Automatic music thumbnailing using localization information of audio object

Hiroyuki Nawata, Noriyoshi Kamado, Hiroshi Saruwatari and Kiyohiro Shikano
Nara Institute of Science and Technology, Nara, 630-0192 Japan
E-mail: {hiroyuki-n, sawatari, shikano}@is.naist.jp Tel: +81-743-72-5287

***Abstract*—In this study, in order to automatically generate thumbnail music that has a main part of the original tune, we propose a new estimation method of structure changes in stereo tunes based on localization information. The proposed method can estimate the main parts of the music tune by analyzing specific timing when localization information changes under the assumption that the changing time of the localization approximately corresponds to timing of the music structure change. We evaluate the effectiveness of the proposed method by the objective assessment in this paper. The experimental result shows that the proposed method can detect correct timing of music structure changes with more than 70 percent accuracy rate.**

I. Introduction

Over the past decade, main means for conveying music to us have changed from optical digital audio discs such as compact discs or digital versatile discs to electronic data such as music files that can be available over the network. Hence we are now able to obtain music easily, regardless of time and place. On the other hand, it is complicated to search an objective music tune from a huge number of music files. Therefore, realization of the system in which we can find the objective music tune easily is a problem requiring urgent attention. Such systems include a function to allow us to preview thumbnail music that has only a main part of the original music tune. By previewing thumbnail music, we can comprehend the image of the tune without listening to it all over and judge easily whether the original music tune is the objective one. However, thumbnail music is generated manually now and it is difficult to generate a huge number of them manually, so the automatic music thumbnailing becomes an essential problem to be recently addressed [1][2]. Therefore, in this paper, we propose a new method to estimate structure changes in multichannel (mostly *stereo*) tunes for achieving the automatic music thumbnailing.

It generally changes which musical instrument group (audio object) is active in each composition section in order to put the intonation in the music tune. For example, a specific audio object in solo parts and almost all the objects in introduction parts become active to make upsurge. Moreover, it is well expected that the number of active audio objects changes before or after these parts, and to be performed by a different composition section of the musical instruments. Thus, by analyzing information on changes of the number of active audio objects and their locations (audio object localization) in each composition section, it would appear that the main composition section of music can be specified.

We have proposed an estimation method of audio objects in the multichannel tunes [3][4] previously. In this paper, we utilize this method for decomposing the mixed audio objects into each active object and capturing the stream of the audio object localization. We propose a new estimation method of musical composition section by analyzing changes of the audio object localization. In addition, we evaluate the effectiveness of the proposed method by the objective assessment.

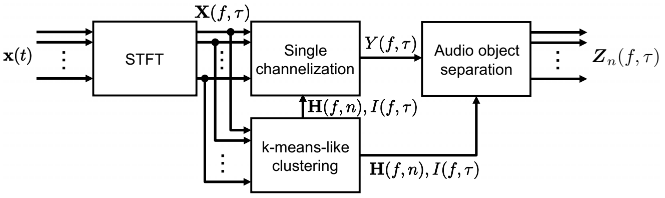

Fig. 1. Block diagram of audio object estimation.

We introduce the estimation method of audio objects in section II. Next, we propose a new estimation method of music structure based on the audio object localization in section III. Next, we evaluate the effectiveness of the proposed method by the objective assessment in section IV. The conclusion of this paper is described in section V at the end.

II. Estimation method of audio object[3][4]

In this section, we describe an estimation method of the audio objects in the music signal. First, the arrangement of the audio objects are assumed so that they do not change in the tune. This method quantizes localization information in the tune by some quantization vectors based on the above-mentioned assumption and then estimates the audio objects. Figure 1 shows the block diagram of the estimation method of audio objects, where we consider that the M-channel input signal has N audio objects.

