Journal of the Audio Engineering Society
Vol. 64, No. 09, September 2016 (© 2016)
DOI: http://dx.doi.org/10.17743/jaes.2016.0033

Transcription of Polyphonic Vocal Music with a Repetitive Melodic Structure

CIRIL BOHAK AND MATIJA MAROLT

University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia

This paper presents a novel method for transcription of folk music that exploits its specifics to improve transcription accuracy. In contrast to most commercial music, folk music recordings may contain various inaccuracies as they are usually performed by amateur musicians and recorded in the field. If we use standard approaches for transcription, these inaccuracies are reflected in erroneous pitch estimates. On the other hand, the structure of western folk music is usually simple as songs are often composed of repeated melodic parts. In our approach we make use of these repetitions to increase transcription robustness and improve its accuracy. The proposed method fuses three sources of information: (1) frame-based multiple F0 estimates, (2) song structure, and (3) pitch drift estimates. It first selects a representative segment of the analyzed song and aligns all the other segments to it considering temporal as well as frequency deviations. Information from all segments is summarized and used in a two-layer probabilistic model based on explicit duration HMMs, to segment frame-based information into notes. The method is evaluated with state-of-the-art transcription methods where we show that significant improvement in accuracy can be achieved.

1 INTRODUCTION

Automatic music transcription is a process of transforming an acoustic music signal into some kind of symbolic music notation [1]. It typically involves the detection of multiple concurrent pitches in a musical signal (multi-pitch detection), the detection of note onsets and offsets, as well as recognition of the instruments present in the audio signal. The problem is considered to be an open problem, even more so for high degree of polyphony and presence of multiple instruments in the musical signal. Automatic music transcription is part of the broader field of music information retrieval and can be considered an example of intelligent audio processing as it extracts a high level semantic description from a music recording. A recent review of automatic music transcription approaches can be found in [2] and the challenges and future directions in [3].

Most transcription approaches are based on estimation of fundamental (F0) frequencies of notes present in the signal. To do so, authors in [4] use probability density functions to estimate the relative dominance of every possible F0 and the shape of harmonic-structure tone models by maximum a-posteriori probability estimation considering their prior distribution. Klapuri [5] presented an iterative method for multiple fundamental frequency estimation that is based on harmonicity and spectral smoothness. The method estimates F0 of the most prominent sound, subtracts the sound from the mixture and repeats the process for the residual signal. A transcription method presented in [6] uses a note event modeling technique for estimating individual note events with a hidden Markov model (HMM). The model uses extracted acoustic features for calculating the likelihood of different notes and a musicological model for modeling note transitions.

Many recent approaches to transcription rely on decomposition of a time-frequency representation of the audio signal with methods such as non-negative matrix factorization (NMF) [7–9], non-negative matrix approximation [10], or probabilistic latent component analysis (PLCA) [11–13]. A state-of-the-art method is presented in [14], which couples PLCA with explicit duration hidden Markov models (EDHMMs). Authors have also presented several adaptations of the method presented in [15–17]. Their most recent method is based on shift-invariant probabilistic latent component analysis (SI-PLCA) with support for spectral templates for individual sound states trained on several orchestral instruments.

According to the Music Information Retrieval Evaluation Exchange (MIREX) [18], current state-of-the-art algorithms are reaching the accuracy of 0.65 on a collection of (mostly synthesized) instrumental music and 0.35 on real-world instrumental music. To improve on transcription results, many authors limit the domain to transcription of specific instruments, e.g., [19, 20] for piano music or [21] for bell chiming.

With ongoing digitization of ethnomusicological archives, the demand for methods for automatic extraction of metadata from folk music field recordings is growing [22–24]. While the aforementioned transcription methods can be applied to folk music recordings, we show in this paper that their accuracy can be improved by considering folk music specifics. In contrast to commercially recorded popular or classical music, field recordings can contain a significant amount of noise, either due to poor recording equipment and noisy environment or due to amateur musicians who make mistakes, sing individual tones inaccurately, or change intonation during a song (pitch drift). On the other hand, in many western folk music traditions songs usually have a simple structure, with several repetitions of the same melodic part (stanza) and with simple harmonic content.

The main contribution of this paper is a novel method that improves transcription of folk song recordings by taking into account folk music specifics. It first analyzes folk song structure and estimates boundaries of melodic repetitions as well as pitch drift. It combines this knowledge with output of a conventional multiple F0 estimation method and processes the combined estimates with a musicological model based on explicit duration HMMs, which yields the final note-based transcription. We show that results of the original F0 estimates are significantly improved with our approach.

In the following sections we first present our method, then evaluate it and discuss its performance. We conclude the paper with suggestions for future improvements.

2 THE PROPOSED METHOD

The goal of the proposed method is to obtain a note-based transcription of a single most representative melodic part (stanza) of a folk song. It is based on the assumption that a song is composed of repeated stanzas, which may, however, differ due to performance variations ranging from small to large tempo deviations, singing inaccuracies, pitch drift, as well as recording conditions such as environmental noises and interruptions. Songs can be polyphonic and the method is especially suited for transcribing vocal songs, although it is also useful in an instrumental context.

