POLYPHONIC MUSIC TRANSCRIPTION USING NOTE EVENT MODELING

Matti P. Ryynänen and Anssi Klapuri

Institute of Signal Processing, Tampere University of Technology
P.O.Box 553, FI-33101 Tampere, Finland
{matti.ryynanen, anssi.klapuri}@tut.fi

ABSTRACT

This paper proposes a method for the automatic transcription of real-world music signals, including a variety of musical genres. The method transcribes notes played with pitched musical instruments. Percussive sounds, such as drums, may be present but they are not transcribed. Musical notations (i.e., MIDI files) are produced from acoustic stereo input files using probabilistic note event modeling. Note events are described with a hidden Markov model (HMM). The model uses three acoustic features extracted with a multiple fundamental frequency (F0) estimator to calculate the likelihoods of different notes and performs temporal segmentation of notes. The transitions between notes are controlled with a musicological model involving musical key estimation and bigram models. The final transcription is obtained by searching for several paths through the note models. Evaluation was carried out with a realistic music database. Using strict evaluation criteria, 39% of all the notes were found (recall) and 41% of the transcribed notes were correct (precision). Taken the complexity of the considered transcription task, the results are encouraging.

1. INTRODUCTION

Transcription of music refers to the process of generating symbolic notations, i.e., musical transcriptions, for musical performances. Conventionally, musical transcriptions have been written by hand, which is time-consuming and requires musical education. In addition to the transcription application itself, the computational transcription methods facilitate automatic music analysis, automatic search and annotation of musical information in large music databases, and interactive music systems.

The automatic transcription of real-world music performances is an extremely challenging task. Humans (especially musicians) are able to recognize time-evolving acoustic cues as musical notes and larger musical structures, such as melodies and chords. A musical note is here defined by a discrete note pitch with a specific onset and an offset time. Melodies are consecutive note sequences with organized and recognizable shape whereas chords are combinations of simultaneously sounding notes.

Transcription systems to date have either considered only limited types of musical genres [1], [2], [3], or attempted only partial transcription [4]. For a more complete review of different systems for polyphonic music transcription, see [5]. To our knowledge, there exists no system transcribing music performances without setting any restrictions on the instruments, musical genre, maximum polyphony, or the presence of percussive sounds or sound effects in the performances. The proposed method transcribes the pitched notes in music performances without limiting the target material in any of the above-mentioned ways. Goto has previously proposed methods for the partial transcription of such complex material [4].

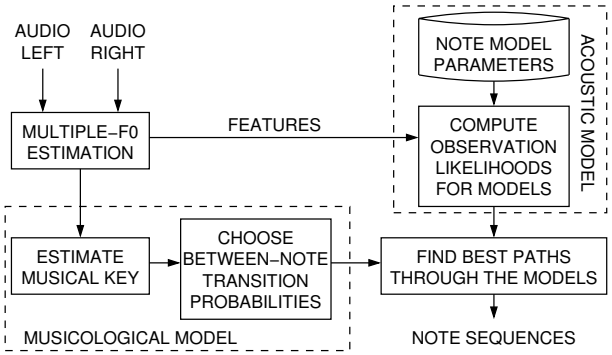

Figure 1: *The block diagram of the transcription method.*

The proposed transcription method is based on probabilistic modeling of note events and their relationships. The approach was previously applied in monophonic singing transcription [6]. Figure 1 shows the block diagram of the method. First, both the left and the right channel of an audio recording are processed frame-by-frame with a multiple-F0 estimator to obtain several F0s and their related features. The F0 estimates are processed by a musicological model which estimates musical key and chooses between-note transition probabilities. Note events are described with HMMs which allow the calculation of the likelihoods for different note events. Finally, a search algorithm finds multiple paths through the models to produce transcribed note sequences.

We use the RWC (Real World Computing) music genre database which consists of stereophonic acoustic recordings sampled at 44.1 kHz from several musical genres, including popular, rock, dance, jazz, classical, and world music [7]. For each recording, the database includes a reference MIDI file which contains a manual annotation of the note events in the acoustic recording. The annotated note events are here referred to as the reference notes. Since there exist slight time deviations between the recordings and the reference notes, all the notes within one reference file are collectively time-scaled to synchronize them with the acoustic signal. In particular, the synchronization can be performed more reliably for the beginning of a recording. Therefore, we use the first 30 seconds of 91 acoustic recordings for training and testing our transcription system. The MIDI notes for drums, percussive instruments, and sound effects are excluded from the set of reference notes.

