

MUSICAL METER ESTIMATION AND MUSIC TRANSCRIPTION

Anssi P. Klapuri

Tampere University of Technology, P.O.Box 553, FIN-33101 Tampere, Finland

klap@cs.tut.fi

ABSTRACT

This paper concerns the segmentation and automatic transcription of acoustic musical signals. First, a method is proposed for estimating the temporal structure of music, the musical meter. The method consists of an acoustic analysis front-end and of a subsequent probabilistic model which performs joint estimation of the meter at three time scales: beat (tactus), measure, and tatum pulse levels. Secondly, musicological models are described which utilize musical information to improve transcription accuracy.

1. INTRODUCTION

Transcription of music aims at discovering the “recipe” of a musical performance which musicians can use to reproduce or modify the original performance. The transcription result is typically given in the form of a musical notation which comprises the times, durations, and pitches of the sounds that constitute the piece. Sometimes, a rougher symbolic representation may suffice, for example the musical chords for a guitar player.

Attempts toward the automatic transcription of polyphonic music date back to 1970s. More recently, Kashino et al. applied psychoacoustic processing together with temporal and musical predictions [1]. Martin utilized musical rules in transcribing four-voice piano compositions [2]. Godsmark and Brown proposed a system which integrates evidence from different auditory organization principles and demonstrated that the model could segregate melodic lines from polyphonic music [3]. Goto introduced the first pitch analysis method that works quite reliably for real-world complex musical signals, finding the melody and bass lines in them [4]. For a review on transcription systems, see [14].

Musical signals exhibit temporal regularity, a meter. Meter estimation is an essential part of understanding musical signals and an innate cognitive ability of humans even without musical education. Perceiving musical meter can be characterized as processing musical events so that underlying periodicities are detected [5]. Musical meter is hierarchical in structure, consisting of pulse sensations at different levels (see Fig. 1). The most prominent metrical level is *beat* (foot tapping rate). *Tatum* (time quantum) refers to the shortest durational values in music that are still more than incidentally encountered. The other durational values, with few exceptions, are integer multiples of the tatum. Musical *measure* pulse is usually related to the harmonic change rate.

Automatic estimation of the meter alone has several applications. It facilitates the cut-and-paste operations and editing of music signals, enables synchronization with light effects or video, and can be used in further signal analysis. Earlier work on meter estimation has concentrated almost solely on beat tracking, with the few exceptions in [6,7,11]. Most earlier systems have attempted to discover periodicities in symbolic input (MIDI)

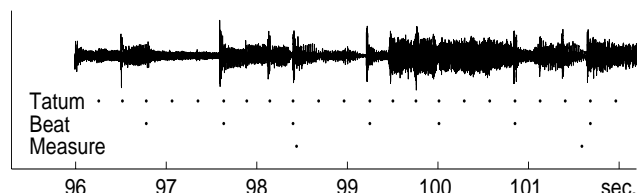


Fig. 1. A musical signal with three metrical levels illustrated.

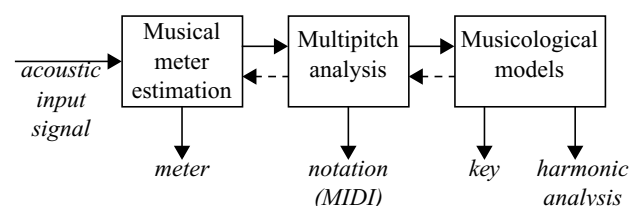


Fig. 2. Block diagram of the transcription system.

[6,7,8, see 8 for a review]. Only a couple of systems have been proposed that process acoustic input signals [9,10,11,8].

The scope of this paper is the musical meter estimation and automatic transcription of acoustic musical signals. In meter estimation, beat, measure, and tatum levels are considered. Figure 2 shows the overview of the transcription system. Meter estimation forms the backbone also in an automatic analysis. It can be done rather robustly, and allows the positioning and sizing of the analysis windows in further analysis. In principle, it might be advantageous to run meter estimation and multipitch analysis in parallel, but that is not computationally very efficient. Beyond multipitch analysis, there are still higher-level musical structures. Musicological models play the role of a “language model” in music. They are used to improve the transcription accuracy or, for example, to perform harmonic analysis based on note data.

2. METER ANALYSIS MODEL

In this section, a method is described which estimates the beat, measure, and tatum pulses in acoustic input signals. The target signals are not limited to popular music, but all main genres (including classical music) are represented in the evaluation set of 478 musical pieces. The time-frequency analysis part of the system (see Fig. 3) is a synthesis and generalization of two earlier proposed methods [9] and [10]. This is followed by a probabilistic model, where beat, measure, and tatum periods are jointly estimated, and temporal continuity of the meter is taken into account. Both causal and non-causal models are presented.

