Matti P. Ryynänen and Anssi P. Klapuri
Department of Signal Processing
Tampere University of Technology
P.O. Box 553
FI-33101 Tampere, Finland
{matti.ryynanen, anssi.klapuri}@tut.fi

Automatic Transcription of Melody, Bass Line, and Chords in Polyphonic Music

This article proposes a method for the automatic transcription of the melody, bass line, and chords in polyphonic pop music. The method uses a frame-wise pitch-salience estimator as a feature extraction front-end. For the melody and bass-line transcription, this is followed by acoustic modeling of note events and musicological modeling of note transitions. The acoustic models include a model for the target notes (i.e., melody or bass notes) and a background model. The musicological model involves key estimation and note bigrams that determine probabilities for transitions between target notes. A transcription of the melody or the bass line is obtained using Viterbi search via the target and the background note models. The performance of the melody and the bass-line transcription is evaluated using approximately 8.5 hours of realistic polyphonic music. The chord transcription maps the pitch salience estimates to a pitch-class representation and uses trained chord models and chord-transition probabilities to produce a transcription consisting of major and minor triads. For chords, the evaluation material consists of the first eight Beatles albums. The method is computationally efficient and allows causal implementation, so it can process streaming audio.

Transcription of music refers to the analysis of an acoustic music signal for producing a parametric representation of the signal. The representation may be a music score with a meticulous arrangement for each instrument or an approximate description of melody and chords in the piece, for example. The latter type of transcription is commonly used in commercial songbooks of pop music and is usually sufficient for musicians or music hobbyists to play the piece. On the other hand, more detailed transcriptions are often employed in classical music to preserve the exact arrangement of the composer.

Computer Music Journal, 32:3, pp. 72–86, Fall 2008

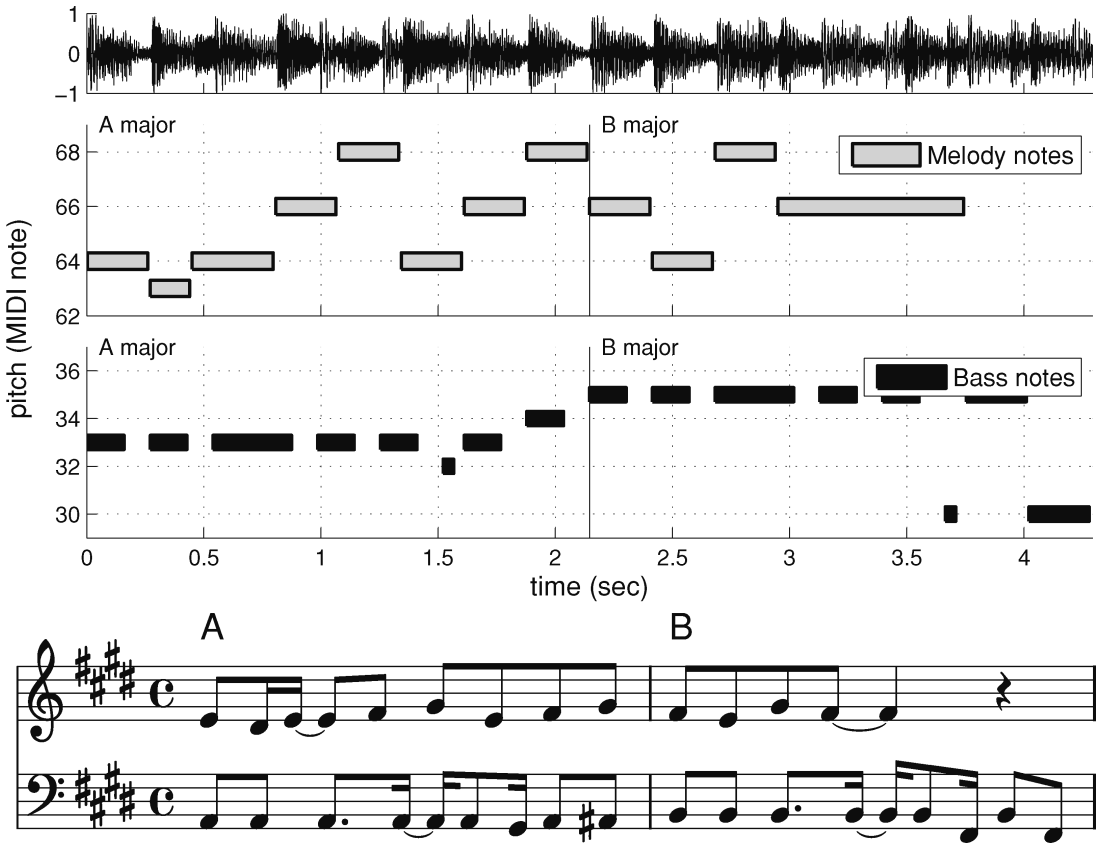
We propose a method for the automatic transcription of the melody, bass line, and chords in pop-music recordings. Conventionally, these tasks have been carried out by trained musicians who listen to a piece of music and write down notes or chords by hand, which is time-consuming and requires musical training. A machine transcriber enables several applications. First, it provides an easy way of obtaining a description of a music recording, allowing musicians to play it. Second, the produced transcriptions may be used in music analysis, music information retrieval (MIR) from large music databases, content-based audio processing, and interactive music systems, for example.

A *note* is here defined by a discrete pitch, an onset time, and duration. The *melody* of a piece is an organized sequence of consecutive notes and rests, usually performed by a lead singer or by a solo instrument. More informally, the melody is the part one often hums when listening to a piece. The *bass line* consists of notes in a lower pitch register and is usually played with a bass guitar, a double bass, or a bass synthesizer. A *chord* is a combination of notes that sound simultaneously or nearly simultaneously. In pop music, these concepts are usually rather unambiguous.

Figure 1 shows the waveform of an example music signal and two different representations of its melody, bass line, and chords. The middle panels show a piano-roll representation of the melody and the bass notes, respectively. Notes in this representation can be compactly saved in a MIDI file. The lowest panel represents the same notes and the chords in the common musical notation where the note onsets and durations are indicated by discrete symbols. The proposed method produces a piano-roll representation of the melody and the bass line together with chord labels. If desired, the note timings can be further quantized to obtain common music notation (Hainsworth 2006; Whiteley, Cemgil, and Godsill 2006).

Work on polyphonic music transcription dates back more than 30 years (Moorer 1977). Nowadays,

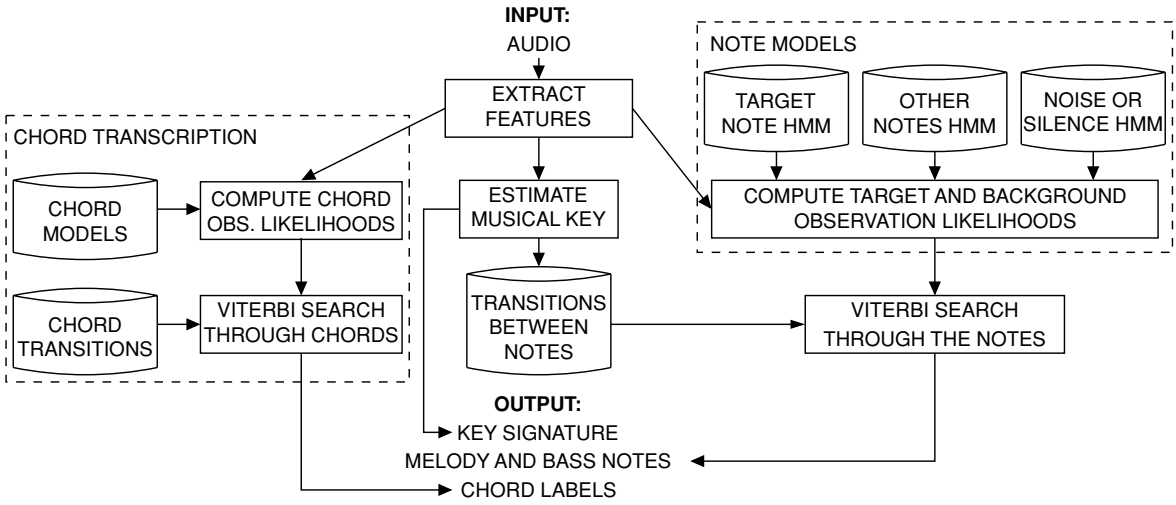
Figure 1. Given an audio recording (top panel), one can represent the melody, bass line, and chords with a piano roll (middle panels) and with a music score (bottom panel).



