EXPRESSIVE ORIENTED TIME-SCALE ADJUSTMENT FOR MIS-PLAYED MUSICAL SIGNALS BASED ON TEMPO CURVE ESTIMATIONS

Yuma Koizumi,

Graduate School of
Computer and Information Sciences,
Hosei University
Tokyo, Japan
12t0005@hosei.ac.jp

Katunobu Itou,

Faculty of
Computer and Information Sciences,
Hosei University
Tokyo, Japan
it@fw.ipsj.or.jp

ABSTRACT

Musical recordings, when performed by non-proficient (amateur) performers, include two types of tempo fluctuations–intended "tempo curves" and non-intended "mis-played components"–due to poor control of instruments. In this study, we propose a method for estimating intended tempo fluctuations, called "true tempo curves," from mis-played recordings. We also propose an automatic audio signal modification that can adjust the signal by time-scale modification with an estimated true tempo curve to remove the mis-played component. Onset timings are detected by an onset detection method based on the human auditory system. The true tempo curve is estimated by polynomial regression analysis using detected onset timings and score information. The power spectrograms of the observed musical signals are adjusted using the true tempo curve. A subjective evaluation was performed to test the closeness of the rhythm, and it was observed that the mean opinion score values of the adjusted sounds were higher than those of the original recorded sound, and significant differences were observed for all tested instruments.

1. INTRODUCTION

User-generated content (UGC) publishing through the internet has been increasing, since this easy method of publishing inspires many musicians. However, most musicians need to fix and/or modify their performance before publishing their creation. Although the performances may be of poor quality, many of the notes are roughly performed as intended but with erroneous deviances in the rhythm (tempo and timing), amplitude, timbre, and pitch of a few notes. In the case of rhythm, the only way by which the error can be automatically modified is by timing with a metronome, making the result too artificial. If we want to modify it more naturally, we have to manually adjust by trial and error.

Automatic tempo detection from a tempo-varying piece of music is one of the most important issues in music information retrieval (MIR). Note onset timings are detected [1, 2], and then the tempo fluctuation, called the "*tempo curve*," is analyzed to detect recurring patterns and quasi-periodic pulse trains [3, 4]. Recently, a novel "tempogram" that uses musically meaningful local pulse information has been presented [5].

These studies for MIR target only recordings that are performed as intended by sophisticated players. Low-proficiency performances have not only fluctuations from the intended tempo but also deviances caused by erroneous notes being played. We propose a tempo-curve estimation method for monophonic audio signals (with included performance errors) by polynomial regression analysis to prevent over-fitting. Here we define the interpreted tempo fluctuation as the "*true tempo curve*."

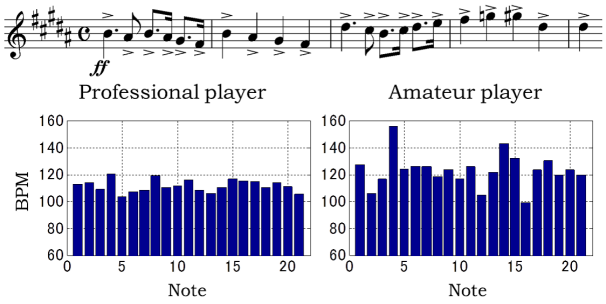

Fig. 1. Examples of actual tempo fluctuation from monophonic violin recordings played by professional and amateur musicians. The musicians played a violin phrase from "Tannhauser Act II 'Grand March' " by R. Wagner.

We also propose an automatic audio signal modification that can *adjust* the signal by time-scale modification with the estimated true tempo curve. The estimation method requires accurate onset detection. We also propose an onset detection method based on the human auditory system to cover the various sound generation mechanisms of musical instruments.

Recently, attempts have been made to analyze intended tempo fluctuation separately into *micro–* and *macro–*, [6]. *Macro–tempo* is the tempo experienced by listeners; it can change rather slowly. *Micro-Tempo* consists of slight anticipations of note events followed by a deferral of the subsequent events in such a way that the result does not contribute to macro-tempo changes. Since this study focus on the separation of the performance error, it does not discuss separation of the intended fluctuations.

2. MODELING AND ADJUSTMENT OF DEVIANCE FROM MUSICAL SCORES

In this paper, the method, which corrects audio that has been recorded as interpreted by the player, is discussed.

In an actual performance, the tempo is not constant. Sophisticated players control the tempo "smoothly" to convey their in-

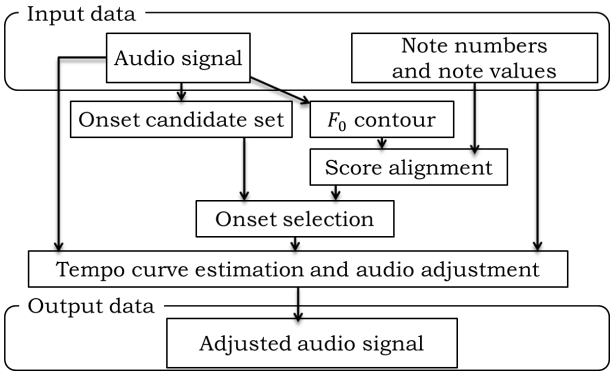

Fig. 2. Flowchart of the proposed method

terpretation of a musical score during a performance (Fig.1 left). This means that a "tempo curve" exists, which has been described in many previous studies. Meanwhile, in the performances of low-proficiency players, the tempo is not smoothly changed; in other words, it is changed "discretely" (Fig. 1 right).

