AUDIO-BASED GUITAR TABLATURE TRANSCRIPTION USING MULTIPITCH ANALYSIS AND PLAYABILITY CONSTRAINTS

Kazuki Yazawa, Daichi Sakaue, Kohei Nagira, Katsutoshi Itoyama, Hiroshi G. Okuno

Graduate School of Informatics, Kyoto University, Japan

ABSTRACT

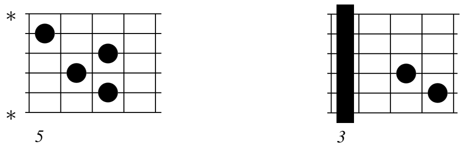
This paper proposes a method of guitar tablature transcription from audio signals. Multipitch estimation and fingering configuration estimation are essential for transcribing tablatures. Conventional multipitch estimation methods, including latent harmonic allocation (LHA), often estimate combinations of pitches that people cannot play due to inherent physical constraints. Unplayable combinations of pitches are eliminated by filtering the results of LHA with three constraints. We first enumerate playable fingering configurations, and use them to suppress any undesirable combination of pitches. The optimal fingering configuration in each time frame is optimized to satisfy the need for temporal continuity by using dynamic programming. We use synthesized guitar sounds from MIDI data (ground truth) for evaluation. Experiments with them demonstrate the improvement of multipitch estimation by 5.9 points on average in F-measure and the transcribed tablatures are playable.

Index Terms— guitar tablature, fingering configuration, multipitch estimation, onset detection, music signal processing

1. INTRODUCTION

The guitar is one of the most popular musical instruments, with a large number of amateur players who typically practice by using their favorite pieces with tablatures because doing so is both pleasable and motivational. Tablature is a musical score format that describes the fingering configurations instead of the pitch combinations. Many musical pieces including CD recordings of guitar playing can be easily bought, but the corresponding tablatures are not easily obtainable. Transcribing guitar tablatures from musical CD recordings is difficult without expertise in doing so. An automatic guitar tablature transcription system would thus be valuable.

Guitar tablature transcription consists of two tasks: estimation of simultaneous multiple pitches and estimation of fingering configuration sequence. If the multipitch estimation is perfect, the latter, fingering configuration estimation, is rather trivial because optimal fingering configurations can be easily determined from the combinations of pitches. However, most multipitch estimation methods [1–6] usually produce noisy results, resulting in ambiguity and unplayable configurations. For example, seven or more simultaneous sounds are unplayable using standard guitars with six strings. We have developed an automatic guitar tablature transcription method that consistently estimates appropriate tablatures, that is, sequences of playable fingering configurations and the corresponding active (plucked) strings. It incorporates three constraints on fingering configurations based on physical and musical constraints that are related to (1) likelihood of fingering configurations, (2) timing of configuration changes, and (3) duration of each configuration. It first estimates the existence probability of the fundamental frequencies in each time frame and then calculates the likelihood of each fingering configuration. Optimal configurations are chosen under these constraints using a newly designed dynamic programming technique. Experimental evaluation showed that our method outperforms a conventional multipitch estimation method due to the three constraints and that the transcribed tablatures were playable.



(a) Open chord (b) Barre chord

Fig. 1. Example fingering configurations: (a) open chord and (b) barre chord. Asterisks represent open strings. Rectangle represents index finger. Digits *5* and *3* indicate fret indices.

2. FINGERING CONFIGURATIONS

Fingering configurations for constraint (1) are described in this section. We assume that an input guitar piece is performed using a standard guitar with six strings and 20 frets with standard tuning (EADGBE) though the method can easily be applied to other kinds of guitars.

To constrain fingering configurations, invalid configurations must be eliminated. The validity of a configuration is defined by two conditions: reach of fingers and number of fingers. Finger reach for valid configurations must be less than or equal to four because we assume that most players can spread their fingers at most the width of four frets. Finger number for valid configurations must be less than or equal to four because guitar players use only their four fingers.