the concept of automatic music transcription includes several topics, such as multi-pitch analysis, beat tracking and rhythm analysis, transcription of percussive instruments, instrument recognition, harmonic analysis and chord transcription, and music structure analysis. For an overview of the topics, see Klapuri and Davy (2006). Fundamental frequency (F0) tracking of the melody and bass lines in polyphonic music signals was first considered by Goto (2000, 2004). Later, either F0 tracking or note-level transcription of the melody has been considered by Paiva, Mendes, and Cardoso (2005); Ellis and Poliner (2006); Dressler (2006); and Ryynänen and Klapuri (2006), and bass-line transcription has been addressed by Hainsworth and Macleod (2001) and Ryynänen and Klapuri (2007). Poliner et al. (2007) reported results on a comparative evaluation of melody transcription methods. Automatic chord transcription from audio has been considered by Sheh and Ellis (2003), Bello and Pickens (2005), and Harte and Sandler (2005).

Figure 2 shows a block diagram of the proposed method. An audio signal is processed frame-wise with two feature extractors: a pitch-salience estimator and an accent estimator that indicates potential note onsets based on signal energy. These features are used to compute observation likelihoods for target notes (i.e., melody or bass notes), other instrument notes, and silence or noise segments, each of which is modeled using a Hidden Markov Model (HMM). (See Rabiner and Juang 1993 for an introduction.) The musicological model estimates the musical key based on the pitch-salience function and then chooses between-note transition probabilities modeled with a note bigram. The Viterbi algorithm (Forney 1973) is used to find the optimal path through the target note models to produce a transcription. To summarize, the method

Figure 2. A block diagram of the proposed method.



incorporates both low-level acoustic modeling and high-level musicological modeling, and it produces discrete pitch labels and the beginning and ending times for the transcribed notes simultaneously.

The chord-transcription method uses a 24-state HMM consisting of twelve states for both major and minor triads. The observation likelihoods are computed by mapping the pitch saliences into a pitch-class representation and comparing them with trained profiles for major and minor chords. Probabilities of between-chord transitions are estimated from training data, and Viterbi decoding is used to find a path through the chord models. The rest of this article explains the proposed method in more detail and presents evaluation results with quantitative comparison with other methods.

Feature Extraction

The front-end of the method consists of two frame-wise feature extractors: a pitch-salience estimator and an accent estimator. Pitch salience $s_t(\tau)$ measures the strength of fundamental period τ in analysis frame t, and the accent a_t measures spectral change from frame $t-1$ to frame t, in practice indicating potential note onsets. Input signals are sampled at $f_s = 44.1$ kHz rate, and stereo signals are downmixed to mono before the feature extraction.

Pitch Salience Estimation

The salience, or strength, of each F0 candidate is calculated as a weighted sum of the amplitudes of its harmonic partials in a spectrally whitened signal frame. The calculations are similar to those of Klapuri (2006) and are briefly explained here.

Spectral whitening, or flattening, is first applied to suppress timbral information and thereby make the estimation more robust against variation in the sound sources. Given one frame of the input signal $x(n)$, the discrete Fourier transform $X(k)$ is calculated after Hamming-windowing and zero-padding the frame to twice its length. Then, a band-pass filterbank is simulated in the frequency domain. Center frequencies of the subbands are distributed uniformly on the critical-band scale, $f_c = 229(10^{(0.5c+1)/21.4} - 1)$, and each subband $c = 1,\ldots,60$ has a triangular power response extending from f_{c-2} to f_{c+2}. Power σ_c^2 of the signal within each subband c is calculated by applying the triangular response in the frequency domain and adding the resulting power spectrum values within the band. Then, bandwise compression coefficients $\gamma_c = \sigma_c^{\nu-1}$ are calculated, where $\nu = 0.33$

is a parameter determining the amount of spectral whitening applied. The coefficients γ_c are linearly interpolated between the center frequencies f_c to obtain compression coefficients $\gamma(k)$ for all frequency bins k. Finally, a whitened magnitude spectrum $|Y(k)|$ is obtained as $|Y(k)| = \gamma(k)|X(k)|$.

The salience $s(\tau)$ of a fundamental period candidate τ is then calculated as

$$s(\tau) = \sum_{i=1}^{I} g(\tau, i) \max_{k \in \kappa_{\tau,i}} |Y(k)| \quad (1)$$

where the set $\kappa_{\tau,i}$ defines a range of frequency bins in the vicinity of the partial number i of the F0 candidate f_s/τ and $I = 20$. More precisely,

$$\kappa_{\tau,i} = \{\langle iK/(\tau + \Delta\tau/2)\rangle, \ldots, \langle iK/(\tau - \Delta\tau/2)\rangle\} \quad (2)$$

where $\langle \bullet \rangle$ denotes rounding to the nearest integer, K is the length of the Fourier transform, and $\Delta\tau$ denotes the spacing between successive period candidates τ. We use $\Delta\tau = 0.5$, that is, a spacing corresponding to half the sampling interval. The partial weighting function $g(\tau, i)$ in Equation 1 is of the form

$$g(\tau, i) = \frac{f_s/\tau + \varepsilon_1}{if_s/\tau + \varepsilon_2} \quad (3)$$

where $\varepsilon_1 = 52\,\text{Hz}$ and $\varepsilon_2 = 320\,\text{Hz}$. Note that f_s/τ is the fundamental frequency corresponding to τ, and Equation 3 reduces to $1/i$ if the moderation terms ε_1 and ε_2 are omitted. For details, see Klapuri (2006).

The salience function in Equation 1 is calculated for F0 values between 35 Hz and 1.1 kHz in overlapping 92.9-msec frames, with a 23.2-msec interval between successive frames. Based on this, the differential salience of a particular period candidate τ is defined as $\Delta s_t(\tau) = s_t(\tau) - s_{t-1}(\tau)$. For convenience, the fundamental frequency candidates are expressed as unrounded MIDI note numbers by

$$F(\tau) = 69 + 12\log_2((f_s/\tau)/440) \quad (4)$$

Figure 3 shows the salience and the differential salience features extracted from the signal shown in Figure 1. Salience values indicate the melody notes quite clearly. Differential salience shows potential note onsets and vibrato, which is commonly applied in singing.

