Automatic transcription of Turkish microtonal music

Emmanouil Benetos[a)]
Centre for Digital Music, Queen Mary University of London, London E1 4NS, United Kingdom

André Holzapfel
Department of Computer Engineering, Boğaziçi University, 34342 Bebek, Istanbul, Turkey

(Received 21 January 2015; revised 18 August 2015; accepted 24 August 2015; published online 14 October 2015)

Automatic music transcription, a central topic in music signal analysis, is typically limited to equal-tempered music and evaluated on a quartertone tolerance level. A system is proposed to automatically transcribe microtonal and heterophonic music as applied to the *makam* music of Turkey. Specific traits of this music that deviate from properties targeted by current transcription tools are discussed, and a collection of instrumental and vocal recordings is compiled, along with aligned microtonal reference pitch annotations. An existing multi-pitch detection algorithm is adapted for transcribing music with 20 cent resolution, and a method for converting a multi-pitch heterophonic output into a single melodic line is proposed. Evaluation metrics for transcribing microtonal music are applied, which use various levels of tolerance for inaccuracies with respect to frequency and time. Results show that the system is able to transcribe microtonal instrumental music at 20 cent resolution with an F-measure of 56.7%, outperforming state-of-the-art methods for the same task. Case studies on transcribed recordings are provided, to demonstrate the shortcomings and the strengths of the proposed method. © *2015 Acoustical Society of America.*
[http://dx.doi.org/10.1121/1.4930187]

[TRM] Pages: 2118–2130

I. INTRODUCTION

Automatic music transcription (AMT) is defined as the process of converting an acoustic music signal into some form of music notation. The problem may be divided into several subtasks, including multiple-F0 estimation, onset/offset detection, instrument identification, and extraction of rhythmic information (Davy *et al.*, 2006). Applications of AMT systems include transcribing audio from musical styles where no score exists (e.g., music from oral traditions, jazz), automatic search of musical information, interactive music systems (e.g., computer participation in live human performances), as well as computational musicology (Klapuri and Davy, 2006). While the problem of automatic pitch estimation for monophonic (single voice) music is considered solved (de Cheveigné, 2006), the creation of a system able to transcribe multiple concurrent notes from multiple instrument sources with suitable accuracy remains open.

The vast majority of AMT systems target transcription of 12-tone equal-tempered (12-TET) Eurogenetic[1] music and typically convert a recording into a piano-roll representation or a MIDI file [cf. Benetos *et al.* (2013b) for a recent review of AMT systems]. Evaluation of AMT systems is typically performed using a quartertone (50 cent) tolerance, as, for instance, in the MIREX Multiple-F0 Estimation and Note Tracking Tasks (MIREX, 2007; Bay *et al.*, 2009). To the authors' knowledge, no AMT systems have been evaluated regarding their abilities to transcribe non-equal tempered or microtonal music, even though there is a limited number of methods that can potentially support the transcription of such music.

Related works on multiple-F0 estimation and polyphonic music transcription systems that could potentially support non-equally tempered music include the systems of Fuentes *et al.* (2013), Benetos and Dixon (2013), and Kirchhoff *et al.* (2013), which are based on *spectrogram factorization* techniques and utilize the concept of shift-invariance over a log-frequency representation in order to support tuning deviations and frequency modulations. The techniques employed include shift-invariant probabilistic latent component analysis (Fuentes *et al.*, 2013; Benetos and Dixon, 2013) and non-negative matrix deconvolution (Kirchhoff *et al.*, 2013). The method of Bunch and Godsill (2011) is also able to detect multiple pitches with high resolution, by decomposing linear frequency spectra using a Poisson point process and by estimating multiple pitches using a sequential Markov chain Monte Carlo algorithm. Other systems that support high-precision frequency estimation for polyphonic music include Dixon *et al.* (2012), which was proposed as a front-end for estimating harpsichord temperament, and the method of Rigaud *et al.* (2013), which is able to detect multiple pitches for piano music, as well as inharmonicity and tuning parameters.

The value of a transcription that takes microtonal aspects into account is illustrated by the history of transcription in ethnomusicology. In the late 19th century Alexander J. Ellis recognized the multitude of musical scales present in the musical styles of the world, and proposed the *cent* scale in order to accurately specify the frequency relations between scale steps (Stock, 2007). In the beginning of the 20th century, Abraham and von Hornbostel (1994) proposed

[a)]Electronic mail: emmanouil.benetos@qmul.ac.uk

 0001-4966/2015/138(4)/2118/13/$30.00

notational methods to transcribe "exotic" melodies, including a multitude of ways to describe microtonal inflections. Seeger (1958) suggested methods for accurately annotating microtonal inflections with an accuracy of 20 cents, a value close to the range of just noticeable differences in musical intervals [see Houtsma (1968), as cited by Thompson (2013, p.124)].

In addition to microtonality, another aspect of music that has been so far ignored in AMT systems is the phenomenon of *heterophony*. Heterophony, as defined by Cooke (2001), is the simultaneous variation of a single melody by several musicians. From a technical perspective, a heterophonic performance could be considered as polyphonic[2] due to the presence of several instruments, but the underlying concept is a monophonic melody. While heterophony is widely absent from European musical styles, it is often associated with the music of the Arab world (Racy, 2003) and encountered in similar ways in the music of the Balkans, Turkey, Iran, and other cultures of the near and Middle East. Not restricted to geographical area, it has also been assigned, for instance, to Javanese Gamelan (Anderson Sutton and Vetter, 2006), Korean music (Lee, 1980), and African American congregational singing, to name but a few. It has even been hypothesized as the origin of all music by Brown (2007), by interpreting polyphony as a later state of organization in pitch space. Because there is an apparent absence of previously published microtonal or heterophonic AMT approaches [see Bozkurt *et al.* (2014)], presumably attributed to a cultural bias toward Eurogenetic music, a consideration of these wide-spread musical traits in an AMT system seems timely. In general this would be advantageous for accommodating newfound access to the diversity of musical styles.

In this work, a system for transcribing heterophonic and microtonal music is proposed and applied to Turkish makam music, following preliminary work presented by Benetos and Holzapfel (2013). A collection of instrumental and vocal recordings has been compiled, along with detailed microtonal reference pitch annotations for quantitative evaluation of the system. The proposed method adapts a previously developed multi-pitch detection algorithm (Benetos and Dixon, 2013) to address the specific challenges of Turkish makam music and includes methods for converting a multi-pitch heterophonic output into a single melodic line. Evaluations are performed using various levels of tolerance for inaccuracies with respect to frequency and time. Results show that the system is able to transcribe microtonal instrumental music with 20-cent resolution. Case studies on transcribed recordings are provided, in order to demonstrate the shortcomings and the strengths of the method.

An outline of the paper is as follows: In Sec. II, motivations for creating technologies for transcribing heterophonic and microtonal music are given. Section III presents the instrumental and vocal music collections that were used for experiments, along with the pitch annotation process. The proposed system is described in Sec. IV. The employed evaluation metrics and results are presented in Sec. V. Finally, the performance of the proposed system is discussed in Sec. VI, followed by conclusions in Sec. VII.

II. MOTIVATIONS

Until recently, AMT approaches were developed and evaluated mainly in the context of Eurogenetic music. A disadvantage of such concentration is that AMT technology may be inadequate when applied to many music styles around the world, whose characteristics are fundamentally different from those of Eurogenetic music.

Regarding timbre as a first property, the authors note that polyphonic performances by piano and other Eurogenetic instruments attract a lot of attention for the development of AMT systems, while the consideration of instrumental timbres from other cultures represent rather an exception (Nesbit *et al.*, 2004). How a wider diversity of instrumental timbres can be transcribed automatically and accurately remains to be explored.

A second property of Eurogenetic music that limits the musical diversity that AMT systems can handle is the assumption that pitch is distributed according to the 12-TET system. Most current AMT approaches aim to produce a so-called "piano-roll" that specifies which note of the equal-tempered system is sounded at what time. Many music traditions, however, make use of very different organization of the tonal space, as for instance, the modal structures used in Turkey, Iran, Iraq, and India.

Finally, AMT systems for Eurogenetic music have been built on the assumption that music signals contain several distinct melody lines, or one melody line with a harmonic accompaniment. However, several music traditions in the world express melodies in a heterophonic way. That means that several instruments play one basic melody with each instrument interpreting it slightly differently, according to the aesthetic concepts of the music tradition. As far as the authors are aware, heterophony, as a combination of apparent polyphony at the signal level and monophony at the conceptual level has so far never been approached systematically with a prior AMT system.

