Scientific Research Publishing

Journal of Signal and Information Processing, 2017, 8, 65-77
http://www.scirp.org/journal/jsip
ISSN Online: 2159-4481
ISSN Print: 2159-4465

Sound-Environment Monitoring Method Based on Computational Auditory Scene Analysis

Mitsuru Kawamoto

Service Sensing, Assimilation, and Modeling Research Group, Human Informatics Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan
Email: m.kawamoto@aist.go.jp

How to cite this paper: Kawamoto, M. (2017) Sound-Environment Monitoring Method Based on Computational Auditory Scene Analysis. *Journal of Signal and Information Processing*, **8**, 65-77.
https://doi.org/10.4236/jsip.2017.82005

Received: February 24, 2017
Accepted: May 9, 2017
Published: May 12, 2017

Abstract

Monitoring techniques are a key technology for examining the conditions in various scenarios, e.g., structural conditions, weather conditions, and disasters. In order to understand such scenarios, the appropriate extraction of their features from observation data is important. This paper proposes a monitoring method that allows sound environments to be expressed as a sound pattern. To this end, the concept of synesthesia is exploited. That is, the keys, tones, and pitches of the monitored sound are expressed using the three elements of color, that is, the hue, saturation, and brightness, respectively. In this paper, it is assumed that the hue, saturation, and brightness can be detected from the chromagram, sonogram, and sound spectrogram, respectively, based on a previous synesthesia experiment. Then, the sound pattern can be drawn using color, yielding a "painted sound map." The usefulness of the proposed monitoring technique is verified using environmental sound data observed at a galleria.

Keywords

Sound-Environment Visualization, Environmental Sounds, Monitoring, Painted Sound Patterns, Synesthesia

1. Introduction

Recently, the analysis of large data sets, so-called "big data," has allowed a variety of information to be extracted, and this information can help create certain services. Further, monitoring techniques can be useful for determining the phenomena that initially generated the recorded data. Thus, monitoring techniques are regarded as those that allow identification of the monitored environment conditions through analysis of the data observed within the area. For example, in the case of structural monitoring, which is known as building health monitoring,

deterioration and damage to buildings can be checked using findings obtained through the analysis of sensor data, e.g., data acquired from acceleration sensors and cameras [1]. In this paper, sound environments are assumed to be the target field of the monitoring problem; that is, sound environment monitoring is addressed.

Various methods for understanding sound environments have been proposed to date. However, almost all researchers have focused on topics related to environmental sound recognition (ESR) [2]. For example, ESR techniques implemented with features such as a zero-crossing rate, Cepstral features, MPEG-7-based features, and autoregression-based features, which are extracted from environmental sounds, have been proposed [3]-[9], along with a method of understanding environmental sounds that employs a matching pursuit algorithm [10]. To the best of the author's knowledge, no studies have focused on determination of sound environments; therefore, such a method is presented here.

This study proposes an unconventional method that allows the analysis of sound environments using color, where the color rules are based on the concept of synesthesia [11]. That is, sound positions can be estimated using a sound position estimation approach, and a color based on three features extracted from the observed environmental sounds can be painted at the estimated position. Hence, painted sound patterns referred to as "painted sound maps" are obtained, from which sound environment scenarios can be recognized. The efficacy of the proposed monitoring method is evaluated using environmental sound data observed at a galleria.

2. Proposed Method

2.1. Overview of Proposed Method

For application of the proposed method, environmental sounds are first collected using a microphone array (**Figure 1**). Using these sounds, various sound environment conditions can be estimated. These scenarios are then expressed

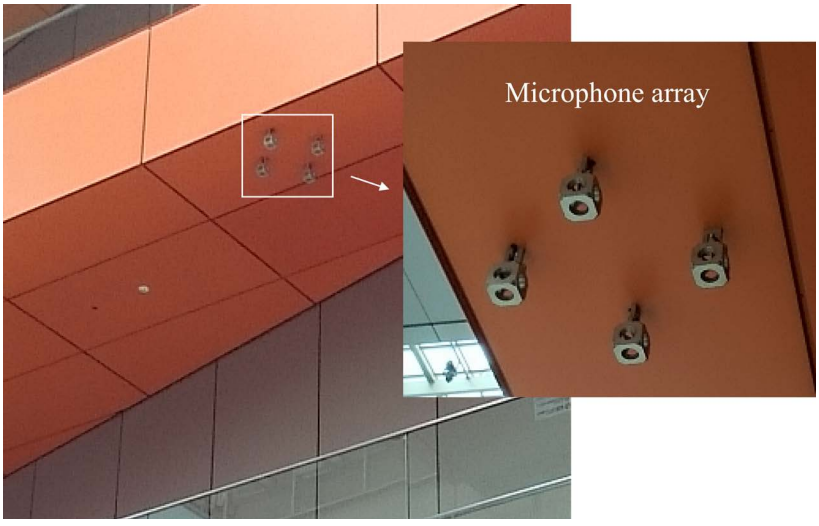

Figure 1. Microphone array utilized to collect environmental sounds at galleria.

using colors, based on the knowledge of synesthesia.