The method takes an audio signal of a song as its input and produces a list of notes with their pitch, onset, and offset times as output. In the process it fuses three types of time-varying information to calculate a robust transcription: (1) multiple F0 estimates, which can be obtained with a transcription method, such as methods by Klapuri [5] or Benetos [17]; (2) segment boundaries that define repeated melodic segments (stanzas); and (3) pitch drift estimates, which describe the change in overall intonation during performance. All three sources of information are fused to obtain a multi-F0 transcription of a single melodic part derived from all stanza repetitions. Finally, a note estimation approach based on musicological knowledge, similar to [6], is used to estimate individual notes. The approach is outlined in Fig. 1 and described in more detail in the following subsections.

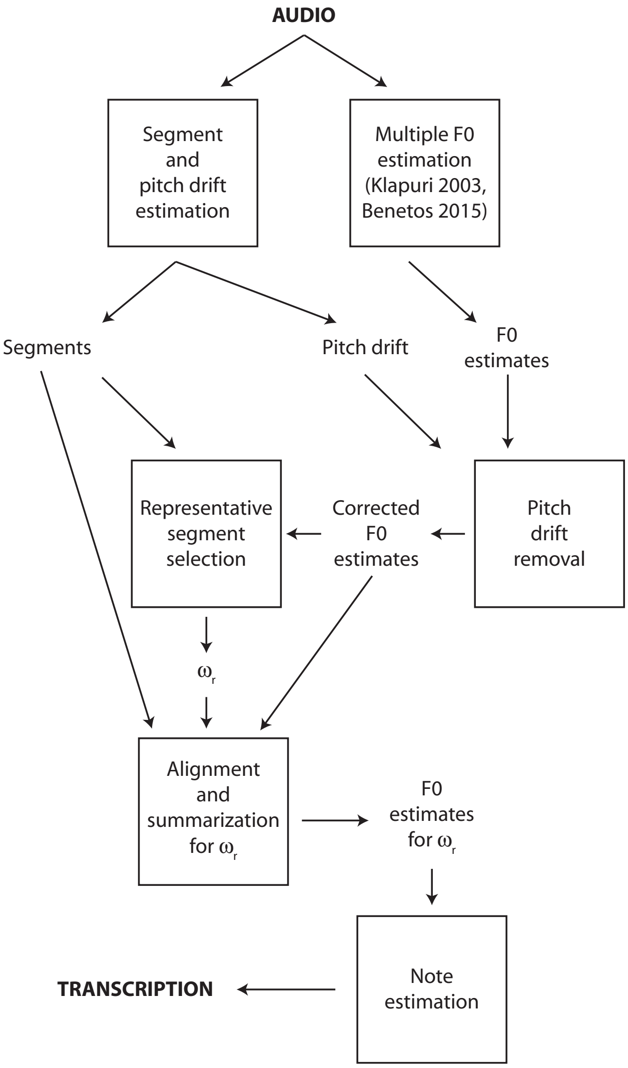
AUDIO
Segment and pitch drift estimation
Multiple F0 estimation (Klapuri 2003, Benetos 2015)
Segments
Pitch drift
F0 estimates
Representative segment selection
Corrected F0 estimates
Pitch drift removal
ω_r
Alignment and summarization for ω_r
F0 estimates for ω_r
Note estimation
TRANSCRIPTION

Fig. 1. Figure shows an outline of the proposed method where ω_i is a selected representative segment.

2.1 Segmentation and Pitch Drift Estimation

Segmentation and pitch drift are obtained with a method presented in [25]. We define segmentation as a set of boundaries between repeated segments $\Omega = \{\omega_i\}$, where ω_i represents the beginning time of the i-th segment. Segmentation is obtained by maximizing segmentation likelihood, defined as:

$$P(\Omega) = P(\omega_1) \prod_{\omega_i \in (\Omega \backslash \omega_1)} P(\omega_i|\omega_{i-1})P(\omega_i).$$

$P(\omega_i|\omega_{i-1})$ is the transition probability between segments and is calculated from the similarity of segments starting at ω_i and ω_{i-1}, also taking the expected duration of segment ω_{i-1} into account. Similarity between segments is calculated with dynamic time warping of chromatic representations of the two segments taking tempo variations, as well as intonation changes (pitch drift) into consideration. An integral part of similarity estimation is thus the estimation of pitch drift $D = [d_t]$, $t = 1\ldots T$, where d_t represents the change of intonation in cents relative to beginning of the song ($d_1 = 0$) and T is the length of the analyzed signal.

$P(\omega_i)$ represents the probability that a segment starts at ω_i. Its calculation is based on the rationale that this value is larger if the boundary is preceded by a low-amplitude signal region—for singing, this often corresponds to breathing pauses before stanza beginnings, while for instrumental parts, this may also correspond to phrase endings.

The method achieves an F1 score of 0.76 on a collection of folk music recordings. For more details on the algorithm and its evaluation, we direct the reader to [25].

2.2 Multiple F0 Estimation

To obtain a set of F0 estimates, our method relies on a multiple F0 estimator such as presented by Klapuri [5] or Benetos [17]. An estimator is expected to return at least two sets of values: per-frame F0 estimates $F = [f_{it}]$ and their saliences $S = [s_{it}]$, $i = 1\ldots M$, $t = 1\ldots T$, where M is the maximum number of simultaneously estimated frequencies and T the length of the analyzed signal. Saliences of F0 estimates represent confidence of the F0 estimator in the found F0 values—a higher salience value means a higher confidence in the found pitch.