We assume that if the composer does not notate extreme tempo changes in a phrase, then the amateur player as described above tries to smoothly control the tempo. However, in reality, these musicians cannot control their instruments as intended; thus, deviances from the interpretation are included, as seen in Fig. 1 right. We consider these deviances to be one of the factors that makes a listener interpret "poor playing" and removes a sense of rhythm. Here we define this deviance as "*mis-played components*."

In the following sections, the "note onset timing" of mis-played performances in monophonic recordings is modeled. The "true tempo curve," which is the interpreted tempo, is fitted as a smooth curve (as undertaken by Takeda et al. [7]) by polynomial regression analysis. The mis-played recordings are adjusted by time-scale expansion and contraction using the true tempo curve.

Fig. 2 shows the overview of the proposed method. First, a mis-played audio signal and the "note numbers [1]" and "note values [2]" are used as score information inputs. Next, onset timing candidates and the fundamental frequency (F_0) contour are analyzed from the audio signal. Then, brief onset timing is detected by score alignment to the F_0 contour. Accurate onset timing is selected from the candidates using the brief onset timing. Finally, the true tempo curve is estimated using accurate onset timing, and the audio signal is adjusted by stretching.

2.1. Onset timing modeling for mis-played audio signals

Although there are various definitions of note duration depending on the type of instrument, here we define note duration as the interval between an onset time of the target note and the next onset time of a following target note; more correctly, we do not consider rest notes. Namely, if the score shows an eighth note and an eighth rest note, these are treated as a quarter note. This enables the note duration to be expressed using a note value and a tempo as follows:

$$\text{duration (s)} = \frac{60\ \text{(s/min)}}{\text{tempo (beats/min)}} \times \text{note value (beats)}. \quad (1)$$

Hence, the n^{th} note onset timing without mis-playing becomes the note duration sum of 1^{st} through $(n-1)^{th}$ as follows.

$$\text{onset timing}[n] = \sum_{m=1}^{n-1} \frac{60}{\text{tempo[m]}} \times \text{note value[m]}. \quad (2)$$

Here the onset timing vector is defined as $\boldsymbol{y} = (y[1], ..., y[N])^T$ whose components denote observed onset times of N notes including mis-played components as follows:

$$y[n] = \sum_{m=1}^{n-1} \frac{60}{b[m]} h[m] + e[n] \quad (3)$$

where $h[m]$ is the note value of the m^{th} note, $b[m]$ is the value of the true tempo curve of the m^{th} note, $e[n]$ is the value of mis-played deviance of the n^{th} note. To Simplify of regression issue, we assume that $e[n]$ has a normal distribution $\mathcal{N}(0, \sigma^2)$.

2.2. Musical signal adjustment

The intended note duration $\hat{z}[n]$ can be written using the estimated true tempo curve $\boldsymbol{b} = (b[1], ..., b[N])^T$ as follows:

$$\hat{z}[n] = \frac{60}{b[n]} h[n]. \quad (4)$$

The observed note duration of the n^{th} note can also be calculated from $y[n+1] - y[n]$. Here $y[N+1] = L_x / f_s$, where L_x denotes the data length of an acoustic signal, and f_s is the sampling rate.

Note duration adjustment is achieved by stretching the note duration using an expansion–contraction factor α, which is defined as follows:

$$\alpha[n] = \frac{\hat{z}[n]}{y[n+1] - y[n]}. \quad (5)$$

3. ONSET DETECTION

Many acoustic features can be used for onset detection, for example, changes in phase [8], complex spectra [9], and spectral flux [10]. An appropriate acoustic feature depends on the type of instrument and the playing style [2]. Especially difficult is onset detection from recordings of bowed string instruments played with a *legato* style.

Other approaches consider the alignment between the observed F_0 contour and note numbers using dynamic time warping (DTW) [11, 12] for aligning the audio signal and the score. However, accurate onset detection by DTW is difficult, because an accurate F_0 contour is often not calculated around the onset time due to an inharmonic spectrum.

In this study, to adjust diverse types of recordings, an acoustic feature that is valid for many types of instruments and playing styles must be used. Moreover, accurate alignment between the onset timing and the score is required.

To cover the various sound generation mechanisms of different musical instruments, complex mel spectra, which can consider aural characteristics, are used to detect candidate onset times. In addition, for accurate onset detection, a DTW alignment method is used to select the best onset time from the candidates.