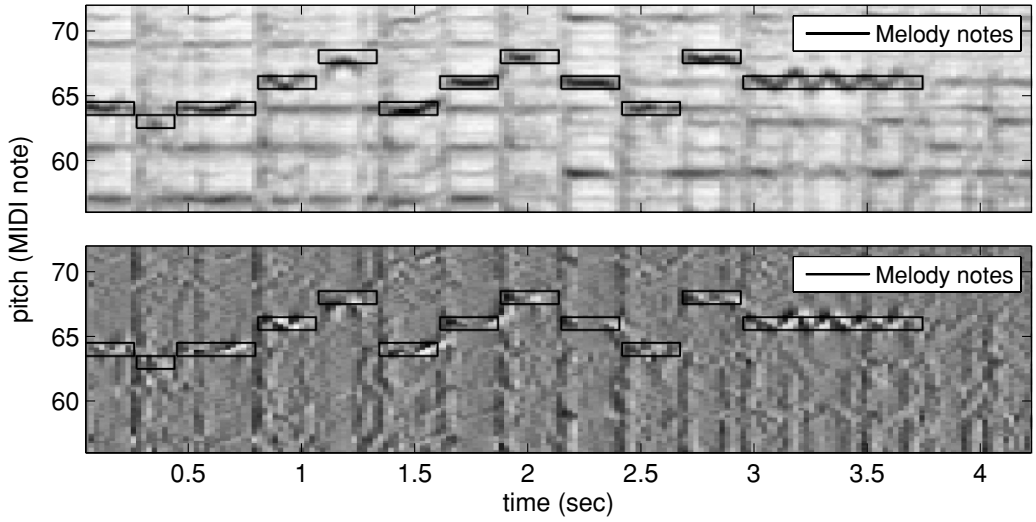
Accent Signal

Accent signal a_t measures the amount of incoming spectral energy in time frame t and is useful for detecting note beginnings. Calculation of the accent feature has been explained in detail in Klapuri, Eronen, and Astola (2006). Briefly, a "perceptual spectrum" is first calculated in analysis frame t by measuring log-power levels within critical bands. Then, the perceptual spectrum in frame $t-1$ is element-wise subtracted from that in frame t, and the resulting positive level differences are added across bands. As a result, we have a_t, which is a perceptually motivated measure of the amount of incoming spectral energy in frame t. The frame rate while calculating the accent signal is four times higher than that of the salience estimation; therefore, we down-sample the accent signal by selecting the maximum value in four-frame blocks to match the frame rate of the pitch-salience function.

Acoustic Modeling of Notes

The basic idea of acoustic modeling is that all possible note pitches n at all times are classified either as target notes (melody or bass), notes from other instruments, or as noise or silence. For this purpose, three acoustic models are trained: (1) target-notes model, (2) other-notes model, and (3) noise-or-silence model. Use of the target-notes and the other-notes models attempts to improve discriminability of the target sound source from other instruments. The target and the other notes are modeled with three-state left-to-right HMMs, where the consecutive states can be interpreted to represent the attack, sustain, and release segments of the notes. The noise-or-silence model is a three-state fully connected HMM, because no similar temporal order can be assumed for these segments. At all times, each candidate note n is in one of the internal states of one of the models. The notes are identified by their discrete MIDI note pitch $n \in \mathrm{N}$, where N is the set

Figure 3. Salience function $s_t(\tau)$ in the top panel and the differential salience function $\Delta s_t(\tau)$ in the bottom panel.



of possible pitches. The set consists of MIDI notes $\{44, \ldots, 84\}$, i.e., A-flat2 to C6, for the melody, and of notes $\{26, \ldots, 59\}$, i.e., D1–B3, for the bass line.

The acoustic models and their parameters do not depend on note pitch n. This has the advantage that only one set of HMM parameters must be trained for each of the three models. However, the observation vectors $\mathbf{o}_{n,t}$ are specific to each note. These are obtained from the extracted features by selecting the maximum-salience fundamental period $\hat{\tau}_{n,t}$ in a ± 1 semitone range around the note n in frame t:

$$\hat{\tau}_{n,t} = \arg\max_i s_t(i), \quad i \in \{\tau \mid |F(\tau) - n| \leq 1\} \qquad (5)$$

The observation vector $\mathbf{o}_{n,t}$ is then defined as

$$\mathbf{o}_{n,t} = [\Delta F, s_t(\hat{\tau}_{n,t}), \Delta s_t(\hat{\tau}_{n,t}), a_t]^{\mathrm{T}} \qquad (6)$$

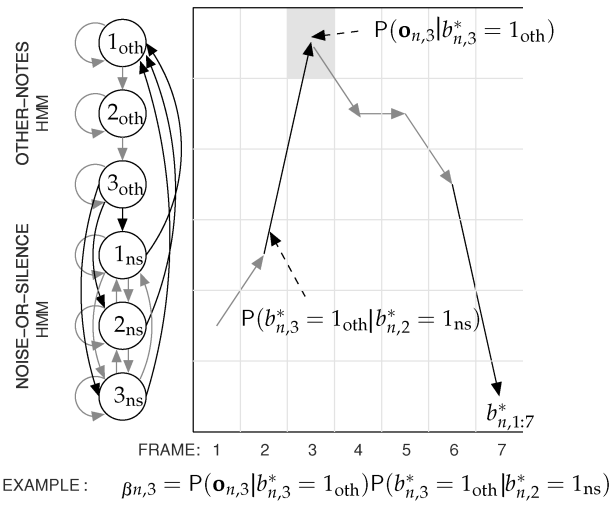
where $\Delta F = F(\hat{\tau}_{n,t}) - n$ is the distance between the pitch of the detected salience peak and the nominal pitch of the note, and $s(\hat{\tau}_{n,t})$ and $\Delta s(\hat{\tau}_{n,t})$ are the salience and the differential salience of $\hat{\tau}_{n,t}$. The accent value a_t does not depend on pitch but is common to all notes. Notice that the pitch-dependent features are obtained directly from the salience function in Equation 5. This is advantageous to the often-used approach of deciding a set of possible pitches within a frame already at the feature-extraction stage; here, the final decision of transcribed pitches is postponed for the probabilistic models, and the feature extraction becomes considerably simpler and computationally more efficient.

Training the Acoustic Models

The acoustic models are trained using the Real World Computing (RWC) database, which includes realistic musical recordings with manual annotations of the melody, the bass line, and the other instruments (Goto et al. 2002, 2003). For the time region of each annotated note n, the observation vectors by Equation 6 constitute a training sequence for either the target-notes or the other-notes model. The HMM parameters are then obtained using the Baum-Welch algorithm (Rabiner 1989) where observation-likelihood distributions are modeled with Gaussian Mixture Models (GMMs). Prior to the training, the features are normalized to have zero mean and unit variance over the training set.

The noise-or-silence model requires training sequences as well. Therefore, we generate random "note events" at positions of the time-pitch plane where there are no sounding notes in the reference annotation. Durations of the generated notes are sampled from a normal distribution with mean and variance calculated from the annotated notes in the song.

Figure 4. Background probability evaluation for one note. See text for details.



After the training, we have the following parameters for each HMM: observation likelihood distributions $P(\mathbf{o}_{n,t} \mid q_t = j)$ for states $j = 1, 2, 3$; state-transition probabilities $P(q_t = j \mid q_{t-1} = i)$, i.e., the probability that state i is followed by state j within a random state sequence $q_1 q_2 \ldots q_T \equiv q_{1:T}$; and initial state probabilities $P(q_1 = j)$. The observation likelihood distributions are modeled with diagonal covariance matrices and three or four GMM components for melody and bass model sets, respectively. In the following, we distinguish the states of the target-notes model, the other-notes model, and the noise-or-silence model by the notation j_{tgt}, j_{oth}, and j_{ns}, respectively.

Using the Acoustic Models

As already mentioned, the sequence of target notes is extracted by classifying all possible note pitches either as a target note, a note from another instrument, or as noise or silence. However, by definition there can be only up to one target note sounding at each time. This constraint is implemented as follows.

First, we find a state sequence $b^*_{n,1:T}$ which best explains the feature vectors $\mathbf{o}_{n,1:T}$ of note n using only the other-notes and the noise-or-silence model:

$$b^*_{n,1:T} = \arg\max_{q_{1:T}} \left[P(q_1) P(\mathbf{o}_{n,1} \mid q_1) \prod_{t=2}^{T} P(q_t \mid q_{t-1}) P(\mathbf{o}_{n,t} \mid q_t) \right] \tag{7}$$

Here, $q_t \in \{i_{\text{oth}}, j_{\text{ns}}\}$, where $i, j \in \{1, 2, 3\}$, meaning that the sequence $b^*_{n,1:T}$ may visit the states of both the other-notes and the noise-or-silence model. Because we have a combination of the two models, we must allow switching between them by defining nonzero probabilities for $P(q_t = j_{\text{oth}} \mid q_{t-1} = i_{\text{ns}})$, where $j = 1$ and $i \in \{1, 2, 3\}$, as well as for $P(q_t = j_{\text{ns}} \mid q_{t-1} = i_{\text{oth}})$ where $i = 3$ and $j \in \{1, 2, 3\}$. The state sequence is found with the Viterbi algorithm (Forney 1973). This procedure is repeated for all considered note pitches $n \in \mathrm{N}$.