Synesthesia is a phenomenon in which one kind of sensory stimulation is expressed as another sensation [12]. In the case of synesthesia relating sound and color, Nagata *et al.* have reported an experimental result in which keys, tones, and pitches were respectively related to hue, saturation, and brightness [13]. Based on this result, this study utilizes information on the keys, tones, and pitches of environmental sounds to draw painted sound maps.

Keys, tones, and pitches are assumed to be detected by the chromagram, sonogram, and sound spectrogram, respectively. Hence, the hue score is calculated using the key histogram yielded by the chromagram. Similarly, the saturation and brightness scores are calculated using the frequency-band histograms produced by the sonogram and sound spectrogram, respectively, with a clustering method then being applied to the environmental-sound spectrogram. Similar frequency components are categorized so that the frequency component dispersion of the environmental sounds is clarified. This dispersion information is then used to calculate the histogram with respect to the spectrogram frequency elements.

Below, the proposed method of sound environment analysis is presented in detail. The sound data $\boldsymbol{y}(t)$ can be obtained using the microphone array shown in **Figure 1**, where a short-term Fourier transform $\boldsymbol{y}(\omega,t)$ is applied to $\boldsymbol{y}(t)$. Then, the amplitude of $\boldsymbol{y}(t)$ has $\max_{\omega} \left|\boldsymbol{y}(\omega,t)\right|^2 > a$, where a is a constant value, and the maximum period of $\boldsymbol{y}(t)$ is 2 s. The sound position estimation is conducted using multiple signal classification (MUSIC) [14], utilizing the microphone array outputs.

2.2. Key Information Extraction from Chromagram

The environmental sound chromagram is calculated using the MATLAB chroma toolbox [15]. First, the environmental-sound pitch features can be computed using the audio_to_pitch_via_FB function. **Figure 2** and **Figure 3** show an environmental sound and its pitch features obtained using audio_to_pitch_via_FB, respectively. Next, a chromagram can be calculated (**Figure 4**) using pitch_to_chroma, based on pitch features such as those shown in **Figure 3**.

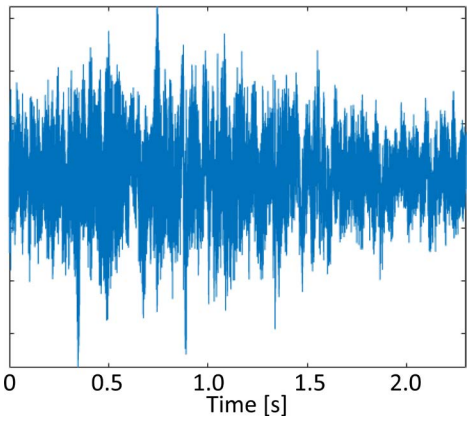

Figure 2. Environmental sound.

