Automatic Musical Thumbnailing Based on Audio Object Localization and Its Evaluation

Hiroyuki Nawata, Noriyoshi Kamado, Hiroshi Saruwatari, kiyohiro Shikano

Nara Institute of Science and Technology, JAPAN

Outline of This Talk

- Research background and motivation
- Spatial information analysis based on cosine-based k-means clustering
- Evaluation experiment and comparison
- Merged approach with conventional method
- Conclusion

Background

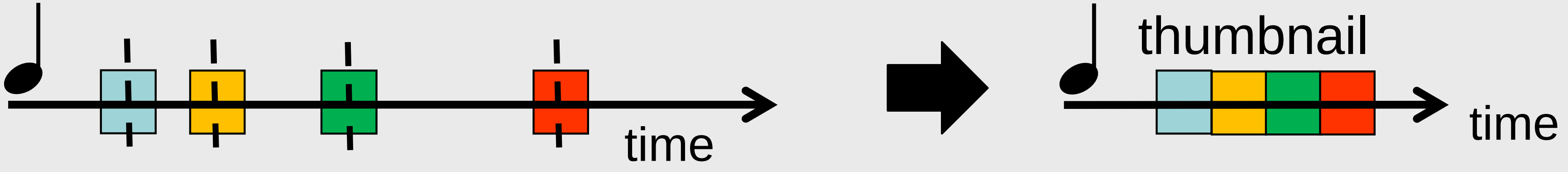
- Main media for delivering music contents are shifted from CD and DVD to digital data via Internet
- ***Musical Thumbnail*** that consists of scraps of a music tune can show abstract information of the tune.
  - We can understand the structure of the music tune by hearing the musical thumbnail, which helps us to judge to buy it or not.

Problem

Currently, musical thumbnails are mainly made manually.

- Difficulty in making thumbnails for tons of music tunes.
- Bad thumbnail provides us a false (negative) advertisement.

The technology for analyzing accurate construction of the musical tune requires an urgent attention.



Related Works

- Conventional methods
  - Construction analysis by beat tracking [M. E. P. Davies, 2005]
  - Construction understanding by main melody analysis [M. Levy, 2006]

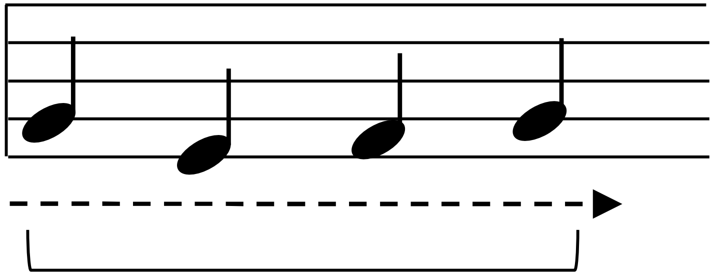
Property Monaural signal Analysis by temporal information

⇒ There has been less studies on multichannel (mostly stereo) tune analysis based on spatial information

Temporal info.

☐Repeat of melody

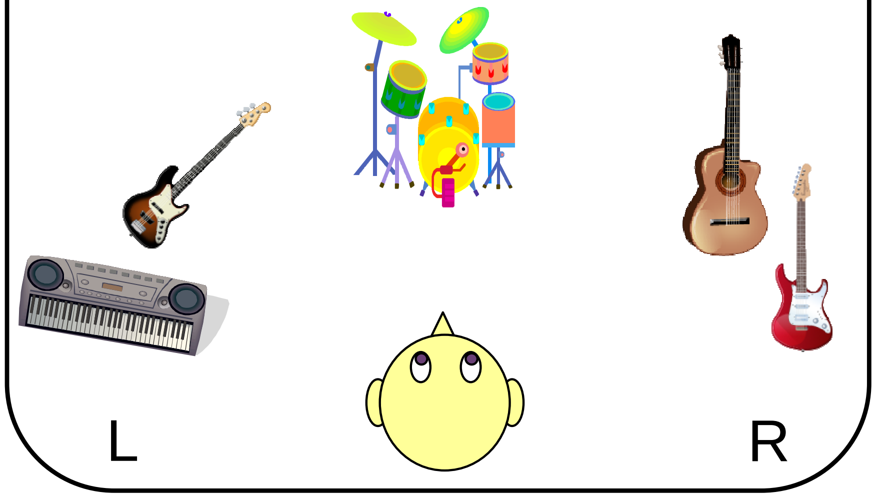
☐Beat pattern of drums



Melody pattern analysis

Spatial info.

☐Instrument localization



Research Aim

- Conventional studies only dealt with monaural music tunes.

It is worth trying to develop an alternative method for musical tune analysis based on spatial information because almost all of the current musical tunes are delivered via stereo format.