First, we apply short-time Fourier analysis to the M-channel time-series input signal $\boldsymbol{x}(t)$ and then obtain the time-frequency input signal $\boldsymbol{X}(f,\tau) = [X_1(f,\tau),\ldots,X_M(f,\tau)]^{\mathrm{T}}$. Moreover, we define the M-dimensional complex unit vector $\boldsymbol{H}(f,n) = [H_1(f,n),\ldots,H_M(f,n)]^{\mathrm{T}}$ (n=1, 2, …, N) as the quantization vector. Also, f denotes the frequency subband, τ is the frame index, n is the class index, N is the number of the quantization vectors, and superscript T is the transposition of a matrix. The decoded signal vector $\boldsymbol{Z}(f,\tau)$ that minimizes the quantization error with the input signal $\boldsymbol{X}(f,\tau)$ on the quantization vector $\boldsymbol{H}(f,n)$ is expressed in terms of the orthogonal projection of $\boldsymbol{X}(f,\tau)$ and $\boldsymbol{H}(f,n)$ as follows:

$$\begin{aligned}\boldsymbol{Z}(f,\tau) &= \frac{\boldsymbol{H}^{\mathrm{H}}(f,I(f,\tau))\boldsymbol{X}(f,I(f,\tau))}{\boldsymbol{H}^{\mathrm{H}}(f,I(f,\tau))\boldsymbol{H}(f,I(f,\tau))}\boldsymbol{H}(f,I(f,\tau)) \\ &= \left\{\boldsymbol{H}^{\mathrm{H}}(f,I(f,\tau))\boldsymbol{X}(f,\tau)\right\}\boldsymbol{H}(f,I(f,\tau)), \qquad (1)\end{aligned}$$

where superscript H denotes the complex conjugate transposition of a matrix and $I(f,\tau)$ is the index of the centroid that minimizes the quantization error between $\boldsymbol{X}(f,\tau)$ and $\boldsymbol{Z}(f,\tau)$ at every time-frequency grid in all channels. The quantization error $E(\boldsymbol{X}(f,\tau),\boldsymbol{H}(f,n))$ between $\boldsymbol{X}(f,\tau)$ and

Proceedings of the Second APSIPA Annual Summit and Conference, pages 915–918,
Biopolis, Singapore, 14-17 December 2010.

$\boldsymbol{Z}(f,\tau)$ is defined as follows:

$$E(\boldsymbol{X}(f,\tau),\boldsymbol{H}(f,n)) = \|\boldsymbol{X}(f,\tau)\| \sin(\boldsymbol{X}(f,\tau),\boldsymbol{H}(f,n)) = \|\boldsymbol{X}(f,\tau)\| \sqrt{1-\cos^2(\boldsymbol{X}(f,\tau),\boldsymbol{H}(f,n))}, \quad (2)$$

where $\|\cdot\|$ denotes the Euclidean norm and $\cos(\boldsymbol{X}(f,\tau),\boldsymbol{H}(f,n))$ is the cosine-distance between $\boldsymbol{X}(f,\tau)$ and $\boldsymbol{H}(f,n)$, as

$$\cos(\boldsymbol{X}(f,\tau),\boldsymbol{H}(f,n)) = \frac{|\boldsymbol{X}^{\mathrm{H}}(f,\tau)\boldsymbol{H}(f,n)|}{\|\boldsymbol{X}(f,\tau)\|}. \quad (3)$$

We optimize the basis vector $\boldsymbol{H}(f,n)$ so that the total error of the signals, E_{total}, is minimized. The total error is given by

$$E_{\mathrm{total}}(f) = \sum_n \sum_{\tau\in\Theta_n} \Big(E(\boldsymbol{X}(f,\tau),\boldsymbol{H}(f,n))\Big)^2, \quad (4)$$

where Θ_n is the nth class of the cluster. Optimization of the quantization vector $\boldsymbol{H}(f,n)$ is equivalent to the k-means clustering problem for the cosine-distance as follows.

[STEP 1] Prototype centroid $\boldsymbol{C}^{(k)}(f,n)$ $(n = 1,\ldots,N)$ is generated. Also, $\boldsymbol{C}^{(k)}(f,n)$ is the centroid of the nth class when the iteration times of k is updated.