Now we have the most likely explanation for all notes at all times without using the target-notes model. We define the "background" probability for note n in frame t as

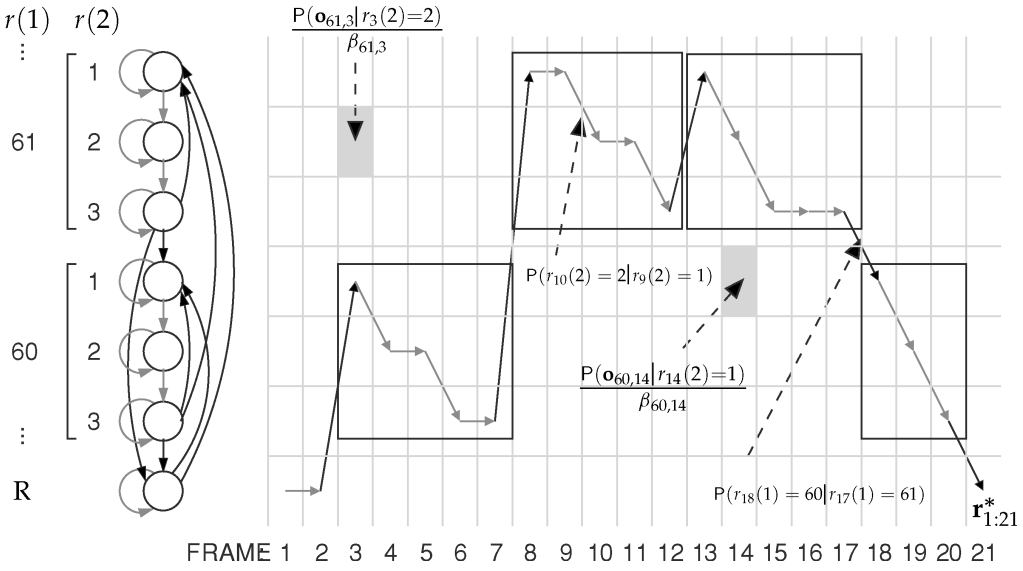
$$\beta_{n,t} = P(\mathbf{o}_{n,t} \mid b^*_{n,t}) P(b^*_{n,t} \mid b^*_{n,t-1}) \tag{8}$$

The use of the other notes and the noise-or-silence model is illustrated in Figure 4 where the continuous arrowed line shows the state sequence $b^*_{n,1:7}$ solved from Equation 7. The example shows the background probability $\beta_{n,3}$ evaluation in frame $t = 3$ using Equation 8.

Next, we try to find a path trough time for which the target-notes model gives a high likelihood and which is not well explained by the other models. The state space of this "target path" is larger than in Equation 7, because the path can visit the internal states j_{tgt} of any note n. We denote the state of the target path by a variable $\mathbf{r}_t$, which determines the note n and the internal state j_{tgt} at time t. More exactly, $\mathbf{r}_t = [r_t(1), r_t(2)]^{\mathrm{T}}$, where $r_t(1) \in \{\mathrm{N} \cup \mathrm{R}\}$ and $r_t(2) \in \{1, 2, 3\}$. Here R denotes a rest state, where no target notes are sounding. In this case, the value of $r_t(2)$ is meaningless.

To find the best overall explanation for all notes at all times, let us first assume that the notes are independent of each other. In this case, the overall probability given by the "background" model in frame t is $Z_t = \prod_n \beta_{n,t}$. When the target path state $\mathbf{r}_t$ visits a note n at time t, the overall probability

Figure 5. Using the target-note models to produce a transcription. See text for details.



becomes

$$Z_t' = \mathrm{P}(\mathbf{r}_t \mid \mathbf{r}_{t-1})\mathrm{P}(\mathbf{o}_{n,t} \mid r_t(2))\frac{Z_t}{\beta_{n,t}} \tag{9}$$

where $Z_t/\beta_{n,t} \equiv \prod_{i \neq n} \beta_{i,t}$. The best explanation for all the notes is given by a target path that maximizes $\prod_t Z_t'$. The factor Z_t in Equation 9 is independent of the target path and can be omitted while searching the optimal path. As a result, we can omit the independence assumption too, because it only concerns the calculation of Z_t. For rest frames, the background explanation applies for all notes, and thus $Z_t' = \mathrm{P}(\mathbf{r}_t \mid \mathbf{r}_{t-1})Z_t$. The transition probabilities $\mathrm{P}(\mathbf{r}_t \mid \mathbf{r}_{t-1})$ must still remain in all frames, because they include the musicological model that controls transitions between notes and between notes and rests.

The best target path is then obtained as

$$\mathbf{r}^*_{1:T} = \arg\max_{\mathbf{r}_{1:T}} \left[\mathrm{P}(\mathbf{r}_1) f(\mathbf{r}_1) \prod_{t=2}^{T} \mathrm{P}(\mathbf{r}_t \mid \mathbf{r}_{t-1}) f(\mathbf{r}_t) \right] \tag{10}$$

where

$$f(\mathbf{r}_t) = \begin{cases} \mathrm{P}(\mathbf{o}_{n,t} \mid r_t(2))/\beta_{n,t}, & r_t(1) = n \in \mathrm{N} \\ 1, & r_t(1) = \mathrm{R} \end{cases} \tag{11}$$

The transition probabilities $\mathrm{P}(\mathbf{r}_t \mid \mathbf{r}_{t-1})$ are defined as

$$\mathrm{P}(\mathbf{r}_t \mid \mathbf{r}_{t-1}) = \begin{cases} \mathrm{P}(r_t(2) = j_{\mathrm{tgt}} \mid r_{t-1}(2) = i_{\mathrm{tgt}}), \\ \quad \text{when inside target HMMs} \\ \mathrm{P}(r_t(1) = u \mid r_{t-1}(1) = u'), \\ \quad \text{when applying musical model} \end{cases} \tag{12}$$

Above, the acoustic model is applied when staying within the same target-note HMM, that is, when $r_t(1) = r_{t-1}(1) \in \mathrm{N}$ and $r_t(2) \geq r_{t-1}(2)$. The musicological model is applied when switching between notes or between notes and rests, that is, when the above condition is not fulfilled, and $u, u' \in \{\mathrm{N} \cup \mathrm{R}\}$. The musicological model is explained in the next section. The path by Equation 10 is found with the Viterbi algorithm. This simultaneously produces the discrete pitch labels and note segmentation, i.e., the note onsets and offsets. A note starts when the path enters the first state of a note model and ends when the path exits its last state. Rests are produced where $r^*_t(1) = \mathrm{R}$.

Figure 5 illustrates this process with two target-note HMMs for MIDI notes 60 and 61 and the rest state. The arrowed line shows the best target path $\mathbf{r}^*_{1:21}$ where the gray arrows show the transitions inside target-note models and the black arrows show the transitions where the musicological model is

applied. The black boxes show the four transcribed notes (i.e., MIDI notes 60, 61, 61, 60) with their onsets and offsets. The figure also shows example calculations for Equations 11–12.

Musicological Model for Note Transitions

Musical context plays an important role in both melodies and bass lines. Consider a target note sequence in the key of C major in which the note E is possibly followed by F-sharp or G, for which the acoustic model gives approximately equal likelihoods. Because F-sharp follows E less often in this key, the note G is musically preferred. We use this feature in the melody and bass-line transcription by employing key-dependent note-transition probabilities to solve ambiguous situations.

Key Estimation

A musical key is roughly defined by the basic scale used in a piece. A major key and a minor key are here referred to as a *relative-key pair* if they consist of scales with the same notes (e.g., C major and A minor). The relative-key pair also specifies the key signature of a musical score.