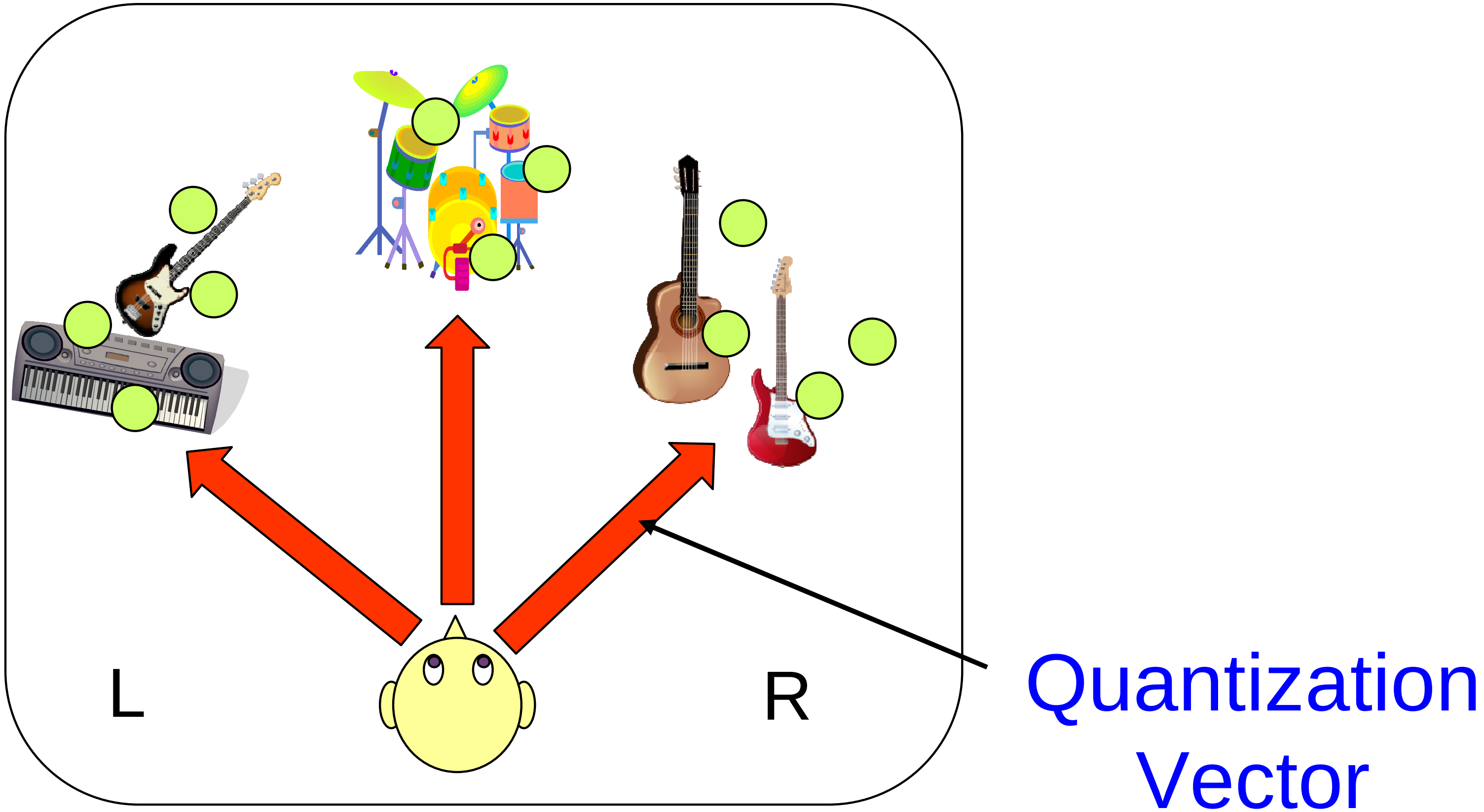
- Aim of this research

We propose a method for analyzing musical tune construction and automatic generation of musical thumbnail based on audio object localization.

Next, we combine the proposed and conventional (temporal-information-based) methods for achieving further performance in making thumbnails.