[STEP 2] Each input signal $\boldsymbol{X}(f,\tau)$ is assigned to the nth class Θ_n based on the error between the input signal $\boldsymbol{X}(f,\tau)$ and the centroid vector $\boldsymbol{C}^{(k)}(f,n)$ as follows:

$$\Theta_{I^{(k)}(f,\tau)} = \{\tau\}, \quad (5)$$

$$I^{(k)}(f,\tau) = \underset{n}{\mathrm{argmin}}\ E(\boldsymbol{X}(f,\tau),\boldsymbol{C}^{(k)}(f,n))^2, \quad (6)$$

where $\mathrm{argmin}_{\mathrm{n}}\ \cdot$ denotes the minimization function, $\{\cdot\}$ denotes the class that corresponds to a set of $\cdot$, and $I^{(k)}(f,\tau)$ is the index function of the kth iteration.

[STEP 3] The optimal basis vector is extracted to minimize the error, as

$$\begin{aligned}
&\boldsymbol{C}^{(k+1)}(f,n)\\
&= \underset{\boldsymbol{C}^{(k)}(f,n)}{\mathrm{argmin}} \sum_{\tau\in\Theta_n} E(\boldsymbol{X}(f,\tau),\boldsymbol{C}^{(k)}(f,n))^2\\
&= \underset{\boldsymbol{C}^{(k)}(f,n)}{\mathrm{argmin}} \sum_{\tau\in\Theta_n} \|\boldsymbol{X}(f,\tau)\|^2\Big(1-\cos^2(\boldsymbol{X}(f,\tau),\boldsymbol{C}^{(k)}(f,n))\Big)\\
&= \underset{\boldsymbol{C}^{(k)}(f,n)}{\mathrm{argmin}} \sum_{\tau\in\Theta_n} \|\boldsymbol{X}(f,\tau)\|^2\left(1-\frac{|\boldsymbol{X}^{\mathrm{H}}(f,\tau)\boldsymbol{C}^{(k)}(f,n)|^2}{\|\boldsymbol{X}(f,\tau)\|^2\|\boldsymbol{C}^{(k)}(f,n)\|^2}\right)\\
&= \underset{\boldsymbol{C}^{(k)}(f,n)}{\mathrm{argmin}} \sum_{\tau\in\Theta_n} -|\boldsymbol{X}^{\mathrm{H}}(f,\tau)\boldsymbol{C}^{(k)}(f,n)|^2\\
&= \underset{\boldsymbol{C}^{(k)}(f,n)}{\mathrm{argmax}} \boldsymbol{C}^{(k)}(f,n)^{\mathrm{H}}\Big(\sum_{\tau\in\Theta_n}\boldsymbol{X}(f,\tau)\boldsymbol{X}^{\mathrm{H}}(f,\tau)\Big)\boldsymbol{C}^{(k)}(f,n).
\end{aligned} \quad (7)$$

Owing to the constraint $\|\boldsymbol{C}^{(k)}(f,n)\| = 1$, the maximization problem on the right-hand side of Eq. (7) is equivalent to finding the maximum eigenvalue of $\sum_{\tau\in\Theta_n}\boldsymbol{X}(f,\tau)\boldsymbol{X}^{\mathrm{H}}(f,\tau)$. Therefore, the basis vector $\boldsymbol{C}^{(k+1)}(f,n)$ is derived as the maximum eigenvector of $\sum_{\tau\in\Theta_n}\boldsymbol{X}(f,\tau)\boldsymbol{X}^{\mathrm{H}}(f,\tau)$.

[STEP 4] If the centroid vector does not change from that obtained by the previous iteration in **STEP 3**, the optimal vector $\boldsymbol{C}^{(k)}(f,n)$ is determined to be the basis vector $\boldsymbol{H}(f,n)$. Then, the class n is determined as follows:

$$I(f,\tau) = I^{(k)}(f,\tau). \quad (8)$$

If the centroid vector changes, the algorithm returns to **STEP 2** with $k = k+1$.