[1] "Note numbers" are the unique numbers assigned to each note. In this study, "Middle C" (261.6 Hz) is defined as note number 60, and note A3 (440 Hz) is number 69. These definitions are the same as the MIDI note numbers.

[2] "Note values" are the normal length of notes. Here for example, a note value of a quarter note is defined as 1, that of half note is 2, and that of eighth note is 0.5

3.1. Detecting candidate onset times

In any performance of a musical score, the onsets of notes are recognizable to the audience. Therefore, the modeling of an acoustic feature for aural characteristics, complex mel-spectrum KL divergence, is used for onset detection. The mel scale is a perceptual scale of pitches based on a log-frequency scale [13].

For the cognitive modeling of sharp changes in the spectra, the difference between an observed mel spectrum $\boldsymbol{S}_{\psi,k}$ at a target time frame k and a predicted spectrum $\hat{\boldsymbol{S}}_{\psi,k}$ from the previous short time frame τ is calculated, where ψ is the mel-log frequency. For a auditory difference scale based on signal features, Kullback-Leibler Divergence (KLD) for onset detection [14] is extended to the complex frequency domain.

Each mel spectrum is calculated using

$$\boldsymbol{S}_{\psi,k} = \mathrm{mel}\left[\frac{|\boldsymbol{X}_{\omega,k}| + C}{\sum_{\omega}|\boldsymbol{X}_{\omega,k}| + C}\right] e^{j\mathrm{mel}[\phi_{\omega,k}]}, \tag{6}$$

$$\hat{\boldsymbol{S}}_{\psi,k} = \mathrm{mel}\left[\frac{|\boldsymbol{X}_{\omega,k-\tau}| + C}{\sum_{\omega}|\boldsymbol{X}_{\omega,k-\tau}| + C}\right] e^{j\mathrm{mel}[\hat{\phi}_{\omega,k}]}, \tag{7}$$

where τ is the time used for spectrum prediction, ω is the linear frequency, $|\boldsymbol{X}_{\omega,k}|$ and $\phi_{\omega,k}$ are an amplitude spectrum and a phase spectrum, respectively, obtained by the short-time Fourier transform (STFT), $\hat{\phi}_{\omega,k}$ is a predicted phase spectrum by a previous method [8], C is a constant that reduces the uncertainty of the white noise spectrum, and mel$[\cdot]$ denotes the spectral transformation to the mel-scale [15]. Here to detect the onset time with high resolution, the step size of the STFT is $0.001 \times f_s$ point (i.e. 1 ms), and the number of data points used in the STFT is $0.01 \times f_s$ point (i.e., 10 ms). A value of $\tau = 10$ ms (i.e., $0.01 \times f_s$ point) was used in accordance with the length of the STFT. The constant $C(= 0.2)$ was decided in accordance with a preliminary experiment.

Complex mel-spectrum KLD (CMKLD) $\mathcal{D}[k]$ at k is calculated using

$$\mathcal{D}[k] = \sum_{\psi}\left|\hat{\boldsymbol{S}}_{\psi,k}\log\frac{\hat{\boldsymbol{S}}_{\psi,k}}{\boldsymbol{S}_{\psi,k}}\right|, \tag{8}$$

$$= \sum_{\psi}\left|\hat{\boldsymbol{S}}_{\psi,k}\right|\sqrt{\left(\log\frac{\left|\hat{\boldsymbol{S}}_{\psi,k}\right|}{|\boldsymbol{S}_{\psi,k}|}\right)^2 + \left(\hat{\phi}_{\psi,k} - \phi_{\psi,k}\right)^2}. \tag{9}$$

As can be seen in Eq. 9, the CMKLD can consider a difference in the amplitude spectra and phase spectra. The first term in the square root $\log(\hat{\boldsymbol{S}}_{\psi,k}/\boldsymbol{S}_{\psi,k})$ is appropriate for the detection of a relative amplitude increase on the harmonic frequency bin due to the note onset, because it is more sensitive to an increase in the denominator than a decrease. The second term in the square root $(\hat{\phi}_{\psi,k} - \phi_{\psi,k})^2$ is also appropriate for the detection of sharp changes in F_0, because almost all phases depend on F_0.

