Abstract

The goal of this study is to detect onsets of singing voice in polyphonic audio recordings from Turkish Makam singing. We show that tracking in a probabilistic way the vocal note onsets simultaneous to the current position in a metrical cycle (usul) could improve the accuracy of vocal onset detection.

1 Introduction

TODO:

2 Related Work

The automatic detection of vocal non-vocal is usually performed in the context of singing voice transcription.

In the case of monophonic singing the more successful approaches are based on discontinuities in pitch contours [7]. The idea is usually referred to as 'island building' groups temporally candidate onsets and has been the basis for other works.

In the case when background instruments are present, essential first step is to extract the predominant pitch curve. A probabilistic note HMM is presented in [8], where a note has 3 states: attack (onset), steady 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 [4]. Recently Mauch et al. suggested to compact the musical knowledge into rules as a way to describe the observation and transition likelihoods, instead of training them [6]. Although being conceptually capable of tracking onsets in singing voice audio with accompaniments, these approaches were tested only on a cappella singing.

In recent work a flamenco-specific approach has been suggested [5]. In flamenco some musical aspects evince significant deviations from the principles of Western music style. From this perspective the authors suggest a tradition-specific approach.

As a primary step of the note transcription stage, notes are segmented by a set of flamenco-specific onset detection rules, based on pitch contour and volume characteristics. Varying only one parameter - the Gaussian filter - we showed that using this method notes from makam singing can be segmented with reasonable recall of 75 and 65 % in the monophonic and polyphonic case [Dzhambazov].

In general vocal note onsets are divided into two types: with a change of pitch and at the same pitch.

3 Dataset

`https://docs.google.com/spreadsheets/d/1f9wyxB6emGHvVGUhIjNQwSxxhOJhPAdNzA2BieyKVXw/edit?usp=sharing`

4 Approach

A hidden Markov model (HMM) detects simultaneously beat and vocal note onsets. It is assumed to emit an onset feature that distinguishes beats frames, and in parallel a pitch contour, from which the note onsets are determined. The vocal *Vs* non-vocal detection is done as a preprocessing step. The knowledge if a frame is vocal or not is integrated in the pitch observation model.

4.1 Model description

We modify the bar-pointer model presented in [3] by adding a hidden state for vocal note state, which depends on the current note position. The model is represented as DynamicBayesian Network in Fig. 1

4.1.1 Hidden states

The tempo and bar position states have the same state space as in [3].