The concentration of prior AMT systems on a limited range of timbres, the equal-tempered system, and restriction to either monophonic or polyphonic music, creates a distinct cultural bias towards Eurogenetic music. This motivates us to present a systematic study of an AMT system which focuses on a music that challenges all three structural biases. Turkish makam music was practiced at the Ottoman court and in religious ceremonies during the times of the Ottoman Empire, and continues to live on in today's music practice in modern Turkey in various forms. The melodies of this music follow the modal framework of the *makam*, which includes the notion of a scale and rules for melodic progression. While a comprehensive overview of the tonal and rhythmic concepts of Turkish makam music is given in Bozkurt *et al.* (2014), some of the properties of this music that are of particular relevance for the development of an AMT system will be emphasized.

(1) The Arel-Ezgi notation and tuning (Arel, 1968) represents the official standard today. While intervals are defined using Pythagorean pitch ratios, they are quantized to integer multiples of the Holderian-Mercator comma, ≈ 22.64 cents (Bozkurt *et al.*, 2014). As shown

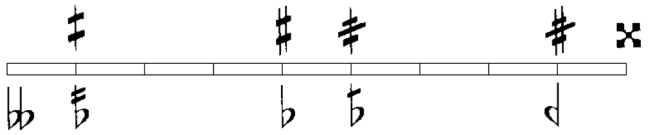


FIG. 1. Visualization of the accidentals used in Turkish music. Only four of the possible eight intermediate steps that divide a whole tone are used. The size of the step is 1 Hc (Holderian-Mercator comma) ≈22.64 cents.

in Fig. 1, according to Arel-Ezgi notation the intervals of 1, 4, 5, 8, and 9 commas are used within a whole tone. The whole tone interval, which is related to the frequency ratio 9/8 or 203.91 cents, is slightly sharp compared to 12-tone equal temperament (200 cents). Musical practice, however, tends to deviate from this notation by using pitch classes systematically different from those defined by the accidentals. This theory-practice mismatch in respect to the underlying tuning system represents a challenge for an AMT system. It is, however, important to point out that the Holderian-Mercator comma defines the smallest pitch difference in Turkish music, further supporting a maximum resolution of about 20 cents in frequency.

(2) Whereas the staff notation in a Eurogenetic context generally relates a certain pitch value to a specific fundamental frequency, this is not the case for Turkish makam music. Here, the musician may choose between one of 12 different transpositions, with the choice usually determined by the type of instrument being played, or the preferred vocal register of a singer.

(3) As a heterophonic music, a melody may be interpreted by instruments in different octaves.

(4) The notated melodies are richly ornamented using a set of idiosyncratic playing or singing techniques. These techniques are not documented in written theory, but are part of oral tradition and therefore only possible to grasp by applying ethnographic approaches. According to insights from fieldwork by the second author, the common practice is to add notes and embellishments during a performance while at the same time maintaining the onsets of the notes in a score. An illustrative example is given in Fig. 2, where the notation of a short phrase in makam *Beyati* is compared with the transcription of a performance of the phrase as performed on the instrument *oud*, a short-necked fretless lute. It is apparent that the density of notes has increased, while the originally notated onsets are maintained.

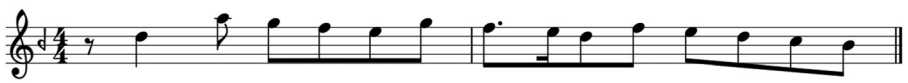
(5) In Turkish makam music, a set of instrumental timbres is encountered that is very characteristic for this music, and that differs from timbres usually encountered in Eurogenetic music. Further detail about the timbral qualities of these instruments is provided in Sec. III.

(6) Turkish makam music possesses no concept of functional harmony, but is a music based on modes called *makam*. The modes define a scale and characteristics of melodic progression. In the melodic progression, central notes of the mode are emphasized, and particular importance has the final note (*karar*) that concludes the progression and is usually referred to as the *tonic* in English language.

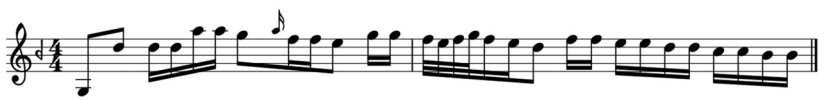
While these traits clearly set Turkish makam music apart from Eurogenetic music, Turkish makam music offers a relatively well-controlled environment for experiments with AMT systems. This is because of the large collections of music recordings and associated notations that are available from the work of the CompMusic project.[3] These controlled conditions significantly facilitate the creation of reference pitch annotations for music performances, which is necessary for the quantitative evaluation of an AMT system. On the other hand, establishing an AMT system for this musical style is an important first step towards automatic transcription of microtonal and heterophonic music throughout the music traditions of the world.

III. MUSIC COLLECTION

Turkish makam music makes use of instrumental timbres that clearly define the acoustic identity of this music. In Sec. III A an overview of the two instrumental timbres that were chosen as representatives of this identity is given, and the recorded material used for acquiring timbral templates is explained. Since the music collection used for the evaluation should cover a large variety, a set of instrumental performances and a set of vocal performances were compiled, which will be described in detail in Secs. III B and III C. Only performances of pieces that are available in the SymbTr collection (Karaosmanoğlu, 2012), which contains microtonal notation for Turkish music in a machine-readable format, were chosen. These annotations are a valuable starting point for the note-to-note alignment between notation and performance, which is needed for the evaluation of the system. The compilation of these reference transcriptions will be detailed in Sec. III D.



(a) Melody as notated



(b) Transcription of *oud* performance

FIG. 2. Two representations of the same melody, comparing the basic form as found in a score with the transcription of the same phrase as performed on the *oud*.

A. Timbral properties

Among the most widespread instruments for Turkish makam music are the *tanbur*, a long-necked lute, and the *ney*, an end-blown flute (constructed from reed material). The tanbur in its common form has seven strings, one bass string and three courses of double strings plucked using a plectrum made of tortoise shell. Melodies are played almost exclusively on the lowest course of double strings, while the other strings are plucked as a drone when the player wishes to give emphasis. Because of this playing technique, the instrument can be considered to have a limited polyphonic capacity. The length of the neck, the number of frets and their placement vary among instrument makers. Because the frets are movable, players frequently change their positions to adapt their tuning to, for instance, the tuning of another player. The main factor that influences the characteristic sound of this instrument is its very thin resonating soundboard, which does not have a sound hole. The open vibrating drone strings amplified by the resonating soundboard lend a very specific timbre to this instrument: the fourth harmonic exceeds the energy of the second harmonic in the radiated sound of the instrument, especially for forcefully plucked notes (Erkut *et al.*, 1999).

The ney has importance in court music ensembles as well as in religious practice. It is an end-blown flute, and as such is strictly monophonic, with a length between 52 and 104 cm depending on which lowest pitch is desired. As described in Sec. II, there are theoretically 12 transpositions in Turkish music, and their names refer to the ney's fundamental pitches. There are variations in the positioning of the finger holes depending on the instrument maker, just as there are variations in the placement of frets on the tanbur. Additionally, natural deviation of the nodes of the reed stalk from being equidistant result in a further source of variance. The pitch of a single fingering is strongly influenced by embouchure adjustments and angle of attack, enabling a player to precisely adjust tuning. The basic tonal range of the instrument is expanded by varying force and angle, reaching up to two and a half octaves. Notes in the higher range in particular, demand greater effort by the player to correct pitch to a desired one. Due to its construction as an end-blown flute, including all variations in positioning the instrument, the ney's timbre always contains a very high noise component.

Further instrumental timbres that are contained in ensemble performances of Turkish makam music are the *kemence*, a small fiddle played with a bow; the *oud*, a short necked lute; and the *kanun*, a type of zither played by plucking the strings. While the kemence can be considered a monophonic instrument, the oud and kanun can express polyphony.

