METRICAL-ACCENT AWARE VOCAL ONSET DETECTION IN POLYPHONIC AUDIO

First Author
Affiliation1
author1@ismir.edu

Second Author
Retain these fake authors in submission to preserve the formatting

Third Author
Affiliation3
author3@ismir.edu

ABSTRACT

The goal of this study is the automatic detection of onsets of the singing voice in polyphonic audio recordings. Starting with a hypothesis that the knowledge of the current position in a metrical cycle (i.e. metrical accent) can improve the accuracy of vocal note onset detection, we propose a novel probabilistic model to jointly track beats and vocal note onsets. The proposed model extends a state of the art model for beat and meter tracking, in which a-priori probability of a note at a specific metrical accent interacts with the probability of observing a vocal note onset. We carry out an evaluation on a varied collection of multi-instrument datasets from two music traditions (English popular music and Turkish makam) with different types of metrical cycles and singing styles. Results confirm that the proposed model reasonably improves vocal note onset detection accuracy compared to a baseline model that does not take metrical position into account.

1. INTRODUCTION

Singing voice analysis is one of the most important topics in the field of music information retrieval because singing voice often forms the melody line and creates the impression of a musical piece. The automatic transcription of singing voice can be considered to be a key technology in computational studies of singing voice. It can be utilized for end-user applications such as enriched music listening and singing education. It can as well enable other computational tasks including singing voice separation, karaoke-like singing voice suppression or lyrics-to-audio alignment [3].

The process of converting an audio recording into some form of musical notation is commonly known as automatic music transcription. Current transcription methods use general purpose models, which are unable to capture the rich diversity found in music signals [2]. In particular, singing voice poses a challenge to transcription algorithms because of its high degree of expressive elements such as soft onsets, portamento and vibrato. One of the core subtasks of singing voice transcription (SVT) is detecting note events with a discrete pitch value, an onset time and an offset time from the estimated time-pitch representation. Detecting the times of vocal note onsets can benefit from automatically detected events from complementary musical facets, such as musical meter. In fact, the accents in a metrical cycle determine to a large extent the temporal backbone of singing melody lines. Studies on symbolic music data showed that the timestamps where vocal note onsets occur are influenced by the their position in a metrical cycle [4, 6].

In this work we propose a novel probabilistic model that tracks simultaneously note onsets of singing voice and instumental energy accents in a metrical cycle. We extend a state of the art model for beat and meter tracking, based on dynamic Bayesian networks (DBN). A model variable is added that models the temporal segments of a note and their interaction with metrical position. The proposed model is applied for the automatic detection of vocal note onsets in multi-instrumental recordings with predominant singing voice. Vocal melody contours are input from an external predominant voice estimation algorithm. Evaluation is carried out on datasets from music traditions, for which there is a clear correlation between metrical accents and the onset times on the vocal line.

2. RELATED WORK

2.1 Singing voice transcription

In the case of monophonic singing the more successful approaches are based on discontinuities in pitch contours [13]. The idea is usually referred to as 'island building' groups candidate onsets in time and has been the basis for other works. A probabilistic note HMM is presented in [17], where a note has 3 states: attack (onset), stable pitch state and silent state. The transition probabilities are learned from data. Later this approach has been extended by adding a separate HMM for regions of transition between the steady notes [10]. Recently [12] suggested to compact the musical knowledge into rules as a way to describe the observation and transition likelihoods, instead of learning them from data. The authors cover a range with distinct pitch from lowest MIDI C2 up to B7. Each MIDI pitch is further divided into 3 sub-pitches, resulting in $n = 207$ notes with different pitch, each having the 3 note states. Although being conceptually capable of track-

© First Author, Second Author, Third Author. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** First Author, Second Author, Third Author. "Metrical-accent aware vocal onset detection in polyphonic audio", 18th International Society for Music Information Retrieval Conference, Suzhou, China, 2017.

ing onsets in singing voice audio with accompaniments, these approaches were tested only on a cappella singing.