2.3 Adjusting for Pitch Drift

In vocal performances, where no explicit stable pitch reference is available, amateur musicians often tend to change intonation during performance. This happens in solo as well as in choir recordings and intonation may change upwards or downwards at various parts of a song, such as between stanzas or at large intervals. To be able to compare and align different parts of a song, we need to adjust the F0 estimates according to the estimated change of intonation. As pitch drift $\mathcal{D}$ is already estimated during segmentation [25], F0 values are adjusted in this step according to the estimated drift, as in: $f'_{it} = f_{it} + d_t$, where all values are given in cents.

2.4 Selecting a Representative Segment

The goal of our transcription method is to obtain the most representative transcription of a melodic segment in a song. As all repetitions of segments are different due to non-mechanical performances, we need to choose a segment that is representative of a song, which we define to be a segment that is similar to most other segments. To cast it differently, we are looking for a segment that is not likely to be an outlier and that thus has little inaccuracies in performance. The representative segment will in further steps be aligned to all others, to obtain the most likely (most often sung) sequences of pitches in the song and thus the most representative transcription.

We could calculate an exhaustive alignment and comparison of all segments, however since we are mostly interested in not picking an outlier, the selection of a representative segment is simplified as follows. We first convert pitch drift corrected F0 and salience values into a piano roll representation $P = [p_{it}]$, $i = 1\ldots F$, $t = 1\ldots T$, where i represents frequency bins on a quantized logarithmic frequency scale. P contains non-negative salience values in all bins that correspond to estimated F0 values. We then calculate the segment pitch profile pp of each segment ω_j by summing the piano roll over time as:

$$pp_i(\omega_j) = \sum_{t=\omega_j}^{\omega_{j+1}} p_{it}, i = 1\ldots F. \tag{1}$$

A representative segment ω_r is then calculated as the one that has the most similar pitch profile to all others and is of an approximately average length:

$$r = \underset{i}{\operatorname{argmax}} \left(\frac{1}{|\Omega|} \sum_{j=1}^{|\Omega|} sim(pp(\omega_i), pp(\omega_j)) + \mathcal{N}(|\omega_i|, \mu_\omega, \sigma_\omega) \right), \tag{2}$$

where sim represents the cosine similarity between two pitch profiles, $\mathcal{N}$ the unnormalized normal distribution, $|\omega_i|$ length of segment ω_i, and μ_ω and σ_ω the average and standard deviation of lengths of all segments. The segment thus obtained has a pitch profile that is similar to the profile of most other segments and an approximately average duration. This guarantees that we are not dealing with an outlier in pitch or time domain and that segment ω_r represents a good candidate for alignment.

2.5 Segment Alignment and Summarization

Since individual segments ω_i are not of same length due to tempo fluctuations during performance, we need to time-align all segments to the representative segment ω_r. Alignment is performed with dynamic time warping (DTW) over segment pairs represented by respective piano roll excerpts P_{ω_i}. We use correlation distance (one minus correlation coefficient) as our local distance measure, which has been shown to perform well for this task (as shown in [25, 26]). To be more robust to individual incorrectly performed notes and other pitch fluctuations, such as vibrato or pitch changes during onset and offset, we smooth the piano roll representation with a Gaussian filter over the frequency axis prior to DTW calculation. Calculating correlations between smoothed pitch estimates yields more robust similarity estimates in places where inaccuracies occur.

Alignment results in a series of optimal alignment paths between segment ω_r and all other segments ω_i: $\{\rho_i: i = 1\ldots|\Omega|\}$. Each path minimizes:

$$\min \sum_{(j,k)\in\rho_i} corr(g(\mathbf{p}_j), g(\mathbf{p}_k)), \mathbf{p}_j \in P_{\omega_i} \wedge \mathbf{p}_k \in P_{\omega_r}, \tag{3}$$

where $corr$ represents the correlation distance and g the Gaussian filter with kernel size 3.

The final result of alignment is a set of segments aligned in the time domain, as well as in the frequency domain due to pitch drift removal. The next step of the algorithm

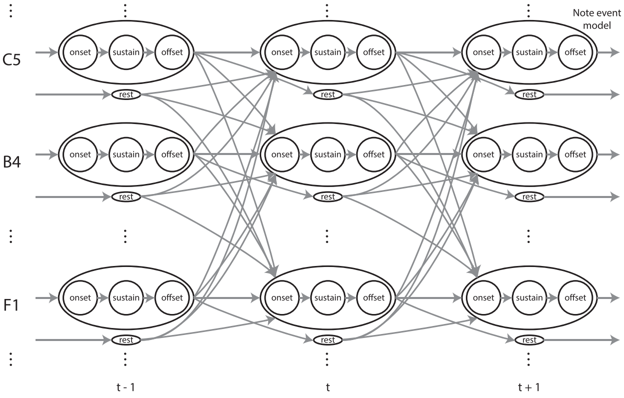

Fig. 2. Note estimation model.

summarizes F0 and salience information across all aligned segments to obtain a more robust estimate of pitches and their saliences in a segment. The rationale is that repetitions provide additional information on performance, and we are thus looking for an *average* performance, where pitches that are more often repeated will be more salient.