The AMT system introduced in this paper offers the possibility to incorporate knowledge about the timbres of instruments targeted for transcription. As described in more detail in Sec. IV A, this knowledge is incorporated by learning typical magnitude spectra for pitches throughout the range of an instrument. In order to learn these templates, solo recordings of the target timbres are needed. To this end, pitches in approximately semitone intervals throughout the whole range of the instruments were recorded from ney, tanbur, kemence and kanun in a quiet environment using a studio quality portable recorder. In addition to these recordings, three ney and four tanbur solo performances from commercially available recordings were included in order to increase the timbral variety of the templates. In order to evaluate the system for vocal performances, vocal timbre templates were derived from solo recordings of those singers included in the collection of performances used for system evaluation. From the recordings of singers and the solo instrument performances, regions with relatively stable pitch were identified manually throughout the vocal range of the singer or instrument. All recordings of stable pitch regions were then used to derive the spectral templates for our system as described in Sec. IV A.

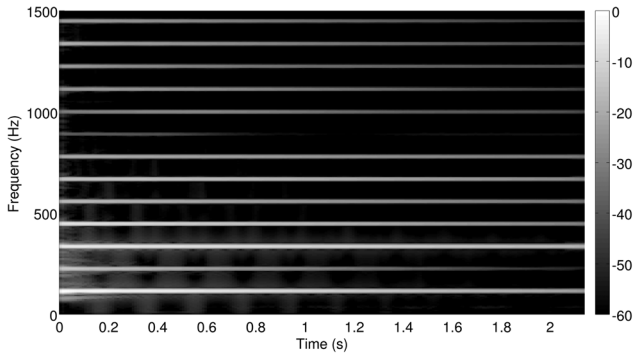
Descriptive examples of the spectral content of these template recordings are contrasted with a piano example in Fig. 3. In Fig. 3(a), the harmonic series for the piano has a clear fundamental, and generally slowly decreasing amplitudes towards higher harmonics, with the second harmonic at 220 Hz having a slightly smaller amplitude than the third harmonic. The tanbur is characterized by a very weak fundamental (at 110 Hz), which is a phenomenon present throughout the range of the instrument, and not restricted to this note. The strongest harmonics are the third to fifth harmonics, and the higher harmonics have less energy than for the piano. Throughout the duration of the note, an increasing focus on the frequency band between 300 and 1000 Hz can be seen. The spectrogram of the ney in Fig. 3(c) displays its noisy character caused by the type of excitation, as well as the practical absence of harmonics beyond the fourth. A harmonic series based on 220 Hz can be detected in the spectrogram. Even so, the actual perceived pitch seems to be about 440 Hz.

B. Instrumental performances

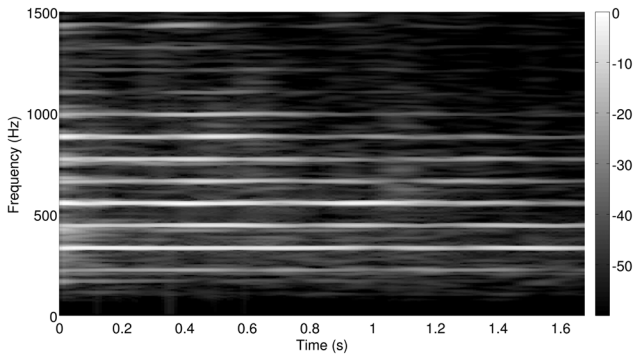
The most common forms of makam instrumental compositions are the *Peşrev* and *Saz Semaisi* forms. They share a similar overall compositional structure, with one repeated section (*Teslim*) interchanging with up to four sections (*Hane*) of new thematic material. Five solo performances for each ney and tanbur were chosen, and six ensemble performances that contain various instruments. Table I gives an overview of the instrumental performances. Horizontal lines in Table I divide the collection amongst groups that represent different recordings of the same composition. The tonic frequencies, obtained by manual annotation, demonstrate the different tonal ranges of the instruments, as well as the diversity of chosen transpositions. The depicted number of notes is obtained from the SymbTr notations, and the notes are edited as described in detail in Sec. III D. For ensemble performances, the instruments are identified as kanun (k), kemence (f), ney (n), oud (o), percussion (p), and tanbur (t).

C. Vocal performances

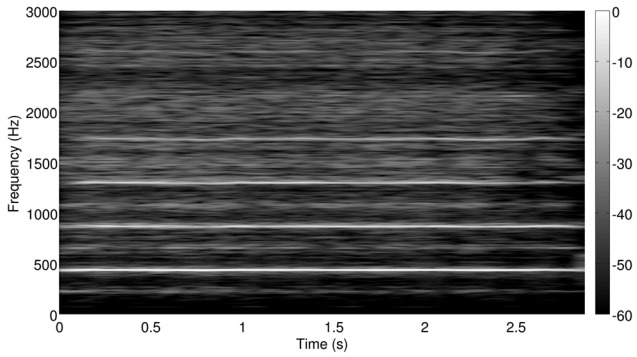
Three renowned vocal performers of Turkish makam music were chosen for vocal performances (see Table II),



(a) Spectrogram for piano note at 110Hz



(b) Spectrogram for tanbur note at 110Hz



(c) Spectrogram of ney note at 440Hz

FIG. 3. Note spectrograms for three instruments. The sidebar values are in dB, with 0 dB denoting the largest magnitude in the depicted power spectrum.

with most of the recordings chosen from the middle of the 20th century. All recordings contain the accompaniment by several instruments and can therefore be considered as heterophonic performances. The choice of singers was influenced by the availability of recorded vocal performances without instrumental accompaniment, a necessary prerequisite in our context, in order to obtain spectral templates per pitch for the specific singers. Solo vocal improvisations (*gazel*), *Quran* recitations and vocal teaching material were used to this end. Commercially or publicly available solo performances could not be found for any of the recent popular singers of Turkish makam music. This led to the vocal collection having an average recording quality that is inferior to the average quality of the instrumental collection.

TABLE I. Collection of instrumental recordings used for transcription.

	Form	Makam	Instr.	Notes	Tonic/Hz
1	Peşrev	Beyati	Ensemble (k,f,n,o,p,t)	906	125
2	Peşrev	Beyati	Ney	233	438
3	Saz S.	Hicazkar	Tanbur	706	147
4	Peşrev	Hüseyni	Ensemble (n,p)	302	445
5	Peşrev	Hüseyni	Ensemble (k,f,n,p,t)	614	124
6	Saz S.	Muhayyer	Ney	560	495
7	Saz S.	Muhayyer	Ensemble (k,f,n,t)	837	294
8	Peşrev	Rast	Tanbur	658	148
9	Peşrev	Rast	Ney	673	392
10	Peşrev	Segah	Ney	379	541
11	Peşrev	Segah	Ensemble (k,f,n)	743	246
12	Saz S.	Segah	Ensemble (k,f,n,p,t)	339	311
13	Saz S.	Segah	Tanbur	364	186
14	Saz S.	Uşşak	Tanbur	943	165
15	Saz S.	Uşşak	Tanbur	784	162
16	Saz S.	Uşşak	Ney	566	499

D. Manual pitch annotation process

For evaluation purposes, it was a necessary step to specify the onset times of notes as well their (microtonal) pitch values. Considering the performance practice described in Sec. II, where a high density of individual notes corresponds to a lower density series of notes in the notated melody, note offset times were not annotated. Machine-readable notations from the SymbTr collection were used as a starting point to derive the desired reference annotations. The semi-automatic approach that was followed had to take into account the micro-tonality and elaboration of melodies that are described in Sec. II. The SymbTr collection includes microtonal information, but all currently available pitch and onset alignment software is restricted to the lower resolution of the 12-TET system. For this reason, time alignment was performed using standard tools in 12-TET resolution, and then microtonal information was re-established. However, the SymbTr notations depict only the notes of the basic melody of a composition, with embellishments in the performances not being included. Hence, the objective of the transcription task is to create a transcription of the basic melody played by all included instruments in the heterophony, rather than a descriptive transcription (Seeger, 1958) of the detailed ornamentations present. The manual annotation process was conducted by the two authors. The first author holds a degree in piano performance, and the second author has five years of

TABLE II. Collection of vocal recordings used for transcription.