Next, by selecting the peak value of $\mathcal{D}$, a set of candidate onset times $\mathcal{Y}$ is generated. To determine the dynamic threshold for peak picking, $\boldsymbol{d}_k$ denotes a time series of CMKLD that closed-interval $[k - \mathcal{T}/2, k + \mathcal{T}/2]$ by centering on the index k as follows:

$$\boldsymbol{d}_k = \left(\mathcal{D}\left[k - \frac{\mathcal{T}}{2}\right], \cdots, \mathcal{D}\left[k + \frac{\mathcal{T}}{2}\right]\right)^T. \tag{10}$$

In this study, $\mathcal{T} = 100$ ms (i.e. $0.1 \times f_s$ point) was selected to ensure sufficient data points for the threshold calculation. The

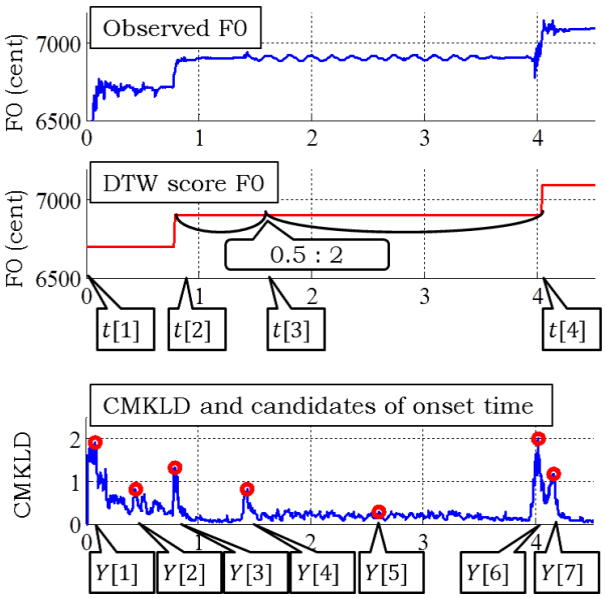

Fig. 3. An example of onset selection in the case of note values $= (0.5, 0.5, 2, 0.5, ...)$ and note numbers $= (64, 69, 69, 71, ...)$. The x-axis of both figures is time (s).

times that have a $\mathcal{D}$ peak larger than the dynamic threshold $\delta[k]$ are selected as onset time candidates $\mathcal{Y}$. The threshold is defined building on a previous study [2] as follows:

$$\mathcal{D}_{th}[k] = \lambda(\sigma_d + \mathrm{Median}(\boldsymbol{d}_k)) + \frac{\mathrm{Median}(\mathcal{D})}{2}, \tag{11}$$

where σ_d is the standard deviation of $\boldsymbol{d}_k$, and λ is set so as to satisfy $|\mathcal{Y}| \geq N$. In the preliminary experiments using the onset detection of a plucked string instrument and a bowed string instrument, a sufficient detection performance of $\lambda = 0.9$ or 1 was obtained.

3.2. Onset selection

The next step was onset selection from the candidate set $\mathcal{Y}$

First, the F_0 contour was analyzed from the audio signal using a F_0 estimation method YIN [16]. Next, the score F_0 contour was created using input note values and note numbers, which is aligned to the observed F_0 contour by the DTW method [11] (Fig. 4 middle figure). Here $t[n]$ was selected as the pitch switch timing in the stretched score F_0 contour by DTW.

If note numbers of successive notes do not change, $t[n]$ was calculated by scaling with respect to the note value ratios of these notes. For example, in Fig. 3, $t[3]$ was not selected by the pitch switch algorithm; hence, $t[2]$ and $t[4]$ are scaled by the note value ratio of 0.5:2 to calculate $t[3]$. If, successively in note values exist above three or more notes, $t[n]$ values were calculated in the same scaling manner. Finally, the onset times were selected as the times that had minimum values of CMKLD weight distance (Fig. 3) determined as follows:

$$y[n] = \underset{Y[i]\in\mathcal{Y}}{\mathrm{argmin}}\frac{|Y[i] - t[n]|}{\mathcal{D}\left[Y[i]\right]}. \tag{12}$$