We define $\mathcal{F}^a$ and $\mathcal{S}^a$ as a concatenation of F0 estimates and their saliences of all segments ω_i, time aligned to ω_r according to ρ_i. We summarize F0 and salience pairs at each time t with a greedy approach, where in each iteration, F0 with the highest salience (f_t^{max}) is selected and then values across all segments are summarized as:

$$\mathcal{F}_t^{max} = [f_t^{max} - \eta, f_t^{max} + \eta] \quad (4)$$

$$s_t^c = \sum_i s_{it}^a \Big|_{s_{it}^a \hat{=} f_{it}^a \in \mathcal{F}_t^{max}} \quad (5)$$

$$f_t^c = \frac{1}{s_t^c} \sum_i f_{it}^a s_{it}^a \Big|_{s_{it}^a \hat{=} f_{it}^a \in \mathcal{F}_t^{max}} \quad (6)$$

Parameter η defines the range of F0 values around f_t^{max}, taken to represent the same fundamental frequency—due to imperfect performances, especially in vocal recordings, we must allow for some tolerance in summarization. We set η to value of 150 cents.

The summarized values f_t^c and s_t^c are added to collections of all summarized values $\mathcal{F}^c$ and $\mathcal{S}^c$, and all values used for their calculation are removed from $\mathcal{F}^a$ and $\mathcal{S}^a$. Summarization is repeated until all values from $\mathcal{F}^a$ have been exhausted.

The resulting sets of F0 and salience estimates $\mathcal{F}^c$, $\mathcal{S}^c$ represent a more robust frame-based transcription of an *average* performance of a segment in comparison to direct estimates of transcription algorithms, as segment repetitions contribute to more stable F0 estimates with less inaccuracies. Improvements are evaluated in Sec. 3.

2.6 Note Estimation

To segment the series of F0 estimates into notes, we use a probabilistic model similar to Ryynänen [6], who based his ideas on connected word models in speech recognition [27]. The model (outlined in Fig. 2) is based on three submodels: a note event model for modeling individual notes, a rest model for modeling rests, and a musicological model that ties both together and models transitions between notes. Transcription is obtained by repeatedly calculating the most likely sequence of states of the model over the entire signal, as described in Sec. 2.6.4.

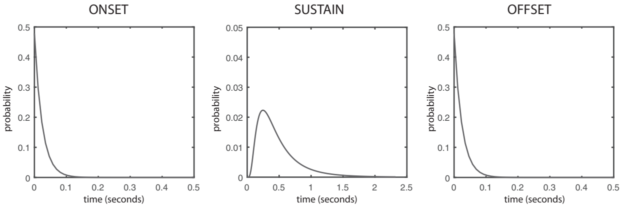

Fig. 3. Distributions used in modeling occupational probabilities for individual note model states (onset, sustain, and offset).

2.6.1 Note Event Model

Individual notes are modeled with a three state hidden Markov model (HMM), where states represent the onset, sustain, and offset of a note. Unlike Ryynänen, and similar to other recent models (e.g., [14]), we use explicit duration HMMs (EDHMMs) to model notes, as they offer a more flexible way of expressing state durations in comparison to regular HMMs. The entire model has one note model for each note we wish to transcribe—the number of models is dynamically set according to the range of F0s detected in the previous steps.

The parameters of EDHMMs are set as follows. The initial state is always the onset state and the final state is always the offset state. Transitions are trivial, as we only allow forward transitions and thus their probabilities are equal to 1 (EDHMMs do not allow for self transitions).

Occupational probabilities $d_j(u) = P(S_{t+u+1} \neq j, S_{t+2}^{t+u} = j | S_{t+1} = j, S_t \neq j)$ for onset and offset states are geometric distributions (as in regular HMMs), which tend to keep state occupancy short. For the sustain state, we use the log-normal probability distribution (as in [28]), which does not favor very short durations, however it has a longer tail, thus allowing for longer durations. Parameters of all distributions were set by hand according to musicological knowledge and are presented in Fig. 3.

Emission probabilities are based on three features: F0 distance from ideal note pitch, salience, and delta salience, which models note dynamics. Each feature is modeled with a separate probability distribution for each state, so in total nine distributions are defined, as shown in Fig. 4. Their parameters were estimated on a small validation set. F0 distance from the ideal note pitch is modeled with a normal distribution and has more tolerance in onset and offset states and less in the sustain state. This is especially true in vocal music, where pitch may fluctuate a lot during onset and offset of a note, while it remains relatively stable in the sustain state (although tolerance is still needed due to vibrato and inaccurate singing). Salience is usually lower at the onset, where pitch fluctuations are large and should

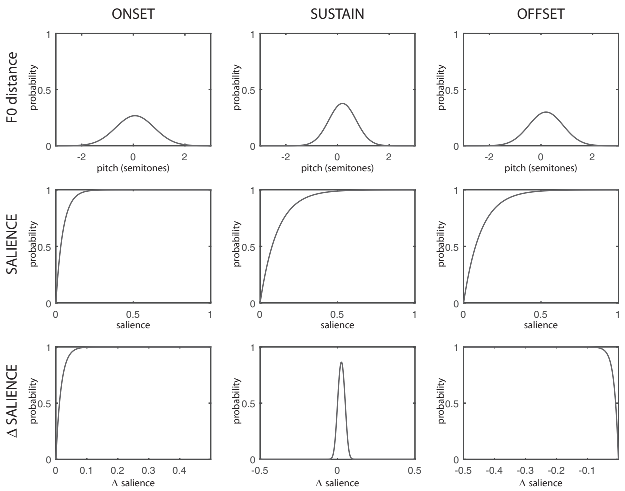

Fig. 4. Distributions used in modeling emission probabilities for individual note model states (onset, sustain, and offset) for each feature (F0 distance, salience, and Δ salience).

be relatively stable in the sustain and offset states, so we model it with an exponential distribution. Dynamics should be positive at the onset, negative at the offset, and fluctuate around zero in the sustain state, which is also reflected in the chosen distributions.