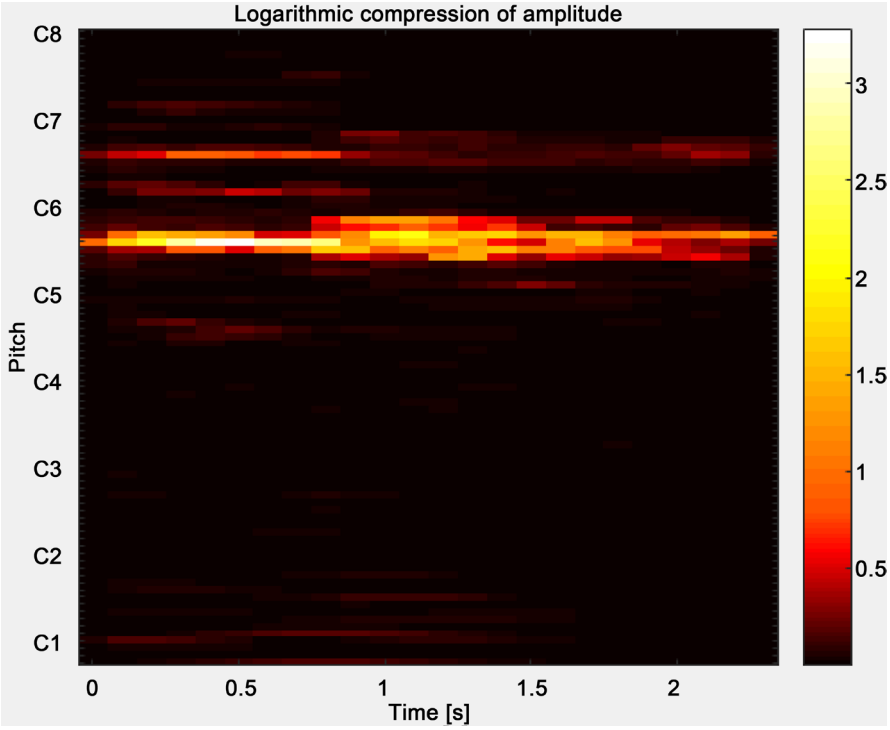

Figure 3. Pitch features of environmental sound shown in Figure 2.

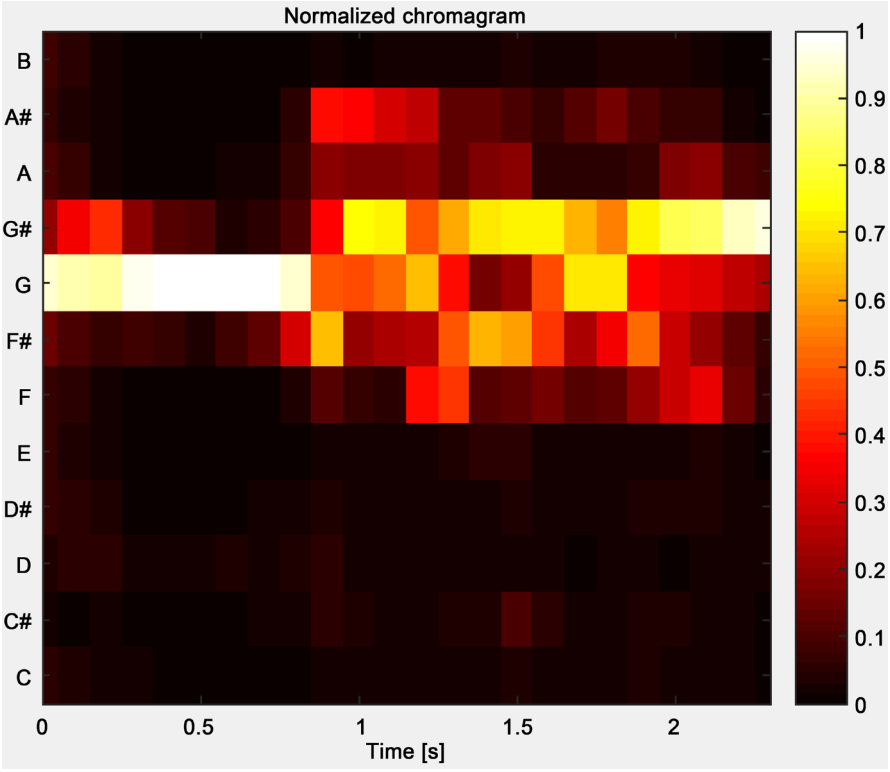

Figure 4. Chromagram of environmental sound shown in Figure 2.

Subsequently, a histogram showing the key information indicated in the chromagram is calculated. For example, Figure 5 shows the histogram calculated based on the chromagram shown in Figure 4, where the ma_sh function of

the MATLAB ma toolbox [16], which can calculate a spectrum histogram from the chromagram, is used. The MATLAB function, hist, is then applied to the spectrum histogram. Depending on the histogram variability, the histogram data can be transformed into an 8-bit binary code as follows: 1) The histogram mean is calculated (dashed line in Figure 5); 2) Values greater or less than the mean are replaced with "1" or "0," respectively; 3) An 8-bit binary code is obtained; the code corresponding to Figure 5 is "00000101." Hence, the hue score is determined by converting the binary code to decimal values. It is apparent that a higher score indicates an environmental sound consisting of some dominant keys.

2.3. Tonal Information Extraction from Sonogram

The sonogram can be calculated using the MATLAB ma toolbox [16], where the loudness sensation per frequency band is estimated using auditory models and the ma_sone function of the ma toolbox. Figure 6 shows the sonogram of the environmental sound shown in Figure 2.

The frequency-band histogram of the sonogram is computed [16], and the saturation score is then determined using the same approach as that used for the hue score. Therefore, it is apparent that a higher score indicates an environmental sound with some characteristic components in the loudness sensation per frequency band.