Fingering configurations are divided into two types: those for open chords and those for barre chords. For an open chord configuration, a finger presses down only one string. The configurations of open chords contain open (unpressed) strings because the number of available fingers, 4, is less than the number of guitar strings, 6. An example of an open chord configuration is shown in Fig. 1(a). Finger reach and number for this configuration are respectively 3 and 4. The first and sixth strings are open. We enumerated all possible configurations for open chords.

For a barre chord configuration, a finger presses down multiple strings simultaneously. The index finger is usually used for barre chords, pressing down all six strings. The other three fingers must be located on the right side of the index finger. We enumerate this type of barre configurations and omit other types because they are rarely used. An example of a barre chord configuration is shown in

Fig. 1(b). Finger reach and number for this configuration are respectively 4 and 3.

A total of 38,119 configurations were enumerated. We use K_p (p = 1, ..., 38,119) to denote the set of pitches that can be performed by the p-th configuration. Each K_p has up to six pitches.

3. TABLATURE TRANSCRIPTION METHOD

Before describing our method of multipitch estimation of guitar sounds, we briefly summarize a conventional multipitch estimation method, latent harmonic allocation (LHA) [6].

3.1. Preprocessing: latent harmonic allocation

LHA is a machine learning method that estimates the multiple pitches from an observed spectrogram. In LHA, the harmonic structures of musical instrument sounds are approximated using a nested mixture of Gaussian distributions, and the parameters of these distributions are estimated. The output is the relative strength of the k-th harmonic mixture component in the t-th time frame. This is equivalent to the effective energy of the k-th pitch in this frame and is denoted as N_{tk}. The higher the N_{tk}, the more the k-th pitch is likely to sound in the t-th time frame. To extract the pitch activity in binary form, we place a threshold on N_{tk}: the threshold parameter is α. All the pitches that satisfy $N_{tk} \geq \alpha \max_{tk} N_{tk}$ are regarded as sounding.

If we try to translate this result into tablature form, we immediately run into three problems.

1. Combinations of sounds that cannot be played on a guitar, such as seven or more simultaneous pitches, are sometimes estimated.
2. Estimated fingering configurations may change at other than the onset times.
3. Estimated fingering configurations may change too frequently due to neglecting the time spent changing the configurations.

Our method does not suffer these problems due to the three constraints that are imposed. It estimates the most likely fingering configuration sequence by examining the estimation results and eliminates sounds that cannot be generated by these configurations.

3.2. Three constraints

1. The optimal configuration in each time frame must be chosen so as to maximize $\sum_t N_{tp_t}$, where p_t is the index of the configuration used in the t-th time frame, and N_{tp} is the likelihood of using the p-th configuration in the t-th time frame.
2. The configuration can only change at onset times.
3. The same configuration must be used during D consecutive time frames.

3.2.1. Configuration likelihood constraint

To determine the likelihood of each configuration in each time frame, the method calculates $N_{tp} = \sum_{k \in \mathbf{K}_{p_t}} N_{tk}$ for all enumerated fingering configurations. The higher N_{tp}, the more likely the configuration is to be used in the t-th time frame.

For this calculation, duplicate note numbers in $\mathbf{K}_p$ are eliminated. That is, if $\mathbf{K}_p$ has the same notes played by different strings, it is counted only once. Otherwise, configurations that have note duplications would unfairly get a higher score.

3.2.2. Configuration change timing constraint

The fingering configuration changes when the player is going to play a note that cannot be played using the current configuration. Therefore, configurations changes must occur only at onset times. The onset time candidates are determined using the spectral flux [7]:

$$SF_t = \sum_f \max(0, X_{tf} - X_{t-1,f}),$$

where X_{tf} represents the spectral power of the f-th frequency in the t-th time frame. The time frames of which SF_t is higher than a threshold, $\beta \max_t SF_t$, are regarded as candidate onset times. Due to this constraint, the fingering configurations changes only at an onset time.