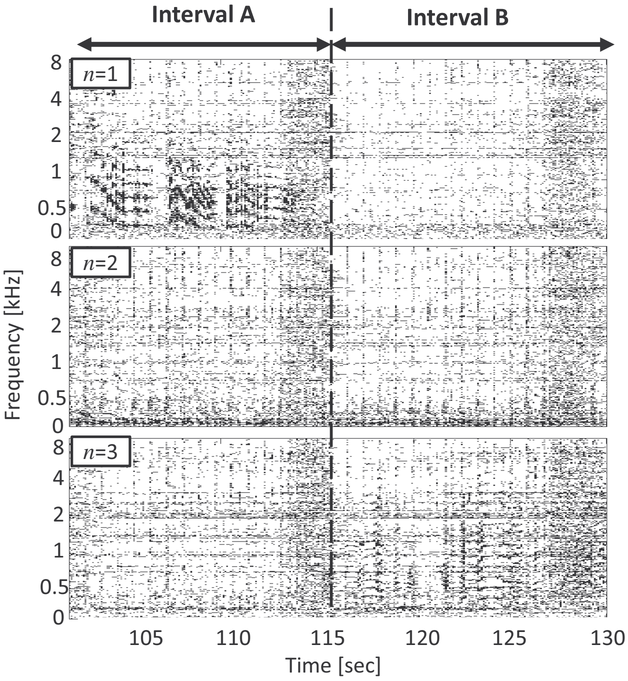

Fig. 2. Audio object localization of audio signal.

The single-channel encoded signal $Y(f,\tau)$ in the Fig. 1 is obtained as follows:

$$Y(f,\tau) = \boldsymbol{H}^{\mathrm{H}}(f,I(f,\tau))\boldsymbol{X}(f,\tau), \quad (9)$$

where $\boldsymbol{H}(f,I(f,\tau))$ denotes the optimal quantization vector at every time-frequency grid. The nth clustered audio object signal included in the tune is expressed by applying $\boldsymbol{H}(f,I(f,\tau))$ to the single-channel signal $Y(f,\tau)$ under the constraint $I(f,\tau) = n$. Therefore, the nth audio object time-frequency signal $\boldsymbol{Z}_n(f,\tau)$ is given by

$$\boldsymbol{Z}_n(f,\tau) = \begin{cases} Y(f,\tau)\boldsymbol{H}(f,I(f,\tau)) & (I(f,\tau)=n)\\ 0 & (otherwise)\end{cases}. \quad (10)$$

III. Analyzing method of music structure based on audio object localization

A. Audio object localization estimation

In this section, we propose a new estimation method of music structure using the audio object localization derived from the audio object signals in multichannel tunes, which are estimated in the previous section. As described in Sect. I, the main composition section of the tune would be specified by analyzing changes of the audio object localization, and in the following, the full detail of the estimation method is shown.

First, we replace the constraint $I(f,\tau) = n$ shown in Eq. (10) with the masking function $W_n(f,\tau)$ as follows:

$$\boldsymbol{Z}_n(f,\tau) = W_n(f,\tau)Y(f,\tau)\boldsymbol{H}(f,I(f,\tau)), \quad (11)$$

$$W_n(f,\tau) = \begin{cases} 1 & (I(f,\tau)=n)\\ 0 & (otherwise)\end{cases}, \quad (12)$$

where $W_n(f,\tau)$ is regarded as a function which shows an existence of the nth audio object at every time-frequency grid, so we call $W_n(f,\tau)$ audio object localization.

Figure 2 shows an audio object localization of a specific tune, which is derived by the proposed estimation method. There are three audio objects, namely, a trumpet (the first audio object; $n = 1$), drums and a bass guitar (the second audio object; $n = 2$) and a saxophone (the third audio object; $n = 3$) in this tune. The interval A shows the trumpet solo part, the interval B shows the saxophone solo part and the second audio object is performed in all intervals. A black dot means $W_n(f,\tau) = 1$ and the other means $W_n(f,\tau) = 0$. The broken line shows timing when the music structure actually changes. Then Fig. 2 shows that the intervals of solo parts of the first and the third audio objects and the low-frequency part of the second audio object are crowded with black dots. Thus, an audio object localization and its density express the existence of an audio object roughly.