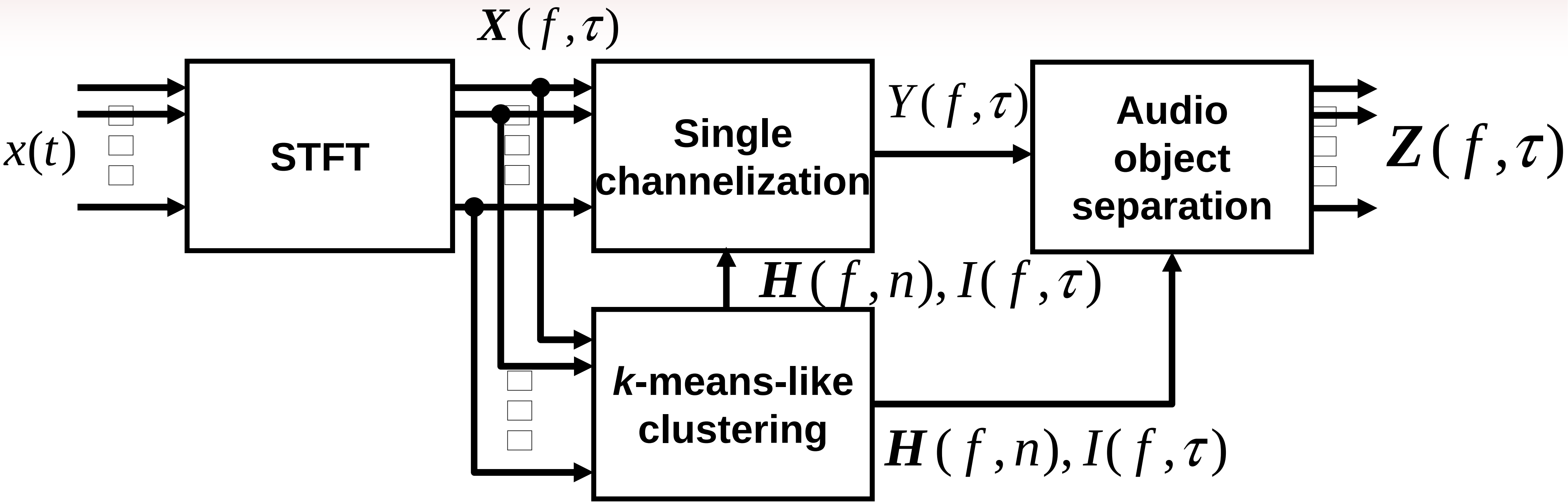
Estimation Method of Audio Object

- Estimation of audio objects in multichannel musical signal
  - We quantize direction of each input signal by n quantization vectors, and then classify them.
  - Each (quantized) direction represents the existence of audio object located around the direction.



Estimation Process of Audio Object

[Miyabe, 2006]



$x(t)$: input time series

$\boldsymbol{X}(f,\tau)$: time-frequency input signal

$\boldsymbol{H}(f,n)$: quantization vector

$Y(f,\tau)$: single-channel encoded signal

$I(f,\tau)$: index function

$\boldsymbol{Z}(f,\tau)$: nth audio object time-frequency signal

τ : time index

f : frequency index

n : class index (1 to N)

Input Signal and Quantization Vector

- Time-frequency input signal (grid)

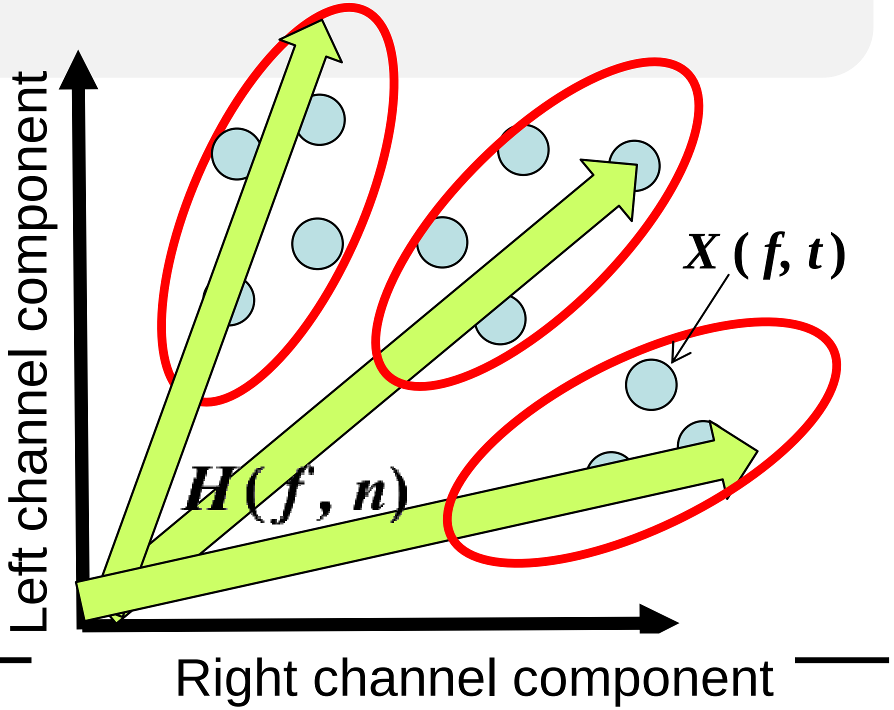
$$\boldsymbol{X}(f,\tau) = [X_1(f,\tau), \ldots, X_M(f,\tau)]^{\mathrm{T}}$$

- Quantization vector

$$\boldsymbol{H}(f,n) = [H_1(f,n), \ldots, H_M(f,n)]^{\mathrm{T}} \qquad (n = 1, 2, \ldots, N)$$