In multi-instrumental recordings, an essential first step is to extract reliably the predominant vocal melody. One of the few works dealing with SVT for polyphonic recordings [11, 14] rely on the algorithm of for predominant melody extraction [18]. Time deviations of sung vocal onsets from the onsets indicated in musical score are modeled in a probabilistic way in [14].

2.2 Beat Detection

Recently a Bayesian approach, referred to as the *bar-pointer* model, has been presented [19]. It describes events in music as being driven by their current position in a metrical cycle (i.e. musical bar). The model represents as hidden variables in a hidden Markov model (HMM) the current position in a bar, the tempo, and the type of musical meter.

The work of [5] applied this model to recordings from non-Western music, in order handle jointly beat and downbeat tracking. The authors showed that the original model can be adapted to different rhythmic styles and time signatures, and an evaluation is presented on Indian, Cretan and Turkish music datasets.

A modification of the bar-tempo state used in this work that optimizes its size, was later suggested by [9].

3. DATASETS

3.1 Turkish makam

The Turkish dataset is a subset of the dataset, presented in [5], including only the recordings with singing voice present. It is divided into training and test dataset. The test dataset comprises 5 1-minute excerpts from recordings with solo singing voice for each of two meter classes, referred to as usuls in Turkish makam: the 9/8-usul aksak and the 8/8-usul düyek. All excerpts are manually annotated with beats, downbeats and vocal note onsets. Interestingly, each usul has a characteristic pattern of beat positions, on which percussive strokes are hit. For example, in aksak the beats 1,3,4,5,7 and 9 have strokes. Percussionists of Turkish Makam tend to observe these patterns rather conservatively.

The training set spans around 7 minutes of audio from each of the two usuls, annotated also manually with beats and downbeats. Due to the scarcity of material with solo singing voice, several excerpts with choir sections were included.

3.2 English pop

The datasets, on which singing voice transcription in popular polyphonic music is evaluated are one or two [2]. To overcome this bias, we compiled a new one: We selected 10 30-second clips of English pop songs, which have been aligned to their corresponding MIDIs in a recent study [16]. The audio recordings are 30-second thumbnails of the Million Song Dataset [1] . All pop songs are in 4/4 meter. Criteria for selecting the songs are the presence of drums accompaniment and predominance of the vocal line. We generated the singing voice transcription from the vocal MIDI channel and the corresponding aligned timestamps in the audio recording. We added annotations of vocal segments. To encourage further studies on singing voice transcription we make available the derived annotations [2] .

4. APPROACH

The proposed approach extends the beat and meter tracking model, presented in [5]. We adopt from that model the variables for the position in a metircal cycle (bar position) ϕ, the instantaneous tempo $\dot{\phi}$ and the rhythmic pattern r, related to the metrical cycle type. We also adopt the observation model, which describes how the metrical accents (beats) are related to an observed onset feature vector y_f. All variables and their conditional dependencies are represented as the hidden variables in a DBN (see Figure 1).

In this paper we study how the *a priori* probability of a note at a specific metrical accent interacts with the probability of observing a vocal note onset. To represent that interaction we add a hidden state for vocal note *n*, which depends on the current position in the metrical cycle. The probability of observing a vocal onset is derived from the emitted pitch y_p of the vocal melody.

In a DBN, an observed sequence of features derived from an audio signal $y_{1:K} = \{y, .., y_K\}$ is generated by a sequence of hidden (unknown) variables $x_{1:K} = \{x_1, ..., x_K\}$, where K is the length of the sequence (number of audio frames in an audio excerpt). The joint probability distribution of hidden and observed variables factorizes as:

$$P(x_{1:K}, y_{1:K}) = P(x_0)\Pi_{k=1}^{K} P(x_k|x_{k-1})P(y_k|x_k) \quad (1)$$

where $P(x_0)$ is the initial state distribution; $P(x_k|x_{k-1})$ is the transition model and $P(y_k|x_k)$ is the observation model.