The proposed key estimator finds the relative-key pair using the pitch salience function $s_t(\tau)$. First, the function values are mapped into a pitch-class representation, where notes in different octaves are considered equivalent. The set of notes that belong to a pitch class $m \in \{0,1,\ldots,11\}$ is defined by $\mathrm{H}_m = \{n \in \mathrm{N} \mid \mathrm{mod}\,(n, 12) = m\}$, where $\mathrm{mod}(x, y) \equiv x - y\lfloor x/y \rfloor$. For key estimation, the note range N consists of the MIDI notes $\{48,\ldots,83\}$. The salience function is mapped to the pitch-class profile $\mathrm{PCP}_t(m)$ by

$$\mathrm{PCP}_t(m) = \frac{1}{W_t} \sum_{n \in \mathrm{H}_m} \max_i s_t(i), \;\; i \in \{\tau \,||\, F(\tau) - n| \leq 0.5\} \tag{13}$$

where W_t is a normalization factor so that $\sum_{\mathrm{m}=0}^{11} \mathrm{PCP}_t(m) = 1$. The pitch-class profile in each frame is then compared to key profiles $K_{\mathrm{maj}}(d)$ and $K_{\mathrm{min}}(d)$, which give the occurrence frequencies of *note degrees* $d \in \{0,1,\ldots,11\}$ in major and minor keys, respectively. For example, $d = 0$ is the tonic note of the key and $d = 7$ is the perfect fifth.

Let $k \in \{0,1,\ldots,11\}$ denote the relative-key pairs [C major/A minor], [D-flat major/B-flat minor], and so forth, until the pair [B major/G-sharp minor], respectively. Given the pitch-class profile $\mathrm{PCP}_t(m)$ and the key profiles, the most probable relative-key pair k_t^* at time t is calculated by

$$\begin{aligned} \mathrm{L}_t(k) = \sum_{d=0}^{11} [&\log[K_{\mathrm{maj}}(d)]\,\mathrm{PCP}_t(\mathrm{mod}(d + k, 12)) \\ &+ \log[K_{\mathrm{min}}(d)]\,\mathrm{PCP}_t(\mathrm{mod}(d + k + 9, 12))], \end{aligned} \tag{14}$$

$$k_t^* = \arg\max_k \left[\sum_{j=1}^{t} \mathrm{L}_j(k) \right]. \tag{15}$$

When calculating note degree for minor keys, the term +9 in Equation 14 shifts the key index k to the relative minor key. As key profiles $K_{\mathrm{maj}}(d)$ and $K_{\mathrm{min}}(d)$, we use those reported by Krumhansl (1990, p. 67).

Note Bigrams

Note bigrams determine the probability of making a transition between notes or between notes and the rest state. This aims at favoring note sequences characteristic to the target material. In Equation 12, these transitions were denoted by $\mathrm{P}(r_t(1) = u \mid r_{t-1}(1) = u')$, where $u, u' \in \{\mathrm{N} \cup \mathrm{R}\}$. Because the target note models are left-to-right HMMs, it is possible to enter or exit a target note HMM only via its first or last state, respectively. In addition, our system includes transitions from target-note models to the rest state and vice versa, as well as the probability to stay within the rest state.

Given the relative-key pair k, a transition from note u' to note u is specified by the degree of the first note, $\mathrm{mod}(u' - k, 12)$ and the interval between the two notes $u - u'$. For note-to-rest and rest-to-note

transitions, only the note degree matters. The probability to stay within the rest state does not depend on the key but is a free parameter that can be used to control the amount of rests in the transcription. In summary, the transitions given by the musicological model are defined by

$$\mathrm{P}(r_t(1) = u \mid r_{t-1}(1) = u') = \begin{cases} \mathrm{P}(\mathbf{r}_t = [u,1]^{\mathrm{T}} \mid \mathbf{r}_{t-1} = [u',3]^{\mathrm{T}}, k), & \text{note to note} \\ \mathrm{P}(\mathbf{r}_t = [u,1]^{\mathrm{T}} \mid r_{t-1}(1) = \mathrm{R}, k), & \text{rest to note} \\ \mathrm{P}(r_t(1) = \mathrm{R} \mid \mathbf{r}_{t-1} = [u',3]^{\mathrm{T}}, k), & \text{note to rest} \\ \mathrm{P}(r_t(1) = \mathrm{R} \mid r_{t-1}(1) = \mathrm{R}), & \text{rest to rest} \end{cases} \tag{16}$$

For melody notes, the note bigram probabilities were estimated by counting the note transitions in the Essen Associative Code and Folksong Database (EsAC) with about 20,000 folksongs with key information (see www.esac-data.org). Krumhansl's profiles were used to give probabilities for rest-to-note and note-to-rest transitions. This is based on the assumption that it is more probable to start or end a phrase with a note that is harmonically "stable" in the estimated key.

For bass lines, the note bigram estimation is done from a collection of over 1,300 MIDI files including bass lines. For each file, we estimate the relative-key pair using the Krumhansl profiles and all MIDI notes from the file. A rest is added between two consecutive bass notes if they are separated by more than 200 msec. Then, we count the key-dependent note transitions similarly to the melody notes. The bigrams for both melody and bass lines are smoothed with the Witten-Bell discounting algorithm (Witten and Bell 1991).

The note bigrams do not take into account the absolute pitch of the notes but only the interval between them. However, it is advantageous to prefer target notes in the typical pitch range of melodies or bass lines: for example, preferring low-pitched notes in bass-line transcription. We implement this by weighting the note bigram probabilities with values from a normal distribution over notes n. For melodies, we found a distribution mean at MIDI note value 62.6 and a variance of 6.8 semitones to perform best in our simulations. For bass lines, we found mean 33.2 and variance 3.7.

Chord Transcription

The proposed chord-transcription method attempts to label each frame with a certain major or minor chord. The reason for introducing the method here is to demonstrate the usefulness of the pitch-salience function in the chord-transcription task. In addition, chords complement the melody and the bass line transcriptions to produce a useful representation of the song under analysis.

Harte et al. (2005) proposed a text-based chord syntax and publicly provided the chord annotations for the first eight albums by the Beatles. This database is used here for training and evaluation of the proposed method. The database also includes complex chords, such as major and dominant-seventh and extended chords (ninths, sixths), which are treated here as major or minor triads. Augmented or diminished chords are left out from the training. Database details are given in the next section.

The chord transcription method is the following. First, we train pitch-class profiles for major and minor triads. The salience function is mapped into two pitch-class profiles by using Equation 13: $\mathrm{PCP}_t^{\mathrm{lo}}(m)$ for low-register MIDI notes $n \in \{26,\ldots,49\}$, and $\mathrm{PCP}_t^{\mathrm{hi}}(m)$ for high-register MIDI notes $n \in \{50,\ldots,73\}$. For each major chord in the training data, $\mathrm{PCP}_t^{\mathrm{lo}}(m)$ and $\mathrm{PCP}_t^{\mathrm{hi}}(m)$ are calculated, the profiles are rotated so that the pitch-class $m = 0$ corresponds to the chord root note, and the profiles are then obtained by averaging over time. A similar procedure is repeated for the minor chords. Figure 6 shows the estimated low-register profiles $C_{\mathrm{maj}}^{\mathrm{lo}}$, $C_{\mathrm{min}}^{\mathrm{lo}}$ and high-register profiles $C_{\mathrm{maj}}^{\mathrm{hi}}$, $C_{\mathrm{min}}^{\mathrm{hi}}$ for both major and minor chords. The low-register profile captures the bass notes contributing to the chord root, whereas the high-register profile has more clear peaks also for the major or minor third and the fifth.

Next, we define a chord HMM with 24 states (twelve states for both major and minor triads). In frame t, a state in this HMM is denoted by $\mathbf{c}_t = [c_t(1), c_t(2)]^{\mathrm{T}}$, where $c_t(1) \in \{0,1,\ldots,11\}$ denotes

Figure 6. Major and minor chord profiles for low and high registers.