The final emission probability of each state is modeled by multiplying probabilities of individual features given the observed values.

2.6.2 Rest Model

The rest model is used for parts of the signal where note likelihood is low. Rests are modeled with a single HMM state, but unlike Ryynänen, we use one rest state per note model, as this gives us the flexibility of maintaining the melodic context over rests in a melodic line. The observation likelihood is defined as in [6] to be the negation of the greatest observation likelihood in any state of any note model; thus if observation likelihood of a note is high, the likelihood of a rest is low, and vice versa.

2.6.3 Musicological Model

The musicological model governs transition probabilities between note and rest models. We use an approach similar to [6], where transition probabilities are calculated on a corpus of folk song melodies.

The first step in applying the musicological model is to estimate the key of the song, which is needed to apply appropriate transition probabilities. We use a simple approach for key estimation: the transcribed F0s are summed and quantized into one octave range according to their salience (only pitch class is retained, octave information is ignored). The pitch profile thus obtained is correlated with the well-known Krumhansl-Kessler key profiles [29, 30] in all possible keys and the maximum correlation value taken as the song key.

Note transition probabilities given key are estimated from bigram probabilities on a corpus of folk song melodies [31] as described in [32]. As we use one rest state per note model, note-to-rest and rest-to-note probabilities are the same as the corresponding note-to-note probabilities. Rest-to-rest transitions are not allowed.

2.6.4 Finding Note Sequences

Transcription is performed by calculating the optimal path through the model given the observed F0s and saliences. We use a token-passing algorithm to calculate the optimal path, as presented by [27], modified to take EDHMM states into account. The modification is done by separately keeping track of the probability with which a token enters an EDHMM state, the duration the token spends in the state, and the cumulative observation probability of the token while in the state. Thus, in each step we can calculate the probability that the token stays in the state and create a new token leaving the state.

A single run of the token-passing algorithm will yield an optimal monophonic note sequence given the observed values. Since we are dealing with polyphonic music, we need to apply the algorithm iteratively, where after each iteration, the found notes are removed from observations and a new melodic line searched for in the next iteration. We can limit the number of iterations and thus maximal transcribed polyphony or let the algorithm run until all possibilities are exhausted and the model returns only the rest state as result.

3 EVALUATION

We evaluated the presented method on a collection of 37 Slovenian vocal polyphonic folk songs (107.7 minutes in total) from field recordings in the EthnoMuse archive [33]. The collection is available at http://lgm.fri.uni-lj.si/ciril/jaes-dataset/. Transcriptions of all songs were made manually by ethnomusicologists and were semi-automatically time-aligned to audio recordings. Songs were chosen so that they reflect the range of problems encountered in folk music field recordings, from inaccurate singing, pitch drifting, to poor recording quality. The average polyphony of the dataset is 2.3, one song contains vocals and instruments, others are vocal only.

The method has a small set of parameters (width of the smoothing kernel, summarization tolerance, parameters of EDHMM distributions), which were set to values estimated on a small validation set. We tested the sensitivity of the method to moderate changes of the parameters and concluded that sensitivity of the method to such changes is low. In case of small changes on the width of the smoothing kernel the method still returns the correct number of pitches per frame, while for much smaller width the method fails to blend corresponding pitch lines from different stanza repetitions into one, and for much bigger width the method joins the non-corresponding pitch lines. The same is true for changes in summarization tolerance. Small changes on the EDHMM distribution parameters also did not affect the results significantly.

Table 1. Evaluation results (mean values)

Method	P	R	F1
Sonic	0.39	0.50	0.43
Klapuri	0.26	0.59	0.36
Benetos 2015	0.21	0.47	0.29
proposed (Klapuri)	0.50	0.68	**0.58**
proposed (Benetos 2015)	0.51	0.55	**0.52**

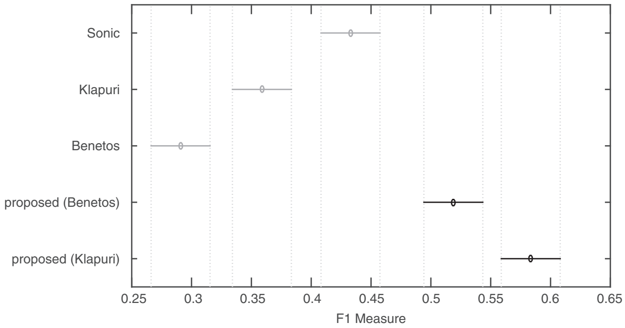

Fig. 5. Performance comparison of transcription methods according to F1 measure.