2.1 Calculation of registral accent signals

Sound onset points (beginnings), sudden changes in intensity or timbre, and harmonic changes are important for the perception of

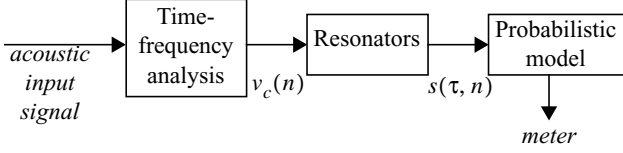


Fig. 3. Overview of the meter estimation method.

musical meter. These can be collectively called as *phenomenal accents*, events that give emphasis to a moment in music [5].

All the mentioned accent types can be observed in the time-frequency representation of a signal. While an auditory-model based front-end would be theoretically better, it did not yield a performance advantage over a straightforward filterbank implementation in our simulations.

Acoustic input signals are sampled at 44.1 kHz rate and 16-bit resolution and normalized to have zero mean and unity variance. A discrete Fourier transform is calculated in successive 23 ms frames which are hanning-windowed and overlap 50%. Then 36 triangular-response bandpass filters having equal bandwidth on the ERB critical-band scale are simulated between 50 Hz and 20 kHz. The power at each band is calculated and stored to $x_b(k)$, where k is the frame index and $b \in [1, B]$, $B=36$.

There are many potential ways of measuring the degree of change or stress, given the power intensity envelopes at critical bands. For humans, the smallest detectable change in intensity is approximately proportional to the intensity of the signal, the same amount of increase being more prominent in a quiet signal. That is, $\Delta I/I$, the Weber fraction, is constant over a wide range of intensities. Thus it is reasonable to normalize the differential of power with power. This leads to $[(d/dt)x_b(k)]/(x_b(k))$, which is equal to $(d/dt)\ln[x_b(k)]$. This formula can be seen as “differentiated loudness”, since the perception of loudness is roughly proportional to the sum of the log-powers at critical bands.

The above-mentioned logarithm and differentiation operations are both given in a more general form. A numerically flexible way of taking the logarithm is the μ -law compression

$$y_b(k) = \ln[1 + \gamma x_b(k)] / \ln(1 + \gamma), \quad (1)$$

which performs mapping from values of $x_b(k)$ between zero and one to values of $y_b(k)$ between zero and one. Constant γ can be used to compromise between a close-to-linear ($\gamma < 0.1$) and a logarithmic ($\gamma > 1000$) transformation. A close-to-log value $\gamma=100$ is employed, but any value in range $[10, 10^6]$ is valid, too.

To achieve better time resolution, the compressed power envelopes $y_b(k)$ are interpolated by factor two, leading to $y_b'(n)$ with $f_r=172$ Hz sampling rate. Sixth-order Butterworth lowpass filter with $f_{LP}=10$ Hz cutoff frequency is applied to smooth the compressed and interpolated power envelopes. The resulting smoothed signal is denoted with $z_b(n)$.

Differentiation of $z_b(n)$ is performed as follows. First, a half-wave-rectified differential of $z_b(n)$ is calculated as

$$z_b'(n) = \text{HWR}\{z_b(n) - z_b(n-1)\} \quad (2)$$

where $\text{HWR}\{\}$ denotes halfwave rectification and is essential to make the differentiation useful. Then a weighted average of $z_b(n)$ and its differential $z_b'(n)$ is formed as

$$u_b(n) = (1-w)z_b(n) + w\beta z_b'(n) \quad (3)$$

where $w=0.9$ and the factor $\beta = f_s/(2f_{LP})$ compensates for the fact that differential of lowpassed signal is small in amplitude.

Finally, each M adjacent bands are linearly summed to get $C = \lceil B/M \rceil$ *registral accent signals* $v_c(n)$:

$$v_c(n) = \sum_{b=(c-1)M+1}^{cM} u_b(n) \quad (4)$$

The registral accent signals $v_c(n)$ serve as a mid-level representation for musical meter estimation. They represent the degree of accentuation (stress) as a function of time at frequency channel c . We used $B=36$ and $M=9$, leading to $C=4$. It should be noted that combining each M adjacent bands at this stage is not an issue of computational complexity, but improves the analysis accuracy.

The presented general form of calculating the registral accent signals is very flexible when varying γ , w , B , and M . A representation similar to that used by Scheirer in [9] is obtained by setting $\gamma=0.1$, $w=1$, $B=6$, $M=1$. A representation similar to that used by Goto [10] is obtained by setting $\gamma=0.1$, $w=1$, $B=36$, $M=6$. By setting $\gamma=0.1$, $w=0$, non-differentiated powers of adjacent channels are summed in Eq. (4). We fix values $\gamma=100$, $w=0.9$, $B=36$, $M=9$.