B. Estimation of music structure based on audio object localization

In the previous section, a possibility that timing of the music structure change is estimated by tracing the changes of the audio object localization was shown. Next, in this subsection, we quantify the amount of changes on the audio object localization, and propose a new method to estimate the music structure based on localization information.

Figure 2 shows that the audio object exists in a specific interval where a time-frequency density of the audio object localization is high. Therefore, we quantify the density by a vector $\boldsymbol{D}(\tau)$ defined as

$$\boldsymbol{D}(\tau) = [D_n, \cdots, D_N(f,\tau)], \tag{13}$$

$$D_n(\tau) = \frac{W_{Sn}(\tau)}{\sum_n W_{Sn}(\tau)}, \tag{14}$$

where $W_{Sn}(\tau)$ denotes the value multiplying the total of each audio object localization by the frequency weighting function $m(f)$ at each frame τ, i.e.,

$$W_{Sn}(\tau) = \sum_f m(f) W_n(f,\tau). \tag{15}$$

We call $\boldsymbol{D}(\tau)$ localization density. Generally speaking, a fundamental tone exists in the low frequency range and the harmonic tone exists in the high frequency range in musical tunes. So we apply the frequency weighting function $m(f)$ to every frequency subband in Eq. (15) as follows:

$$m(f) = \begin{cases} 1 & (\log_2(F_f) \le 0) \\ \dfrac{1}{2^{\lceil \log_2(F_f) \rceil}} & (0 < \log_2(F_f) \le 3) \\ 0 & (3 < \log_2(F_f)) \end{cases}, \tag{16}$$

$$F_f = 10^{-3} f \frac{f_s}{f_n} \text{ [kHz]}, \tag{17}$$

where F_f denotes the frequency subband which expresses f with kHz, f_s is the sampling rate of the music signal, f_n is fast Fourier transform (FFT) points and $\lceil \cdot \rceil$ is the ceiling function. Equation (16) shows that $m(f)$ logarithmically decreases as the frequency increases, and is set to 0 if the frequency is above 8 kHz.

Figure 3 shows the localization density $\boldsymbol{D}(\tau)$ of a certain music tune. The broken line shows timing when the music structure actually changes. This music tune has three audio objects in the same arrangement as that described in the previous section. This figure indicates that the localization density drastically changes according to timing of the music structure change. However, the fine structure of the localization density behaves almost randomly regardless of the music structure change. This implies that the *raw* density would lead to miss estimation of timing. In order to prevent a wrong decision caused by the fine behavior of $\boldsymbol{D}(\tau)$, we quantize the localization density into some states. Those are classified by k-means clustering into some states which have prototype centroids $\boldsymbol{C}^{(0)}(l)$ of emphasized localization density (for example, $D_1(\tau) = 1$ and the others are set to 0) for the error $E(\tau)$ given by

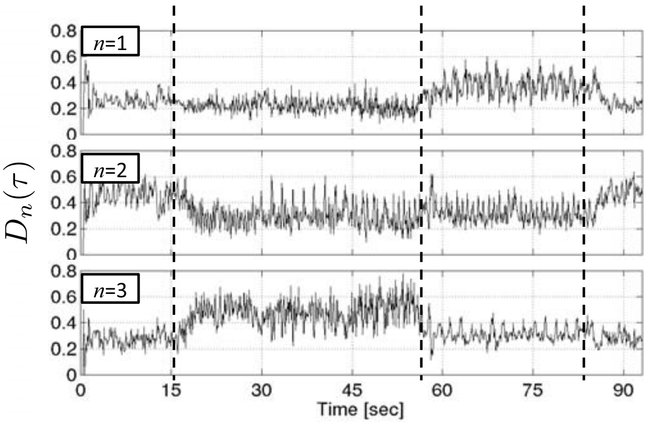

Fig. 3. Localization density of audio signal.