4.1 Hidden variables

At each audio frame k, the hidden variables describe the state of a hypothetical bar pointer $x_k = [\dot{\phi}_k, \phi_k, n_k, r_k]$, representing the instantaneous tempo, the bar position, the note state, and a rhythmic pattern indicator, respectively.

4.1.1 Tempo state $\dot{\phi}$ and bar position state ϕ

The bar position ϕ points to the current position in the metrical cycle (bar). The instantaneous tempo $\dot{\phi}$ encodes how much bar positions the pointer advances from the current to the next time instant. To assure feasible computational time we relied on the combined bar-tempo efficient state space, presented in [9]. To keep the size of the bar-tempo

[1] http://colinraffel.com/projects/lmd/

[2] suppressed for anonymity

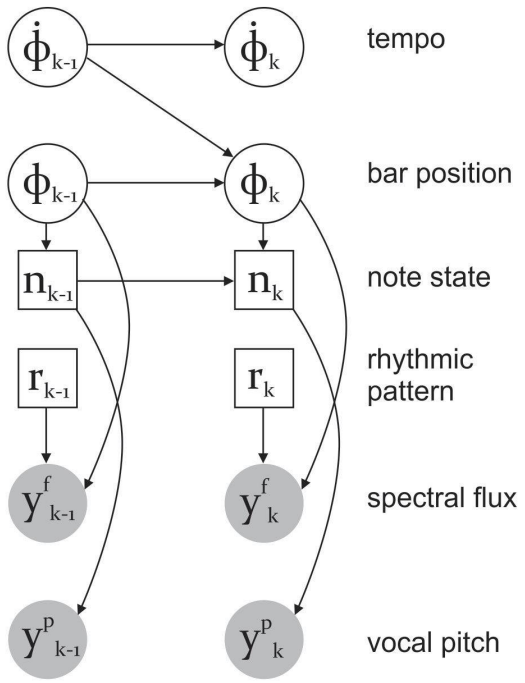

Figure 1: A dynamic Bayesian network for the proposed beat and vocal onset detection model. Circles and squares denote continuous and discrete variables, respectively. Gray nodes and white nodes represent observed and hidden variables, respectively.

state space small, we input the ground truth tempo for each recording, allowing $\dot{\phi}$ to deviate within ± 10 bpm from it. Another motivation to limit the tempo in such a way is avoiding possible octave errors in the beat tracking, which would not be desirable for beat-aware note onset detection. This yields around 100-1000 states for the bar positions within a single beat (around 10K for usuls and around 5K for 4/4) .

4.1.2 Note state n_k

The note states represent the temporal segments of a sung note. They are a modified version of these suggested in the note transcription model of [12]. We adopted the first two segments: attack region (A), stable pitch region (S). We replaced the silent segment with non-vocal state (N). Because full-fledged note transcription is outside the scope of this work, instead of 3 steps per semitone, we used for simplicity only a single one, which deteriorated just slightly the note onset detection accuracy. Also, to reflect the pitch range in the datasets, on which we evaluate, we set as minimal MIDI note E3 covering almost 3 octaves up to B5 (35 semitones). This totals to 105 note states.

4.1.3 Rhythmic pattern r_k

A rhythmic patterns indicates the pattern of accents in a metrical cycle. For simplicity we use only one rhythmic pattern for each metric cycle. Let also $\theta(r)$ denote the number of beats in a rhythmic pattern r. Since the metrical type for each recording from the dataset is known a priori, a hidden state for the rhythm pattern is not modeled explicitly.

To be able to represent the DBN as a hidden Markov model, the bar-tempo efficient state space is combined with the note state space into a joint state space x. The joint state space is a cartesian product of the two state spaces, resulting in up to 10K x 105 = 1M states.