2.2 Bank of comb-filter resonators

Periodicity of the registral accent signals $v_c(n)$ is analyzed to estimate the strength of different metrical pulse periods. This somewhat resembles the idea of “registral inter-onset-interval” computation for MIDI data in [7]. Four different periodicity estimation algorithms were evaluated: enhanced cross-correlation, enhanced *YIN* method of de Cheveigné and Kawahara, comb-filter resonators [9], and phase-locking resonators [6]. As an important observation, however, three of the four methods performed equally well after a thorough optimization. The method presented in following is the least complex among the three best-performing algorithms, a modification from [9].

Using a bank of comb-filter resonators with constant half-time has been originally proposed for beat tracking by Scheirer [9]. Output of a comb filter with delay τ for input $v_c(n)$ is given as

$$r_c(\tau, n) = \alpha_\tau r_c(\tau, n - \tau) + (1 - \alpha_\tau)v_c(n) \quad (5)$$

where the feedback gain $\alpha_\tau = 0.5^{\tau/T_0}$ is calculated for a selected half-energy time T_0 . We used $T_0 = 3f_r$ i.e., a half-time of three seconds which is short enough to react to tempo changes but long enough to reliably estimate musical-measure periods of up to four seconds. Scheirer used halftimes of 1.5–2.0 seconds.

The filters implement a frequency response where frequencies kf_r/τ , $k = 0, \dots, \lfloor \tau/2 \rfloor$ have a unity response and maximum attenuation between the peaks is $[(1 - \alpha_\tau)/(1 + \alpha_\tau)]^2$. Overall power $g(\alpha_\tau)$ of the comb filters can be calculated by squaring and summing their impulse responses, which yields

$$g(\alpha_\tau) = (1 - \alpha_\tau)^2 / (1 - \alpha_\tau^2). \quad (6)$$

A bank of such resonators was applied, with τ getting values from 1 to τ_{max} , where $\tau_{max} = 688$ corresponds to four seconds. Computational complexity of one resonator is $O(1)$ per each input sample, and complexity of the overall resonator filterbank is $O(Cf_r\tau_{max})$, which is not computationally too demanding.

Powers $\hat{r}_c(\tau, n)$ of each comb filter are calculated as

$$\hat{r}_c(\tau, n) = \frac{1}{\tau} \sum_{i=n-\tau+1}^n [r_c(\tau, i)]^2. \quad (7)$$

These are then normalized as

$$s_c(\tau, n) = \frac{1}{1 - g(\alpha_\tau)} \left[\frac{\hat{r}_c(\tau, n)}{\hat{v}_c(n)} - g(\alpha_\tau) \right], \quad (8)$$

where $\hat{v}_c(n)$ is the power of the registral accent signal $v_c(n)$, calculated by squaring $v_c(n)$ and by applying a leaky integrator, i.e., a resonator which has $\tau=1$ and the same half-time as the other resonators. Normalization with $g(\alpha_\tau)$ is applied to compensate for the differences in the overall power responses. The proposed normalization is advantageous because it preserves a unity power response for the peak frequencies kf_r/τ and at the same time removes the τ -dependent trend in $s_c(\tau, n)$ for a white noise input.

Finally, a function $s(\tau, n)$ which represents the strengths of different metrical pulses at time n is obtained as

$$s(\tau, n) = \sum_{c=1}^C s_c(\tau, n). \quad (9)$$

This function acts as the *observation* for the probabilistic models.

For tatum period estimation, a discrete Fourier transform $S(f, n)$ of $s(\tau, n)$ is calculated as

$$S(f, n) = \sqrt{f} \sum_{\tau=1}^{\tau_{max}} [s(\tau, n) w(\tau) e^{-i2\pi f(\tau-1)/\tau_{max}}] \quad (10)$$

where \sqrt{f} removes spectral trend and $w(\tau)$ is half-Hanning

$$w(\tau) = 0.5 \{1 - \cos[\pi(\tau_{max} + \tau - 1)/\tau_{max}]\}. \quad (11)$$

The rationale behind calculating the DFT is that, by definition, other pulse periods are integer multiples of the tatum period. Thus the overall function $s(\tau, n)$ contains information about tatum and is conveniently gathered using Eq. (10).

2.2.1 Phase estimation

As will be discussed in Sec. 2.3.3, *phase* is estimated only for one beat and measure estimate at successive time instants, after the *period* of these has been decided. As proposed by Scheirer in [9], we use the last τ outputs $r_c(\tau, n)$, $n \in [n_0 - \tau + 1, n_0]$ of a resonator to determine the phase of the corresponding pulse.