3.2.3. Configuration duration constraint

To ensure that the time spent changing the fingering configuration is explicitly considered, a minimum duration is imposed on each configuration. This limitation means that a performance with configuration changes that occur too frequently is regarded as unsuitable. The minimum duration is assumed to be the same for all configurations.

3.3. Formulation of constraints

Optimal configuration index $\hat{p}_t$ for each time frame is estimated using the three constraints. The procedure for estimating an optimal configuration sequence is illustrated in Fig. 2.

This procedure comes down to the longest path problem on a weighted directed acyclic graph, as shown in Fig. 3. This problem can be solved using dynamic programming [8]. Let v_{tp} be the vertex corresponding to the p-th configuration in the t-th time frame and e_{tupq} be a directed edge from v_{tp} to v_{uq}. An edge e_{tupq} exists only when either of the following conditions is satisfied:

1. $p = q$ and $u = t + 1$.
2. $p \neq q$, $u = t + D$, and t is an onset time.

The first condition means that the configuration does not change at the t-th time frame. The weight of this edge $e_{t,t+1,p,p}$ is set to N_{tp}. The second condition means that the configuration used changes from the p-th to the q-th one at the t-th time frame. In accordance with the configuration change timing constraint, the configuration can be changed only at an onset time frame. Since the player must hold the q-th configuration for D frames due to the duration constraint, the weight of edge $e_{t,t+D,p,q}$ is determined to be $\sum_{t'=t}^{t+D} N_{t'q}$.

The longest path in this graph is equivalent to the optimal configuration sequence. It reflects the configuration likelihood constraint. The optimal configuration indexes $\hat{p}_1, \ldots, \hat{p}_T$ can thus be obtained by tracking this path.

3.4. Modified multipitch estimation

Finally, the N_{tk} corresponding to the k-th pitch that cannot be produced in the $\hat{p}_t$-th configuration is nullified.

$$\tilde{N}_{tk} = \begin{cases} N_{tk} & (\text{if } k \in \mathbf{K}_{\hat{p}_t}) \\ 0 & (\text{otherwise}) \end{cases} \quad (1)$$

The multipitch estimation accuracy should be improved by placing the threshold on $\tilde{N}_{tk}$ rather than the original N_{tk}. The result is illustrated in Fig. 4.

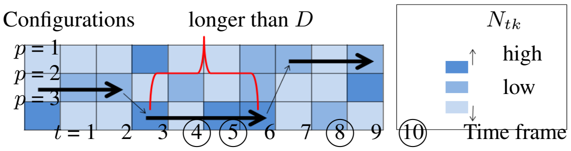


Fig. 2. Illustration of estimating optimal configuration sequence. Arrowed lines represent estimated sequence. Encircled numbers indicate onset candidates. Configuration can change only at onset times and must be used for more than D frames (here D = 3). Optimal sequence maximizes summation of N_{tp_t} under these constraints.

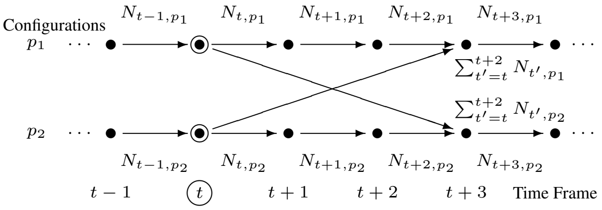


Fig. 3. Graphical illustration of finding optimal configuration. Arrowed lines represent configuration sequence candidates. Onset occurs at t-th time frame, and configuration may be changed. Here $D = 3$ and there are two example configurations.

4. EVALUATION

4.1. Multipitch estimation

We experimentally evaluated the ability of the proposed method to estimate the F0s in each time frame and qualitatively compared the performance with that of the conventional one.

4.1.1. *Experimental conditions*

We used 9 guitar solo parts from the jazz genre and 70 from the popular one. They were retrieved from 9 jazz pieces and 52 popular ones in the RWC music database [9]. Some pieces were performed using several instruments, sometimes with two or more guitars. The guitar parts containing pitch bend messages and those containing less than two simultaneous notes on average with respect to all the sounding times were removed. This is because our method uses the durations and fingering configurations, so the removed parts were inappropriate for evaluating the characteristics of our method.