We tested the performance of three transcription methods on our dataset: Klapuri [5], Sonic [19], and Benetos-2015 [17]. Since Klapuri's and Benetos' approaches return per-frame F0s and saliences, we used both as underlying methods for our algorithm and evaluated the improvement obtained by using our method. Evaluation was performed frame-wise with up to half a semitone tolerance in pitch. Mean values of precision (P), recall (R), and F1 scores (F1) for all methods are presented in Table 1 and a multiple comparison analysis in Fig. 5. Precision represents the fraction of correctly found pitches over all found pitches, recall the fraction of correctly found pitches over all ground truth pitches, and F1 measure the harmonic mean of precision and recall values. Each evaluation was conducted on a single representative stanza selected from the ground truth, its transcription by the selected methods (Benetos-2015 and Klapuri) and its improved transcription by the presented method.

3.1 Performance of Transcription Algorithms

It is interesting to see that from the compared transcription approaches, Benetos's approach, which achieves good results in MIREX evaluations, performs the worst (F1 score 0.29). A possible explanation would be that the method relies on a factorization approach that assumes that the timbre of instruments is stable during performance, which is not true for vocal performances, and, thus, factorization into different instrument sources does not succeed. Klapuri's iterative algorithm performs somewhat better, although it seems that iterative subtraction removes too little energy from the signal and thus the algorithm finds too many notes, so precision is low. Surprisingly, Sonic, which is tuned to transcription of piano music, performs the best of the three. This could be attributed to the partial tracking approach with oscillator networks, which diminish the effect of timbre, so note recognition with neural networks still performs well.

3.2 Improvements with the Proposed Method

Results show that our approach significantly improves the results of both underlying methods (over 50% increase in F1 measure). Improvement is higher for precision (it almost doubles), although recall is also increased by approximately 15%. If we ignore octave errors, both approaches achieve the same F1 score of 0.59, meaning that Benetos's approach produces more octave errors. We must point out that our results cannot be directly compared to results achieved by the underlying methods only, as the latter do not take repetitions into account. The comparison is given so that the amount of improvement that may be achieved by our approach can be assessed.

Detailed analysis showed that improvement is achieved in both main parts of the algorithm: alignment and summarization of F0 and salience values, as well as note estimation. Contribution of both parts is quite balanced, summarization increases the F1 measure by 25%, and note estimation by an additional 25%. This shows that our method gathers a lot of additional information from repetitions but also achieves significant improvement with the note estimation model.

Repetitions provide additional information that we exploit to reduce the number of errors, produced by the underlying methods either due to noise, poor timbral modeling or imperfect singing. As the salience of F0 values increases with the number of segments they were found in, and simultaneously, F0s are adapted to an average value over all segments, false positives that occur due to noise are reduced, while false positives and negatives that occur due to imperfect singing are corrected. The note estimation model additionally reduces errors due to its musicological model. The latter is all the more successful as folk song melodies from our collection are usually simple, so the model, also estimated on a set of folk song melodies (different to ours), is very relevant and significantly improves results. In Fig. 6, which shows a part of song #2 from our dataset, we outline how the proposed method works.

The input multiple F0 estimations (black dots) and ground truth transcription (gray bars) are presented in Fig. 6(a), the summarized multiple F0 information (black dots) and ground truth (gray bars) in Fig. 6(b), and note estimations (white bars with black outline) and ground truth (gray bars) in Fig. 6(c). Ground truth represents the ethnomusicological transcription of a stanza. The figure also contains annotations of specific issues addressed by the presented method such as vibrato, missed tone, or inaccurate offset and onset times.

Analysis of songs where the underlying methods achieved the worst results showed that our method can improve transcription accuracy (F1 measure) from 0.18 (the worst case for Klapuri) to 0.45, or for Benetos-2015 from 0.15 to 0.64. Only in a single case our method returned worse results, where the difference was –0.01 and –0.04

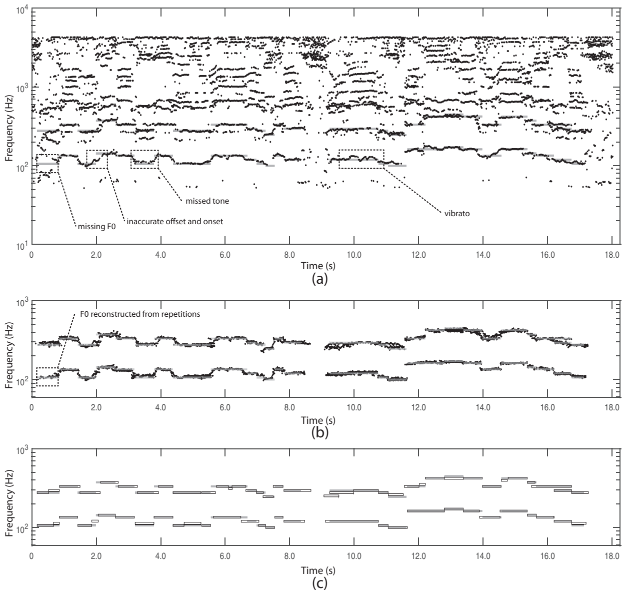

Fig. 6. The figure shows: (a) the comparison between input multiple F0 estimations by Benetos-2015 (dots) and ground truth (gray bars); and (b) summarization of information from multiple segments (dots) and ground truth (gray bars); and (c) final note estimations (white bars with black outline) and ground truth (gray bars). (a) represents a single segment, while in (b) and (c) the cumulative information from all repeated segments is gathered.

for Klapuri and Benetos-2015 respectively. In both of these cases precision is improved by 0.06 and 0.03 but recall decreased by 0.09 and 0.1 for Klapuri and Benetos-2015 respectively, decreasing the final F1 measure values.