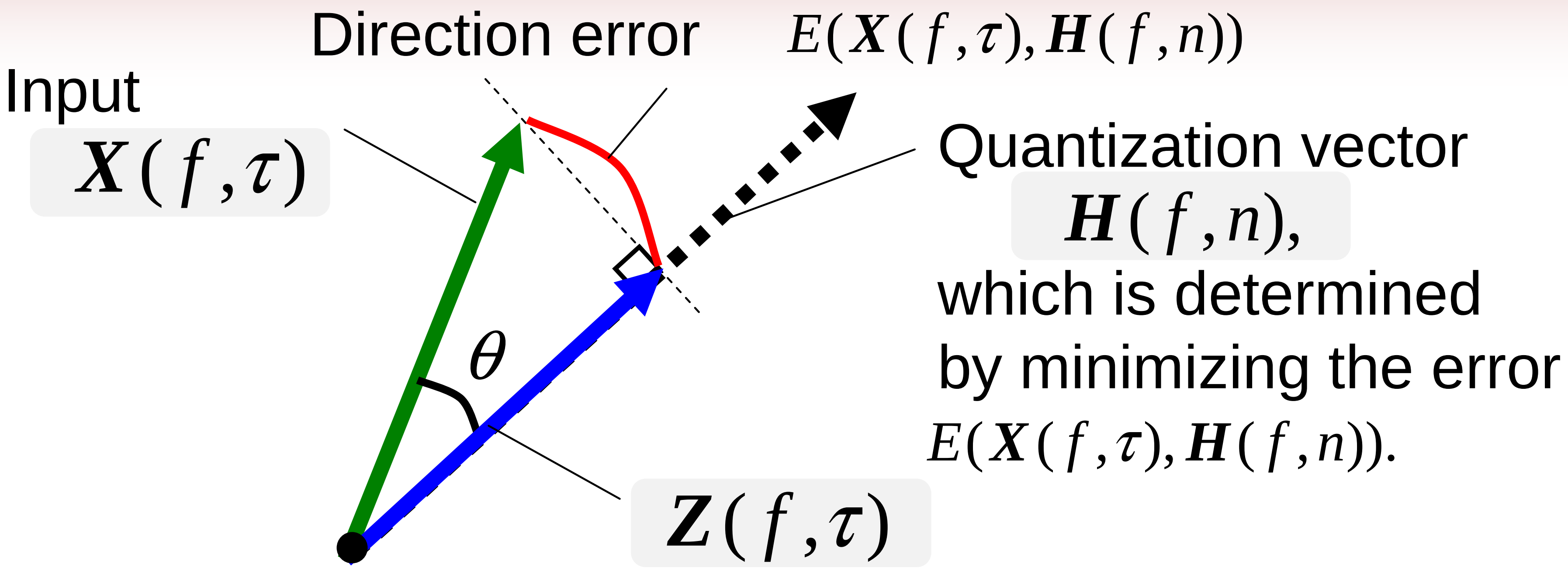
where $\|\boldsymbol{H}(f,n)\| = 1$

M : the number of channels
N : the number of quantization vectors
n : class index of quantization vector



Example of M=2 (stereo), N=3

Error Function to Be Minimized



Direction error $E(\boldsymbol{X}(f,\tau),\boldsymbol{H}(f,n))$

Input $\boldsymbol{X}(f,\tau)$

θ

Quantization vector $\boldsymbol{H}(f,n)$, which is determined by minimizing the error $E(\boldsymbol{X}(f,\tau),\boldsymbol{H}(f,n))$.

$\boldsymbol{Z}(f,\tau)$

$$E(\boldsymbol{X}(f,\tau),\boldsymbol{H}(f,n)) = \|\boldsymbol{X}(f,\tau)\|\sin(\boldsymbol{X}(f,\tau),\boldsymbol{H}(f,n))$$

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

$$\cos(\boldsymbol{X}(f,\tau),\boldsymbol{H}(f,n)) = \frac{\left|\boldsymbol{X}^{\mathrm{H}}(f,\tau)\boldsymbol{H}(f,n)\right|}{\|\boldsymbol{X}(f,\tau)\|\|\boldsymbol{H}(f,n)\|}$$

Cosine-Based *K*-means Clustering

STEP 1: Each signal grid is assigned to the nth class Θ_n with its centroid $\boldsymbol{C}^{(k)}(f,n)$