Figure 7. The estimated chord transition probabilities from major and minor chords. The text boxes show examples for transitions from a C-major chord and an A-minor chord.

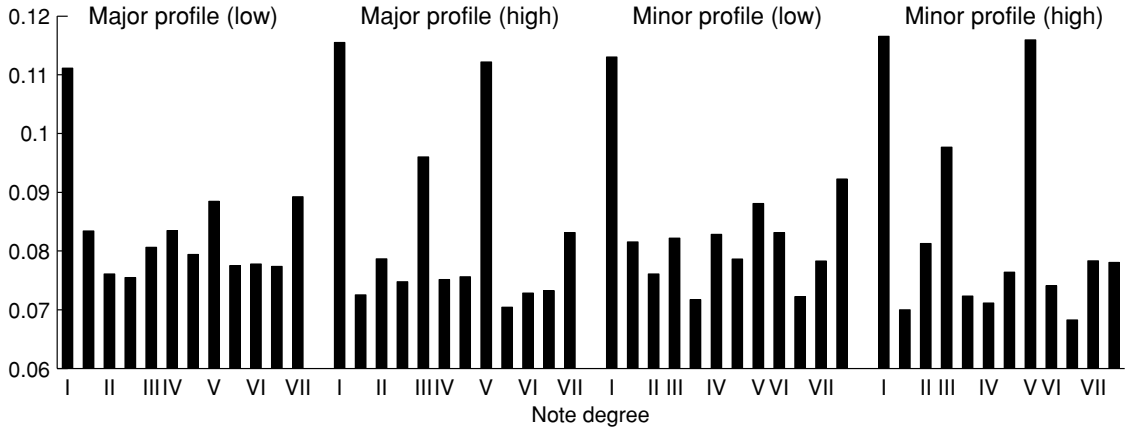


Figure 6

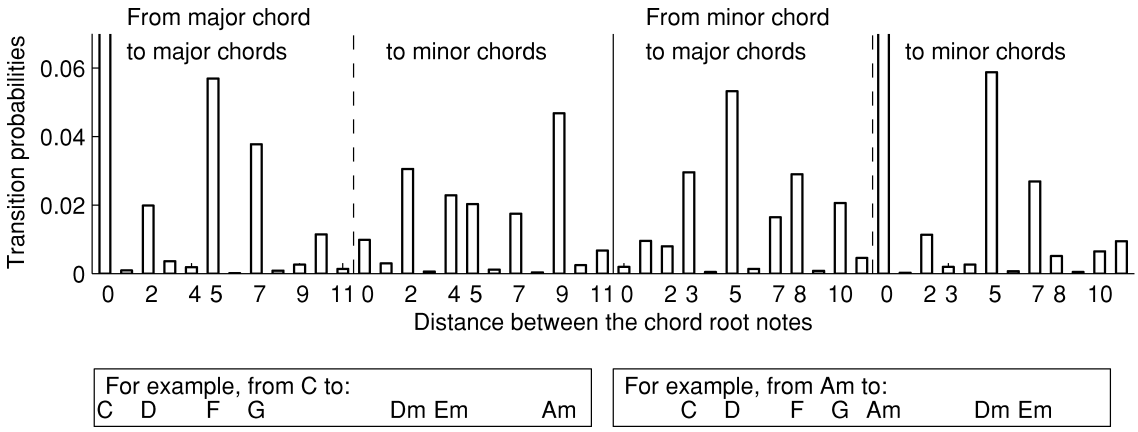


Figure 7

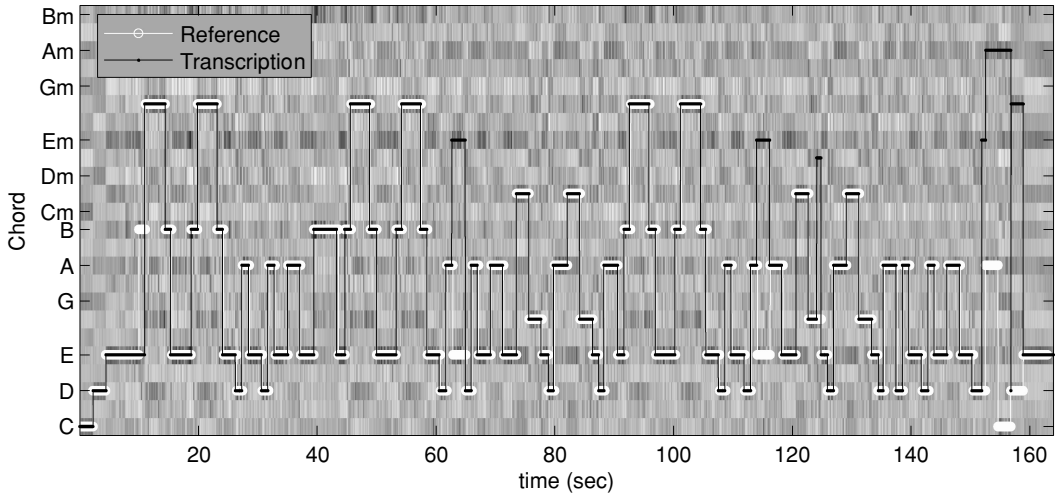
the chord roots C, D-flat, and so forth, and $c_t(2) \in \{\text{maj, min}\}$ denotes the chord type. We want to find a path $\mathbf{c}_{1:T}$ through the chord HMM. For the twelve major chord states $\mathbf{c}_t$, $c_t(2) = \text{maj}$, the chord observation log-likelihoods $\mathrm{L}(\mathbf{c}_t)$ are calculated in a manner analogous to the key estimation:

$$\mathrm{L}(\mathbf{c}_t) = \sum_{d=0}^{11} \left[\log\left[C^{\text{lo}}_{\text{maj}}(d)\right]\text{PCP}^{\text{lo}}_t(\text{mod}(d + c_t(1), 12)) + \log\left[C^{\text{hi}}_{\text{maj}}(d)\right]\text{PCP}^{\text{hi}}_t(\text{mod}(d + c_t(1), 12))\right]. \quad (17)$$

The calculation is exactly similar for the minor triads $\mathrm{L}(\mathbf{c}_t)$, $c_t(2) = \text{min}$, except that the major profiles are replaced with the minor profiles.

We also train a chord-transition bigram $\mathrm{P}(\mathbf{c}_t \mid \mathbf{c}_{t-1})$. The transitions are independent of the key so that only the chord type and the distance between the chord roots, $\text{mod}(c_t(1) - c_{t-1}(1),12)$, matters. For example, a transition from A minor, $\mathbf{c}_{t-1} = [9, \text{min}]^{\text{T}}$, to F major, $\mathbf{c}_t = [5, \text{maj}]^{\text{T}}$, is counted as a transition from a minor chord to a major chord with distance $\text{mod}(5 - 9,12) = 8$. Figure 7 illustrates the estimated chord-transition probabilities. The probability to

Figure 8. Chord transcription of With a Little Help From My Friends *by the Beatles. Likelihoods from Equation 17 are shown for all major and minor triads. The reference chords are indicated by the white line and the transcription* $\mathbf{c}^*_{1:T}$ *by the black line.*



stay in the chord itself is a free parameter that controls the amount of chord changes.

Now, we have defined the log-likelihoods $\mathrm{L}(\mathbf{c}_t)$ for all chords in each frame t and the chord-transition bigram $\mathrm{P}(\mathbf{c}_t \mid \mathbf{c}_{t-1})$. The chord transcription is then obtained by finding an optimal path $\mathbf{c}^*_{1:T}$ through the chord states:

$$\mathbf{c}^*_{1:T} = \arg\max_{\mathbf{c}_{1:T}} \left[\mathrm{L}(\mathbf{c}_1) + \sum_{t=2}^{T} (\mathrm{L}(\mathbf{c}_t) + \log \mathrm{P}(\mathbf{c}_t \mid \mathbf{c}_{t-1})) \right], \tag{18}$$

which is again found using the Viterbi algorithm. The initial probabilities for chord states are uniform and are therefore omitted. The method does not detect silent segments but produces a chord label in each frame. Figure 8 shows the chord transcription for *With a Little Help From My Friends* by the Beatles.