4 CONCLUSIONS

The presented method shows that making use of song structure has potential for improving transcription of polyphonic, especially vocal folk music. We presented two novel contributions: (1) exploiting repetitions, typical of folk songs, aligned in time and pitch domain for improving F0 estimates and (2) a probabilistic model based on EDHMMs to estimate notes from F0 estimates. In evaluation, we show that both significantly improve transcription accuracy on a collection of vocal polyphonic folk music field recordings.

Future work will include the development of a new F0 estimation method adapted to transcribe vocal music, as most current methods do not yield satisfactory results. Also, current parameters of the note estimation method were either manually determined or estimated on a small validation set, so we plan to gather a larger collection that could be used to estimate the parameters. Detecting tempo could also improve the duration model and consequently note estimation. While Krumhansl-Kessler profiles follow the principles of Western tonality we are also planning on test other key finding models (e.g., [34]) as part of our future work. Finally, we plan to test and further develop the method also on non-vocal folk music recordings.

5 ACKNOWLEDGMENTS

This work could not been done without field recordings provided by the Institute of Ethnomusicology at Research Centre of Slovenian Academy of Sciences and Arts.

6 REFERENCES

[1] A. Klapuri and M. Davy *Signal Processing Methods for Music Transcription* (Springer US, New York, 2006). http://dx.doi.org/10.1007/0-387-32845-9.

[2] P. Grosche, B. Schuller, M. Müller, and G. Rigoll "Automatic Transcription of Recorded Music," *Acta Acustica united with Acustica*, vol. 98, no. 2, pp. 199–215 (2012). http://dx.doi.org/10.3813/AAA.918505.

[3] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri "Automatic Music Transcription: Challenges and Future Directions," *J. Intel. Inform. Sys.*, vol. 41, no. 3, pp. 407–434 (2013). http://dx.doi.org/10.1007/s10844-013-0258-3.

[4] M. Goto "A Predominant-f0 Estimation Method for Real-World Musical Audio Signals: MAP Estimation for Incorporating Prior Knowledge about f0s and Tone Models," *Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis*, Aalborg, Denmark (2001). http://dx.doi.org/10.1109/ICASSP.2001.940380.

[5] A. P. Klapuri "Multiple Fundamental Frequency Estimation Based on Harmonicity and Spectral Smoothness," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 6, (Nov. 2003). http://dx.doi.org/10.1109/TSA.2003.815516.

[6] M. P. Ryynänen and A. Klapuri "Polyphonic Music Transcription Using Note Event Modeling," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005*, pp. 319–322 (Oct. 2005). http://dx.doi.org/10.1109/ASPAA.2005.1540233.

[7] A. Cont "Realtime Multiple Pitch Observation Using Sparse Non-Negative Constraints," *Proceedings of the 7th International Conference on Music Information Retrieval*, Victoria (BC), Canada (Oct. 2006).

[8] B. Niedermayer "Non-Negative Matrix Division for the Automatic Transcription of Polyphonic Music," *Proceedings of the 9th International Conference on Music Information Retrieval*, pp. 544–549, Philadelphia, PA, USA (Sep. 2008).

[9] A. Dessein, A. Cont, and G. Lemaire "Real-Time Polyphonic Music Transcription with Non-Negative Matrix Factorization and Beta-Divergence," *Proceedings of the 11th International Society for Music Information Retrieval Conference*, pp. 489–494, Utrecht, The Netherlands (Aug. 2010).

[10] S. A. Raczyński, N. Ono, and S. Sagayama "Multipitch Analysis with Harmonic Nonnegative Matrix Approximation," *Proceedings of the 8th International Conference on Music Information Retrieval*, pp. 381–386, Vienna, Austria (Sep. 2007).

[11] G. Grindlay and D. P. W. Ellis "Probabilistic Subspace Model for Multi-Instrument Polyphonic Transcription," *Proceedings of the 11th International Society for*

Music Information Retrieval Conference, pp. 21–26, Utrecht, The Netherlands (Aug. 2010).

[12] T. Cheng, S. Dixon, and M. Mauch "Deterministic Annealing EM Algorithm for Automatic Music Transcription," *Proceedings of the 14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil (Nov. 2013).

[13] B. Fuentes, R. Badeau, and G. Richard "Harmonic Adaptive Latent Component Analysis of Audio and Application to Music Transcription," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1854–1866 (Sep. 2013). http://dx.doi.org/10.1109/TASL.2013.2260741.

[14] E. Benetos and T. Weyde "Explicit Duration Hidden Markov Models for Multiple-Instrument Polyphonic Music Transcription," *Proceedings of the 14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil (Nov. 2013). http://dx.doi.org/10.13140/2.1.3459.5209.

[15] E. Benetos and S. Dixon "Multiple-Instrument Polyphonic Music Transcription Using a Temporally Constrained Shift-Invariant Model," *J. Acous. Soc. Amer. (JASA)*, vol. 133, no. 3, pp. 1727–1741 (2013). http://dx.doi.org/10.1121/1.4790351.