STEP 2: Update the centroid

$$\begin{aligned}
\boldsymbol{C}^{(k+1)}(f,n) &= \mathop{\arg\min}_{\boldsymbol{C}^{(k)}(f,n)} \sum\nolimits_{\tau\in\Theta_n} E(\boldsymbol{X}(f,t),\boldsymbol{C}^{(k)}(f,n))^2 \\
&= \mathop{\arg\min}_{\boldsymbol{C}^{(k)}(f,n)} \sum\nolimits_{\tau\in\Theta_n} \left\|\boldsymbol{X}(f,t)\right\|^2 \left(1-\frac{\left|\boldsymbol{X}^{\mathrm{H}}(f,\tau)\boldsymbol{C}^{(k)}(f,n)\right|^2}{\left\|\boldsymbol{X}(f,t)\right\|^2\left\|\boldsymbol{C}^{(k)}(f,t)\right\|^2}\right) \\
&= \mathop{\arg\min}_{\boldsymbol{C}^{(k)}(f,n)} \sum\nolimits_{\tau\in\Theta_n} -\left|\boldsymbol{X}^{\mathrm{H}}(f,\tau)\boldsymbol{C}^{(k)}(f,n)\right|^2 \\
&= \mathop{\arg\min}_{\boldsymbol{C}^{(k)}(f,n)} \boldsymbol{C}^{(k)}(f,n)^{\mathrm{H}} \left(\sum\nolimits_{\tau\in\Theta_n} \boldsymbol{X}(f,\tau)\boldsymbol{X}^{\mathrm{H}}(f,\tau)\right)\boldsymbol{C}^{(k)}(f,n)
\end{aligned}$$

The solution is given by finding maximum eigenvalue of $\left(\sum\nolimits_{\tau\in\Theta_n} \boldsymbol{X}(f,\tau)\boldsymbol{X}^{\mathrm{H}}(f,\tau)\right)$

STEP 3: Re-assign the signal grid based on the new centroid, then Return step 1.

Quantization Vector and Audio Object Localization

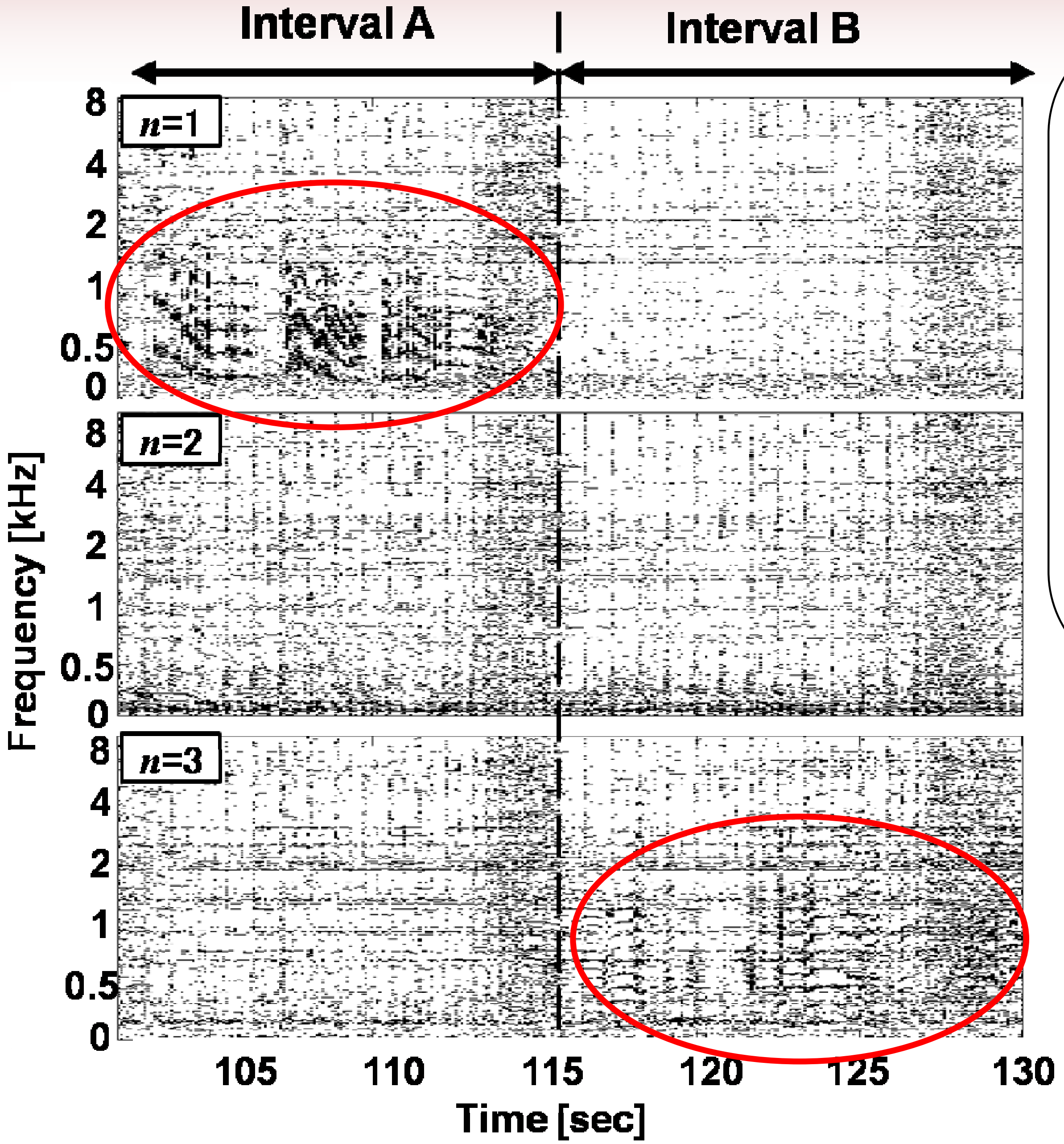
- Quantization vector is estimated by minimizing the error $E(\boldsymbol{X}(f,\tau),\boldsymbol{H}(f,n))$ via cosine-based *k*-means clustering.