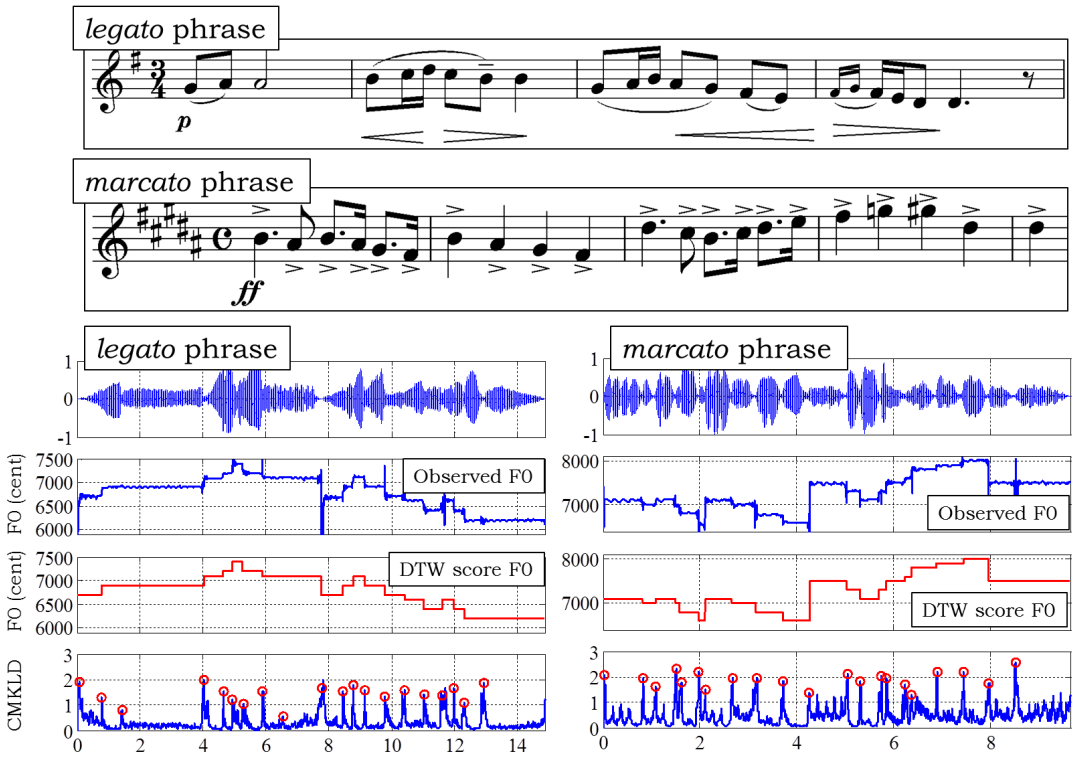

Fig. 4. Detected onsets in *legato* (left) and *marcato* (right) phrases. The top figures show the waveforms, the middle figures depict the F_0 contour and aligned score F_0 contour, and the bottom figures show the complex mel spectra KLD and selected onset times (circles).

Fig. 4 shows examples of onset detection from monophonic violin recordings of a *legato* phrase and a *marcato* phrase. Onset detection in the *legato* phrase, which has been traditionally difficult, is resolved by the proposed onset detection method. Onset timing in the *marcato* phrase was also successful using this method.

4. AUDIO SIGNAL ADJUSTMENT USING AN ESTIMATED TEMPO CURVE

In this section, the true tempo curve $\boldsymbol{b}$, which is the tempo fluctuation intended by the performer, is estimated by the estimated onset timing vector $\boldsymbol{y}$. In addition, the automatic audio signal adjusting method using the estimated true tempo curve is also described.

4.1. Tempo curve estimation

Polynomial regression was undertaken on the true tempo curve to fit the curve to a model. This modeling is an extension of the curve fitting presented by Takeda et al. [7]. The proposed model is defined as follows:

$$b[n]^{-1} = \sum_{p=0}^{P} w_p g[n]^p, \quad g[n] = \sum_{m=1}^{n-1} h[m], \tag{13}$$

where P is the degree of polynomial regression. Here $g[n]$ denotes the accumulated note value. Thus, from Eq. 1, 2, 3 and 13, n^{th} note duration $\Delta y[n]$ can be written as

$$\begin{aligned}\Delta y[n] &= y[n+1] - y[n] = \frac{60}{b[n]} h[n] + e[n+1] - e[n], \\ &= 60 \sum_{p=0}^{P} w_p g[n]^p h[n] + e[n+1] - e[n]. \end{aligned} \tag{14}$$

Here the explanatory variable matrix $\boldsymbol{G}$ is defined as

$$\boldsymbol{G} = \begin{pmatrix} h[1] & g[1]h[1] & \dots & g[1]^P h[1] \\ h[2] & g[2]h[2] & \dots & g[2]^P h[2] \\ \vdots & \vdots & \ddots & \vdots \\ h[N] & g[N]h[N] & \dots & g[N]^P h[N] \end{pmatrix}, \tag{15}$$

and then the observed note duration vector $\Delta\boldsymbol{y} = (\Delta y[1], ..., \Delta y[N])^T$ is rewritten as

$$\Delta\boldsymbol{y} = 60\boldsymbol{G}\boldsymbol{w} + \Delta\boldsymbol{e}, \tag{16}$$

where $\boldsymbol{w}$ is the regression coefficient vector $\boldsymbol{w} = (w_0, ..., w_P)^T$ and $\Delta\boldsymbol{e}$ is the delta vector of mis-played components $\Delta\boldsymbol{e} = (e[2] - e[1], e[3] - e[2], ..., -e[N])^T$. From the regeneration of normal distribution, each component of $\Delta\boldsymbol{e}$ is also normally distributed.

Accordingly, by calculating the regression coefficient vector $\boldsymbol{w}$ by the least-squares method (LSM), the true tempo curve can be calculated using Eq. 13. Selecting the optimal polynomial degree P is an issue when using LSM, and in this study, the polynomial degree was determined the by minimization of Akaike information criterion (AIC) [17].