0-7803-9154-3/05/$20.00 ©2005 IEEE

2. MULTIPLE-F0 ESTIMATION

The front-end of the transcription method is a multiple-F0 estimator proposed in [5], [8]. The estimator applies an auditory model where an input signal is passed through a 70-channel bandpass filterbank and the subband signals are compressed, half-wave rectified, and lowpass filtered with a frequency response close to $1/f$. Short-time Fourier transforms are then computed within the bands and the magnitude spectra are summed across channels to obtain a summary spectrum where all the subsequent processing takes place. Periodicity analysis is carried out by simulating a bank of comb filters in the frequency domain. F0s are estimated one at a time, the found sounds are canceled from the mixture, and the estimation is repeated for the residual. In addition, the method performs detection of the onsets of pitched sounds by observing positive changes in the estimated strengths of different F0 values.

Here we used the estimator to analyze audio signal in 92.9 ms frames overlapped with 11.6 ms interval between the beginnings of successive frames. In each frame, the estimator produces five distinct fundamental frequency values. Both the left and the right channels are independently analyzed from the input stereo signal, resulting in ten F0 estimates at each analysis frame t. As an output, the estimator produces four matrices X, S, Y, and D of size $10 \times t_{\max}$ ($t_{\max}$ is the number of analysis frames):

- F0 estimates X and their salience values S. For a F0 estimate $x_{it} = [X]_{it}$, the salience value $s_{it} = [S]_{it}$ roughly expresses how prominent x_{it} is in the analysis frame t.
- Onsetting F0 estimates Y and their onset strengths D. If a sound with F0 estimate $y_{it} = [Y]_{it}$ sets on in frame t, the onset strength value $d_{it} = [D]_{it}$ is high.

The F0 values in both X and Y are expressed in units of unrounded MIDI note numbers by

$$\text{MIDI note number} = 69 + 12 \log_2 \left(\frac{\text{F0}}{440\,\text{Hz}} \right) . \qquad (1)$$

Logarithm is taken from the elements of S and D to compress their dynamic range. If an onset strength value d_{it} is small, the onsetting F0 estimate y_{it} is random valued. Therefore, those y_{it} values are set to zero for which the onset strength d_{it} is below a fixed threshold. We empirically chose a threshold value of $\ln 3$.

3. PROBABILISTIC MODELS

The transcription system applies three probabilistic models: a note event HMM, a silence model, and a musicological model. The note HMM uses the output of the multiple-F0 estimator to calculate likelihoods for different notes, and the silence model corresponds to time regions where no notes are sounding. The musicological model controls transitions between note HMMs and the silence model, analogous to a "language model" in automatic speech recognition. Transcription is done by searching for disjoint paths through the note models and the silence model.

3.1. Note event model

Note events are described with a three-state HMM. The note HMM is a state machine where state $q_i, 1 \le i \le 3$, represents the typical values of the features in the i:th temporal segment of note events. The model allocates one note HMM for each MIDI note number $n = 29, \ldots, 94$, i.e., from note F1 to B♭6. Given the matrices X, S, Y, and D, the observation vector $\mathbf{o}_{n,t} \in \mathbb{R}^3$ is defined for a note HMM with nominal pitch n at frame t as

$$\mathbf{o}_{n,t} = (\Delta x_{n,t}, s_{jt}, d_{n,t}) , \qquad (2)$$

where

$$\Delta x_{n,t} = x_{jt} - n , \qquad (3)$$

$$d_{n,t} = \begin{cases} d_{kt} , & \text{if } |y_{kt} - n| \le 1 \\ 0 , & \text{otherwise} \end{cases} , \qquad (4)$$

$$j = \arg\min_i \{|x_{it} - n|\} , \; 1 \le i \le 10 , \qquad (5)$$

$$k = \arg\min_i \{|y_{it} - n|\} , \; 1 \le i \le 10 . \qquad (6)$$

The observation vectors thus consist of three features: the pitch difference $\Delta x_{n,t}$ between the measured F0 and the nominal pitch n of the modeled note, the salience s_{jt}, and the onset strength $d_{n,t}$. For a note HMM n, the nearest F0 estimate and its salience values s_{jt} are associated with the note by (3), (5). The onset strength feature is used only if its corresponding F0 value is closer than one semitone to the nominal pitch n of the model (4), (6). Otherwise, onset strength feature value $d_{n,t}$ is set to zero, i.e., no onset measurement is available for the particular note.