2.4. Pitch Information Extraction from Spectrogram

A spectrogram can also be calculated. Figure 7 shows the spectrogram of the

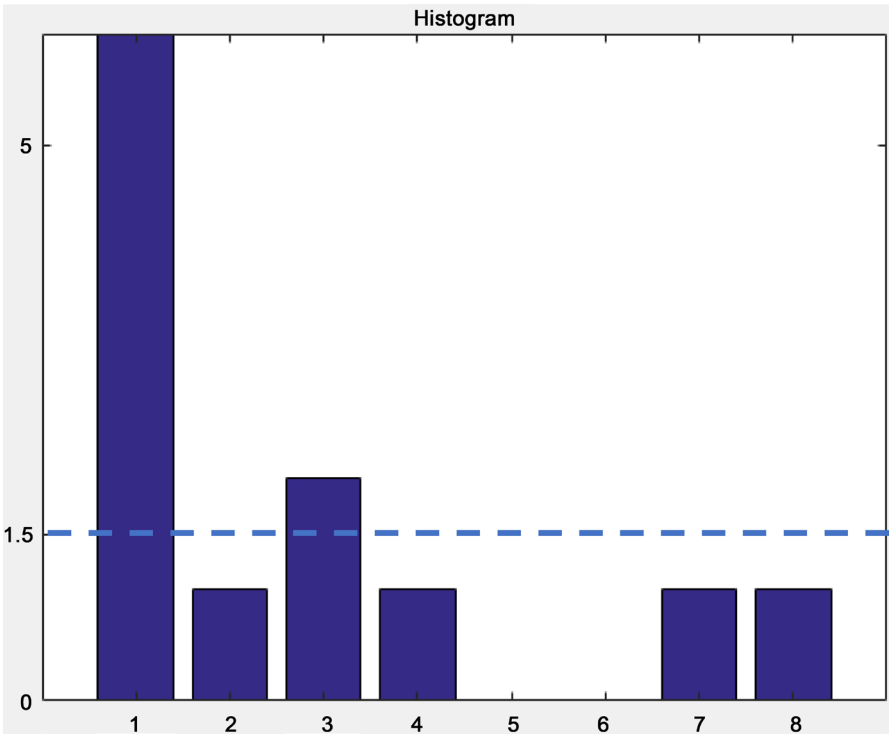

Figure 5. Key histogram obtained from Figure 4 chromagram, with mean.

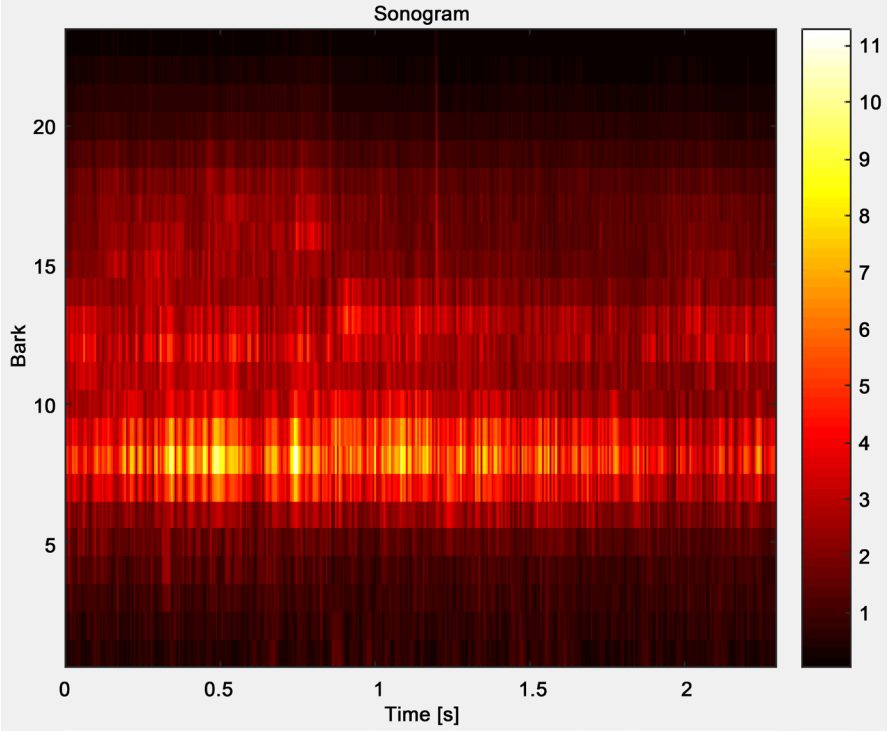

Figure 6. Sonogram of environmental sound shown in Figure 2.