The phase of a beat pulse (after its period is solved) is decided by finding n , $n \in [n_0 - \tau + 1, n_0]$, which maximizes

$$d_B(\tau, n) = \sum_{c=1}^C (C - c + 2) r_c(\tau, n), \quad (12)$$

i.e., a bass-weighted sum over the resonator outputs at channels c .

The phase of a winning measure period is decided in a same manner. A function $d_M(\tau, n)$ similar to that in Eq. (12) is constructed, but so that different channels are weighted and delayed in a more complex way. It does not make sense to present the deterministic rule here, since equally good results were obtained by sampling $r_c(\tau_M, n)$ (τ_M is measure period) at the 2-8 time points which correspond to beat pulses, and using these features from different channels to train a Gaussian mixture model classifier which selects one of the beat pulse points to be a time anchor for measure. For tatum, the beat phase determines tatum phase.

2.3 Probabilistic model for meter estimation

To enable probabilistic calculations, we evaluate the probability of metrical pulse periods $\tau_i^{(n)}$ given an observation $s(\tau, n)$. The index $i \in \{B, M, T\}$ represents beat, measure, or tatum period, respectively. For the sake of convenience, we first omit the time index n , and denote observation vector $s = s(\tau)$.

According to Bayes's formula, we can write

$$P(\tau_i | s) = \frac{P(s | \tau_i) P(\tau_i)}{P(s)}, \quad (13)$$

where $P(\tau_i)$ and $P(s)$ are prior probabilities for pulse period and observation, respectively.

The probability $P(s | \tau_i)$ can be written as

$$P(s | \tau_i) = P(s(\tau_i) | \tau_i) \prod_{\tau \neq \tau_i} P(s(\tau) | -\tau), \quad (14)$$

where $P(s(\tau) | \tau)$ is the probability of value $s(\tau)$ given τ is a pulse period, and $P(s(\tau) | -\tau)$ is the probability of $s(\tau)$ given that τ is not a pulse period. These two conditional probability distributions were approximated by discretizing the value range of $s(\tau) \in [0, 1]$ and by calculating a histogram of $s(\tau)$ values in the cases that τ is or is not an annotated metrical pulse period. Then, by defining

$$\beta = \prod_{\tau=1}^{\tau_{max}} P(s(\tau) | -\tau), \quad (15)$$

Equation (14) can be rewritten as

$$P(s | \tau_i) = \beta \frac{P(s(\tau_i) | \tau_i)}{P(s(\tau_i) | -\tau_i)}, \quad (16)$$

where the scalar β does not depend on τ_i . By using the two approximated histograms for the nominator and denominator, a discrete histogram which models the underlying distribution $P(s | \tau_i)$, $i \in \{B, M\}$ is obtained. For tatum, a distribution $P(s | \tau_T)$ is approximated using the same procedure. These histograms can be accurately modeled with a third-order polynomial, but since there was no performance advantage compared to a linear model, we employ the simple model

$$P(s^{(n)} | \tau_i^{(n)}) = \begin{cases} \alpha_i s(\tau_i, n), & i = B \\ \alpha_i s(\tau_i, n), & i = M \\ \alpha_i S(1/\tau_i, n), & i = T \end{cases}, \quad (17)$$

where the scalars α_i do not depend on τ_i .

Prior probabilities $P(\tau_i)$ for beat period lengths have been measured from actual data by several authors [12,13]. As suggested by Parncutt in [12], we apply a lognormal prior distribution

$$P(\tau_i) = \frac{1}{\tau_i \sigma_i \sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma_i^2} \left[\ln \left(\frac{\tau_i}{m_i} \right) \right]^2 \right\}, \quad (18)$$

where m_i and σ_i are the scale and shape parameters, respectively. The following values for beat, measure, and tatum periods were estimated from data: $m_B = 0.55$, $\sigma_B = 0.65$, $m_M = 2.1$, $\sigma_M = 0.60$, $m_T = 0.18$, $\sigma_T = 0.90$. However, the dynamics of these true priors have to be matched with the dynamics of $P(s | \tau_i)$. For this reason, we use $P(\tau_i) = P'(\tau_i)^{e_i}$ in the following, with $e_B = 1/3$, $e_M = 1/3$, $e_T = 1/6$.

2.3.1 Joint meter estimation

The prior probabilities $P(\tau_i)$ can be intuitively understood as smooth boundaries for searching beat and measure periods in $s(\tau, n)$ and tatum period in $S(f, n)$. Indeed, rather reliable estimates for the three pulse periods can be obtained by finding τ_i which maximizes $P(s | \tau_i) P(\tau_i)$. However, even better results are obtained by estimating beat, measure, and tatum periods jointly, and by placing continuity constraints on temporally successive meter estimates.