	Singer	Title (Makam)	Notes	Tonic/Hz
1	Bekir Sıdkı Sezgin	Bekledim Yıllarca Lâkin Gelmedin (Hüzzam)	451	141
2	Bekir Sıdkı Sezgin	Yandıkça Oldu Sûzân (Suzidil)	243	111
3	Kani Karaca	Bir Nigâh Et Ne Olur Halime (Hicaz)	306	123
4	Kani Karaca	Ülfet Etsem Yâr ile Ağyâre Ne (Hicaz-Uzzal)	425	123
5	Safiye Ayla	Vars in Gönül Aşkınla (Nişaburek)	339	335
6	Safiye Ayla	Bu Akşam Ayışığ inda (Saba)	333	328

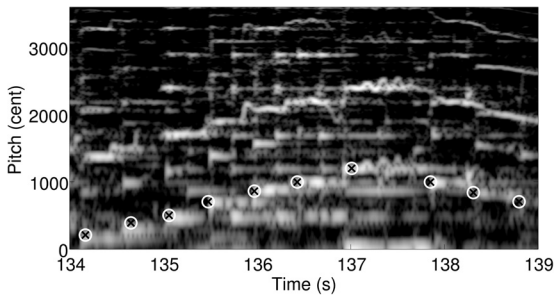


FIG. 4. Screenshot of an aligned sequence: the spectrogram for a short phrase in recording 5, with the aligned note onsets. Four instruments interpret the melody heterophonically: while the kanun ornaments strongly, the other instruments play closer to the basic melody.

practice of Turkish oud and took lessons from several masters of Turkish makam music.

As a first step in compiling the reference annotations, the SymbTr notations were manually edited in order to reflect exactly the sequence of sections as chosen in the performance. This is necessary because performers may omit sections of a piece or their notated repetitions. In the next step the (microtonal) SymbTr notation was converted to standard MIDI, and the approach presented by Macrae and Dixon (2010) was applied in order to get a rough estimate of the temporal alignment between the recording and the notes in a MIDI representation of the music. The resulting pre-aligned MIDI file was then loaded into Sonic Visualiser[4] as a notation layer on top of the spectrogram of the recording, and the timing of the alignment was corrected manually.

The manual alignment resulted in a list of notes with an accurate temporal alignment and a frequency resolution of 1 semitone. Micro-tonal information was then recovered from the edited SymbTr notations, obtaining a list of pitch values with 1 Hc resolution. The pitch values are normalized with respect to the tonic (*karar*) of the piece, so that the tonic was assigned a value of 0 cent. In Fig. 4 an example of the pitch annotation output is depicted. In this example, the ornamentations resulting in additional notes between 135.5 and 137 s (vertical lines caused from onsets of the kanun can be recognized) were not annotated, resulting in an alignment of the basic melody to this heterophonic ensemble performance. The annotation process resulted in reference annotations containing a total of 11 704 notes, consisting of 2411 for ney, 3455 for tanbur, 3741 for ensemble, and 2097 for vocal pieces. The annotations are available on the second author's website,[5] while the audio recordings can be obtained by using the provided identifiers.

In order to compare a reference pitch annotation with the output of the system for a given performance, the tonic frequency from this performance is needed. As described in Sec. II, this frequency depends on the chosen transposition and on tuning inaccuracies. The tonic frequencies for all performances were manually annotated. However, experiments on automatic tonic frequency estimation (Bozkurt, 2008) were conducted, and errors due to automatic estimation were monitored.

IV. SYSTEM

The proposed system takes as input an audio recording and information about the melodic mode (in this case, the makam). Multi-pitch detection with 20 cent resolution is performed based on the systems of Benetos and Dixon (2013) and Benetos *et al.* (2013a), which were originally proposed for transcribing Eurogenetic music [they ranked first in the MIREX 2013 Multiple F0 Estimation & Note Tracking public evaluation (MIREX)]. The original systems' abilities to support multi-pitch detection in a resolution finer than a semitone, which had not been exploited nor evaluated in Benetos and Dixon (2013), have been utilized. In addition, a note template dictionary using instruments and vocal performances from Turkish makam music is included in the proposed system. Finally, in order to support the transcription of heterophonic music, post-processing steps are included that convert a multi-pitch output into a single melodic line, and center the cent-scale output around the detected tonic. A diagram of the proposed transcription system can be seen in Fig. 5.

A. Spectral template extraction

In dictionary-based transcription systems, spectral templates per pitch are typically extracted from isolated note samples (Dessein *et al.*, 2010). Since to the authors' knowledge such a database of isolated note samples for Turkish instruments and vocals does not exist, recordings were performed and appropriate available solo performances were selected in order to obtain material from which to extract spectral templates, as detailed in Sec. III A.

For the ney and tanbur solo performances, each note segment is identified and manually labeled, and the probabilistic latent component analysis (PLCA) method (Smaragdis *et al.*, 2006) with one component was employed per segment in order to extract a single spectral template per pitch. The time/frequency representation used was produced by a

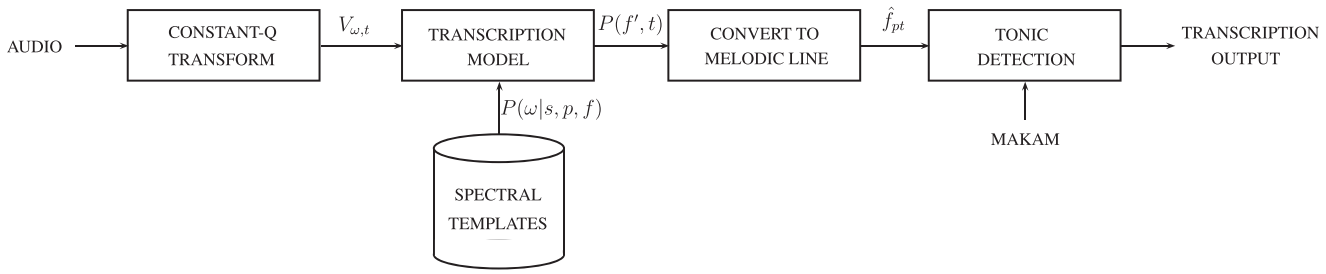


FIG. 5. Proposed transcription system diagram.

constant-Q transform (CQT) with a spectral resolution of 60 bins/octave (corresponding to 20 cent resolution), with 27.5 Hz as the lowest bin, and a 20 ms time step (Brown, 1991). Since in log-frequency representations like CQT inter-harmonic spacings are consistent for all pitches, spectral templates for missing pitches in the training set were created by shifting the CQT spectra of neighboring pitches. The same PLCA-based process was used for extracting templates from the set of isolated notes for ney, tanbur, kemence, and kanun. This resulted in an instrumental dictionary consisting of 5 ney models (spanning notes 60–88 in the MIDI scale), 5 tanbur models (spanning notes 39–72), 2 kanun models (spanning notes 53–88), and one kemence model (spanning notes 56–88).

For creating vocal templates, a training dataset of six solo voice recordings of Turkish makam music was used, covering the singers listed in Table II. Given the non-stable nature of the singing voice, a semi-supervised method was employed in order to speed up the annotation/template extraction process. A spectrogram of each recording was displayed using Sonic Visualiser;[4] stable pitch areas were manually annotated, and these annotated segments were concatenated to a new recording exclusively containing stable pitches. The aforementioned recording was used as input to the supervised PLCA algorithm, where the pitch activations were fixed (using the aforementioned user annotations) and the dictionary was estimated. The resulting vocal templates span MIDI notes 46 to 80.

B. Transcription model

For performing multi-pitch detection the model of Benetos and Dixon (2013), originally developed for computationally efficient transcription of Eurogenetic music, was employed and adapted. This model expands PLCA techniques by supporting the use of multiple pre-extracted templates per pitch and instrument source, as well as shift-invariance over log-frequency; the latter is necessary for performing multi-pitch detection at a frequency resolution higher than the semitone scale, as in the present work.

The transcription model takes as input a normalized log-frequency spectrogram $V_{\omega,t}$ (ω is the log-frequency index and t is the time index) and approximates it as a bivariate probability distribution $P(\omega,t)$. $P(\omega,t)$ is decomposed into a series of log-frequency spectral templates per pitch, instrument, and log-frequency shifting (which indicates deviation from the 12-TET system), as well as probability distributions for pitch activation, instrument contribution, and tuning.