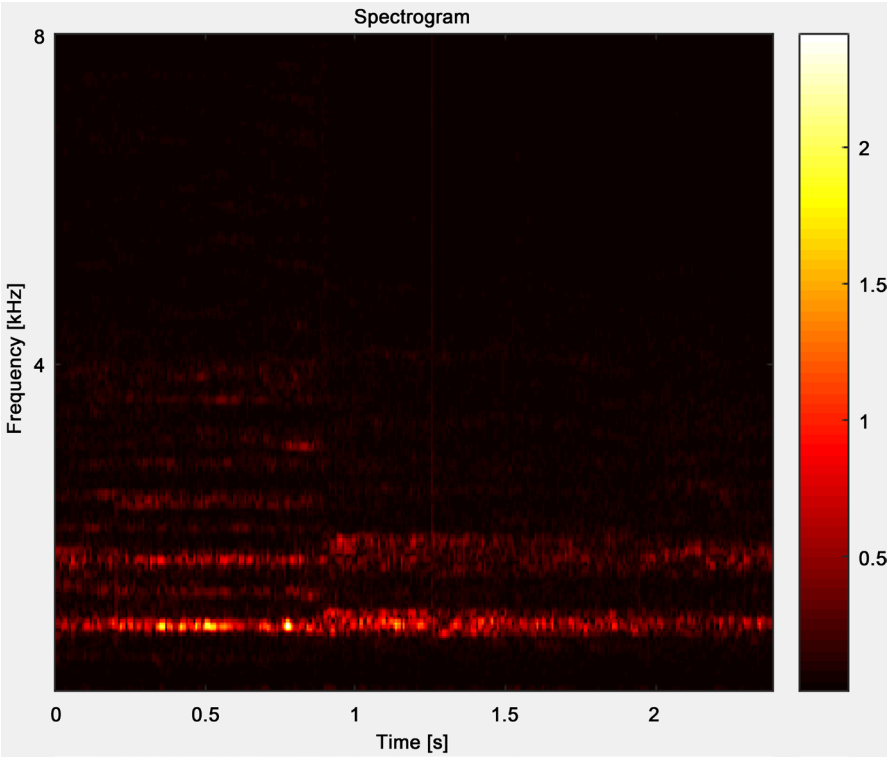

Figure 7. Spectrogram of environmental sound shown in Figure 2.

environmental sound shown in Figure 2. An edge-extraction image processing technique is applied to the spectrogram, and the number of pixels in its frequency characteristic areas and their centroid frequencies are then computed.

The frequency characteristic areas of the spectrogram detected by the edge extraction technique are categorized using an improved affinity propagation (IAP) method (see Appendix). For details of the affinity propagation, see [17].

Each exemplar centroid frequency obtained by the IAP is classified into a low-, medium-, or high-frequency group; then, the frequency-group histogram can be acquired. The brightness score is obtained from the histogram in a similar manner to the case of the hue score. However, when the 8-bit binary code is obtained, the threshold determining "1" or "0" values is set to zero. Therefore, a low score indicates that the dominant frequency of the environmental sound is low, while a higher score indicates that the environmental sound consists of various frequencies.

2.5. Painted Sound Map from Three Scores

The hue, saturation, and brightness scores are used to draw the painted sound map, where the hue-saturation-brightness color model obtained using these three scores is converted to a red-green-blue (RGB) color model.

3. Experimental Results and Discussion

In this section, the efficacy of the painted sound map method is demonstrated using environmental sounds observed in the sound environment shown in **Figure 1**. In each demonstration, the painted sound map is drawn using the environmental sounds generated in a single day. Further, in each figure shown below, the position of the microphone array is indicated by a red circle.

3.1. Painted Sound Map of Sound Environment on Typical Day

The sound environment shown in **Figure 1** is a shopping-center galleria, the painted sound map of which is shown in **Figure 8**. A train station is located near the galleria, outside the left side of **Figure 8**. Therefore, train sounds are generated intermittently. Further, rattling sounds from chairs and desks are generated during the galleria preparation time, along with the voices of children and students visiting the galleria.

Figure 9 shows the painted sound map for a different day. It is notable that the generated sound patterns are similar to those in **Figure 8**. From these two maps, it can be concluded that painted sound maps can be utilized to determine similarities in the sound patterns of sound environments.

3.2. Painted Sound Map on Windy Day

Figure 10 shows a painted sound map obtained on a windy day. Comparing **Figures 8-10,** it is apparent that the painted sound map varies with the state of the sound environment; hence, painted sound maps can be utilized to detect variations in a sound environment.