The three period estimates are combined into a parameter set

$$\theta^{(n)} = \{\tau_B^{(n)}, \tau_M^{(n)}, \tau_T^{(n)}\} \quad (19)$$

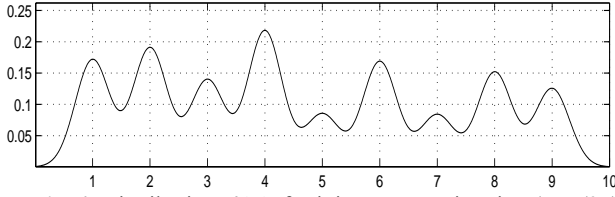


Fig. 4. Distribution $f(x)$ for joint meter estimation (Eq. (25)).

Again, according to Bayes's formula,

$$P(\theta|s) = \frac{P(s|\theta)P(\theta)}{P(s)}. \quad (20)$$

The conditional probability $P(s|\theta)$ can be estimated through a procedure similar to that shown above for $P(s|\tau_i)$. It turns out that $P(s^{(n)}|\theta^{(n)})$ can be reasonably approximated as

$$P(s^{(n)}|\theta^{(n)}) = \alpha s(\tau_B, n) s(\tau_M, n) S(1/\tau_T, n) \quad (21)$$

where the scalar α varies with $s^{(n)}$ but does not depend on τ_i .

The prior probability $P(\theta)$ can be written as

$$P(\theta) = P(\tau_B)P(\tau_M|\tau_B)P(\tau_T|\tau_B, \tau_M). \quad (22)$$

It is safe to assume that $P(\tau_T|\tau_B, \tau_M) = P(\tau_T|\tau_B)$, i.e., given the beat period, the measure period does not give additional information regarding the tatum period.

It is advantageous to represent the probabilities $P(\tau_M|\tau_B)$ and $P(\tau_T|\tau_B)$ as a product of two terms as

$$P(\tau_M|\tau_B) = P(\tau_M)f(\tau_M/\tau_B) \quad (23)$$

$$P(\tau_T|\tau_B) = P(\tau_T)f(\tau_B/\tau_T) \quad (24)$$

where $P(\tau_i)$ are the already-introduced priors and

$$f(x) = \sum_{l=1}^9 w_l N(x; l, \sigma_1) \quad (25)$$

is a Gaussian mixture density with component weights w_l , component means l and common variances $\sigma_1 = 0.3$. The exact weight values are not critical, but are designed to realize a tendency towards binary or ternary integer relationships between concurrent pulses. For example, it is relatively probable that one beat period consists of 2, 4, or 6 tatum periods, but multiples 5 and 7 are much less likely and thus have lower weights. The distribution $f(x)$ is shown in Fig. 4.

The probabilities $P(\theta^{(n)}|s^{(n)})$ are computed once in a second. At these points, five candidates for beat, measure, and tatum periods are selected by finding $\tau_i^{(n)}$ which maximize Eq. (13). Then, Eq. (20) is evaluated for each of the 5^3 combinations of these, i.e., for 125 different meter candidates.

2.3.2 Temporal continuity and retrospection

Musical meter cannot be assumed to be static over the duration of a piece. It has to be estimated causally at successive time instants and there must be some tying between temporally successive meter estimates. We define the probability that a meter $\theta^{(n)}$ follows meter $\theta^{(n-N_0)}$ one second ($N_0 = f_s$ samples) earlier as

$$P(\theta^{(n)}|\theta^{(n-N_0)}) = \prod_{i \in \{B, M, T\}} P(\tau_i^{(n)}|\tau_i^{(n-N_0)}) \quad (26)$$

where

$$P(\tau_i^{(n)}|\tau_i^{(n-N_0)}) = \frac{1}{\sigma_2 \sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma_2^2} \left[\ln \left(\frac{\tau_i^{(n)}}{\tau_i^{(n-N_0)}} \right) \right]^2 \right\} \quad (27)$$

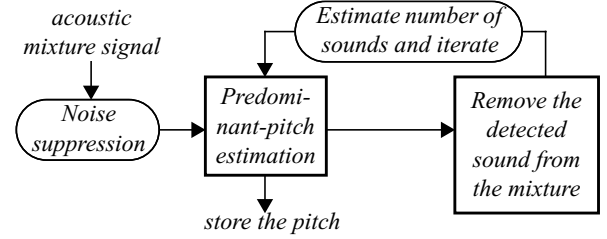


Fig. 5. Parts of the multipitch estimation method.

implements a normal distribution as a function of the relative change in period and parameter $\sigma_2 = 0.2$.