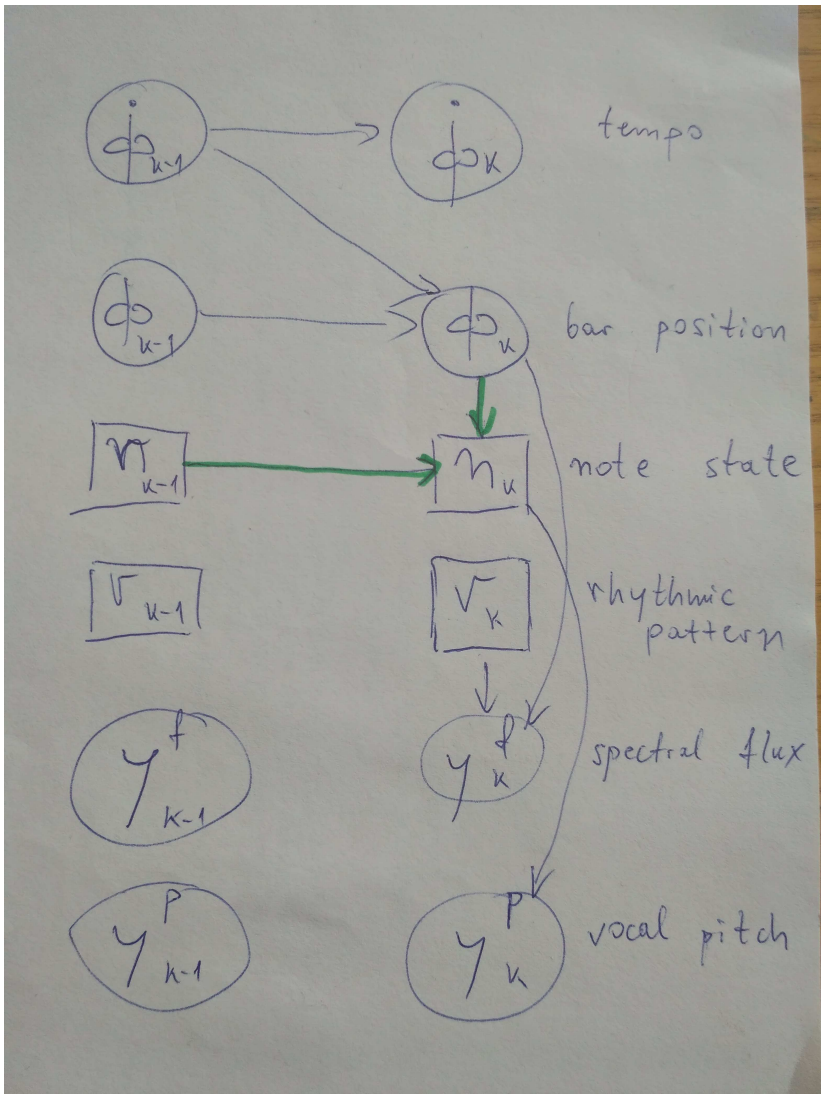

Figure 1: DBN for the simultaneous beat and note onset detection

Note state n_k The note states are a modified version of these suggested in [6] for note transcription: Each musical note is represented by three states: attack (A), stable pitch part (S) and non-vocal state (N). The non-vocal state covers time intervals where the singing voice is not active, e.g. instrumental interludes. We let notes cover a range with distinct pitch from lowest MIDI pitch .. Hz up to .. Hz. To reflect the fine-grained microtones in Makam, each MIDI pitch is further divided into 3 sub-pitches, resulting in $n = 207$ notes with different pitch, each having the 3 note states described above.

Rhythmic pattern r_k We use only one rhythmic pattern per usul.

All hidden states can be agglomerated into a mega-variable x with state space the cartesian product of the individual state spaces.

4.1.2 Transition model

The tempo transition probability $p(\dot{\phi}_k|\dot{\phi}_{k-1})$ and bar position probability $p(\phi_k|\phi_{k-1})$ are same as in [3].

Note transition probability $p(n_k|n_{k-1})$ Let $p_{nn'}$ be the probability for transition from non-vocal state N_n at note n to attack state $A_{n'}$ of note n'. TODO: describer p_nn. Let also P_{sa} be a global probability of transiting from silence to any of the attack states. In [6] the note transition is defined as:

$$p(n_k|n_{k-1}) = \begin{cases} \frac{p_{nn'}}{\sum_{n'} p_{nn'}} P_{sa} & n_{k-1} = N_n, \quad n_k = A_{n'} \\ 1 - P_{sa} & n_{k-1} = n_k = N_n \end{cases}$$

so that in the first case all transitions sum up to P_{sa}, which is set to a constant 0.0001. We modify P_{sa} to depend on current bar position ϕ_k:

$$P_{sa} = 0.0001 + [N_{0,\sigma}(d(\phi_k))]^w p_b \tag{1}$$

where $N_{0,\sigma}$ is a normal distribution, assigning probability, depending on the time interval $d(\phi_k)$ to closest beat for frame k. We set $\sigma = 0.03$ s.

w : weights the contribution of beat

p_b : probability of note onset co-occurring with the beat number b $\in$ (0,|B|). Note that in equation 1 b is the number of the closest beat to frame k.

This means essentially, that $p_{nn'}$ is kept, but scaled varyingly when close in time to a beat (also scaled more at down beats than on beats with less accents). .

4.1.3 Beat Observation Model

In this paper for the beat observation $P(y_k^f|x_k)$ we use the same observation model proposed in [3]. A Spectral Flux-like feature, y^f, that represents note onsets, is extracted from the audio signal.

4.1.4 Pitch Observation

We adopt the idea proposed in [6] that a vocal note state emits pitch y^p according to normal distribution, centered around the note pitch. The standard deviation for all stable states is set to ..., whereas the one of the attack states is ... The pitch contour is extracted with *PredominantMelodyMakam*, whereby each frame k is assigned a pitch value and probability of being voiced v_k [1]. Since the voicing detection of the contour extraction method is not optimal we ran a separate singing voice detection method [ref] and assigned $v_k = 0$ for the detected non-vocal frames.

The observation probability $P(y_k^p|x_k)$ 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 $^{1-v_k}/_n$.

4.2 Learning model parameters

4.2.1 Beat Observation model

TODO: learned fitting a GMM

4.2.2 Probability of note onset

The probability of a vocal note onset co-occurring at a given bar position p_b is learned separately for each usul. We adopt the probabilities learned in [2].

4.3 Inference

The optimal hidden state sequence $x_{1:K}$ that incorporates the tempo, bar position and note state sequence can de estimated by combining the observation models:

$$P(X_{1:K}|Y^f_{1:K}, Y^p_{1:K}) = P(X_1)\Pi^K_{k=2}P(x_k|x_{k-1})P(y^f_k|x_k)P(y^p_k|x_k)$$

We obtain $x_{1:K}$ by decoding with the well-known Viterbi algorithm. A vocal note onset is detected when the state path enters an attack note state after being in non-vocal state.

5 Experiments

5.1 Evaluation metrics

5.2 Experiment 1: annotated beats

5.3 Experiment 2: full model

6 Conclusions

TODO

References

[1] Hasan Sercan Atlı, Burak Uyar, Sertan Şentürk, Barış Bozkurt, and Xavier Serra. Audio feature extraction for exploring Turkish makam music. In *3rd International Conference on Audio Technologies for Music and Media*, Ankara, Turkey, 2014. Bilkent University, Bilkent University.

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

[3] Andre Holzapfel, Florian Krebs, and Ajay Srinivasamurthy. Tracking the" odd": Meter inference in a culturally diverse music corpus. In *ISMIR*, pages 425–430, 2014.

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

[5] 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.

[6] 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.

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

[8] Matti Ryynänen. Probabilistic modelling of note events in the transcription of monophonic melodies. 2004.