We use the pitch difference as a feature instead of the absolute F0 value so that only one set of note-HMM parameters needs to be trained. In other words, we have a distinct note HMM for each nominal pitch $n = 29, \ldots, 94$ but they all share the same trained parameters. This can be done since the observation vector itself is tailored to be different for each note model (2). The HMM parameters include (i) state-transition probabilities $P(q_j|q_i)$, i.e., the conditional probability that state q_i is followed by state q_j, and (ii) the observation likelihood distributions, i.e., the likelihoods $P(\mathbf{o}_n|q_i)$ that observation $\mathbf{o}_n$ is emitted by the state q_i from note model n.

For the time region of a reference note with nominal pitch n, the observation vectors by (2) form an observation sequence for training the acoustic note event model. The observation sequence is accepted for the training only if the median of the absolute pitch differences $|\Delta x_{n,t}|$ is smaller than one semitone during the reference note. This type of selection is necessary, since for some reference notes there are no reliable F0 measurements available in X. The note HMM parameters are then obtained using the Baum-Welch algorithm [9]. The observation likelihood distributions are modeled with a four-component Gaussian mixture model.

Figure 2 shows the parameters of the trained note HMM. On top, the HMM states are connected with arrows to show the HMM topology and the state-transition probabilities. Below each state, the figure shows the observation likelihood distributions for the three features. The first state is interpreted as the attack state where pitch difference has a larger variance, the salience feature gets lower values, and the onset strength has a prominent peak at 1.2, thus indicating note onsets. The second state, here called as the sustain state, includes small variance of the pitch difference feature and large salience values. The final state is a noise state, where F0s with small salience values dominate. The sustain and the noise states are full-connected, thus allowing note HMM to visit the noise state and switch back to the sustain state, if a note event contains noisy regions.

3.2. Silence model

We use a silence model to skip the time regions where no notes are sounding. The silence model is a 1-state HMM for which the

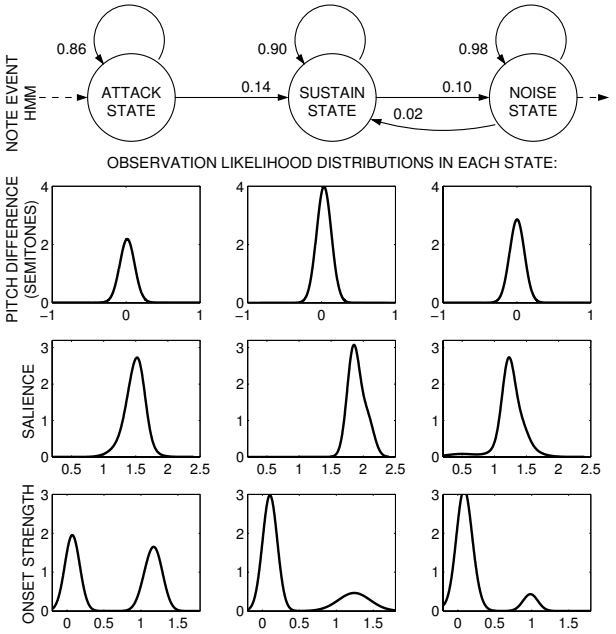

Figure 2: *The parameters of the trained note event HMM.*

observation likelihood at time t is defined as

$$P(\text{silence})_t = 1 - \max_{n,j}\{P(\mathbf{o}_{n,t}|q_j)\} \ , \qquad (7)$$

i.e., the observation likelihood for the silence model is the negation of the greatest observation likelihood in any state of any note model at time t. Here, the observation likelihood value $P(\mathbf{o}_{n,t}|q_j)$ is on linear scale, and its maximum value is unity. If a note state has a large observation likelihood, the observation likelihood for the silence model is small at that time.