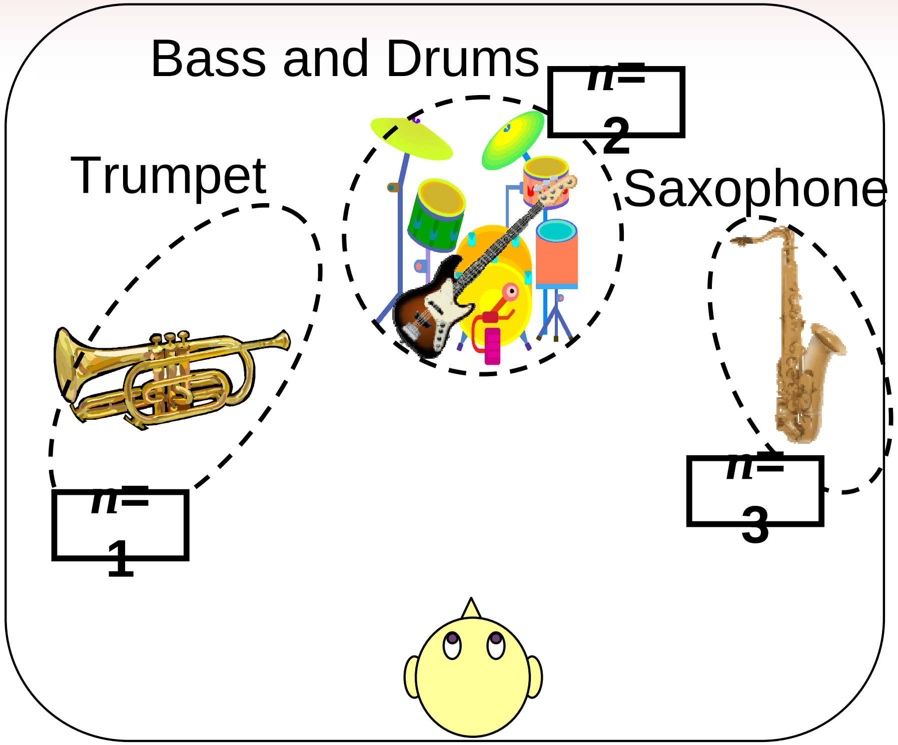
We finally obtain the optimal quantization vector $\boldsymbol{H}(f,n)=\boldsymbol{C}(f,n)$ that classify each component to N audio objects and those class index $I(f,\tau)$.

- **Audio Object Localization**: The function which shows the existence of the nth audio object at every time-frequency grid. Its dense pattern in time-frequency domain represents the instruments’ construction change.