The model is formulated as

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

where p denotes pitch, s denotes instrument source, and f denotes log-frequency shifting. $P(t)$ is equal to $\Sigma_\omega V_{\omega,t}$, which is a known quantity. All factors in the right-hand side of Eq. (1) are matrices (or tensors) containing values which vary from 0 to 1, indexed by their respective integer random variables. $P(\omega|s,p,f)$ denotes pre-extracted log-spectral templates per pitch p and source s, which are also pre-shifted across log-frequency according to index f. The pre-shifting operation is made in order to account for pitch deviations, without needing to formulate a convolutive model across log-frequency, as was the case for Smaragdis (2009). $P_t(f|p)$ is the time-varying log-frequency shifting distribution per pitch, $P_t(s|p)$ is the time-varying source contribution per pitch, and finally, $P_t(p)$ is the pitch activation, which is essentially the resulting transcription. The shifting index f is constrained to a semitone range with respect to an ideally tuned pitch according to 12-TET; given the CQT resolution (20 cents), $f \in [1,...,5]$, with 3 indicating no deviation from 12-TET (this represents tuning values of −40, −20, 0, 20, and 40 cents).

The unknown model parameters $P_t(f|p)$, $P_t(s|p)$, and $P_t(p)$ are estimated using iterative update rules based on the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977). For the expectation step, an intermediate distribution (i.e., the model posterior) is computed,

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

For the maximization step, the unknown model parameters are updated using the posterior from Eq. (2):

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

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

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

Equations (2)–(5) are iterated, with the number of iterations set to 30. The various matrices are initialized with random values; from EM theory, convergence to a local maximum is guaranteed (Dempster *et al.*, 1977). The templates $P(\omega|s,p,f)$ are kept fixed using the pre-extracted and pre-shifted spectral templates from Sec. IV A. The output of the transcription model is a pitch activation matrix and a pitch shifting tensor, which are, respectively, given by

$$P(p,t) = P(t)P_t(p), \quad (6)$$

$$P(f,p,t) = P(t)P_t(p)P_t(f|p). \quad (7)$$

By stacking slices of $P(f,p,t)$ for all pitch values, a time-pitch representation with 20 cent resolution can be created,

$$P(f', t) = [P(f, p_{low}, t) \cdots P(f, p_{high}, t)], \quad (8)$$

where f' denotes pitch in 20 cent resolution, with $p_{low} = 39$ being the lowest MIDI-scale pitch value, and $p_{high} = 88$ the highest pitch value considered. In Fig. 6 the time-pitch representation for a ney recording (piece no.2 from Table I) can be seen.

C. Post-processing

The outputs of the transcription model of Sec. IV B are non-binary and need to be converted into a list of note events, listing onset, offset, and pitch (the latter relative to the tonic frequency). First, median filtering is performed on $P(p, t)$, which is subsequently thresholded (i.e., matrix elements below a certain value are set to zero), and followed by minimum duration pruning (i.e., removing note events with durations less than 120 ms).

Since a significant portion of the transcription dataset consists of ensemble pieces where instruments (and in some cases, voice) are performing in octave unison, the heterophonic output of the multi-pitch detection algorithm needs to be converted into a monophonic output that will be usable as a final transcription. Thus, a simple "ensemble detector" is created by measuring the percentage of octave intervals in the detected transcription. If the percentage is above 15%, the piece is considered an ensemble piece. Subsequently, for each ensemble piece each octave interval is processed by merging the note event of the higher note with that of the lower one.

In order to convert a detected note event into the cent scale, information from the pitch shifting tensor $P(f, p, t)$ is used. For each detected event with pitch p and for each time frame t, the value of pitch deviation f that maximizes $P(f, p, t)$ is found,

$$\hat{f}_{p,t} = \arg\max_{f} P(f, p, t). \quad (9)$$

The median of $\hat{f}_{p,t}$ for all time frames belonging to each note event is selected as the tuning that best represents that note. Given the CQT resolution (60 bins/octave), the value in cent scale for the lowest frequency bin of the detected pitch is simply $20(\hat{f} - 1)$, where $\hat{f}$ is the pitch shifting index ($f \in [1, \ldots, 5]$) of the detected note.

D. Tonic detection

Because of the unknown transposition of the performance, we need to determine the frequency of the tonic in Hz in order to compare the automatic transcription with the reference pitch annotations. To this end, the procedure described by Bozkurt (2008) is applied. The method computes a histogram of the detected pitch values and aligns it with a template histogram for each makam using a cross-correlation function. The peak value of the pitch histogram is then assigned to the tonic that is closest to the peak of the tonic in the template, and all detected pitches are centered on this value. Finally, after centering the detected note events by the tonic, note events that occur more than 1700 cents or less than −500 cents from the tonic are eliminated, since such note ranges are rarely encountered in Turkish makam music.

V. EVALUATION

A. Metrics

For assessing the performance of the proposed system in terms of microtonal transcription, a set of metrics is proposed, by adapting the onset-based transcription metrics used for the MIREX Note Tracking evaluations (Bay *et al.*, 2009). In onset-based transcription evaluation of Eurogenetic music, an automatically transcribed note is assumed to be correct if its F0 deviates less than 50 cents from the annotated reference pitch and its onset is within either a 50 or 100 ms tolerance from the ground truth onset.

For the proposed evaluations, an automatically transcribed note is considered to be correct if its F0 is within a ±20 cent tolerance around the annotated reference pitch and its onset is within a 100 ms tolerance. The ±20 cent and ±100 ms tolerance levels were considered as "fair margins for an accurate transcription" by Seeger (1958). The

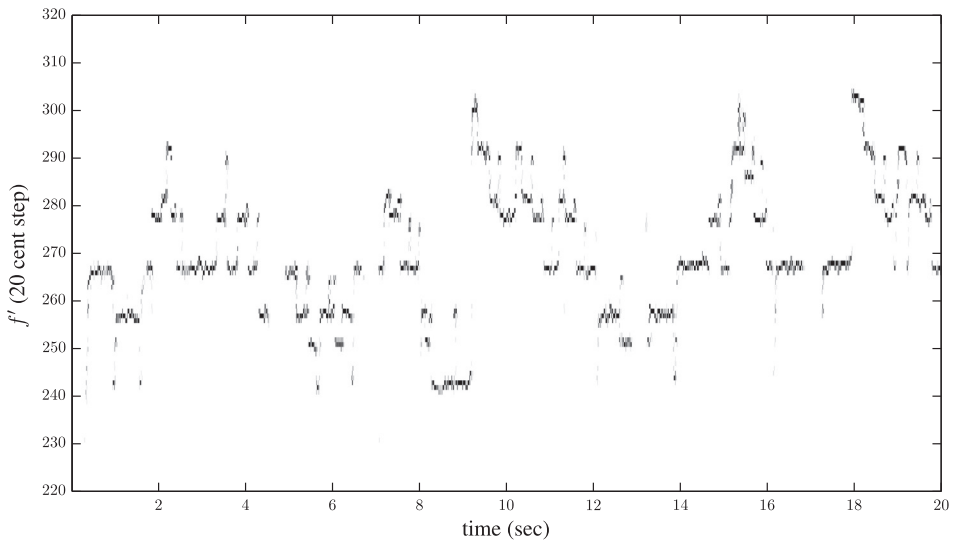


FIG. 6. The time-pitch representation $P(f', t)$ for the ney piece No. 2 from Table I.

following onset-based Precision, Recall, and F-measure are subsequently defined:

$$\mathcal{P}_{ons} = \frac{N_{tp}}{N_{sys}}, \quad \mathcal{R}_{ons} = \frac{N_{tp}}{N_{ref}}, \quad \mathcal{F}_{ons} = \frac{2\mathcal{R}_{ons}\mathcal{P}_{ons}}{\mathcal{R}_{ons} + \mathcal{P}_{ons}}, \quad (10)$$

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 positives. In all results, we display the metrics averaged across groups of recordings. It is important to point out that notes in octave distance are not considered as equal, since the differentiation of the octaves is important for a correct transcription of the melodic progression.

B. Results—Instrumental transcription

Two types of instrumental transcription evaluations were performed. The first tested the automatically detected tonic produced by the system of Bozkurt (2008). In the second evaluation, a manually annotated tonic was used. The proposed method was able to transcribe the entire 75 min instrumental dataset in less than one hour, i.e., less than real time. The instrumental transcription system included templates from the ney, tanbur, kanun, and kemence dictionaries.

Results using manually annotated tonic are shown in Table III, for the complete dataset as well as for individual instrument families. Results using the automatically detected tonic for the same dataset can be seen in Table IV. Using a manually annotated tonic, the proposed system reached $\mathcal{F}_{ons} = 56.75\%$ with a 20 cent tolerance [preliminary experiments in Benetos and Holzapfel (2013) reached 51.24% using a smaller dictionary]. All instrument subsets exhibited transcription performance above 50%, with a best performance of 58.5% being reached by the tanbur subset.