Only the first 60 seconds of each part was evaluated to reduce the heavy computational time. There were some silent sections, and the average sounding time of each part was 37.7 seconds. A MIDI version of each piece was used to enable quantitative evaluation. The audio signals were recorded using a MIDI synthesizer (YAMAHA MOTIF-XS). The signals were transformed into wavelet spectrograms with 20-ms time resolution. We constructed the ground truth from the MIDI files.

Several β settings, which determines the sensitivity of onset detection, were used. The threshold parameter for LHA, α, was optimized for both methods and each piece because it is easily optimized by a system user. Minimum duration D was set to ten time frames (200 ms) for all pieces because most amateur guitar players cannot change fingering configurations faster than this. The standard LHA was evaluated using the same data set for comparison. The metric was the F-measure.

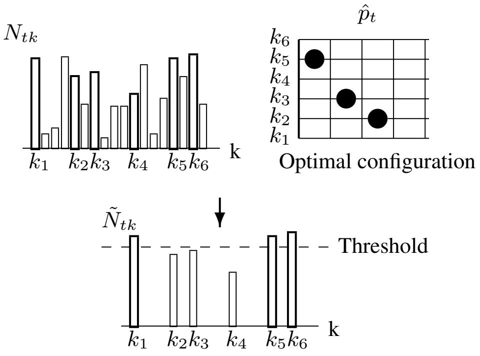


Fig. 4. Illustration of revision of multipitch estimation using constraints. *Optimal configuration* illustrates configuration estimated for t-th time frame; k_1 – k_6 are note numbers that can be played with optimal configuration.

Table 1. F-measures of fundamental frequency estimation. Average # means average number of simultaneous notes in all sounding time. *LHA* represents conventional method with no constraints. β represents threshold parameter for configuration change timing constraint. Bold values indicate maximal performance.

Genre	Average #	LHA	$\beta = 0.2$	$\beta = 0.3$	$\beta = 0.4$	$\beta = 0.5$
Jazz	$2 < x \leq 3$	0.701	**0.721**	0.719	0.713	0.691
	$3 < x \leq 4$	0.583	0.645	**0.652**	**0.652**	0.648
	$4 < x \leq 5$	0.515	0.648	0.643	**0.664**	0.640
Popular	$2 < x \leq 3$	0.615	0.649	**0.651**	0.644	0.638
	$3 < x \leq 4$	0.604	0.669	0.672	**0.674**	0.667
	$4 < x \leq 5$	0.712	0.808	0.813	0.820	**0.823**
	$5 < x \leq 6$	0.736	0.837	**0.840**	0.837	0.821

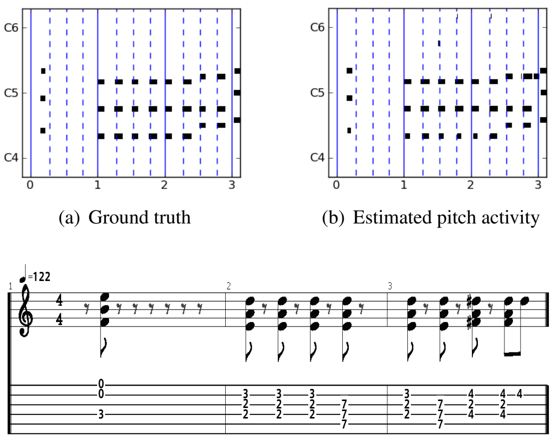
4.1.2. *Experimental results*

The experimental results are shown in Table 1. The proposed method outperformed the conventional one for both genres. On average, over all guitar parts, the F-measure was better by 5.9 points for $\beta = 0.3$ compared to that for the conventional method. Moreover, the larger the average note number, the better the proposed method against the conventional one.