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

Example of Audio Object Localization in Real Tune





Interval A Trumpet solo

↓

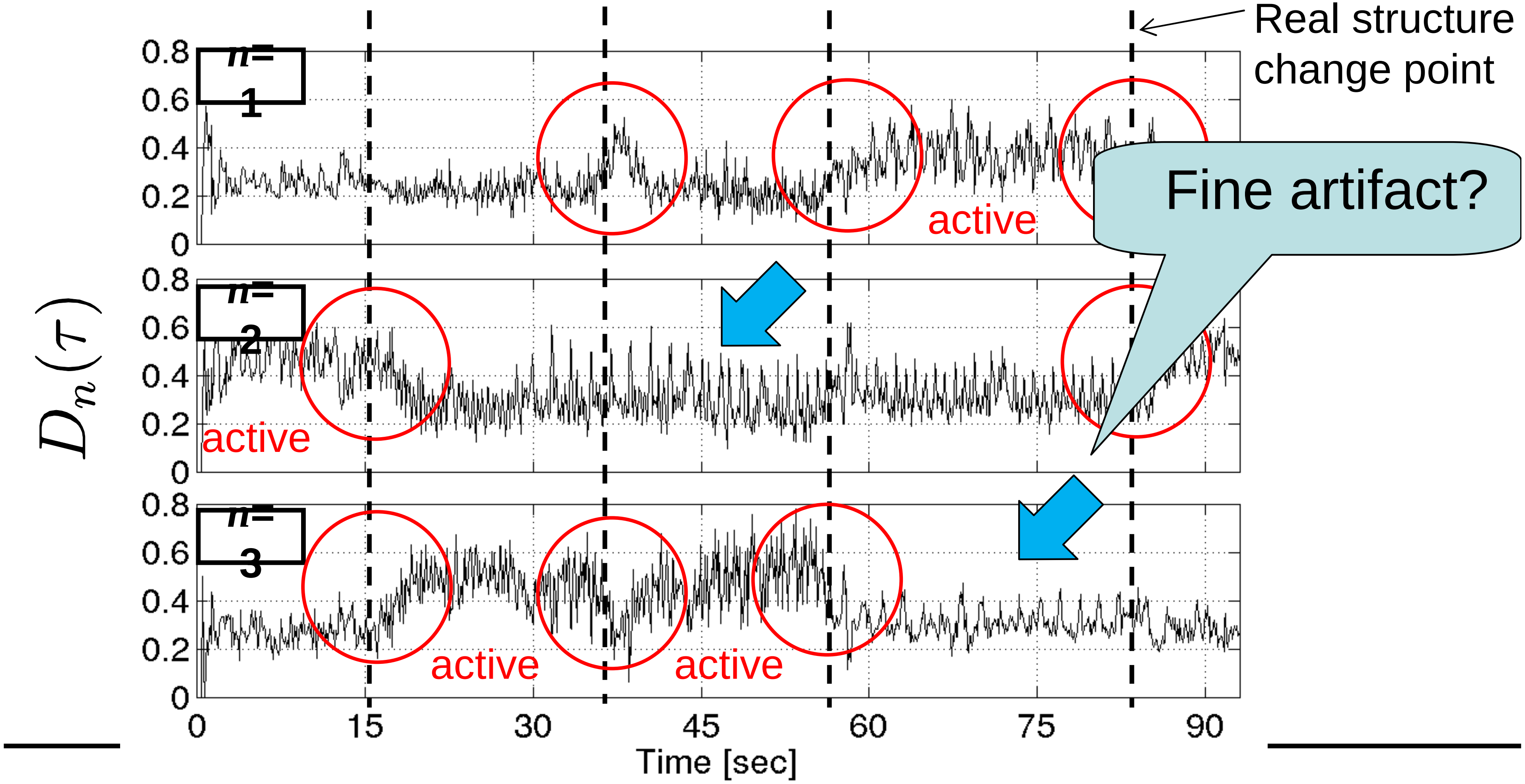
Interval B Saxophone solo

Localization Density at a Time

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

$$W_{Sn}(\tau) = \sum_f o(f)\hat{W}_n(f,\tau)$$

$O(f)$: frequency weighting function



Real structure change point

Fine artifact?

Clustering of Localization Density

- Set of localization density at a time

$$\boldsymbol{D}(\tau) = [D_n, \cdots, D_N(f,\tau)] \quad (N=3)$$

- We quantize the localization density set into some states which have prototype centroids by *k*-means clustering.

$$\begin{aligned} C_l^{(0)} =& \{C_1^{(0)}, C_2^{(0)}, C_3^{(0)}, C_4^{(0)}\} \\ =& \{[1,0,0],[0,1,0],[0,0,1],[\frac{1}{3},\frac{1}{3},\frac{1}{3}]\} \end{aligned}$$

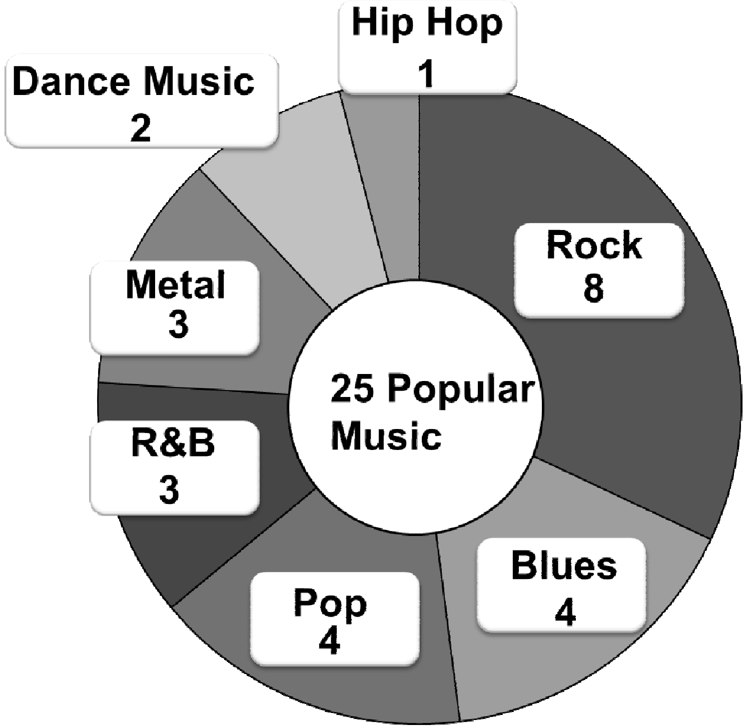
l : class index of state

- Change of limited number of state classes show global change in audio object structure sequence.
- Changing timing is regarded as musical structure change.