3.3. Musicological model

The musicological model controls transitions between note HMMs and the silence model in a manner similar to [6]. The musicological model is based on the fact that some note sequences are more common than others in a certain musical key. A musical key is roughly defined by the basic note scale used in a song. A major key and a minor key are called a relative-key pair if they consist of scales with the same notes (e.g., the C major and the A minor).

The musicological model first finds the most probable relative-key pair using a musical key estimation method proposed and evaluated in [10]. The method produces likelihoods for different major and minor keys from those F0 estimates x_{it} (rounded to the nearest MIDI note numbers) for which salience value is larger than a fixed threshold. Based on empirical investigation of the data, we chose a threshold value of $\ln 2$. The most probable relative-key pair is estimated for the whole recording and this key pair is then used to choose transition probabilities between note models and the silence model.

The transition probabilities between note HMMs are defined by note bigrams which were estimated from a large database of monophonic melodies, as reported in [6]. As a result, given the previous note and the most probable relative-key pair k, the note bigram probability $P(n_t = j|n_{t-1} = i, k)$ gives a transition probability to move from note i to note j. This means that a possible visit in the silence model is skipped and only the previous note accounts.

The musicological model assumes that it is more probable both to start and to end a note sequence with a note which is frequently occurring in the musical key. A silence-to-note transition (if there is no previously visited note, e.g., in the beginning of the piece) corresponds to starting a note sequence and a note-to-silence transition corresponds to ending a note sequence. Krumhansl reported the occurrence distributions of different notes with respect to musical key estimated from a large amount of classical music [11, p. 67]. The musicological model applies these distributions as probabilities for the note-to-silence and the silence-to-note transitions so that the most probable relative-key pair is taken into account. A transition from the silence model to itself is prohibited.

3.4. Finding several note sequences

The note models and the silence model constitute a network of models. The optimal path through the network can be found using the Token-passing algorithm [12] after calculating the observation likelihoods for note model states and the silence model, and choosing the transition probabilities between different models. The Token-passing algorithm propagates tokens through the network. Each model state contributes to the overall likelihood of a token by the observation likelihood and the transition probabilities between the states. When a token is emitted out of a model, a note boundary is recorded. Between the models, the musicological model contributes to the likelihood of the token by considering the previous note, or the silence model. Eventually, the token with the greatest likelihood defines the optimal note sequence.

In order to transcribe polyphonic music, we need to find several paths through the network. We apply the Token-passing algorithm iteratively as follows. As long as the desired number of paths has not yet been found and the found paths contain notes and not just silence, (i) find the optimal path with Token-passing, and (ii) prohibit the use of any model (except the silence model) on the found path during the following iterations. After each iteration, recalculate the observation likelihoods for the silence model with (7) by discarding the observation likelihoods for note models on the found paths. As a result, several disjoint note sequences have been transcribed. In the simulations, the maximum number of iterations was set to 10, meaning that the system can transcribe at most 10 simultaneously sounding notes. Figure 3 shows a transcription example from the beginning of a jazz ballad, including piano, a contrabass, and drums.

4. SIMULATION RESULTS

The proposed transcription system was evaluated using a threefold cross validation. Three evaluation criteria were used: the recall rate, the precision rate, and mean overlap ratio. Given that $c(\text{ref})$ is the number of reference notes, $c(\text{trans})$ is the total number of transcribed notes, and $c(\text{cor})$ is the number of correctly transcribed notes, the rates are defined as

$$\text{recall} = \frac{c(\text{cor})}{c(\text{ref})} \ , \quad \text{precision} = \frac{c(\text{cor})}{c(\text{trans})} \ . \qquad (8)$$

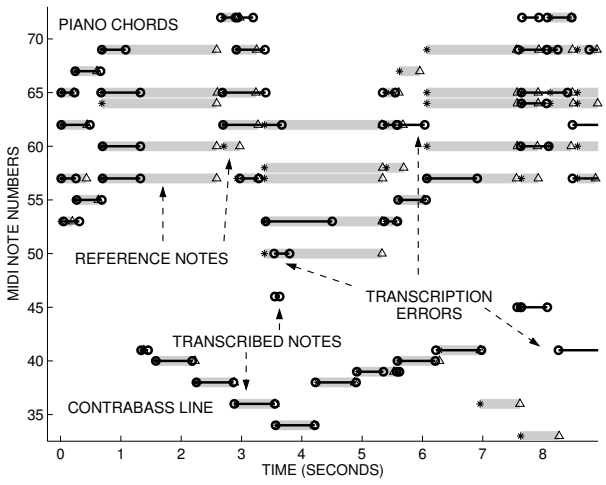

Figure 3: *A transcription from the beginning of a jazz ballad, including piano, a contrabass, and drums. The grey bars indicate the reference notes and the black lines are the transcribed notes.*