4.2. Transcription of tablature

One of the key advantages of this method is that it can be used to transcribe guitar tablatures with a only few modifications. That is, we can use each $\hat{p}_t$ and $\tilde{N}_{tk}$, i.e., the optimal fingering configuration index and the likelihood of the k-th note in the t-th time frame, to estimate the strings plucked in each time frame. An example tablature transcribed with our method is shown in Fig. 5. The tablature contains only playable guitar configurations because configurations are chosen from the playable configurations. It does not contain overly frequent changes in configuration since the method takes into account the time spent changing configurations.

In the transcription of this tablature, we assume that measures and beats are perfectly estimated by another estimator. In real applications, we can use a beat tracker [10] to do this.



(a) Ground truth

(b) Estimated pitch activity

(c) Tablature

Fig. 5. Ground truth and example outputs of proposed method: (a) ground truth pitch activity, (b) estimated pitch activity, and (c) tablature. In tablature, notes appearing for short time have been removed. Measures and bars are assumed to be estimated correctly. They correspond to first three bars (about five seconds) of *RM-J038*.

5. DISCUSSION

5.1. Validity of fingering configuration enumeration

The enumerated configurations (i.e., "collection") seems to cover most configurations generally played. Indeed, it contains almost all of (1186 out of 1202) the configurations in a common guitar chord chart [11]. For improvement, we should consider the following two points: (1) more strict way of playability definition rather than the reach and number of fingers, (2) including other playable configurations (such as configurations using the thumb, barre chord configurations with a ring finger, etc.) in the collection. Also, some configurations are very difficult to perform or rarely used, so it is desirable to restrict their occurrences. These are the remaining tasks of this research.

5.2. Fingering constraint

In our algorithm, we did not examine the priority of configurations if there are two or more possible configurations that contain all the candidate pitches estimated using LHA. Indeed, a combination of five or less sounds can be played in multiple ways (Fig. 5 (c)). This point should be discussed in future research by considering the following two points: (1) the ease of each fingering configuration and (2) that of changing the configuration from one to another. The minimum duration, D, seems to have an important role.

5.3. Characteristics of experimental data

The evaluation was conducted using signals generated with a MIDI synthesizer. Since such signals contain less fluctuation and noise than of a guitar performance recording using microphones, an evaluation using the actual signals would enrich the discussion. However, MIDI data is the same to real audio data on the point that recording guitar performances are subject to the constraint of the harmonic structures and the way of performances. Therefore, the proposed method is expected to work toughly for real audio recordings.

5.4. Related work

The transcription of guitar tablatures has already been reported [12]. Their method uses hidden Markov models to capture the temporal relationship of the fingering configurations. The main objective of their method is to accurately estimate the musical chords and plucked strings. By restricting 330 configurations (major, minor, major 7th, and minor 7th) of fingerings, the proposed system outperformed a non-guitar-specific reference chord transcription method. Since real guitar performance needs much more configurations, the range of applicability of the proposed system may be limited. Since our method covers more than 30,000 configurations, it may be expected to generate more physically plausible tablatures. Moreover, our method works toughly for non-chord pieces (like solo guitar) and ones by guitar arpeggio, while their method is based on the assumption that pieces are composed strictly of chords.

There are other researches related to the transcription of guitar tablature [13–18]. Some researches [14, 16, 17] use the visual information to obtain a precise tablature, that may be useful for the analysis of the concert videos and guitar practice. Another researcher [13] has pointed out that the inharmonicity of the guitar sounds may be useful for obtaining a tablature. Although pitches and guitar fingerings can be effectively estimated with these approaches, complementary knowledge or training data is needed to determine the model parameters. In contrast, our method estimates the pitches and fingerings by using only audio signals from such sources as audio CDs. It thus has wider application for supporting amateur guitar players.

For many of the most famous pieces, a different approach of obtaining a guitar tablature has been proposed [19]. That is, to collect handmade guitar tablatures of several formats and correct the errors by comparing them. This idea works fine if a number of tablatures of the target piece are available.