4.2 Transition model

Due to the conditional dependence relations in Figure 1 the transitional model factorizes as

$$\begin{array}{ll} P(x_k|x_{k-1}) = & P(\dot{\phi}_k|\dot{\phi}_{k-1}) \times \\ P(\phi_k|\phi_{k-1},\dot{\phi}_{k-1}) \times & P(n_k|n_{k-1},\phi_k) \end{array} \quad (2)$$

The tempo transition probability $p(\dot{\phi}_k|\dot{\phi}_{k-1})$ and bar position probability $p(\phi_k|\phi_{k-1},\dot{\phi}_{k-1})$ are the same as in [5]. Transition from one tempo to another is allowed only at bar positions, at which the beat changes. This is a reasonable assumption for the local tempo deviations in the analyzed datasets, which can be considered to occur relatively beat-wise.

4.2.1 Note transition probability

The probability of advancing to a next note state is based on the transitions of the note-HMM, introduced in [12]. Let use briefly review it: From a given note segment the only possibility is to progress to its following note segment. To ensure continuity each of the self-transition probabilities is rather high, given by constants c_A, c_S and c_N for A, S and N segments respectively (c_A=0.9; c_S=0.99; $c_N = 0.9999$). Let $P_{N_iA_j}$ be the probability of transition from non-vocal state N_i after note i to attack state A_j of its following note j. The authors assume that it depends on the difference between the pitch values of notes i and j and the diffrence can be approximated by a normal distribution centered at change of zero ([12], Figure 1.b). This implies that small pitch changes are more likely than larger ones. Now we can formalize their note transition as:

$$p(n_k|n_{k-1}) = \begin{cases} P_{N_iA_j}, & n_{k-1} = N_i \quad n_k = A_j \\ c_N, & n_{k-1} = n_k = N_i \\ 1 - c_A, & n_{k-1=}A_i \quad n_k = S_j \\ c_A, & n_{k-1} = n_k = A_i \\ 1 - c_S & n_{k=1} = S_i \quad n_k = N_j \\ c_S, & n_{k-1} = n_k = S_i \\ 0 & else \end{cases} \quad (3)$$

Note that the outbound transitions from all non-vocal states N_i should sum to 1, meaning that

$$c_N = 1 - \sum_i P_{N_iA_j} \quad (4)$$

In this study, we modify $P_{N_iA_j}$ to allow varation in time, depending on the current bar position ϕ_k.

$$p(n_k|n_{k-1},\phi_k) = \begin{cases} P_{N_iA_j}\Theta(\phi_k), & n_{k-1} = N_i, n_k = A_j \\ c_N, & n_{k-1} = n_k = N_i \\ ... \end{cases} \quad (5)$$

where

$\Theta(\phi_k)$: function weighting the contribution of a beat adjacent to current bar position ϕ_k

and

$$c_N = 1 - \Theta(\phi_k) \sum_i P_{N_iA_j} \quad (6)$$

The transition probabilities in all the rest of the cases remain the same. We explore two variants of the weighting function $\Theta(\phi_k)$:

4.2.1.1 Time-window redistribution

In performance singers often advance or delay slightly note onsets in relation to the beats. The work [14] presented an idea of how to handle vocal onsets, time-shifted from a beat, by stochastic distribution. Similarly, we introduce a normal distribution $\mathcal{N}_{0,\sigma}$, centered around 0 to re-distribute the importance of beats over a time window around a beat. Let b_k be the beat, closest in time to a current bar position ϕ_k. Now:

$$\Theta(\phi_k) = [\mathcal{N}_{0,\sigma}(d(\phi_k, b_k))]^w e(b_k) \quad (7)$$

where

$e(b)$: probability of a note onset co-occurring with the b^{th} beat in the metrical cycle (b $\in\theta(r)$)

w : sensitivity of vocal onset probability to beats

$d(\phi_k, b_k)$: the distance from current bar position ϕ_k to closest beat position b_k

Equation 5 means essentially that the original $P_{N_iA_j}$ is scaled accordingly to how close in time to a beat it is.