The described temporal model can be straightforwardly computed using the Viterbi algorithm. At each time point, there are 125 different "states", corresponding to the 125 meter candidates (see Sec. 2.3.1). Probabilities of being in these states are calculated according to Eq. (20). Transition probabilities between the states are obtained from Eq. (26). In a causal model, the meter estimate $\theta^{(n)}$ at time n is determined according to the maximum-likelihood estimate up to that point in time. A *retrospective* (non-causal) estimate after seeing a complete sequence of observations can be computed using the forward-backward decoding where the most likely sequence of meter estimates is decided by backtracking after all the observations have been seen.

2.3.3 Phase estimation

Phase estimation is done separately for beat and measure pulses after the period-lengths of these pulses have been decided at a given point in time. Probability for phase $\phi_i^{(n)}$ given the latest few outputs of resonator $\tau_i^{(n)}$ is modeled as

$$P(\phi_i^{(n)}|d_i(\tau_i^{(n)}, n)) = \beta_i d_i(\tau_i^{(n)}, n) \quad (28)$$

where $i \in \{B, M\}$ and β_i is a normalizing constant. No joint estimation is attempted. However, temporal processing for phase is analogous to that described for period estimation. Phase is expressed as a *phase anchor*, absolute time value when a pulse point occurs. Given two such anchors, the conditional probability

$$P(\phi_i^{(n)}|\phi_i^{(n-N_0)}) = [1/(\sqrt{2\pi}\sigma_3)] \exp\{-e^2/(2\sigma_3^2)\} \quad (29)$$

where the error e is relative to the period length and calculated as

$$e = [\text{mod}\{\phi_i^{(n-1)} - \phi_i^{(n)}\} + \tau_i/2, \tau_i] - \tau_i/2 / \tau_i \quad (30)$$

where $\text{mod}\{x, y\}$ is modulus-after-division operator, $\sigma_3 = 0.1$.

Using Eqs. (28) and (29), causal and retrospective computation of phase is performed in the way described in Sec.2.3.2.

3. MULTIPITCH ANALYSIS

The multiple-F0 estimation method applied here has been thoroughly explained earlier in [14,15,16]. The descriptions are not repeated where. Figure 5 shows the overview of the method, where the two main parts are applied in an iterative fashion. The first part, predominant-pitch estimation, finds the pitch of the most prominent sound. In the second part, the spectrum of the detected sound is estimated and linearly subtracted from the mixture. The two steps are then repeated for the residual signal.

4. MUSICOLOGICAL MODELS

Although the draft transcription produced after meter estimation

and multiple-F0 estimation contains several errors, the overall harmonic progression and much of the melodic features are preserved. Thus, harmonic analysis is a realistic objective. Also, musicological models (like language models in speech recognition) are indispensable to improve the transcription accuracy.

4.1 Key signature estimation

Key signature is a note system where some notes have greater probability of occurrence than others. Here we consider the most common key signatures in Western music, the twelve major keys and the twelve minor keys. Occurrence probabilities of different notes m given the key, i.e., $P(m|k)$, have been estimated from large amounts of classical music by several authors [17]. Using the Bayes formula, the probability of different key signatures given a note value can be calculated as

$$P(k|m) = \frac{P(m|k)P(k)}{P(m)}. \quad (31)$$

It is reasonable to assume equal prior probabilities for all keys and notes, i.e., $P(k) = 1/24$ and $P(m) = 1/12$, where notes are represented as *pitch classes* (12 different pitch classes). Probabilities $P(m|k)$ are obtained from [17,p.67].

Probability of a key signature k given several notes is

$$P(k|m_1^T) = \prod_{t=1}^T P(k|m_t), \quad (32)$$

where we have denoted note sequence m_1, m_2, \dots, m_T as m_1^T .

4.2 Harmonic analysis

Here, harmonic analysis means dividing an incoming signal into segments and assigning a chord label to each segment. The core of the probabilistic model for that purpose is essentially the same as that for key signature estimation (substitute chord c for k in Eqs. (31) and (32)). The probabilities $P(m|c)$ can be deduced from earlier harmonic analysis programs (e.g. [7]). Also, prior probabilities for chords given a key signature, and probabilities for chords given a preceding chord, $P(c^{(k)}|c^{(k-1)})$, can be deduced from [17,p.181,195]. Because we do not have quantitative results on harmonic analysis, this is not elaborated more.

