A HMM-Based Pitch Tracker for Audio Queries

Nicola Orio and **Matteo Sisti Sette**
Department of Information Engineering, University of Padova
Via Gradenigo, 6/B
35131 Padova, Italy
{orio, msistis}@dei.unipd.it

Abstract

In this paper we present an approach to the transcription of musical queries based on a hidden Markov model (HMM). The HMM is used to model the audio features related to the singing voice, and the transcription is obtained through Viterbi decoding. We report our preliminary work on evaluation of the system.

1 Introduction

Potential users of a Music Information Retrieval (MIR) system are likely to express their information needs through the *query-by-humming* paradigm. The user's query usually needs to be transcribed for segmenting the audio signal in a sequence of events and describing each event with its features, normally the pitch. The effectiveness of a MIR system depends also on the quality of the query transcription. Research on automatic pitch tracking has been extensively carried out for many years, applying different methods, including autocorrelation (Cheveigné and Kawahara, 2002), auditory models (Clarisse et al., 2002), and classical frequency or mel-cepstrum analyses; moreover, a number of commercial pitch trackers is already available. Yet, there are some peculiarities of queries sung by non expert users that require more research on pitch tracking . The query can be systematically or locally out of tune and notes can start with or be connected by glissandi and be recorded in a noisy environment. Moreover, queries can be whistled, hummed with different syllables, or sung with the lyrics. All these problems need to be taken into account during the development of a MIR front-end.

The presented approach has been developed as a part of an existing MIR system (Melucci and Orio, 1999), which is based on the extraction of relevant musical phrases as content descriptors of music documents. Different levels of normalizations are used to deal with errors in the sung queries. The output of the HMM-based pitch tracker is a sequence of MIDI-like music events, which is used as the input for the musical phrase extraction and normalization of musical queries.

2 Segmentation and pitch tracking with a Hidden Markov Model

Hidden Markov models (HMMs) have been applied in many different areas, from speech recognition and biological sequence analysis, to text and music information retrieval. An extensive description of HMMs is given in (Rabiner and Juang, 1993). HMMs have been applied to the pitch tracking of speech by Wu et al. (2001), while an application of HMMs to the alignment of performances with scores was presented by Raphael (1999).

From a statistical point of view, a sung query can be considered as the observation of an unknown process, which is the melody the user has in mind. What is observed is a set of audio features, which has to be chosen. The process can be modeled with a HMM, where transition probabilities concern both the evolution of a given note (e.g., attack, sustain, release, and pauses) and the possible sequences of notes (e.g., the probability that the user sings a given interval), and emission probabilities refer to the observed audio features. The transcription reduces then to the discovery of the evolution along states in the HMM that most likely generated the query, which can be obtained through Viterbi decoding.

We propose the use of a two-level HMM. The *event level* is a set of states, each one labeled with a different pitch in the chromatic scale inside a given range. At this level a melody is modeled as a path among the states. The network is fully connected, as shown in Figure 1(a), but the probability to go from a state to another one depends on the melodic interval between their labels. Transition probabilities can be set to the a priori probability of a musical interval in a given repertoire. The *audio level* is a set of states connected with a left-to-right topology, with labels related to the evolution of a musical event. Figure 1(b) shows a simple topology of a note event at the audio level, with three states representing attack, sustain, and a final rest which may be skipped. This is the audio level topology we have used in our preliminary tests, but we believe that the great potential of the HMM-based approach lies in the possibility to implement more complex topologies. In Figure 2 we propose an extension to the model of Figure 1(b), with multiple parallel attack states (with different emission probability densities) representing different kinds of attacks, e.g. vowel and consonant attacks, from silence and from the steady state of the previous note. Multiple consecutive sustain states (possibly with identical emission densities) allow to impose a lower bound to the note duration. Additional states, marked with '+' and '-' in the figure, are added in order

to model slight detunings during the sustain of a note. Other events, such as glissandi, vibrati, or thrills, can be represented by introducing other topologies and states at both levels.

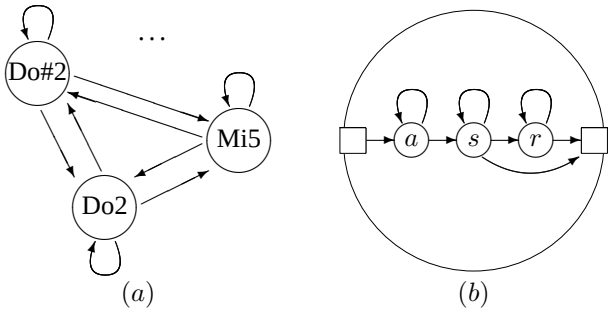

Figure 1: The HMM topology. (a) The event level. (b) The audio level used in our experiments: *a*, *s* and *r* stand for *attack*, *sustain*, and *rest* respectively; squares represent null states.