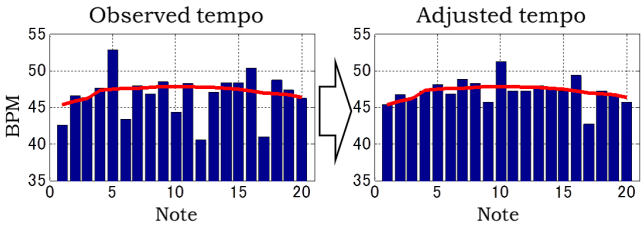

Fig. 5. An example of an adjusted tempo fluctation. Observed tempo (left bar plot), estimated tempo curve (line), and adjusted tempo (right bar plot).

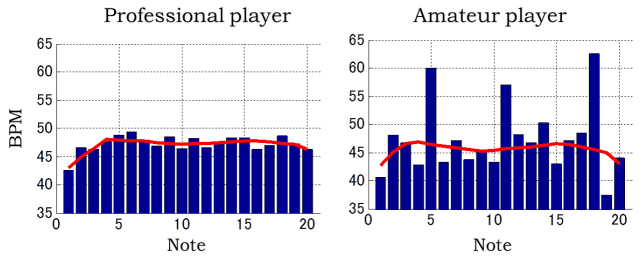

Fig. 6. Observed tempos (bar plot) and estimated true tempo curves (line). The tempo of the left figure was estimated from a professional recording and that of the right figure was estimated from an amateur recording that mimicked the professional one.

4.2. Musical signal adjustment by expansion and contraction

By using the true tempo curve estimated as described in the previous sections, each note is stretched. Expansion and contraction were performed by a time-scale modification method based on the expansion and contraction of the step size of an inverse Fourier transform of the power spectrogram [18]. Phase inconsistency caused by varying the width of the frame shift is removed using a phase reconstruction method, as proposed by Griffin et al. [19]. The stretch factor $\alpha[n]$ is calculated using Eq. 5.

In this study, the acoustic signals were expanded and contracted by multiplying the $\alpha[n]$ obtained for each note with the inverse Fourier transform step size (which is obtained by subtracting the overlap size of overlap-add from FFT length) of each note.

Fig. 5 shows an example of a tempo that has been adjusted. The left figure shows a tempo fluctuation including the mis-played components, and the right figure shows this same tempo fluctuation adjusted using the method described. It can be seen that the adjusted tempo fluctuation more closely matches the true tempo curve (red line).

4.3. Discussion

Fig 6 shows observed tempo curves and estimated true tempo curves of a professional recording (left) and an amateur recording (right). The amateur recording was played by mimicking the professional recording.

The true tempo curve of the professional recording is quite close to the observed tempo curve. This indicates that the tempo is smoothly varying, as mentioned in many previous studies, and that the professional player controls the tempo well. Meanwhile, the observed tempo of the amateur performance deviates from the true tempo curve. In addition, because the amateur mimicked the professional recording, the true tempo curve is quite similar to that of the professional recording. This result denotes suggests that a low-proficiency amateur player cannot control the tempo well.

5. EVALUATION

5.1. Evaluation of onset detection

This section presents the experimental results of the onset detection method described in Chapter 3. In general, for the evaluation of the onset detection method, the detected onset times are compared with the true onset points that were manually selected and labeled. Correct matches imply that the target and detected onsets were within a 50-ms window. The detected result is evaluated on precision, recall, and F-measure.

For the evaluation of the proposed method, these criteria are not valid, because the method is aligned to the score information. Accordingly, the accuracy is evaluated by the mean absolute error (MAE) between the detected onset times and the manually selected ones.

The experiments were performed on a database of solo violin recordings. There were six phrases that included some expression marks, for example *legato*, *marcato*, and *feroce* (wildly). All signals were processed as monaural signals sampled at 48 kHz and 24 bit. There were 152 onsets in total. These phrases included simple phrases (e.g., almost all notes being quarter notes) and complex phrases (e.g., including many shorter notes), and the assigned BPM of the score was 50–150. The performance duration of the phrases was 14–30 s.

The resulting MAE was 18 ms. This error is less than 1/15 of a quarter note duration in the tempo of "Allegro (BPM $\approx$ 120 , fast, quickly and bright)." In addition, the sound duration that is required for pitch perception is 20–30 ms; thus, this accuracy can be considered to be adequate.

5.2. Motion experiment for tempo adjustment

In this section, the accuracy of the proposed adjustment method was evaluated. The BPM including the mis-played components $\tilde{\boldsymbol{b}}$ was calculated from pre-adjusted and post-adjusted sound onset timings as follows:

$$\tilde{b}[n] = \frac{60h[n]}{\Delta y[n]}. \tag{17}$$

The rhythm-weighted mean absolute error (weighted-MAE) between the observed BPM $\tilde{\boldsymbol{b}}$ and the true tempo curve $\boldsymbol{b}$ was calculated as

$$\text{weighted-MAE} = \frac{\sum_n h[n] \left| b[n] - \tilde{b}[n] \right|}{\sum_n h[n]}. \tag{18}$$

If the proposed method is successful in removing the mis-played components, then the weighted MAE will decrease.