4.3 N -grams

N -grams have been found to be a convenient way of modeling the sequential dependencies in natural languages. An N -gram uses $N - 1$ previous events to predict what the next event would be. Whereas harmonic analysis models the “vertical” grouping of notes in a composition, N -grams can be used to model the horizontal (melodic) grouping of notes into streams. Using an N -gram model, the probability of a note sequence can be calculated as

$$P(m_1^T) = \prod_{t=1}^T P(m_t|m_{t-1}^{t-1}). \quad (33)$$

We have examined the usability of N -grams together with Viitanen and Eronen in [19]. Bigram probabilities $P(m_t|m_{t-1})$ were estimated from the 5983 monophonic melodies in the ESaC database [18], and these were applied to the transcription of sung melodies. Using N -grams in *polyphonic* transcription raises several questions. It is difficult to limit to N preceding events when several concurrent notes occur. However, some phenomena can be elegantly modeled with N -grams, e.g., the likely event that a detected note continues ringing in the following analysis frames.

Table 1: Statistics of the evaluation database.

Genre	# Pieces with annotated metrical levels.		
	beat	measure	tatum
Classical	85	0	54
Electronic / dance	66	62	35
Hip hop / rap	37	36	12
Jazz / blues	95	72	45
Rock / pop	124	101	80
Soul / RnB / funk	56	46	34
Unclassified	15	4	10
Total	478	321	270

4.4 Interaction between musicological and pitch model

A guiding principle in using musicological models to improve transcription accuracy is to estimate key signature and chords based on a longer sequence of observations, and to let these affect the probabilities of individual notes. Using the key signature model, the probability of a note m_t after observing sequence m_1^T

$$P_k(m_t|m_1^T) = \sum_{j=1}^{24} P(m_t|k=j)P(k=j|m_1^T). \quad (34)$$

The usability of this model was evaluated in the abovementioned task of transcribing sung melodies [19]. Application in a polyphonic context is straightforward and similar.

5. RESULTS

5.1 Meter estimation

Table 1 shows the statistics of the database used to evaluate the accuracy of the proposed meter estimation model. A database of musical pieces was collected from CD recordings. Beat and measure positions were manually annotated for approximately one-minute-long representative excerpts per piece by tapping along with the pieces, recording this input, and by detecting the tapped times semiautomatically. Beat positions could be annotated for 478 of a total of 505 pieces. Musical measure could be reliably marked by listening for 321 pieces. Tatums were annotated by myself (a nonmusician) by determining the integer ratio between beat and tatum period lengths.

Evaluating a meter estimation system is not trivial. According to criteria proposed by Goto in [11], we use the longest *continuous* correctly estimated segment as a basis for monitoring the performance (one inaccuracy in the middle of a piece leads to 50 % rate). The longest continuous sequence of correct pulse estimates in each piece was sought and compared to the length of the segment which was given to be analyzed (prior to analysis, estimator got four seconds of input to fill resonator states). The relation between these two lengths determines the correct-estimation rate for one piece, which are then averaged over all pieces. Performance measures are given for three different criteria:

- “Correct”: Each pulse estimate is accepted if its period and phase are correct. A correct period is defined to deviate less than 10 % from the reference. A correct phase deviates from a manually annotated value less than 15 % of the correct period length.
- “Accept d/h”: Pulse estimate is accepted if its phase is correct and period is either the correct or exactly doubled or halved.
- “Period correct”: Period must be correct, phase is ignored. This

Table 2: *Beat tracking* performance (%) for different systems using various evaluation criteria.

System	Require continuity			Individual estimates		
	correct	accept d/h	period correct	correct	accept d/h	period correct
Causal	54	65	(70)	61	76	(74)
Noncausal	57	69	(71)	63	78	(74)
Scheirer [9]	26	30	(28)	46	65	(55)
Dixon [8]	5	17	(7)	13	46	(21)
O + Dixon	10	26	(12)	21	58	(28)

Table 3: *Meter estimation* performance for the proposed model. Continuity is required, criteria are “accept d/h” and “period”.

System	Beat	Measure	Tatum
Causal	65 (70)	48 (78)	44 (62)
Noncausal	69 (71)	54 (79)	47 (64)

has the interpretation as *tempo estimation* performance.

When listening to meter estimation results it was quite clear that (a) continuity is aurally very important but (b) period doubling or halving is not disturbing and does not prevent further analysis.

Table 2 compares the beat tracking performance of the proposed causal and noncausal models with two reference systems [8,9]. For the reference systems, the implementation and parameters were those of the original authors. Dixon developed his system particularly for MIDI-input, and provided only a simple front-end for analyzing acoustic input. Therefore, a third system “O + Dixon” was developed where our onset detector was used prior to Dixon’s beat analysis. As the first observation, it was noticed that those earlier state-of-the-art systems do not produce *continuous* good estimates. For this reason, the performance rate is also given for individual good estimates (right half of Table 2).