4.2.1.2 Simple weighting

The transition probability $P_{N_iA_j}$ is modified only at beat positions, i.e. the weighting function is set to the peak of $N_{0,\sigma}$ only at bar positions corresponding to beat positions, and to 1 elsewhere.

$$\Theta(\phi_k) = \begin{cases} [N_{0,\sigma}(0)]^w e(b_k), & d(\phi_k, b_k) = 0 \\ 1 & else \end{cases} \quad (8)$$

4.3 Observation models

The observation probability $P(y_k|x_k)$ describes the relation between the hidden states and the (observed) audio signal. In this work we make the assumption that the observed vocal pitch and the observed metrical accent are conditionally independent from each other. This assumption may not hold in cases when energy accents of singing voice, which contribute to the total energy of the signal, are correlated to changes in pitch. However, for music with percussive instruments the importance of singing voice accents is diminished to a significant extent by percussive accents. Now we can rewrite Eq. 1 as

$$P(x_{1:K}, y^f_{1:K}, y^p_{1:K}) = P(x_0)\Pi_{k=1}^K P(x_k|x_{k-1})P(y^f_k|x_k)P(y^p_k|x_k) \quad (9)$$

This means essentially that the observation probability can be represented as the product of the observation probability of a metrical accent $P(y^f_k|x_k)$ and the observation probability of vocal pitch $P(y^p_k|x_k)$.

4.3.1 Accent observation model

In this paper for $P(y^f_k|x_k)$ we train GMMs on the spectral flux-like feature y^f, extracted from the audio signal using the same parameters as in [8] and [5]. The feature y^f summarizes the energy changes (accents) that are likely to be related to the onsets of all instruments together. The probability of observing an energy change depends on the position in the bar and the rhythmic pattern, $P(y^f_k|x_k) = P(y^f_k|\phi_k, r_k)$

4.3.2 Pitch observation model

The pitch probability $P(y^p_k|x_k)$ reduces to $P(y^p_k|n_k)$, because it depends only the current note state. We adopt the idea proposed in [12] that a vocal note state emits pitch y^p according to a normal distribution, centered around its average pitch. The standard deviation of stable states and the one of the onset states are kept the same as in the original model, respectively 0.9 and 5 semitones. The melody contour of singing is extracted in a preprocessing step. We utilized for the English pop a method for predominant melody extraction [18]. For the Turkish makam dataset, we instead utilized an algorithm, extended from [18] and tailored to Turkish makam. In both algorithms, each audio frame k gets assigned a pitch value and probability of being voiced v_k [1]. Based on frames with zero probabilities, one can infer which segments are vocal and which not. Since correct vocal segments is crucial for the sake of this study and the voicing estimation of these melody extraction algorithms are not state of the art, we preferred to rely on manual vocal annotations and thus assigned $v_k = 0$ for all frames, annotated as non-vocal.

For each state the observation probability $P(y^p_k|n_k)$ of vocal states is normalized to sum to v_k (unlike the original model which sums to a global constant v). This leaves the probability for each non-vocal state be $^{1-v_k}/_n$.

4.4 Learning model parameters

4.4.1 Accent observation model

We trained the accent probability patterns $P(y^f_k|\phi_k, r_k)$ for Turkish makam on the training subset of the dataset (see section 3.1). For each usul we trained one rhythmic pattern by fitting a 2-mixture GMM on the spectral-flux-like feature vector y^f. Analogously to [5] we pooled the bar

positions down to 16 patterns per beat. For English pop we used the 4/4 pattern trained by [8] on ballroom dances. The feature vector is normalized to zero mean, unit variance and taking moving average. Normalization is done per song.

4.4.2 Probability of note onset

The probability of a vocal note onset co-occurring at a given bar position $e(b)$ is obtained from studies on sheet music. Many notes are aligned with a beat in the music score, meaning a higher probability of a note at beats compared to inter-beat bar positions. A separate distribution $e(b)$ is applied for each different metrical cycle. For the Turkish usuls $e(b)$ has been inferred from a recent study [4, Figure 5. a-c]. The authors used a corpus of music scores, on data from the same corpus, from which we derived the Turkish dataset. The patterns reveal that notes are expected to be located with much higher likelihoods on those beats with percussive strokes than on the rest.