Evaluation Experiment (Condition)

Music tunes for assessment	25 popular music tunes included in *The Real World Computing Music Database*
Number of channels M	2
Number of quantization vecters N	3
Average length of tunes	241 [s]
Sampling frequency	44.1 [kHz]

Genre of 25 music tunes used in experiment



We manually put **267 structure change tags** into the database, which are regarded as “correct answesr” in experiment.

Experimental Result

Number of all correct answers	267
Number of detections corresponding to correct answers	193
Number of correct answers that is not able to be detected	74
Number of false detections	60
Precision	0.764
Recall	0.723
$F_{measure}$	**0.742**

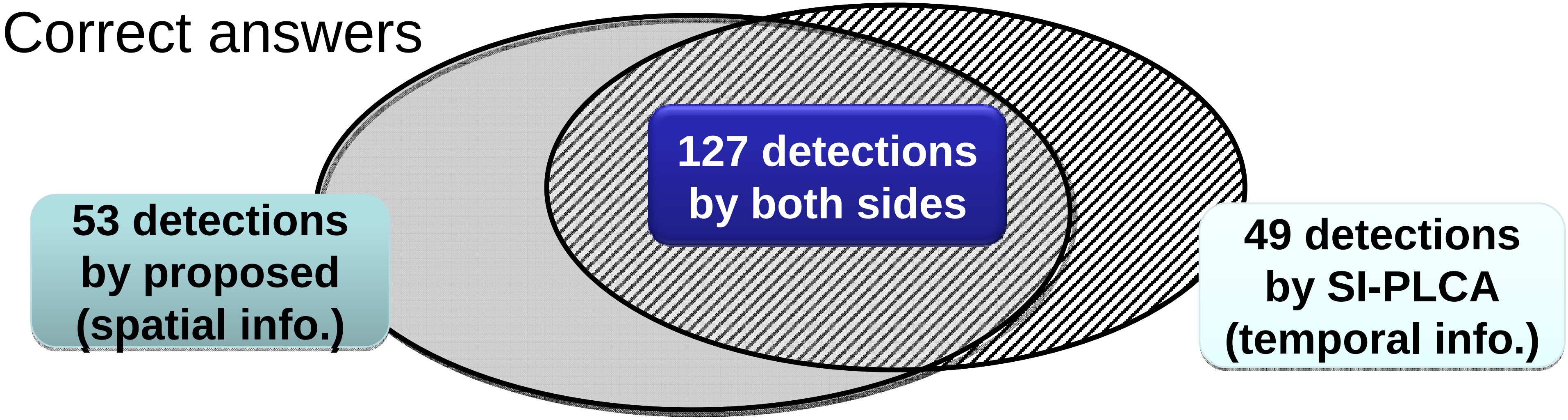
Comparison with Conventional Method

- Competitive method: SI-PLCA [R. J. Weiss et al., 2010]
  - NMF-based method
  - Automatically detect the repeat duration in music tune
  - The method that utilizes **temporal** information
    (cf. proposed method is based on **spatial** information).

	Proposed	SI-PLCA
Precision	0.764	0.677
Recall	0.723	0.768
$F_{measure}$	**0.742**	**0.719**

Proposed Method and SI-PLCA; What's Difference?

- Investigation on difference and correlation between the proposed method (spatial-based) and SI-PLCA (temporal-based).
- We regard detections within 2.5 sec duration as "same detection."



Correct answers

53 detections by proposed (spatial info.)

127 detections by both sides

49 detections by SI-PLCA (temporal info.)

- The correlation among the detections in both methods is not so high. Thus, their detections are **complementary**.
- Therefore, a merged approach would be more effective.
- Next, we simply adopt ***OR operation*** between both.

Experimental Result by *Merged* Approach

Number of all correct answers	267
Number of detections corresponding to correct answers	229
Number of correct answers that is not able to be detected	38
Number of false detections	75
Precision	0.753
Recall	0.858
$F_{measure}$	**0.802**

Spatial-temporal info. gives better F measure (0.742 → 0.802).

Conclusions

- We proposed a new method to detect the structure change in music tunes for automatic thumbnail generation.
- Spatial information included in multichannel (mostly stereo) format signal can be used for detecting the music tune structure. In order to accurately detect them, we newly introduced a cosine-based k-means clustering technique.
- From the evaluation experiment, we can detect correct answers with more than 70% accuracy. In addition, we can find that these detections are complementary with those by the conventional temporal-information-based method.
- We can obtain more better result (80% accuracy) by merging both methods, showing the efficacy of merging spatial and temporal information.

Thank you for your attention!