text{recall} = \frac{TP}{TP + FN}, \quad \text{precision} = \frac{TP}{TP + FP}, \quad F\text{-measure} = \frac{2 \cdot \text{recall} \cdot \text{precision}}{\text{recall} + \text{precision}}
\]

TP (True Positives), FN (False Negatives) and FP (False Positives) to corresponds number of correctly transcribed notes, the number of notes not transcribed and the number of notes not present but transcribed, in turn.

Overall overlap ratio is calculated by the following measure:

\[
\text{overlap ratio} = \frac{\min\{\text{offsets}\} - \max\{\text{onsets}\}}{\max\{\text{offsets}\} - \min\{\text{onsets}\}}
\]

Offsets and onsets in the overlap ratio measure correspond to offsets and onsets of a correctly transcribed note and its reference note. Therefore, for each note pair, overlap ratio is calculated based on the minimum and maximum of offsets and onsets. Mean of the overlap ratios gives the overall overlap ratio. F-measure in the note-based approach is exactly the same measure used for onset detection. However, overlap ratio is also used in note-based approach which indicates the overlap of correctly detected note durations which is not used in evaluation of onset detection. Finally, a simplification proposed by Jiang et al. (2007) for note-based evaluation is applied: Silence in the transcription is deleted and adjacent notes with the same tone are merged as one note.

Frame-based evaluation is found by the following measure:

\[
\text{Acc} = \frac{TP}{(FP + FN + TP)}
\]

Again, threshold of 3Hc is used for the classification of correctly transcribed notes for frame-based evaluation.

Finally, Table 2 presents evaluation results for both evaluation approaches. We will first consider the note-based evaluation results. Firstly, the success rate of automatic transcription for 150ms tolerance value outperforms manual transcriptions for piece #2 uşşak and is within the confidence interval of manual transcriptions for piece #4 hüseyni. The success rates of automatic tran-
scriptions are lower than manual transcriptions for the rest of three recordings, piece #1 hüzzam, piece #3 hicaz and piece #5 saba. Secondly, the success rates of automatic transcriptions for 25ms tolerance value are considerably lower than the success rates of automatic transcriptions for 150ms tolerance value as shown in Table 2. On the contrary, the success rates of most of the manual transcriptions almost remain the same for the onset tolerance values 25 ms and 150 ms. Only the success rate of manual transcription #2 for piece #2 uşşak and piece #4 hüseyni is considerably lower than its success rates for 150ms tolerance. Therefore, the success rates of automatic transcriptions for 25ms tolerance value only outperforms manual transcription #2 for piece #2 uşşak and piece #4 hüseyni.

Table 2: Evaluation results for 3 kinds of transcriptions for 5 recordings. Manual 1 and 2 corresponds to the transcriptions of two musicians.

<table>
<thead>
<tr>
<th>Piece #</th>
<th>Makam</th>
<th>Transcription</th>
<th>Note-based evaluation</th>
<th>Frame-based evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Onset tolerance (msec)</td>
<td>Accuracy</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>F-measure</td>
<td>Overlap Ratio</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>25</td>
<td>150</td>
</tr>
<tr>
<td>1</td>
<td>Hüzzam</td>
<td>Manual 1</td>
<td>0.3111</td>
<td>0.3111</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Automatic</td>
<td>0.0421</td>
<td>0.1263</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manual 2</td>
<td>0.8298</td>
<td>0.8298</td>
</tr>
<tr>
<td>2</td>
<td>Uşşak</td>
<td>Manual 1</td>
<td>0.1047</td>
<td>0.1163</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Automatic</td>
<td>0.0629</td>
<td>0.1887</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manual 2</td>
<td>0.0814</td>
<td>0.1771</td>
</tr>
<tr>
<td>3</td>
<td>Hicaz</td>
<td>Manual 1</td>
<td>0.4867</td>
<td>0.4867</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Automatic</td>
<td>0.0657</td>
<td>0.2300</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manual 2</td>
<td>0.6766</td>
<td>0.6766</td>
</tr>
<tr>
<td>4</td>
<td>Hüseyni</td>
<td>Manual 1</td>
<td>0.6364</td>
<td>0.8909</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Automatic</td>
<td>0.3689</td>
<td>0.6990</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manual 2</td>
<td>0.0730</td>
<td>0.5255</td>
</tr>
<tr>
<td>5</td>
<td>Saba</td>
<td>Manual 1</td>
<td>0.2646</td>
<td>0.2646</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Automatic</td>
<td>0.0524</td>
<td>0.2097</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manual 2</td>
<td>0.5512</td>
<td>0.5748</td>
</tr>
</tbody>
</table>

If we consider frame-based evaluation results, the success rates of automatic transcription for piece #2 uşşak and piece #4 hüseyni, the success rates of manual transcriptions are within the confidence interval of the success rates of manual transcriptions. While the success rates of automatic transcription are out of the confidence interval for the rest of other 3 recordings, they are closer to the success rates of manual transcriptions in comparison to the success rates found in note-based evaluation. This fact can be seen from Table 2. While the mean success rate of automatic transcription is much lower in note-based evaluation, it is much higher in frame-based evaluation as can be seen from Table 3. Another fact is that changes in threshold values result in less changes in evaluation results for manual transcriptions compared to that of automatic transcriptions.
In fact, frame-based evaluation for all transcriptions gives more optimistic results in comparison to note-based evaluation. Since the same pitch interval value, $3Hc$, is used as a threshold for correct transcribed notes, the main difference should result from the difference of approach to note-onsets in two evaluation metrics. In other words, while note-based evaluation has note-onset thresholds, 25 ms and 150 ms, frame-based evaluation does not use any condition for note-onsets.

4. Conclusion

In this paper, we comprehensively presented an AMT system designed for Turkish music for the first time in the literature. We also proposed a new evaluation approach to overcome the lack of a ground truth: the use of 2 different manual transcriptions for the evaluation of automatic transcriptions in comparison to the original scores.

The proposed AMT system consists of several blocks which estimate $f_0$ from audio recordings, automatically recognize makam of the recording and its tonic pitch, segment the $f_0$ and quantize both $f_0$ segments and its durations and finally label them with note names. While the overall success rate of our system for frame-based evaluation is found as %60 in terms of accuracy, the success rates of 2 manual transcriptions are found as %70 and %81. While the overall success rate of our system for note-based evaluation is found as %12 for 25 ms threshold value and %29 for 150 ms threshold value in terms of F-measure, the success rates of 2 manual transcriptions are found as %36 and %44 for 25 ms threshold value and as %41 and %65 for 150 ms threshold value in terms of F-measure.

The limitations of our study can be listed as follows:

i. Our system accepts only monophonic instrumental audio recordings of ney. A more realistic system should cover various kinds of Turkish instruments, as well as singing.

ii. Rhythm analysis is missing in our system. Usul (rhythmic mode) is one of the fundamental components of Turkish music. Therefore, a more realistic system should include rhythm analysis.

iii. Although our system automatically recognizes makam of the recordings, it is unable to use this knowledge for the use of accidentals in the transcription due to the possible modulations (geçki) in a given recording.

iv. Evaluation of our system is based on small number of test recordings. A more realistic system should be evaluated by much bigger number of recordings such as tens or hundreds of recordings.

v. Similarly, training of our system is based on small number of recordings. A more realistic system should be trained by much bigger number of recordings.

We target addressing these limitations as future work.
Acknowledgement
We thank The Scientific and Technological Research Council of Turkey, TÜBİTAK for their financial support (project number 107E024).

References