3.3. Mini-Concert Event

Figure 11 shows the painted sound map obtained for the same area during a

mini-concert event. Blue tones are emphasized at the event location, which is on the left-hand side of the map. In addition, the painted sound map obtained on the same day is shown in Figure 12. At a glance, the painted sound map is simi-

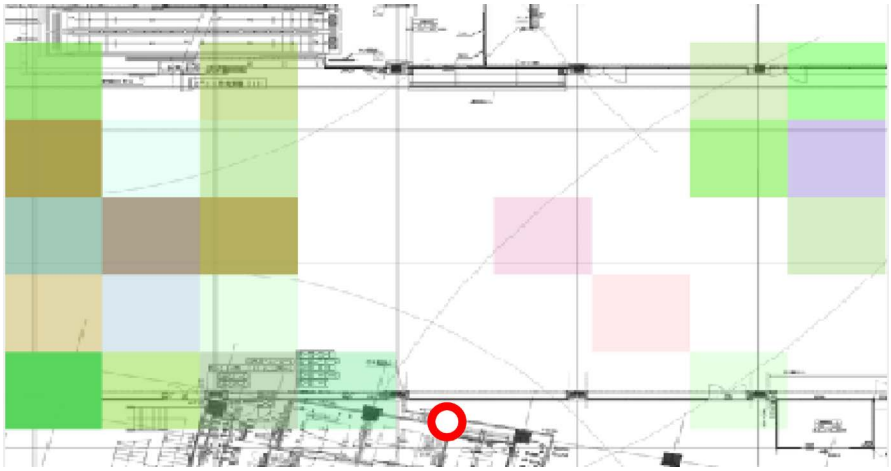

Figure 8. Painted sound map of Figure 1 sound environment.

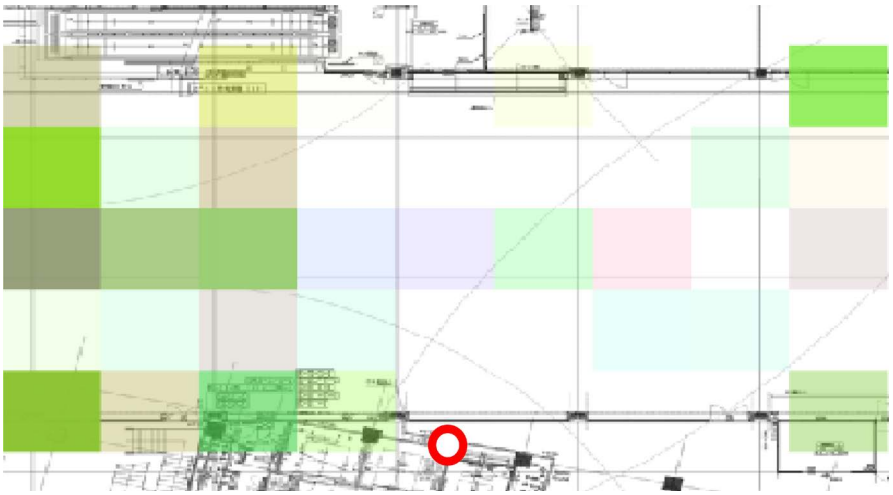

Figure 9. Painted sound map showing similar sound patterns to those of Figure 8.

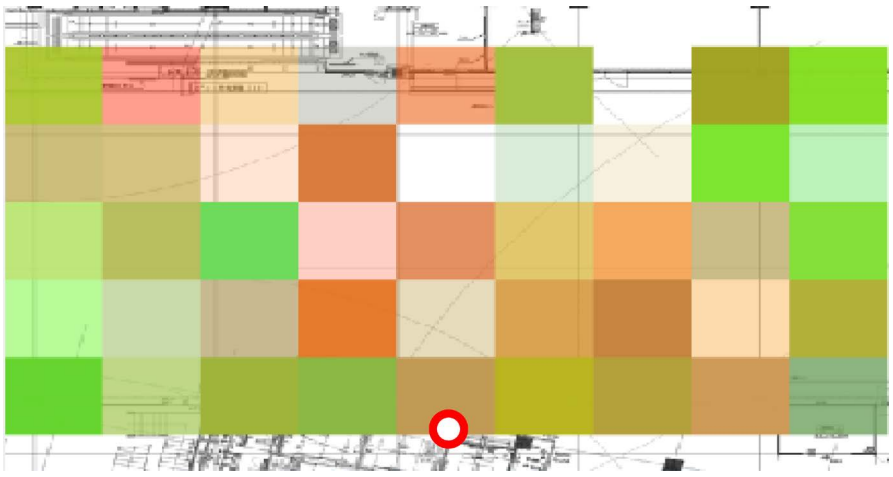

Figure 10. Painted sound map obtained on windy day.