In comparison to a classical tradition like makam, in modern pop music the most likely positions of vocal accents in a bar are arguably much more heterogeneous, due to the big diversity of time-deviations from one singing style to another [6]. Because we do not dispose of a distribution pattern $e(b)$ for English pop, we set it manually with probability of onsets of 0.8 at beats 1 and 3 and 0.6 on beats 2 and 4.

4.5 Inference

4.5.1 With manually annotated beats

We explored the option that beats are given as input from a preprocessing step (i.e. when they are manually annotated). In this case, the detection of vocal onsets can be carried out by a reduced model with a single hidden variable: the note state. The observation model is then reduced to the pitch observation probability. The transition model is reduced to bar-position aware transition probability $a_{ij}(k) = p(n_k = j|n_{k-1} = i, \phi_k)$(see Eq. 5). To represent this time-dependent self-transition probabilities we we utilize time-varying transition matrix. It falls in the general category of variable-time HMMs (VTHMMs) [7]. The standard transition probabilities in the Viterbi maximization step are substituted for the bar-position aware transitions $a_{ij}(k)$

$$\delta_k(j) = \max_{i \in (j,\, j-1)} \delta_{k-1}(i)\, a_{ij}(k)\, b_j(O_k) \qquad (10)$$

Here $b_j(O_k)$ is the observation probability for state i for feature vector O_k and $\delta_k(j)$ is the probability for the path with highest probability ending in state j at time k (complying with the notation of [15, III. B]

4.5.2 Full model

We obtain the most optimal state sequence $x_{1:K}$ by decoding with the well-known Viterbi algorithm. A Beat is detected when the bar position variable hits one of $\theta(r)$ positions of beats within the metrical cycle. A vocal note onset is detected when the state path enters an attack note state after being in non-vocal state.

Note that the size of the state space poses a memory requirement. A recording of 1 minute has around 10K frames at a hopsize of $5.8\,ms$. To use Viterbi thus requires to store in memory pointers to up to 4G states, which amounts to 40G RAM (with uint32 python data type).

5. EXPERIMENTS

The hopsize of computing the spectral flux feature, which resulted in most optimal beat detection accuracy in [5] is $h_f = 20\,ms$. In comparison, the hopsize of predominant vocal melody detection is usually of smaller order i.e. $h_p = 5.8\,ms$ (correspondging to 256 frames at sampling rate of 44100). Preliminary experiments showed that extracting pitch with values of h_p bigger than this values reasonably deteriorated the vocal onset accuracy. Therefore in this work we used hopsize of 5.8 ms for the extraction of both features. The time difference parameter for the spectral flux computation remains unaffected by this change in hopsize, because it can be set separately.

As a baseline we run the algorithm of [12] with the 105 note states, we introduced in Section 4.1.2 [3]. The note transition probability is the original as presented in Eq. 3, i.e. not aware of beats. Note that in [12] the authors introduce a post-processing step, in which onsets of consecutive sung notes with same pitch are detected considering their intensity difference. We excluded this step in all system variants presented, because it could not be integrated in the proposed observation model in a trivial way. This means that, essentially, in this paper cases of consecutive same-pitch notes are missed, which decreases somewhat the recall compared to the original algorithm.

5.1 Evaluation metrics

5.1.1 Beat detection

Since improvement of the beat detector is outside the scope of this study, we report accuracy of detected beats only in terms of their f-measure. This serves solely as reference to existing work [4]. The f-measure can take a maximum value of 1, while beats tapped on the off-beat relative to annotations will be assigned an f-measure of 0. We used the default tolerance window of $70\,ms$, also applied in [5].