Table 3 gives the meter estimation performance for the proposed causal and noncausal models. Estimating the phase of measure-level pulses turned out to be very difficult. This presses down the performance factor for measure-pulse. Measure period is well estimated: the fact that measure-pulses could not be annotated for classical pieces explains the performance difference between beat and measure-period performance. Tatum annotation could not be done reliably which partly explains the low rate.

5.2 Musicological models

Validity of the musicological models has been evaluated in [19] by applying the key signature model and bigrams to the transcription of single-voice melodies. A database was collected where 11 amateur singers perform folk songs by singing, humming, and whistling. The total size of the database is 120 minutes.

The seemingly easy task of transcribing these single-voice melodies is complicated by inaccuracies in the performance, note transitions, and the presence of vibrato. The key signature model reduced the rate of erroneously transcribed analysis frames from 20 % to 17 %. Using a bigram model dropped the error rate to 15 %. The two models together attained 13 % error rate [19].

6. ACKNOWLEDGEMENT

Timo Viitaniemi conducted the experiments on the transcription

of single-voice melodies using musicological models.

7. REFERENCES

- [1] Kashino, K., Nakadai, K., Kinoshita, T., and Tanaka, H. (1995). “Organization of Hierarchical Perceptual Sounds: Music Scene Analysis with Autonomous Processing Modules and a Quantitative Information Integration Mechanism,” Proc. International Joint Conf. on Artificial Intelligence, Montréal.
- [2] Martin, K. D. (1996a). “A Blackboard System for Automatic Transcription of Simple Polyphonic Music,” Massachusetts Institute of Technology Media Laboratory Perceptual Computing Section Technical Report No. 385.
- [3] Godsmark, D., and Brown, G. J. (1999). “A blackboard architecture for computational auditory scene analysis,” Speech Communication 27, 351–366.
- [4] Goto, M. (2000). “A robust predominant-F0 estimation method for real-time detection of melody and bass lines in CD recordings,” Proc. IEEE International Conf. on Acoust., Speech, and Signal Processing, Istanbul, Turkey.
- [5] Lerdahl, F., Jackendoff, R. *A Generative Theory of Tonal Music*. MIT Press, Cambridge, MA, 1983.
- [6] Large, Kolen. “Resonance and the perception of musical meter”. Connection science, 6 (1), 1994, pp. 177-208.
- [7] Temperley, D. *Cognition of Basic Musical Structures*. MIT Press, Cambridge, MA, 2001.
- [8] Dixon, S. “Automatic Extraction of Tempo and Beat from Expressive Performances,” *J. New Music Research* 30 (1), 2001, pp. 39-58.
- [9] Scheirer, E. D. “Tempo and beat analysis of acoustic musical signals,” *J. Acoust. Soc. Am.* 103 (1), 1998, pp. 588-601.
- [10] Goto, M., Muraoka, Y. . “Beat Tracking based on Multiple-agent Architecture — A Real-time Beat Tracking System for Audio Signals,” *In Proc. Second International Conference on Multiagent Systems*, pp.103–110, 1996.
- [11] Goto, M., Muraoka, Y. “Issues in Evaluating Beat Tracking Systems,” *IJCAI-1997 Workshop on Issues in AI and Music*.
- [12] Parncutt, R., “A Perceptual Model of Pulse Salience and Metrical Accent in Musical Rhythms,” *Music Perception*, Summer 1994, Vol. 11, No. 4, 409-464.
- [13] van Noorden, L., Moelants, D. “Resonance in the perception of musical pulse,” *Journal of New Music Res.*, 1999, Vol. 28.
- [14] Klapuri, A., Virtanen, T., Holm, J.–M. “Robust multipitch estimation for the analysis and manipulation of polyphonic musical signals,” in Proc. COST-G6 Conference on Digital Audio Effects, Verona, Italy, 2000.
- [15] Klapuri, A. “Multipitch estimation and sound separation by the spectral smoothness principle,” in Proc. *IEEE International Conf. on Acoust., Speech, and Signal Processing*, 2001.
- [16] Klapuri, A. “Automatic transcription of musical recordings,” in Proc. *Consistent and Reliable Acoustic Cues Workshop*, D. P. W. Ellis, M. Cooke (Chs.), Aalborg, Denmark, Sep.2001.
- [17] Krumhansl, C. *Cognitive Foundations of Musical Pitch*. Oxford University Press, 1990.
- [18] Essen Associative Code and Folksong Database (ESaC). <http://www.esac-data.org/>
- [19] Viitaniemi, T., Klapuri, Eronen, A. “A probabilistic model for the transcription of single-voice melodies,” Finnish Signal Processing Symposium, Tampere, Finland, 2003, *to appear*.