The experiments were performed on five solo violin recordings. The assigned BPM of these phrases were 50–150, and the total number of notes was 154. The performance duration of the phrases was 10–30 s. All signals were recorded as monaural signals sampled at 48 kHz and 16 bit with an IC recorder.

Table 1. Weighted-MAE of BPM.

Before adjusting	After adjusting
5.8677	2.6834

Table 2. Phrases used in subjective evaluations.

Violin	A. Dvorak, "Symphony No. 8"
	1. 1st mov. bar 244–250 1st Violin
	R.Wagner, "Tannhauser Act.II 'Grand March' "
	2. bar 40–44 1st Violin
	3. bar 64–68 1st Violin
Cello	A. Dvorak, "Symphony No. 8"
	1. 1st mov. bar 1–6
	2. 1st mov. bar 165–169
	3. 4th mov. bar 26–33
E.guitar	1. LUNKHEAD "ENTRANCE" 5–12 bars
	2. MONKEY MAJIK "Aishiteru" bar 52–56
	3. T. Matsumoto "Thousand Dreams" bar 2–9

Table 1 shows the weighted-MAE values before and after adjusting. The deviation from the true tempo curve as a result of the mis-played components is seen to be reduced by half. From this result, it was confirmed that the proposed method successfully decreased the mis-played components by time-scale audio signal stretching for violin recordings.

5.3. Subjective evaluation

Finally, an opinion test was conducted. Target instruments were the violin, cello, and electric guitar (clean tone). These experiments used sound data from performances by university students with over three years of experience playing instruments. After the student players had practiced for 30 min while listening to the performance of professionals, they tried to mimic the intended representation of the performance without using a metronome. Therefore, the true tempo curve should resemble the mimicked recording tempo curve, and the adjusted signals should also resemble professional recordings.

There were two players for each instrument, and they played three phrases. All sounds were recorded at a sampling rate of 48 kHz and bit quantization of 16 bit. Sounds from the violin and cello were recorded with an IC recorder and those from the guitar were recorded with a line input. The assigned BPM of these phrases was 60–180, and the average number of notes per phrase was 22. The performance duration of the phrases was 9–16 s.

For a subjective evaluation, five musicians each with over five years of experience playing instruments served as evaluators. These musicians were different people in addition to the recording musicians. The evaluators compared the sound of the professional performances to the original pre-adjusted sounds (ORG) and the sounds adjusted using our method (PRO). The sound pressure was adjusted in advance to facilitate listening to the samples. For the recorded and adjusted sounds, the closeness of rhythm to the professional performance was evaluated. The mean opinion score (MOS) for each sound was used as the metric, with a scale of 1 (very far) to 5 (very close).

Figure 7 shows the MOS values and standard error for each instrument. The MOS values of the method proposed here are higher than those of the recorded sound for all instruments. There were significant differences between the recorded sounds and the adjusted sounds for all tested instruments as ascertained by the Student's t-test (significance levels were 1 % (violin and E. guitar) and 5 % (cello)). The musicians had played as intended for the professional performance tempo fluctuation, and the adjustment using the proposed method brought the sounds significantly nearer the intended performance. Thus, our method can be concluded to clarify the intent of the player from the acoustic signal.

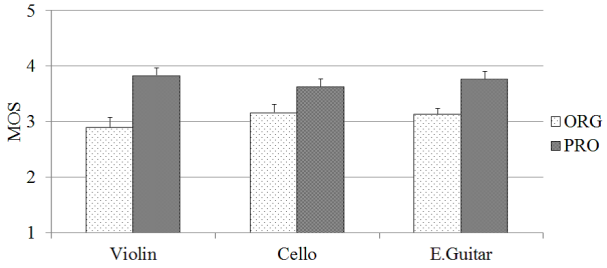

Fig. 7. Results of the subjective evaluation.

6. CONCLUSION

In this paper, a "true tempo curve" estimation method was proposed using monophonic audio signals including performance errors. True tempo curves were estimated by polynomial regression analysis of observed onset timings. In addition, an automatic audio signal modification was proposed, which can adjust the signal by time-scale modification with the estimated true tempo curve. In a subjective evaluation, amateur performances, which tried to mimic professional recordings, were adjusted. In terms of the closeness of rhythm, the MOS values of these adjusted sounds were higher than those of the pre-adjusted original sounds, and significant differences were observed for all tested instruments. The musicians had played with the intention of expressing a professional performance tempo fluctuation, and the adjustment using the proposed method brought the sounds significantly nearer the intended performance. Hence, it can be concluded that the proposed method can estimate a musician's expressive intentions.

Evaluation of an audio signal adjustment method requires a mis-played signal dataset whose performance intention is known. However, such a dataset does not currently exist. In this evaluation, a small quantity of performance expression data for violin, cello and, guitar were used. For future experiment, it is necessary to develop a large dataset.