5.1.2 Vocal onset detection

We measured vocal onset accuracy in terms of precision and recall. Unlike a cappella singing, the exact onset times of singing voice accompanied by instruments, might be much more ambiguous. To accommodate this fact, we adopted the tolerance of $t = 50\,ms$, used for vocal onsets in accompanied flamenco singing by [11]. Note transcription accuracy remains outside the scope of this study.

[3] We ported the original VAMP plugin implementation to python, which is available at suppressed for anonymity

[4] Note that f-measure is agnostic to the phase of the detected beats, which is clearly not optimal.

meter		beat Fmeas	P	R	Fmeas
düyek	Mauch	-	33.1	31.6	31.6
	Ex-1	-	40.4	39.5	39.0
	Ex-2	86.4	37.8	36.1	36.1
aksak	Mauch	-	42.1	36.9	37.9
	Ex-1	-	48.4	39.1	43.0
	Ex-2	72.9	45.0	39.0	40.3

meter		beat Fmeas	P	R	Fmeas
4/4	Mauch	-	23.5	61.2	32.4
	Ex-1	-	29.5	56.2	36.4
	Ex-2	94.2	26.6	58.4	33.4
total	Mauch	-	32.8	43.1	35.2
	Ex-1	-	38.3	45.3	37.9
	Ex-2	84.3	37.3	45.4	37.2

Table 1: Evaluation results for Erperiment 1 (shown as Ex-1) and Experiment 2 (shown as Ex-2). Mauch stands for the baseline, following the approach of [12]. P, R and Fmeas denote the precision, recall and f-measure of detected vocal onsets. Results are averaged per meter type.

5.2 Experiment 1: With manually annotated beats

As a precursor to evaluating the full-fledged model, we conducted an experiment with manually annotated beats. This is done to test the general feasibility of the proposed note transition model (presented in 4.2.1), unbiased from errors in the beat detection.

We did apply both the simple and the time-redistribution weighting schemes for $\Theta(\phi_k)$, presented respectively in Eq. 8 and in Eq. 7. In preliminary experiments we saw that with annotated beats the simple weighting results in much worse onset accuracy than the time-redistributed one. Therefore the experimental results reported are conducted with the latter weighting scheme.

We have tested different pairs of values for w and σ from Eq. 5. For Turkish makam the onset detection accuracy peaked at w=1.2 and $\sigma = 30\,ms$, whereas for the English pop optimal are w=1.1 and $\sigma = 45\,ms$. Table 1 presents metrics compared to the baseline. Inspection of detections showed that the proposed model added some onsets around beats, which are missed by the baseline.

5.3 Experiment 2: Full model

To assure computational efficiency, we did an efficient implemenation of the joint state space [5]. The average f-measure of detected beats for the different metrical cycles can be seen in Table 1 [6]. The beat tracking accuracy for the Turkish usuls is on par with the results reported in [5, Table 1.a-c, R=1]. The results reported are only with the simple weighting scheme for the vocal note onset transition model. Table 1 shows a reasonable improvement of vocal onset detection accuracy for both music traditions.

For simple weigthing, adding the automatic beat tracking results in improvement over the baseline, whereas this was not the case with manual beats. This suggests that the concurrent tracking of beats and vocal onsets is a flexible strategy and can accommodate some vocal onsets, sligthly time-shifted from a beat. We observed also that the vocal onset accuracy is on average almost the same as that with manual beat annotations (done with the time-redistribution weighting). Despite the higheset beat detection accuracy for 4/4, the contribution of automatic beat tracking is the least. One reason for that may be that the pattern note probability pattern $e(b)$ used for 4/4/ is not well representative for the singing style differernces.

Due to time constraints we did not test the time-redistribution weighting in this experiment.

[5] We extended the python toolbox for beat tracking https://github.com/CPJKU/madmom/, which we make available at suppressed for anonymity

[6] per-recoding results can be found in suppressed for anonymity

6. CONCLUSIONS

In this paper we presented a Bayesian approach for the simultaneous tracking of beats and vocal onsets of singing voice in polyphonic music recordings. The main contribution is that the approach integrates in one coherent model two existing probabilistic approaches for different tasks: beat tracking and note transcription. Results confirm that the knowledge of the current position in a metrical cycle can improves the accuracy of vocal note onset detection. We believe that the biggest potential of the model lies in its generasability - applying it to singing material with different singing style and meter is as easy as tuning its parameters.