Figure 11. Painted sound map obtained during mini-concert event.

Figure 12. Painted sound map obtained on event day.

lar to those shown in Figure 8 and Figure 9. This indicates that the snapshot provided by the painted sound maps is important as regards detection of sound environmental changes.

From all the above results, it can be concluded that the proposed painted sound map drawn using the three scores discussed above is effective for sound environment analysis. In particular, this approach is useful for visually detecting and determining the sound environment conditions and their variations, and it should be noted that the snapshots provided by the painted sound maps work effectively in this regard.

4. Conclusions

This paper has proposed a method of monitoring sound environments based on computational auditory scene analysis. The proposed visualization technique allows sound environment conditions to be determined and represented using colors.

As future research work, the proposed monitoring technique can be applied to the monitoring of superannuated building structural conditions.

Acknowledgements

The author thanks Dr. Sashima and Dr. Kurumatani for helpful discussions. This work was partly supported by a JSPS KAKENHI Grant (Number 16H02911).

References

[1] Hamamoto, T. (2015) Structural Health Monitoring of Buildings. *Transactions of Foundation Engineering & Equipment*, **43**, 17-20. (In Japanese)

[2] Chachada, J.S. and Kuo, C.-C.J. (2014) Environmental Sound Recognition: A Survey. SIP (2014), Vol. 3, e14, 1-15. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/S2048770314000122

[3] Mitrovic, D., Zeppelzauer, M. and Breiteneder, C. (2010) Features for Content-Based Audio Retrieval. In: *Advances in Computers*, Vol. 78, Elsevier, Amsterdam, 71-150.

[4] Deng, J.D., Simmermacher, C. and Cranefield, S. (2008) A Study on Feature Analysis for Musical Instrument Classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, **38**, 429-438. https://doi.org/10.1109/TSMCB.2007.913394

[5] Peltonen, V., Tuomi, J., Klapuri, A., Huopaniemi, J. and Sorsa, T. (2002) Computational Auditory Scene Recognition. 2002 *IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), Orlando, FL, 13-17 May 2002, II-1941-II-1944.

[6] Potamitis, I. and Ganchev, T. (2008) Generalized Recognition of Sound Events: Approaches and Applications. In: Tsihrintzis, G.A. and Jain, L.C., Eds., *Multimedia Services in Intelligent Environments*, Springer, Berlin, Heidelberg, 41-79.

[7] Wang, J.-C., Wang, J.-F., He, K.W. and Hsu, C.-S. (2006) Environmental Sound Classification Using Hybrid SVM/KNN Classifier and MPEG-7 Audio Low-Level Descriptor. *International Joint Conference on Neural Networks*, Vancouver, 16-21 July 2006, 1731-1735.

[8] Muhammad, G., Alotaibi, Y.A., Alsulaiman, M. and Huda, M.N. (2010) Environment Recognition Using Selected MPEG-7 Audio Features and Mel-Frequency Cepstral Coefficients. 2010 5*th International Conference on Digital Telecommunications* (*ICDT*), Athens, 13-19 June 2010, 11-16. https://doi.org/10.1109/ICDT.2010.10

[9] Tsau, E., Kim, S.-H. and Kuo, C.-C.J. (2011) Environmental Sound Recognition with CELP-Based Features. 2011 10*th International Symposium on Signals, Circuits and Systems* (*ISSCS*), lasi, 30 June-1 July 2011, 1-4. https://doi.org/10.1109/ISSCS.2011.5978729

[10] Chu, S., Narayanan, S. and Kuo, C.-C.J. (2009) Environmental Sound Recognition with Time-Frequency Audio Features. *IEEE Transactions on Audio, Speech, and Language Processing*, **17**, 1142-1158. https://doi.org/10.1109/TASL.2009.2017438

[11] The Color Science Association of Japan (2011) Handbook of Color Science. 3rd Edition, University of Tokyo Press, Japan. (In Japanese)

[12] Cytowic, R.E. (2003) The Man Who Tasted Shapes. MIT Press, Cambridge, MA.

[13] Nagata, N., Iwai, D., Tsusa, M., Wake, S.H. and Inokuchi, S. (2003) Non-Verbal Mapping between Sound and Color-Mapping Derived from Colored Hearing Possessors and Its Applications. *IEICE*, **J86-A**, 1219-1230. (In Japanese)

[14] Schmidt, R.O. (1986) Multiple Emitter Location and Signal Parameter Estimation.