[16] E. Benetos, S. Ewert, and T. Weyde "Automatic Transcription of Pitched and Unpitched Sounds from Polyphonic Music," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2014* (May 2014). http://dx.doi.org/10.1109/ICASSP.2014.6854172.

[17] E. Benetos and T. Weyde "An Efficient Temporally-Constrained Probabilistic Model for Multiple-Instrument Music Transcription," *Proceedings of 16th International Society for Music Information Retrieval Conference*, pp. 701–707, Malaga, Spain (Oct. 2015).

[18] J. S. Downie "The Music Information Retrieval Evaluation eXchange (2005–2007): A Window Into Music Information Retrieval Research," *Acous. Sci. Tech.*, vol. 29, no. 4, pp. 247–255 (2008). http://dx.doi.org/10.1250/ast.29.247.

[19] M. Marolt "A Connectionist Approach to Automatic Transcription of Polyphonic Piano Music," *IEEE Transactions on Multimedia*, vol. 6, no. 3, pp. 439–449 (Aug. 2004). http://dx.doi.org/10.1109/TMM.2004.827507.

[20] T. Berg-Kirkpatrick, J. Andreas, and D. Klein "Unsupervised Transcription of Piano Music," *Advances in Neural Information Processing Systems 27*, pp. 1538–1546 (2014).

[21] M. Marolt "Automatic Transcription of Bell Chiming Recordings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 844–853 (Mar. 2012). http://dx.doi.org/10.1109/TASL.2011.2166957.

[22] G. Tzanetakis "Computational Ethnomusicology: A Music Information Retrieval Perspective," *Music Technology Meets Philosophy—From Digital Echos to Virtual Ethos: Joint Proceedings of the 40th International Computer Music Conference, ICMC 2014, and the 11th Sound and Music Computing Conference SMC 2014*, Athens, Greece (Sep. 2014).

[23] E. Gómez, P. Herrera, and F. Gomez-Martin "Computational Ethnomusicology: Perspectives and Challenges," *J. New Music Res.*, vol. 42, no. 2, pp. 111–112 (2013). http://dx.doi.org/10.1080/09298215.2013.818038.

[24] M.-P. Baumann, J. P. J. Stock, and M. Marian-Bălaşa *Notation, Transcription, Visual Representation* (World of music. VWB - Verlag für Wissenschaft und Bildung, 2005).

[25] C. Bohak and M. Marolt "Probabilistic Segmentation of Folk Music Recordings," *Mathematical Problems in Engineering*, Article ID 8297987 (2016). http://dx.doi.org/10.1155/2016/8297987.

[26] J. Serrà, E. Gómez, P. Herrera, and X. Serra "Chroma Binary Similarity and Local Alignment Applied to Cover Song Identification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, pp. 1138–1151 (Aug. 2008). http://dx.doi.org/10.1109/TASL.2008.924595.

[27] S. J. Young, N. H. Russell, and J. H. S. Thornton "Token Passing: A Simple Conceptual Model for Connected Speech Recognition Systems," Technical report, Cambridge University Engineering Department (1989).

[28] H. Takeda, T. Nishimoto, and S. Sagayama "Rhythm and Tempo Analysis Toward Automatic Music Transcription," *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP) 2007*, vol. 4, pp. IV–1317–IV–1320 (Apr. 2007). http://dx.doi.org/10.1109/ICASSP.2007.367320.

[29] C. L. Krumhansl and E. J. Kessler "Tracing the Dynamic Changes in Perceived Tonal Organization in a Spatial Representation of Musical Keys," *Psychological Rev.*, vol. 89, no. 4, pp. 334–368 (1982). http://dx.doi.org/10.1037/0033-295X.89.4.334.

[30] C. L. Krumvansl *Cognitive Foundations of Musical Pitch* (Oxford University Press, 1990). http://dx.doi.org/10.1093/acprof:oso/9780195148367.001.0001.

[31] E. Dahlia "EsAC Database: Essen Associative Code and Folksong Database," available at www.esacdata.org (1994).

[32] M. P. Ryynänen and A. P. Klapuri "Modelling of Note Events for Singing Transcription," *Proceedings of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio*, pp. 6 (MIT Press, 2004).

[33] G. Strle and M. Marolt "Conceptualizing the Ethnomuse: Application of CIDOC CRM and FRBR," *Proceedings of CIDOC2007* (2007).

[34] D. Temperley and E. W. Marvin "Pitch-Class Distribution and the Identification of Key," *Music Perception: An Interdisciplinary J.*, vol. 25, no. 3, pp. 193–212 (2008). http://dx.doi.org/10.1525/mp.2008.25.3.193.

THE AUTHORS

Ciril Bohak

Matija Marolt

Ciril Bohak, Ph.D., is a teaching assistant at the Faculty of Computer and Information Science, University of Ljubljana, where he has been working since 2007. He is member of Laboratory for Computer Graphics and Multimedia. He defended his Ph.D. thesis in 2016. His research includes music information retrieval, computer graphics, e-learning, game technology, and human-computer interaction.

Matija Marolt, Ph.D., is an assistant professor with Faculty of Computer and Information Science, University of Ljubljana, where he has been working since 1995. He is head of Laboratory for Computer Graphics and Multimedia, where his research interests include music information retrieval, specifically semantic description and understanding of audio signals, retrieval and organization in music archives, and human-computer interaction.