$$E(\tau) = E(\boldsymbol{D}(\tau), \boldsymbol{C}^{(k)}(l)), \tag{18}$$

where l is the class index, error $E(\tau)$ is the sine-distance between vectors, which is the same as Eq. (2). Moreover, the centroid which minimizes the error is denoted by $\boldsymbol{C}(l)$ and the class index of the centroid which minimizes the error at each frame is $J(\tau)$, then $J(\tau)$ is given by

$$J(\tau) = \underset{l}{\operatorname{argmin}}\ E(\boldsymbol{D}(\tau), \boldsymbol{C}(l)). \tag{19}$$

This clustering problem can be solved by the same manner in Sect. II. Also we define the existence function of $J(\tau)$ as $Q_l(\tau)$ similar to Eq. (12) and the smoothed function $\hat{Q}_{l,K}(\tau)$ averaging k-point samples before/after τ, as

$$\hat{Q}_{l,K}(\tau) = \frac{1}{2K+1} \sum_{k=-K}^{K} Q_l(\tau + k), \tag{20}$$

$$Q_l(\tau) = \begin{cases} 1 & (J(\tau) = l) \\ 0 & (otherwise) \end{cases}. \tag{21}$$

Figure 4 illustrates $\hat{Q}_{l,50}(\tau)$ of the localization density shown in Fig. 3, where the prototype centroids $\boldsymbol{C}^{(0)}(l)$ are $C^{(0)}(1) = [1,0,0]$, $C^{(0)}(2) = [0,1,0]$, $C^{(0)}(3) = [0,0,1]$ and $C^{(0)}(4) = [1/3,1/3,1/3]$.

Figure 4 shows that timing of the music structure change corresponds to the time when the maxima of $\hat{Q}_{l,50}(\tau)$ in the first and the second states are crossing; i.e., the following difference function $d_K(\tau)$ is zero,

$$d_K(\tau) = \left| \max{}_1(\hat{Q}_{l,K}(\tau)) - \max{}_2(\hat{Q}_{l,K}(\tau)) \right|, \tag{22}$$

where $\max_x(\cdot)$ denotes the maximum value in the xth of $\cdot$. Timing of the music structure change, T, is estimated by $d_K(\tau)$ as follows:

$$T = \{\tau \mid d_K(\tau) = 0\}. \tag{23}$$

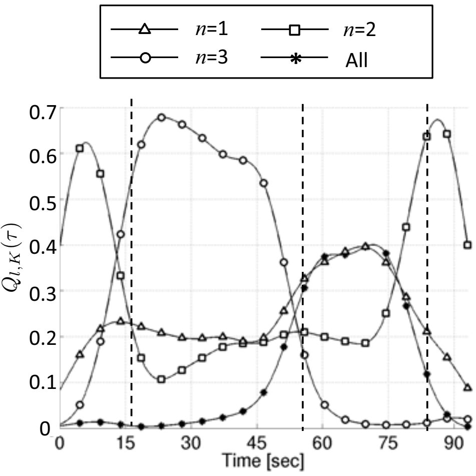

Fig. 4. Ratio of each localization density cluster.

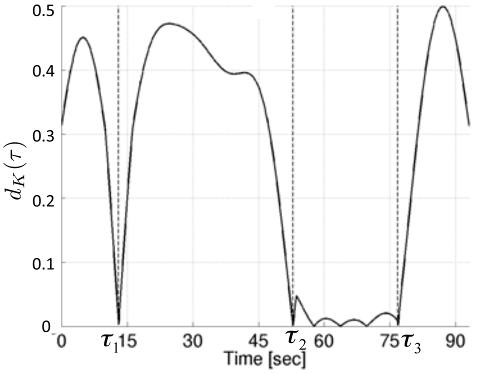

Fig. 5. Differences between main cluster and second cluster densities.