A reference note is correctly transcribed by a note in the transcription if (i) their MIDI notes are equal, and (ii) the absolute difference between their onset times is smaller than or equal to a given maximum onset interval δ, and (iii) the transcribed note is not already associated with another reference note. In other words, one transcribed note can transcribe only one reference note.

The overlap ratio is defined for an individual correctly transcribed note as

$$\text{overlap ratio} = \frac{\min\{\text{offsets}\} - \max\{\text{onsets}\}}{\max\{\text{offsets}\} - \min\{\text{onsets}\}}, \qquad (9)$$

where "onsets" refers to the onset times of both the reference and the corresponding transcribed note, and "offsets" accordingly to the offset times. The mean overlap ratio is then calculated as the average of overlap ratios for all the correctly transcribed notes within one transcription.

Because of the timing deviations between the recordings and the manually annotated reference notes, $\delta = 150$ ms was used. Although this is a rather large value, it is acceptable in this situation. For example, the reference note 50 in Fig. 3 is not correctly transcribed due to the $\delta = 150$ ms criterion. The criteria are otherwise very strict. In addition, the reference notes with colliding pitch and onset (on the average, 20% of reference notes) were required to be transcribed, although our transcription system performs no instrument recognition and is thus incapable of transcribing such unison notes.

The recall rate, the precision rate, and the mean overlap ratio are calculated separately for the transcriptions of each recording. The average over all the transcriptions for each criterion are: recall 39%, precision 41%, and the mean overlap ratio 40%.

5. CONCLUSIONS

This paper described a method for transcribing realistic polyphonic audio. The method was based on the combination of an acoustic model for note events, a silence model, and a musicological model. The proposed transcription method and the presented evaluation results give a reliable estimate of the accuracy of state-of-the-art music-transcription systems which address all music types and attempt to recover all the notes in them. The results are very encouraging and serve as a baseline for the further development of transcription systems for realistic music signals. Transcription examples are available at

`http://www.cs.tut.fi/sgn/arg/matti/demos/polytrans.html`

6. REFERENCES

[1] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka, "Organization of hierarchial perceptual sounds: Music scene analysis with autonomous processing modules and a quantitative information integration mechanism," in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 1, Aug. 1995, pp. 158–164.

[2] K. D. Martin, "Automatic transcription of simple polyphonic music: Robust front end processing," Massachusetts Institute of Technology Media Laboratory Perceptual Computing Section, Tech. Rep. 399, 1996.

[3] M. Davy and S. J. Godsill, "Bayesian harmonic models for musical signal analysis," in *Seventh Valencia International meeting (Bayesian Statistics 7)*. Oxford University Press, 2002.

[4] M. Goto, "A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, no. 4, pp. 311–329, 2004.

[5] A. Klapuri, "Signal processing methods for the automatic transcription of music," Ph.D. dissertation, Tampere University of Technology, 2004.

[6] M. P. Ryynänen and A. Klapuri, "Modelling of note events for singing transcription," in *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio*, Oct. 2004.

[7] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proc. 4th International Conference on Music Information Retrieval*, Oct. 2003.

[8] A. Klapuri, "A perceptually motivated multiple-F0 estimation method," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2005.

[9] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–289, Feb. 1989.

[10] T. Viitaniemi, A. Klapuri, and A. Eronen, "A probabilistic model for the transcription of single-voice melodies," in *Proc. 2003 Finnish Signal Processing Symposium*, May 2003, pp. 59–63.

[11] C. Krumhansl, *Cognitive Foundations of Musical Pitch*. Oxford University Press, 1990.

[12] S. J. Young, N. H. Russell, and J. H. S. Thornton, "Token passing: a simple conceptual model for connected speech recognition systems," Cambridge University Engineering Department, Tech. Rep., July 1989.