In order to demonstrate the robustness of the pitch-activation threshold parameter, an ROC curve for recall-vs-precision is shown in Fig. 7, where the thresholded values are varied from 1.0 to 10.0 [determined by the values in $P(t)$]. It can be seen that the system is fairly robust to threshold changes, with the lowest values reached by Precision and Recall being around 45%. From Fig. 7 it is apparent that the curve for tanbur reaches higher precision values than the ney and ensemble curves, which implies that the system detects fewer spurious onsets for the tanbur. One reason for the maximum possible precision being the lowest for the ensemble pieces is due to a heterophonic performance practice. In the presence of several instruments, usually at least one instrument will strongly ornament the basic melody, which adds additional notes to the automatic transcription. An example of this process is given in Fig. 8, which shows the same excerpt as Fig. 4, but visualizes the automatically transcribed notes. The presence of ornamentations led to four false positive detections (according to the reference pitch annotation) between 135 and 138 s.

TABLE III. Instrumental transcription results using manually annotated tonic.

	$\mathcal{P}_{ons}$	$\mathcal{R}_{ons}$	$\mathcal{F}_{ons}$
Ney recordings	55.89%	54.71%	55.00%
Tanbur recordings	64.88%	53.41%	58.52%
Ensemble recordings	51.44%	65.71%	57.06%
All recordings	57.31%	57.74%	56.75%

TABLE IV. Instrumental transcription results using automatically detected tonic.

	$\mathcal{P}_{ons}$	$\mathcal{R}_{ons}$	$\mathcal{F}_{ons}$
Ney recordings	53.22%	51.12%	51.91%
Tanbur recordings	44.51%	36.26%	39.91%
Ensemble recordings	42.72%	54.24%	47.34%
All recordings	47.22%	47.45%	46.73%

For comparing the proposed method with a recently published transcription algorithm, the method of Vincent *et al.* (2010) was employed. This method performs multi-pitch detection using adaptive non-negative matrix factorization and expresses an audio spectrogram as a series of weighted narrowband harmonic spectra. To ensure a fair comparison with the proposed method, the output of the aforementioned multi-pitch detection system (a list of note onsets and corresponding pitches) is post-processed in the same way as described in Sec. IV C, resulting in a list of onsets and pitches in cent value centered by a tonic. Results for the complete instrumental set using a manually annotated tonic show that the Vincent *et al.* (2010) method reaches $\mathcal{F}_{ons} = 38.52\%$ with 20 cent tolerance and $\mathcal{F}_{ons} = 49.84\%$ with 50 cent (i.e., semitone scale) tolerance, indicating that the proposed method, which reached 56.75%, is more suitable for the task of transcribing Turkish makam music, both in a microtonal setting and using a semitone resolution (cf. Table V).

Another comparison is carried out with respect to monophonic transcription, using the benchmark YIN pitch detection algorithm (de Cheveigné and Kawahara, 2002). Since YIN returns a continuous series of pitch values without identifying onsets/offsets, an "oracle" approach was employed, by using the ground truth (i.e., manually derived) onsets and offsets as additional information. Thus, for each annotated note event (defined by its ground truth

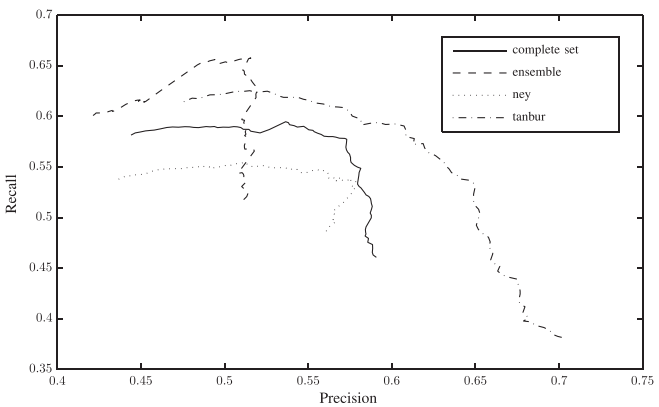


FIG. 7. ROC curves for recall-vs-precision using the instrumental dataset, as pitch-activation threshold values are varied.

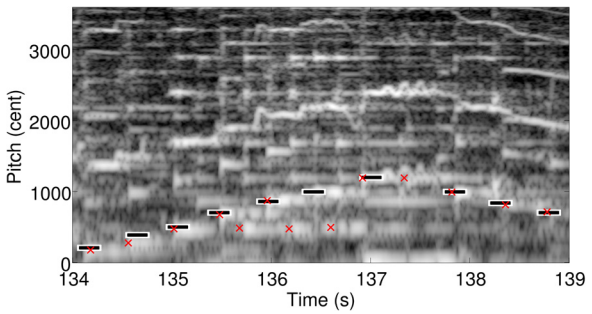


FIG. 8. (Color online) Excerpts from an ensemble transcription: Piece 5 from Table I, F-measure: 42.3%. The pitch axis is normalized to have the tonic frequency at 0 cent. The log-frequency spectrogram is depicted, overlaid with the automatic transcription as crosses, and the reference annotation indicated by black rectangles, framed by white color for better visibility. Width and height of the black rectangles are confined to an allowed tolerance of 100 ms and 20 cents.

onset-offset), its pitch is estimated by selecting computing the median pitch returned from YIN for that segment; this process is followed by the same post-processing steps described in Sec. IV C, again returning a list of onsets and pitches in cent value centered by a tonic. For the instrumental set, $\mathcal{F}_{ons} = 51.71\%$ (cf. Table V), while for the monophonic recordings $\mathcal{F}_{ons} = 55.41\%$ (the latter compared to 56.60% for the proposed system).

Experiments on the robustness of the proposed method to degradations of the audio input were also carried out, using the Audio Degradation Toolbox of Mauch and Ewert (2013). Pre-defined "vinyl" degradations were used, which are relevant for the collection selected for evaluation. The toolbox adds impulse responses, LP surface crackle, wow-and-flutter irregularities in playback speed, and pink noise. Using the degraded audio recordings, transcription performance reached 46.42%, which shows that the proposed system is relatively robust to reduction in recording quality (cf. Table V).

As in our preliminary experiments (Benetos and Holzapfel, 2013), there is a performance drop (10% in terms of F-measure) when the automatically detected tonic was used compared to the manually supplied one. This is attributed to the fact that with a 20 cent F0 evaluation tolerance, even a slight tonic miscalculation might lead to a substantial decrease in performance. Major tonic misdetections were observed for instrumental recordings 3 and 5 (described in Table I), leading to F-measures close to zero for those cases.

The impact of F0 and onset time tolerance on $\mathcal{F}_{ons}$ is shown in Table VI. With a 50 cent tolerance (corresponding to a standard semitone-scale transcription tolerance) the F-measure reaches 66.95%. This indicates that the proposed system is indeed successful at multi-pitch detection, and that a substantial part of the errors stems from detecting pitches at a precise pitch resolution.

TABLE V. Instrumental transcription results using various system configurations, compared with state-of-the-art approaches.

System	$\mathcal{F}_{ons}$
Proposed method	56.75%
Proposed method—added "vinyl" degradation	46.42%
Proposed method—using piano templates	53.28%
(Vincent *et al.*, 2010)—20 cent evaluation	38.52%
(Vincent *et al.*, 2010)—50 cent evaluation	49.84%
YIN (de Cheveigné and Kawahara, 2002)	51.71%

TABLE VI. Instrumental transcription results (in $\mathcal{F}_{ons}$) using different F0 and onset tolerance values.

F0 tolerance	10 cent	20 cent	30 cent	50 cent
$\mathcal{F}_{ons}$	38.90%	56.75%	62.68%	66.95%
Onset tolerance	50 ms	100 ms	150 ms	200 ms
$\mathcal{F}_{ons}$	42.75%	56.75%	60.66%	62.95%

In order to demonstrate the need for using instrument-specific templates for AMT, comparative experiments were made using piano templates extracted from three piano models taken from the MAPS database (Emiya *et al.*, 2010). Using the piano templates, the system reached $\mathcal{F}_{ons} = 53.28\%$, indicating that a performance decrease occurs when templates are applied that do not match the timbral properties of the source instruments (cf. Table V). This best performance with piano templates was obtained for the tanbur recordings (which might be attributed to those instruments having similar excitation and sound production); the ney recording performance was close to the average (53.4%), while the worst performance (of 51.2%) is observed for the ensemble recordings.