Results

The proposed melody-, bass-, and chord-transcription methods are quantitatively evaluated using databases described subsequently. For all evaluations, a two-fold cross validation is used. With a C++ implementation running on a 3.2 GHz Pentium 4 processor, the entire method takes about 19 sec to process 180 seconds of stereo audio. The feature extraction takes about 12 sec, and the melody and bass-line transcription take about 3 sec each. The key estimation and chord transcription take less than 0.1 seconds. In addition, the method allows causal implementation to process streaming audio in a manner described in Ryynänen and Klapuri (2007).

Audio Databases

For the development and evaluation of the melody and bass-line transcription, we use the Real World Computing "popular music" and "genre" databases (Goto et al. 2002, 2003). The databases include a MIDI file for each song, which contains a manual annotation of the melody, the bass, and other instrument notes, collectively referred to as the reference notes. MIDI notes for drums, percussive instruments, and sound effects are excluded from the reference notes. Some songs in the databases were not used due to an unreliable synchronization of the MIDI annotation and the audio recording. Also, some songs do not include the melody or the bass line. Consequently, we used 130 full acoustic recordings for melody transcription: 92 pop songs (the RWC "popular" database) and 38 songs with varying styles (the RWC "genre" database). For bass-line

Table 1. Melody and Bass-Line Transcription Results (%)

Melody	*R*	*P*	*F*	$\bar{\rho}$	*Bass line*	*R*	*P*	*F*	$\bar{\rho}$
RWC "popular"	60.5	49.4	53.8	61.1	RWC "popular"	57.7	57.5	56.3	61.9
RWC "genre"	41.7	50.3	42.9	55.8	RWC "genre"	35.3	57.5	39.3	57.6
Total	55.0	49.6	50.6	59.6	Total	50.1	57.5	50.6	60.4

transcription, we used 84 songs from RWC "popular" and 43 songs from RWC "genre"—altogether, 127 recordings. This gives approximately 8.7 and 8.5 hours of music for the evaluation of melody and bass-line transcription, respectively. There are reference notes outside the reasonable transcription note range for both melody (<0.1 percent of the melody reference notes) and bass lines (1.8 percent of the notes). These notes are not used in training but are counted as transcription errors in testing.

As already mentioned, the chord-transcription method is evaluated using the first eight Beatles albums with the chord annotations provided by Harte and colleagues. The albums include 110 songs with approximately 4.6 hours of music. The reference major and minor chords cover approximately 75 percent and 20 percent of the audio, respectively. Chords that are not recognized by our method and the no-chord segments cover about 3 percent and 1 percent of the audio.

Melody and Bass-Line Transcription Results

The performance of melody and bass-line transcription is evaluated by counting correctly and incorrectly transcribed notes. We use the recall rate R and the precision rate P defined by

$$R = \frac{\#(\text{correctly transcribed notes})}{\#(\text{reference notes})}$$
$$P = \frac{\#(\text{correctly transcribed notes})}{\#(\text{transcribed notes})} \quad (19)$$

A reference note is correctly transcribed by a note in the transcription if their MIDI note numbers are equal, the absolute difference between their onset times is less than 150 msec, and the transcribed note is not already associated with another reference note. We use the F-measure $F = 2RP/(R+P)$ to give an overall measure of performance. Temporal overlap ratio of a correctly transcribed note with the associated reference note is measured by $\rho = (\min\{E\} - \max\{B\})/(\max\{E\} - \min\{B\})$, where sets B and E contain the beginning and ending times of the two notes, respectively. The mean overlap ratio $\bar{\rho}$ is obtained by averaging ρ values over the correctly transcribed notes. The recall rate, the precision rate, the F-measure, and the mean overlap ratio are calculated separately for each recording, and then the average over all the recordings is reported.

Table 1 shows the melody and bass-line transcription results. Both the melody and the bass-line transcription achieve over 50 percent average F-measure. The performance on pop songs is clearly better than for the songs from various genres. This was expected since the melody and bass lines are usually more prominent in pop music than in other genres, such as heavy rock or dance music, for example. In addition, the RWC "popular" database includes only vocal melodies, whereas the RWC "genre" database also includes melodies performed with other instruments. The musicological model plays an important role in the method: The total F-measures drop to 40 percent for both melody and bass-line transcription if the note bigrams are replaced with uniform distributions.

For comparison, Ellis and Poliner kindly provided the pitch tracks produced by their melody-transcription method (Ellis and Poliner 2006) for the recordings in the RWC databases. Their method decides note pitch for the melody in each frame whenever the frame is judged to be voiced. Briefly, the pitch classification in each frame is conducted using one-versus-all, linear-kernel support vector machine (SVM). The voiced-unvoiced decision is based on energy thresholding, and the pitch track is

Table 2. Frame-Level Melody Transcription Results (%)

Method	*Data*	*Overall acc.*	*Raw pitch*	*Vc det.*	*Vc FA*	Vc *d′*
Proposed	RWC "popular"	63.0	63.6	87.0	39.4	1.40
	RWC "genre"	62.4	40.5	58.8	17.9	1.14
	Total	**62.8**	56.9	78.7	33.1	**1.23**
Ellis and Poliner	RWC "popular"	42.9	50.4	93.4	66.3	1.09
	RWC "genre"	38.5	42.8	91.6	64.7	1.00
	Total	**41.6**	48.2	92.9	65.9	**1.06**

Table 3. Chord Transcription Results, Frame Classification Proportions Averaged Over the Songs (%)

	Proposed	*Bello-Pickens*
Exactly correct	70.6	69.6
Root correct (major and minor are confused)	5.7	3.4
Relative chord (e.g., C was labeled as Am or vice versa)	2.7	5.8
Dominant (e.g., C/Cm was labeled as G/Gm)	4.5	4.0
Subdominant (e.g., C/Cm was labeled as F/Fm)	3.7	2.2
III or VI (e.g., C was labeled with Em or vice versa)	1.6	2.4
Other errors	6.4	7.7
Chord not recognized by the methods (e.g., C dim)	4.8	4.8

smoothed using HMM post-processing. Because the Ellis-Poliner method does not produce segmented note events but rather a pitch track, we compare the methods using frame-level evaluation metrics adopted for melody-extraction evaluation in the Music Information Retrieval Evaluation eXchange (Poliner et al. 2007). For this, the RWC reference MIDI note values are sampled every 10 msec to obtain a frame-level reference. Similar conversion is made for the melody notes produced by our method.

Table 2 shows the results for the proposed method and for the Ellis-Poliner method on the RWC databases. The overall accuracy denotes the proportion of frames with either a correct pitch label or correct unvoiced decision, where the pitch label is correct if the absolute difference between the transcription and the reference is less than half a semitone. The raw pitch accuracy denotes the proportion of correct pitch labels to voiced frames in the reference. Voicing detection (Vc det) measures the proportion of correct voicing in transcription to voiced frames in the reference. Voicing false alarm (Vc FA) measures the proportion of frames that are labeled as voiced in the transcription but are unvoiced in the reference. The voicing *d′* (Vc *d′*) combines the voicing detection and voicing false alarm rates to describe the system's ability to discriminate the voiced and unvoiced frames. High voicing detection and low voicing false alarm give good discriminability with a high voicing *d′* value (Duda, Hart, and Stork 2001). According to the overall accuracy and the voicing *d′*, the proposed method outperforms the Ellis-Poliner method in this evaluation. Their method classifies most of the frames as voiced, resulting in a high voicing-detection rate but also high voicing false-alarm rates.