In this study, rest notes were not considered. In actual musical performances, rest notes can be a prominent part of the interpretation. In addition, "*trill*" and extreme "*fermata*" are also important parts of the interpretation, but the DTW of the F_0 contour could not be successfully estimated for these expressions. In the future, this method will be adapted to include these issues, for example, by considering offset detection.

For excitation-continuous musical instruments, a musical tone generally has three possible states: *onset*, *steady*, and *offset*. In particular, the duration of the *onset-state* is closely related to the perception of musical expression [20]. In future, we need to con-

sider these states, and notes should be stretched only in the *steady-state*.

Moreover, intended tempo fluctuations exist; for example, in "*Swing*" and "*Viennese waltz style*", the tempo does not change smoothly. In future, true tempo curve and these fluctuations should be treated separately by considering the musical genre.

As future prospects, the estimated true tempo curve presented here can be regarded as an extraction of the performance expression feature from recordings including background " noise " during the performance. Performance characterization is applied to many applications such as automatic performer identification [21] and musical sound synthesis [22]. We will consider to apply this method to these applications.

7. REFERENCES

[1] S. Dixon, "Onset detection revisited," in *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx-06)*, 2006, pp. 133–137.

[2] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M.B. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 13, no. 5, pp. 1035–1047, Sept. 2005.

[3] E. D. Scheirer, "Tempo and beat analysis of acoustical musical signals," *J. Acoust. Soc. Amer.*, vol. 103, no. 1, pp. 588–601, 1998.

[4] D. P. W. Ellis, "Beat tracking by dynamic programming," *J. New Music Res.*, vol. 36, no. 1, pp. 51–60, 2007.

[5] Peter Grosche and Meinard Muller, "Extracting predominant local pulse information from music recordings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1688–1701, 2011.

[6] M. Marchini, P. Papiotis, and E. Maestre, "Timing synchronization in string quartet performance: a preliminary study," in *International Workshop on Computer Music Modeling and Retrieval (CMMR12)*, 2012, pp. 117–185.

[7] Haruto Takeda, Takuya Nishimoto, and Shigeki Sagayama, "Rhythm and tempo recognition of music performance from a probabilistic approach," in *Proceedings of 5th International Conference on Music Information Retrieval (ISMIR)*, Oct. 2004, pp. 357–364.

[8] J. P. Bello and M. Sandler, "Phase-based note onset detection for music signals," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-03)*, 2003, pp. 49–52.

[9] J. P. Bello, C. Duxbury, M. Davies, and M. Sandler, "On the use of phase and energy for musical onset detection in the complex domain," *IEEE Signal Processing Letters*, vol. 11, no. 6, pp. 553–556, 2004.

[10] P. Masri, "Computer modelling of sound for transformation and synthesis of musical signal," *Ph.D. dissertation, University of Bristol, UK,*, 1996.

[11] Norman H. Adams, Mark A. Bartsch, Jonah B. Shifrin, and Gregory H. Wakefield, "Time series alignment for music information retrieval," in *Proceedings of 5th International Conference on Music Information Retrieval (ISMIR)*, 2004, pp. 303–310.

[12] Norman H. Adams, "Note segmentation and quantization for music information retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 131–141, 2006.

[13] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Am.*, vol. 8, no. 3, pp. 185–190, 1937.

[14] S. Hainsworth and M. Macleod, "Onset detection in musical audio signals," in *Proceedings of ICMC*, 2003.

[15] S. S. Stevens and J. Volkmann, "The relation of pitch to frequency: A revised scale," *The American Journal of Psychology*, vol. 53, no. 3, pp. 329–353, 1940.

[16] Alain de Cheveigne and Hideki Kawahara, "Yin, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[17] H. Akaike, "Information theory and an extention of the maximum likelihood principle," in *Proceedings of the 2nd International Symposium on Information Theory*, 1973, vol. 1, pp. 267–281.

[18] Yuu Mizuno, Nobutaka Ono, Takuya Nishimoto, and Shigeki Sagayama, "Real-time modification of time-scale and pitch based on expansion/contraction of power spectrogram of polyphonic signals," in *Trans. Tech. Comm. Psychol. Physiol. Acoust. (in Japanese)*, 2009, vol. 39, pp. 447–452.

[19] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[20] Knut Guettler and Anders Askenfelt, "Acceptance limits for the duration of pre-helmholtz transients in bowed string attacks," *Journal of the Acoustical Society of America*, vol. 101, no. 5, pp. 2903–2913, 1997.

[21] Rafael Ramirez, Esteban Maestre, and Xavier Serra, "Automatic performer identification in commercial monophonic jazz performances," *Pattern Recognition of Non-Speech Audio*, vol. 31, no. 12, pp. 1514–1523, 2010.

[22] Tomoyasu Nakano and Masataka Goto, "Vocalistener: A singing-to-singing synthesis system based on iterative parameter estimation," in *Proceedings of the 6th Sound and Music Computing Conference (SMC 2009)*, 2009, pp. 343–348.