The impact of system sub-components can also be seen by disabling the "ensemble detection" procedure, which leads to an F-measure of 51.94% for the ensemble pieces, corresponding to about 5% decrease in performance. By removing the minimum duration pruning process, the reported F-measure with manually annotated tonic is 54.54%, which is a performance decrease of about 2%. Finally, by disabling the sub-component which deletes note events that occur more than 1700 cents or less than −500 cents from the tonic, system performance drops to 54.55%; this decrease is more apparent for the ensemble pieces (which were performed in an octave unison, spanning a wider note range), leading to an F-measure of 51.45%.

C. Results—singing transcription

For transcribing the vocal dataset, evaluations were also performed using the automatically detected and manually annotated tonics. The dictionary used for transcribing vocals consisted of a combination of vocal, ney, and tanbur templates.

Results are shown in Table VII; as with the instrumental dataset, there is a drop in performance (7% in terms of $\mathcal{F}_{ons}$) when using the automatically detected tonic. Performance is quite consistent across all recordings, with the best performance of $\mathcal{F}_{ons} = 72.2\%$ achieved for recording No. 4 from

TABLE VII. Singing transcription results using manually annotated and automatically detected tonic.

	$\mathcal{P}_{ons}$	$\mathcal{R}_{ons}$	$\mathcal{F}_{ons}$
Manually annotated	39.70%	44.71%	40.63%
Automatically detected	33.71%	36.53%	33.41%

Table II and the worst performance of $\mathcal{F}_{ons} = 21.2\%$ for recording No. 1 (which suffers from poor recording quality).

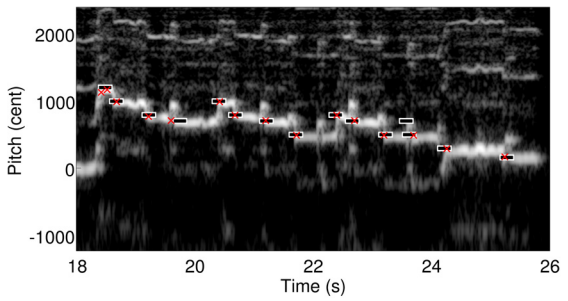
When using only vocal templates, system performance reaches $\mathcal{F}_{ons} = 34.8\%$, while when using only the instrumental templates an F-measure of 39.5% is achieved. This indicates that the instrumental templates contribute more to system performance than the templates extracted from the vocal training set, although including the vocal templates leads to an improvement over using only the instrumental templates. For comparison, using the multi-pitch detection method of Vincent *et al.* (2010) as in Sec. V B, with 20 cent tolerance yields $\mathcal{F}_{ons} = 22.8\%$, while 50 cent tolerance gives $\mathcal{F}_{ons} = 36.6\%$.

In general, these results indicate the challenge of transcribing mixtures of vocal and instrumental music, in particular, in cases of historic recordings. However, the results are promising, and indicate that the proposed system can successfully derive transcriptions from vocal and instrumental ensembles, which can serve as a basis for fixing transcription errors in a user-informed step. Detailed discussion on the instrumental and vocal systems will be made in Sec. VI.

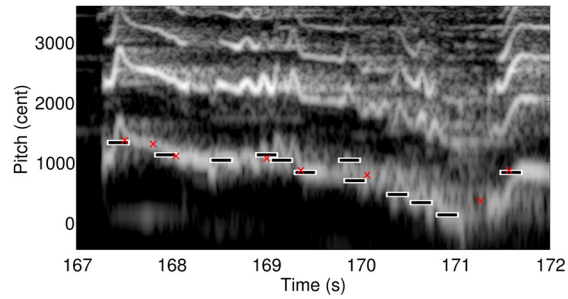
VI. DISCUSSION

The results obtained from the proposed AMT system indicate lower performance for vocal pieces compared to results for instrumental recordings. As pointed out in Sec. III C, the recording quality of the vocal recordings is generally lower than the quality of most instrumental performances, which is reflected in a higher noise level and the absence of high-frequency information due to low-quality analog-to-digital conversion. In order to assess the impact of the low recording quality, an informal experiment was carried out, in which six new vocal recordings were chosen for transcription. Since for those recordings no time-aligned reference pitch annotations exist, a qualitative evaluation was performed by an aural comparison of an original vocal recording with a synthesizer playback of a transcription of the recording. This experiment did not indicate a clear improvement of vocal transcription for the newer recordings.

An insight can be obtained into what was identified as the main reason for the low transcription performance for vocal pieces by comparing the depicted spectrograms in Figs. 9(a) and 9(b). The instrumental example in Fig. 9(a) is characterized by pitch that remains relatively stable for the duration of a note, and by note onsets that can be identified by locating changes in pitch. However, the vocal example in Fig. 9(b) is completely different. Here, the pitch of the voice is clearly distinguishable but characterized by a wide vibrato. For instance, in the downward movement starting at about 170 s, the notation contains a progression through subsequent notes of the Hüzzam-makam scale, which appears in the singing voice with a vibrato of the almost constant range of five semitones. Such characteristics are typical of Turkish vocal performances, and it seems hard to imagine a method based purely on signal processing that could correctly interpret such a performance in terms of the underlying implied note sequence. It is an open question whether this difficulty



(a) Good transcription: Piece 2 from Table I, F-measure: 70.8%



(b) Bad transcription: Piece 1 from Table II, F-measure: 20.0%

FIG. 9. (Color online) Excerpts from transcriptions. Axes and symbols follow the principle of Fig. 8.

of transcribing vocal performances is unique to this form of music, or if AMT systems would exhibit similar performance deficits for other styles of music as well. Based on our own observations of musical practice in Turkey, instrumental music education more frequently explains ornamentations in terms of notes than in vocal education, where teachers tend to teach ornamentations such as the vibrato in Fig. 9(b) purely in terms of performance demonstrations.

One aspect important to point out is that the system performance values displayed in Sec. V contain an over-pessimistic bias. As explained in Sec. II, Turkish makam music practice deviates from the pitch values implied by notation, due to a mismatch between theory and practice. However, our reference annotations contain pitch values that are following the most common theoretical framework to explain the tonal concepts of Turkish makam music, while the performances contain pitches that will deviate from the theoretical values at least for some cases. For instance, the step from the tonic to the fourth note in makam Segah is usually notated as a perfect fourth. However, within performances this interval tends to be larger because the tonic is typically played (by instruments) at a lower pitch. For piece 10 in Table I, a clear increase of this interval compared to the annotated one is observed. For this piece, correcting this interval from 500 to 530 cents changes the F-measure from 30.6% to 39.7%, a substantial improvement. Similar phenomena are very likely to occur for other pieces, but a systematic evaluation would require manual correction of all individual pitch values in our reference annotations.

VII. CONCLUSIONS

In this paper, a system for transcribing microtonal makam music of Turkey is proposed, based on spectrogram factorization models relying on pre-extracted spectral templates per pitch. A collection of instrumental and vocal recordings was compiled and annotated, and evaluation metrics suitable for microtonal transcription were proposed. Results show that the system is able to transcribe both instrumental and vocal recordings with variable accuracy ranging from approximately 40% to 60% for 20 cent resolution, depending on several factors. Results are substantially better using manually determined tonic values as compared with an automatic method. We also observed a discrepancy between music theory and practice, as observed through the reference pitch annotations that followed a theoretical framework. The code for the proposed system is available online.[6]

A logical extension of this work is to combine acoustic models with music language models suitable for microtonal and heterophonic music, in order to both improve transcription performance and quantify the gap between theory and practice in Turkish makam music. Finally, following work in Benetos and Dixon (2013), another suggested extension is to annotate the various sound states observed in typical Turkish makam music instruments (such as attack, sustain, decay), which the authors believe will result in a more robust and accurate AMT system for microtonal music.

ACKNOWLEDGMENTS

The authors would like to thank Bariş Bozkurt for his advice and for providing us with software tools, as well as Robert Reigle and Simon Dixon for proofreading. E.B. was supported by a Royal Academy of Engineering Research Fellowship (RF/128) and by a City University London Research Fellowship. A.H. was supported by a Marie Curie Intra European Fellowship (Grant No. PIEF-GA-2012-328379).