Chord Transcription Results

The chord-transcription method is evaluated by comparing the transcribed chords with the reference chords frame-by-frame. For method comparison, Bello and Pickens kindly provided outputs of their chord transcription method (Bello and Pickens 2005) on the Beatles data. As a framework, their method

also uses a chord HMM where states correspond to major and minor triads. Our method resembles theirs in this sense, but the methods differ in details.

The results with error analysis are given in Table 3. On the average over the songs, both methods decide exactly correct chord in about 70 percent of the frames, that is, the transcribed chord root and type are the same as in the reference. The proposed method makes more major/minor chord confusions, whereas the Bello-Pickens method more often labels a reference chord with its relative chord. Because neither of the methods detects augmented or diminished chords, or no-chord segments, those frames are always treated as transcription errors (4.8 percent of all the frames).

Conclusions

We proposed a method for the automatic transcription of melody, bass line, and chords in polyphonic music. The method consists of frame-wise pitch-salience estimation, acoustic modeling, and musicological modeling. The transcription accuracy was evaluated using several hours of realistic music, and direct comparisons to state-of-the-art methods were provided. Using quite straightforward time quantization, common musical notation such as that shown in Figure 1 can be produced. In addition, the statistical models can be easily retrained for different target materials. Future work includes using timbre and metrical analysis to improve melody and bass-line transcription, and a more detailed chord analysis method.

The transcription results are already useful for several applications. The proposed method has been integrated into a music-transcription tool with a graphical user interface and MIDI editing capabilities, and the melody transcription has been successfully applied in a query-by-humming system (Ryynänen and Klapuri 2008), for example. Only a few years ago, the authors considered the automatic transcription of commercial music recordings as a very difficult problem. However, rapid development of transcription methods and the latest results have demonstrated that feasible solutions are possible. We believe that the proposed method with melody, bass line, and chord transcription takes a step toward more complete and accurate music transcription. Audio examples of the transcriptions are available online at www.cs.tut.fi/sgn/arg/matti/demos/mbctrans.

Acknowledgments

This work was supported by the Academy of Finland, project no. 5213462 (Finnish Centre of Excellence Program 2006–2011). The authors would like to thank Dan Ellis, Graham Poliner, and Juan Pablo Bello for providing outputs of their transcription methods for comparison.

References

Bello, J. P., and J. Pickens. 2005. "A Robust Mid-level Representation for Harmonic Content in Music Signals." *Proceedings of the 6th International Conference on Music Information Retrieval.* London: Queen Mary, University of London, pp. 304–311.

Dressler, K. 2006. "An Auditory Streaming Approach on Melody Extraction." *MIREX Audio Melody Extraction Contest Abstracts.* MIREX06 extended abstract. London: Queen Mary, University of London. Available online at www.music-ir.org/evaluation/MIREX/2006_abstracts/AME_dressler.pdf.

Duda, R. O., P. E. Hart, and D. G. Stork. 2001. *Pattern Classification*, 2nd ed. New York: Wiley.

Ellis, D., and G. Poliner. 2006. "Classification-Based Melody Transcription." *Machine Learning Journal* 65(2–3):439–456.

Forney, G. D. 1973. "The Viterbi Algorithm." *Proceedings of the IEEE* 61(3):268–278.

Goto, M. 2000. "A Robust Predominant-F0 Estimation Method for Real-Time Detection of Melody and Bass Lines in CD Recordings." *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*. New York: Institute for Electrical and Electronics Engineers, pp. 757–760.

Goto, M. 2004. "A Real-Time Music-Scene-Description System: Predominant-F0 Estimation for Detecting Melody and Bass Lines in Real-World Audio Signals." *Speech Communication* 43(4):311–329.

Goto, M., et al. 2002. "RWC Music Database: Popular, Classical, and Jazz Music Databases." *Proceedings of*

the 3rd International Conference on Music Information Retrieval. Paris: IRCAM, pp. 287–288.
Goto, M., et al. 2003. "RWC Music Database: Music Genre Database and Musical Instrument Sound Database." *Proceedings of the 4th International Conference on Music Information Retrieval*. Baltimore, Maryland: Johns Hopkins University. Available online at ismir2003.ismir.net/papers/Goto1.PDF.
Hainsworth, S. 2006. "Beat Tracking and Musical Metre Analysis." In A. Klapuri and M. Davy, eds. *Signal Processing Methods for Music Transcription*. Berlin: Springer, pp. 101–129.
Hainsworth, S. W., and M. D. Macleod. 2001. "Automatic Bass Line Transcription from Polyphonic Music." *Proceedings of the 2001 International Computer Music Conference*. San Francisco, California: International Computer Music Association, pp. 431–434.
Harte, C., et al. 2005. "Symbolic Representation of Musical Chords: A Proposed Syntax for Text Annotations." *Proceedings of the 6th International Conference on Music Information Retrieval*. London: Queen Mary, University of London, pp. 66–71.
Harte, C. A., and M. B. Sandler. 2005. "Automatic Chord Identification Using a Quantised Chromagram." *Proceedings of the 118th Audio Engineering Society Convention*. New York: Audio Engineering Society, paper number 6412.
Klapuri, A. 2006. "Multiple Fundamental Frequency Estimation by Summing Harmonic Amplitudes." *Proceedings of the 7th International Conference on Music Information Retrieval*. Victoria: University of Victoria, pp. 216–221.
Klapuri, A., and M. Davy, eds. 2006. *Signal Processing Methods for Music Transcription*. Berlin: Springer.
Klapuri, A. P., A. J. Eronen, and J. T. Astola. 2006. "Analysis of the Meter of Acoustic Musical Signals." *IEEE Transactions on Audio, Speech, and Language Processing* 14(1):342–355.
Krumhansl, C. 1990. *Cognitive Foundations of Musical Pitch*. Oxford: Oxford University Press.
Moorer, J. A. 1977. "On the Transcription of Musical Sound by Computer." *Computer Music Journal* 1(4): 32–38.
Paiva, R. P., T. Mendes, and A. Cardoso. 2005. "On the Detection of Melody Notes in Polyphonic Adio." *Proceedings of the 6th International Conference on Music Information Retrieval*. London: Queen Mary, University of London, pp. 175–182.
Poliner, G., et al. 2007. "Melody Transcription from Music Audio: Approaches and Evaluation." *IEEE Transactions on Audio, Speech, and Language Processing* 15(4):1247–1256.
Rabiner, L., and B.-H. Juang. 1993. *Fundamentals of Speech Recognition*. Upper Saddle River, New Jersey: Prentice Hall.
Rabiner, L. R. 1989. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." *Proceedings of the IEEE* 77(2):257–289.
Ryynänen, M., and A. Klapuri. 2006. "Transcription of the Singing Melody in Polyphonic Music." *Proceedings of the 7th International Conference on Music Information Retrieval*. Victoria: University of Victoria, pp. 222–227.
Ryynänen, M., and A. Klapuri. 2007. "Automatic Bass Line Transcription from Streaming Polyphonic Audio." *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing*. New York: Institute of Electrical and Electronics Engineers, pp. 1437–1440.
Ryynänen, M., and A. Klapuri. 2008. "Query by Humming of MIDI and Audio Using Locality Sensitive Hashing." *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Available online at www.cs.tut.fi/~mryynane/ryynanen_icassp08.pdf.
Sheh, A., and D. P. Ellis. 2003. "Chord Segmentation and Recognition Using EM-Trained Hidden Markov Models." *Proceedings of the 4th International Conference on Music Information Retrieval*. Baltimore, Maryland: Johns Hopkins University. Available online at ismir2003.ismir.net/papers/Sheh.PDF.
Whiteley, N., A. T. Cemgil, and S. Godsill. 2006. "Bayesian Modeling of Temporal Structure in Musical Audio." *Proceedings of the 7th International Conference on Music Information Retrieval*. Victoria: University of Victoria, pp. 29–34.
Witten, I. H., and T. C. Bell. 1991. "The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression." *IEEE Transactions on Information Theory* 37(4):1085–1094.