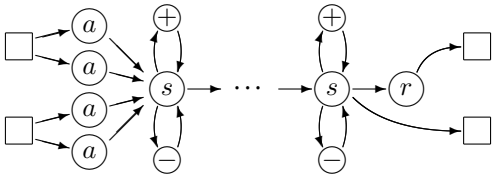

Figure 2: A topology for the audio level modeling different kinds of attacks, durations, and slight detunings; self transition arcs are not shown.

The complete HMM is given by the definition of the emission probabilities. These probabilities need to be set and computed only for states of the audio level, because states at the event level are simply formed of states at the audio level. The features we have considered and tested so far are: the log-energy of the signal, for discriminating between silences and sounds; the sum of the energy in the bands of the first harmonics for each note in the melodic range, for distinguishing different pitches; the first derivatives of both these features, for a more precise detection of note attacks; the maximum difference between energies in adjacent harmonic bands, to reduce the common problem of doubling and halving octaves.

The chosen representation of the spectrum, simply based on the overall energy on the harmonic bands, helps dealing with problems like missing harmonics or distortions due to possible user's low-quality audio equipment; moreover the size of each band is chosen to model temporary detunings of at most a quarter-tone. Since users may not have absolute pitch, a preprocessing step is carried out to estimate the most probable distance between the user's reference and the typical 440 Hz to set accordingly the harmonic bands. We are working on adding other features, such as the difference between the energies in the bands of contiguous pitches in the chromatic scale, in order to model glissandi and small variations of pitch inside a note, and to improve the modelling of attacks.

All the features are modeled with unilateral exponential probability density functions (pdfs); preliminary tests with Gaussians gave slightly worse results. The exponential pdfs are trained using a statistical analysis on a set of labeled audio examples. States of the same kind (e.g., all sustain states) belonging to different notes share the same parameters; they differ in the way the harmonic bands are computed.

The transcription can be obtained through decoding, considering that the HMM has a transition each time a new audio frame is analyzed and a new set of features is observed. The path along states at the event level gives both the segmentation and the pitch tracking, because each state is associated to an event through its label.

3 Evaluation

We carried out a preliminary evaluation of our approach, using the HMM shown in Figure 1. We tested the system with 18 audio queries sung by untrained users, using a low quality microphone and a common digital audio interface. Queries had an overall number of 300 notes.

Results are summarized in Table 1. Errors are mainly due to glissandi, which were not taken into account as possible expressive gestures and so were tracked by the system as chromatic scales preceding notes. Other common errors are the insertion or the wrong recognition of notes a halfstep from the correct one. These results highlight some improvements that can be applied to the simple model, which will be part of our future work.

omitted notes	total	4%
	repetition of same note	3%
pitch errors (whole notes)	total	10%
	± 1 semitone	8%
	$>$ 1fifth	2%
notes inserted during notes	total	13%
	repetition of same note	3%
	1-2 semitones from correct note	8%
	$>$ 1fifth from correct note	2%
notes inserted during silence		1%
multiple notes detected during attacks		27%

Table 1: Preliminary test results. All percentages refer to the total number of notes.

References

Cheveigné, A. & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustic Society of America*, 111(4), 1917–1930.

Clarisse, L. P. et al. (2002). An auditory model based transcriber of singing sequences. In *Proceedings of the International Conference on Music Information Retrieval*, (pp. 116–123).

Melucci, M. & Orio, N. (1999). Musical Information Retrieval Using Melodic Surface. In *Proceedings of IV ACM Conference on Digital Libraries*, (pp. 152–160).

Rabiner, L. & Juang, B.H. (1993). *Fundamentals of speech recognition*. (pp. 321–389). Prentice Hall, Englewood Cliffs.

Raphael, C. (1999). Automatic segmentation of acoustic musical signals using Hidden Markov Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(4), 360–370.

Wu, M. et al. (2001). Pitch Tracking Based on Statistical Anticipation. In *Proceedings of the Internetional Joint Conference on Neural Networks*, Volume 2, (pp. 866–871). DC.