[1]Term used to avoid the misleading dichotomy of Western and non-Western music, proposed by Reigle (2013).

[2]The term "polyphony" in the context of AMT does not necessarily refer to a polyphonic style of composition. It rather refers to music signals that contain either several instruments, or one instrument that is capable of playing several individual melodic voices at the same time, such as the piano. On the other hand, "monophonic" refers to signals that contain one instrument that is capable of playing at most one note at a time (e.g., flute). The two terms are used with this technical interpretation in the paper.

[3]http://compmusic.upf.edu (Last viewed August 6, 2015).

[4]http://www.sonicvisualiser.org/ (Last viewed August 6, 2015).

[5]Please follow the links provided at http://www.rhythmos.org/Datasets.html (Last viewed August 6, 2015) in order to obtain the annotations as two archives. Lists that identify all performances using their MusicBrainz ID (musicbrainz.org) or a YouTube-link if no ID is available, are included.

[6]Code for proposed system: https://code.soundsoftware.ac.uk/projects/automatic-transcription-of-turkish-makam-music (Last viewed August 6, 2015).

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

Anderson Sutton, R., and Vetter, R. R. (**2006**). "Flexing the frame in Javanese gamelan music: Playfulness in a performance of Ladrang Pangkur," in *Analytic Studies in World Music*, edited by M. Tenzer (Oxford University Press, Oxford, UK), Chap. 7, pp. 237–272.

Arel, H. S. (**1968**). *Türk Musikisi Nazariyat i* (*The Theory of Turkish Music*) (Hüsnütabiat matbaas i, Istanbul, Turkey), Vol. 2.

Bay, M., Ehmann, A. F., and Downie, J. S. (**2009**). "Evaluation of multiple-F0 estimation and tracking systems," in *International Society for Music Information Retrieval Conference*, Kobe, Japan, pp. 315–320.

Benetos, E., Cherla, S., and Weyde, T. (**2013a**). "An efficient shift-invariant model for polyphonic music transcription," in *6th International Workshop on Machine Learning and Music*, Prague, Czech Republic, pp. 7–10.

Benetos, E., and Dixon, S. (**2013**). "Multiple-instrument polyphonic music transcription using a temporally-constrained shift-invariant model," J. Acoust. Soc. Am. **133**, 1727–1741.

Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H., and Klapuri, A. (**2013b**). "Automatic music transcription: Challenges and future directions," J. Intell. Inf. Syst. **41**, 407–434.

Benetos, E., and Holzapfel, A. (**2013**). "Automatic transcription of Turkish makam music," in *International Society for Music Information Retrieval Conference*, Curitiba, Brazil, pp. 355–360.

Bozkurt, B. (**2008**). "An automatic pitch analysis method for Turkish maqam music," J. New Mus. Res. **37**, 1–13.

Bozkurt, B., Ayangil, R., and Holzapfel, A. (**2014**). "Computational analysis of makam music in Turkey: Review of state-of-the-art and challenges," J. New Mus. Res. **43**, 3–23.

Brown, J. C. (**1991**). "Calculation of a constant Q spectral transform," J. Acoust. Soc. Am. **89**, 425–434.

Brown, S. (**2007**). "Contagious heterophony: A new theory about the origins of music," Musicae Scientiae **11**, 3–26.

Bunch, P., and Godsill, S. (**2011**). "Point process MCMC for sequential music transcription," in *International Conference on Acoustical Speech and Signal Processing*, Prague, Czech Republic, pp. 5936–5939.

Cooke, P. (**2001**). "Heterophony," Oxford Music Online, Grove Music Online, http://grovemusic.com/ (Last accessed August 6, 2015).

Davy, M., Godsill, S., and Idier, J. (**2006**). "Bayesian analysis of western tonal music," J. Acoust. Soc. Am. **119**, 2498–2517.

de Cheveigné, A. (**2006**). "Multiple F0 estimation," in *Computational Auditory Scene Analysis, Algorithms and Applications*, edited by D. L. Wang and G. J. Brown (IEEE Press/Wiley, New York), pp. 45–79.

de Cheveigné, A., and Kawahara, H. (**2002**). "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am. **111**, 1917–1930.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (**1977**). "Maximum likelihood from incomplete data via the EM algorithm," J. R. Stat. Soc. **39**, 1–38.

Dessein, A., Cont, A., and Lemaitre, G. (**2010**). "Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence," in *International Society for Music Information Retrieval Conference*, Utrecht, Netherlands, pp. 489–494.

Dixon, S., Mauch, M., and Tidhar, D. (**2012**). "Estimation of harpsichord inharmonicity and temperament from musical recordings," J. Acoust. Soc. Am. **131**, 878–887.

Emiya, V., Badeau, R., and David, B. (**2010**). "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," IEEE Trans. Audio, Speech Lang. Proc. **18**, 1643–1654.

Erkut, C., Tolonen, T., Karjalainen, M., and Välimäki, V. (**1999**). "Acoustic analysis of Tanbur, a Turkish long-necked lute," in *International Congress on Sound and Vibration (IIAV)*, pp. 345–352.

Fuentes, B., Badeau, R., and Richard, G. (**2013**). "Harmonic adaptive latent component analysis of audio and application to music transcription," IEEE Trans. Audio Speech Lang. Proc. **21**, 1854–1866.

Houtsma, A. (**1968**). "Discrimination of frequency ratios," J. Acoust. Soc. Am. **44**, 383.

Karaosmanoğlu, K. (**2012**). "A Turkish Makam music symbolic database for music information retrieval: Symbtr," in *International Society for Music Information Retrieval Conference*, Porto, Portugal, pp. 223–228.

Kirchhoff, H., Dixon, S., and Klapuri, A. (**2013**). "Missing template estimation for user-assisted music transcription," in *International Conference on Acoustical Speech and Signal Processing*, Vancouver, Canada, 26–30.

Klapuri, A., and Davy, M. (**2006**). *Signal Processing Methods for Music Transcription* (Springer-Verlag, New York).

Lee, K. (**1980**). "Certain experiences in Korean music," in *Musics of Many Cultures: An Introduction*, edited by E. May (University of California Press, Oakland, CA), pp. 32–47.

Macrae, R., and Dixon, S. (**2010**). "Accurate real-time windowed time warping," in *International Society for Music Information Retrieval Conference*, Utrecht, Netherlands, pp. 423–428.
Mauch, M., and Ewert, S. (**2013**). "The audio degradation toolbox and its application to robustness evaluation," in *International Society for Music Information Retrieval Conference*, Curitiba, Brazil, pp. 83–88.
MIREX (**2007**). "Music Information Retrieval Evaluation eXchange (MIREX)," http://music-ir.org/mirexwiki/ (Last accessed August 6, 2015).
Nesbit, A., Hollenberg, L., and Senyard, A. (**2004**). "Towards automatic transcription of Australian aboriginal music," in *Proceedings of the 5th International Conference on Music Information Retrieval*, Barcelona, Spain.
Racy, A. J. (**2003**). *Making Music in the Arab World: The Culture and Artistry of Tarab* (Cambridge University Press, Cambridge, UK), Chap. Heterophony, pp. 80–96.
Rigaud, F., David, B., and Daudet, L. (**2013**). "A parametric model and estimation techniques for the inharmonicity and tuning of the piano," J. Acoust. Soc. Am. **133**, 3107–3118.
Reigle, R. (**2013**). (personal communication).
Seeger, C. (**1958**). "Prescriptive and descriptive music-writing," Music Quart. **64**, 184–195.
Smaragdis, P. (**2009**). "Relative-pitch tracking of multiple arbitary sounds," J. Acoust. Soc. Am. **125**, 3406–3413.
Smaragdis, P., Raj, B., and Shashanka, M. (**2006**). "A probabilistic latent variable model for acoustic modeling," in *Advances in Models for Acoustic Processing Workshop (NIPS'06)*, Whistler, Canada.
Stock, J. P. J. (**2007**). "Alexander j. Ellis and his place in the history of ethnomusicology," Ethnomusicology **51**, 306–325.
Thompson, W. F. (**2013**). "Intervals and scales," in *The Psychology of Music*, edited by D. Deutsch (Elsevier, Amsterdam, the Netherlands), Chap. 4.
Vincent, E., Bertin, N., and Badeau, R. (**2010**). "Adaptive harmonic spectral decomposition for multiple pitch estimation," IEEE Trans. Audio Speech Lang. Process. **18**, 528–537.