6. CONCLUSION

Our proposed method modifies the results of the conventional multipitch estimation method, LHA, by using a collection of playable fingering configurations. The optimal fingering configuration in each time frame is estimated on the basis of three constraints that suppress unplayable notes. These constraints enable the proposed system to output guitar tablatures consisting of only playable configurations. Testing showed that the proposed system had more precise multipitch estimation than the conventional method: there was a 5.9% improvement in the F-measure on average. Future work includes application of method to musical pieces played used other instruments by changing the method of enumerating playable configurations. This research was partially supported by KAKENHI (S) No. 24220006 and (B) No. 24700168.

7. REFERENCES

[1] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Trans. on ASLP*, vol. 18, no. 6, pp. 1643–1654, 2010.

[2] 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.

[3] H. Kameoka, T. Nishimoto, and S. Sagayama, “A multipitch analyzer based on harmonic temporal structured clustering,” *IEEE Trans. on ASLP*, vol. 15, no. 3, pp. 982–994, 2007.

[4] A. Klapuri, “Multipitch analysis of polyphonic music and speech signals using an auditory model,” *IEEE Trans. on ASLP*, vol. 16, no. 2, pp. 255–266, 2008.

[5] D. Sakaue, K. Itoyama, T. Ogata, and H. G. Okuno, “Initialization-robust multipitch estimation based on latent harmonic allocation using overtone corpus,” in *Proc. ICASSP*, 2012, pp. 425–428.

[6] K. Yoshii and M. Goto, “A nonparametric Bayesian multipitch analyzer based on infinite latent harmonic allocation,” *IEEE Trans. on ASLP*, vol. 20, no. 3, pp. 717–730, 2012.

[7] C. Duxbury, M. Sandler, and M. Davies, “A hybrid approach to musical note onset detection,” in *Proc. DAFx*, 2002, pp. 33–38.

[8] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, The MIT Press, 1990.

[9] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, “RWC music database: Popular, classical, and jazz music database,” in *Proc. ISMIR*, 2002, pp. 287–288.

[10] A. Maezawa, H. G. Okuno, T. Ogata, and M. Goto, “Polyphonic audio-to-score alignment based on Bayesian latent harmonic allocation hidden Markov model,” in *Proc. ICASSP*, 2011, pp. 185–188.

[11] “Guitargearheads – gear up your sound – tiny content,” 2012, http://www.guitargearheads.com/modules/tinycontent/index.php?id=5.

[12] A. M. Barbancho, A. Klapuri, L. J. Tardón, and I. Barbancho, “Automatic transcription of guitar chords and fingering from audio,” *IEEE Trans. on ASLP*, vol. 20, no. 3, pp. 915–921, 2012.

[13] I. Barbancho, L. J. Tardón, S. Sammartino, and A. M. Barbancho, “Inharmonicity-based method for the automatic generation of guitar tablature,” *IEEE Trans. on ASLP*, vol. 20, no. 6, pp. 1857–1868, 2012.

[14] A. Burns and M. Wanderley, “Visual methods for the retrieval of guitarist fingering,” in *Proc. NIME*, 2006, pp. 196–199.

[15] P. D. O’Grady and S. T. Rickard, “Automatic hexaphonic guitar transcription using non-negative constraints,” in *Proc. ISSC*, 2009, pp. 1–6.

[16] A. Hrybyk and Y. Kim, “Combined audio and video analysis for guitar chord identification,” in *Proc. ISMIR*, 2010, pp. 159–164.

[17] M. Paleari, B. Huet, A. Schutz, and D. Slock, “A multimodal approach to music transcription,” in *Proc. ICIP*, 2008, pp. 93–96.

[18] H. Penttinen, J. Siiskonen, and V. Välimäki, “Acoustic guitar plucking point estimation in real time,” *IEEE Trans. on ASLP*, vol. 3, pp. 209–212, 2005.

[19] R. Macrae and S. Dixon, “Guitar tab mining, analysis and ranking,” in *Proc. ISMIR*, 2011, pp. 453–458.