Figure 5 shows $d_K(\tau)$ obtained from Fig. 4 and the estimated time T that almost accurately corresponds to the correct answer. But, T is wrong in 50~75 s because $d_K(\tau)$ shows the oscillational behavior. Therefore, while $d_K(\tau)$ is lower than a threshold α we regard the music structure does not change to prevent the mistake, and then we dismiss τ estimated in this interval. The compensated time T_α is defined as

$$T_\alpha = \{\tau_i \ |d_K(\tau_i) = 0 \cap d_K(\tau) > \alpha \ (\tau_{i-1} < \tau < \tau_i)\}, \qquad (24)$$

where i denotes the index number of the time which is estimated in order with time progress. In this paper, this T_α is the result of estimations.

IV. Evaluation Experiment and Result

A. Experimental Conditions

In this section, we evaluate the effectiveness of the proposed method via the objective assessment. The music signals for the assessment are 27 popular music tunes included in *The Real World Computing Music Database*. All tunes are stereo-recorded, the sampling rate is 44.1 kHz, the quantization bits is 16 bits and the average length of the music signals is 236 s. We evaluate the timing-detection accuracy for the correct answer we manually supplied. We adopt the *F-measure* as an evaluation score, defined by

$$F_{measure} = \frac{2PR}{P+R}, \qquad (25)$$

TABLE I
Result of assessment

Number of all correct answers	300
Number of detections corresponding to correct answers	204
Number of correct answers that is not able to be detected	96
Number of false detections	56
Precision	0.78
Recall	0.68
F-measure	0.72

where P denotes the *precision* and R is the *recall* given by

$$P = \frac{\text{Number of detections corresponding to correct answers}}{\text{Number of all detections}}, \qquad (26)$$

$$R = \frac{\text{Number of detections corresponding to correct answers}}{\text{Number of all correct answers}}. \qquad (27)$$

The precision P becomes a large value when the number of false detections is small. In contrast, the recall R becomes a large value when the number of the undetected correct answers is small. Both have a relation of a trade-off, that is, if either increases, the other decreases. $F_{measure}$ is a harmonic mean of P and R and becomes high in the case that many correct answers are detected in few detections. In this assessment, the detection within 5 seconds before and after the correct answer is regarded as the correct detection.

B. Experimental Result

Table I shows the experimental result. In this assessment, 204 points which are about 70 percent of the whole correct answers (300 points) are able to be detected. Also, the false detections are 56 points out of all 260-point detections. The result shows that the precision is larger than the recall in this case. Finally, the *F-measure* in the proposed method results in 0.72. Thereby, more than 70 percent of the structure changes in music tunes are estimated sufficiently, and this suggests that the proposed method can be applicable to the automatic generation of thumbnail music.

V. Conclusion

In this paper, in order to automatically generate thumbnail music, we proposed a new estimation method of structure changes in stereo tunes based on audio object localization and evaluated the effectiveness by the objective assessment. As a result, the proposed method can detect about 70 percent of the correct answers which are manually given. Consequently, it was shown that the proposed method is effective in the music structure analysis for generating thumbnail music automatically.

Acknowledgement

This work was partly supported by MIC Strategic Information Communications R&D Promotion Programme in Japan.

References

[1] M. E. P. Davies, and M. D. Plumbley, “Beat tracking with a two state model,” Proc. of ICASSP, pp.18–23, 2005.

[2] M. Levy, M. Sandler, and M. Casey, “Extraction of high-level musical structure from audio data and its application to thumbnail generation,” Proc. of ICASSP, pp.13–16, 2006.

[3] S. Miyabe, K. Masatoki, H. Saruwatari, K. Shikano, and T. Nomura, “Temporal quantization of spatial information using directional clustering for multichannel audio coding,” Proc. of WASPAA, pp.261–264, 2009.

[4] S. Suzuki, S. Miyabe, N. Kamado, H. Saruwatari, K. Shikano, and T. Nomura, “Audio object individual operation and its application to earphone leakage noise reduction,” Proc. of ISCCSP, Th. 5.6, 2010.