7. REFERENCES

[1] Hasan Sercan Atlı, Burak Uyar, Sertan Şentürk, Barış Bozkurt, and Xavier Serra. Audio feature extraction for exploring Turkish makam music. In *Proceedings of 3rd International Conference on Audio Technologies for Music and Media (ATMM 2014)*, pages 142Ã–153, Ankara, Turkey, 2014.

[2] Emmanouil Benetos, Simon Dixon, Dimitrios Giannoulis, Holger Kirchhoff, and Anssi Klapuri. Automatic music transcription: challenges and future directions. *Journal of Intelligent Information Systems*, 41(3):407–434, 2013.

[3] Masataka Goto. Singing information processing. In *12th International Conference on Signal Processing (ICSP)*, pages 2431–2438. IEEE, 2014.

[4] André Holzapfel. Relation between surface rhythm and rhythmic modes in turkish makam music. *Journal of New Music Research*, 44(1):25–38, 2015.

[5] André Holzapfel, Florian Krebs, and Ajay Srinivasamurthy. Tracking the "odd": Meter inference in a culturally diverse music corpus. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, pages 425–430, Taipei, Taiwan, October 2014.

[6] David Brian Huron. *Sweet anticipation: Music and the psychology of expectation*. MIT press, 2006.

[7] Michael T Johnson. Capacity and complexity of HMM duration modeling techniques. *Signal Processing Letters, IEEE*, 12(5):407–410, 2005.

[8] Florian Krebs, Sebastian Böck, and Gerhard Widmer. Rhythmic pattern modeling for beat and downbeat tracking in musical audio. In *Proceedings of the 14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil, November 4-8 2013.

[9] Florian Krebs, Sebastian Böck, and Gerhard Widmer. An Efficient State-Space Model for Joint Tempo and Meter Tracking. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)*, pages 72–78, Malaga, Spain, October 2015.

[10] Willie Krige, Theo Herbst, and Thomas Niesler. Explicit transition modelling for automatic singing transcription. *Journal of New Music Research*, 37(4):311–324, 2008.

[11] Nadine Kroher and Emilia Gómez. Automatic transcription of flamenco singing from polyphonic music recordings. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 24(5):901–913, 2016.

[12] Matthias Mauch, Chris Cannam, Rachel Bittner, George Fazekas, Justin Salamon, Jiajie Dai, Juan Bello, and Simon Dixon. Computer-aided melody note transcription using the tony software: Accuracy and efficiency. In *Proceedings of the First International Conference on Technologies for Music Notation and Representation (TENOR 2015)*, pages 23–30, 2015.

[13] Rodger J McNab, Lloyd A Smith, and Ian H Witten. Signal processing for melody transcription. 1995.

[14] Ryo Nishikimi, Eita Nakamura, Katsutoshi Itoyama, and Kazuyoshi Yoshii. Musical note estimation for F0 trajectories of singing voices based on a bayesian semi-beat-synchronous HMM. In *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016, New York City, United States, August 7-11, 2016*, pages 461–467, 2016.

[15] Lawrence Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[16] Colin Raffel. *Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching*. PhD thesis, COLUMBIA UNIVERSITY, 2016.

[17] Matti Ryynänen. Probabilistic modelling of note events in the transcription of monophonic melodies. Master's thesis, 2004.

[18] Justin Salamon and Emilia Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770, 2012.

[19] Nick Whiteley, Ali Taylan Cemgil, and Simon Godsill. Bayesian modelling of temporal structure in musical audio. In *Proceedings of the 7th International Society for Music Information Retrieval Conference (ISMIR 2006)*, pages 29–34, Victoria, Canada, October 2006.