IEEE Transactions on Antennas and Propagation, **34**, 276-280. https://doi.org/10.1109/TAP.1986.1143830

[15] Muller, M. and Ewert, S. (2011) Chroma Toolbox: Matlab Implementations for Extracting Variants of Chroma-Based Audio Features. *Proceedings of the* 12*th International Society for Music Information Retrieval Conference* (*ISMIR* 2011), Miami, 24-28 October 2011, 215-220.

[16] http://www.pampalk.at/ma/documentation.html

[17] Frey, B.J. and Dueck, D. (2007) Clustering by Passing Messages between Data Points. *Science*, **315**, 972-976. https://doi.org/10.1126/science.1136800

[18] Wang, R., Zhang, J., Li, D., Zhang, X. and Guo, T. (2007) Adaptive Affinity Propagation Clustering. *Acta Automatica Sinica*, **33**, 1242-1246.

[19] https://jp.mathworks.com/matlabcentral/fileexchange/18244-adaptive-affinity-propagation-clustering

[20] Calinski, T. and Harabaz, J. (1974) A Dendrite Method for Cluster Analysis. *Communications in Statistics*, **3**, 1-27. https://doi.org/10.1080/03610927408827101

Appendix

1.1. Improved Affinity Propagation

The method proposed in this paper employs a message exchange clustering algorithm based on an affinity propagation (AP) method [17]. The AP performs clustering such that the data can be categorized by modifying the messages $r(i,k)$ and $a(i,k)$ according to

$$r(i,k) \leftarrow s(i,k) - \max_{k',\ \text{s.t.}\ k' \neq k} \left\{ a(i,k') + s(i,k') \right\}, \tag{1}$$

$$a(k,k) \leftarrow \min \left\{ 0, r(k,k) + \sum_{i',\ \text{s.t.}\ i' \notin i,k} \max \left\{ 0, r(i',k) \right\} \right\}, \tag{2}$$

$$a(k,k) \leftarrow \sum_{i',\ \text{s.t.}\ i' \notin i,k} \max \left\{ 0, r(i',k) \right\}, \tag{3}$$

where $r(i,k)$ is a message being sent from data point i in a cluster to a centroid candidate k (exemplar) in the cluster, indicating the appropriateness of data point k becoming the exemplar of i. $a(i,k)$ is a message being sent from an exemplar candidate k to data point i, indicating the appropriateness of i becoming a cluster member of k. $s(i,k)$ is the similarity between data points i and k, where, in each iterative step l, $r(i,k)$ and $a(i,k)$ are updated with those of the previous iteration, *i.e.*, $r_l(i,k) = (1-\text{lam}) r_l(i,k) + \text{lam} r_{l-1}(i,k)$ and $a_l(i,k) = (1-\text{lam}) a_l(i,k) + \text{lam} a_{l-1}(i,k)$. The parameter "lam" denotes a damping factor and is set to $0 < \text{lam} < 1$.

In the AP method, the exemplar is the data point k satisfying the inequality;

$$r(k,k) + a(k,k) > 0. \tag{4}$$

Then, the exemplar satisfying condition (4) can be altered by the preference $s(k,k)$ [17]. That is, $s(k,k)$ influences the output clusters and the number of clusters. The $s(k,k)$ values are set before modification of $r(i,k)$ and $a(i,k)$. In the original AP method, the $s(k,k)$ values were set to the median of all $s(i,k)$ values.

Here, it should be noted that the $s(k,k)$ values can be modified during the updates of $r(i,k)$ and $a(i,k)$. Wang *et al.* have proposed an adaptive scanning method of preferences applicable to the $s(k,k)$ space to determine the optimal clustering solution [18]. They have also proposed a damping-factor adaptive adjustment method to improve the AP method convergence. This study proposes an $s(k,k)$ modification algorithm using the similarity $s(k'',k)$ and satisfying the condition,

$$\text{abs}\left(Xk'' - \text{mean}(\boldsymbol{X}) \right) > \text{astd}(\boldsymbol{X}). \tag{5}$$

That is, based on the $s(k'',k)$ values with respect to the data point k'' satisfying condition (5), all $s(k,k)$ values are updated using

$$s(k,k) = \beta s(k,k) + (1-\beta) \max_{k',\ \text{s.t.}\ k' \neq k''} \left\{ s(k'',k') \right\}. \tag{6}$$

This means that the $s(k,k)$ values are updated such that k'' does not become an outlier in its cluster. The parameters abs(x), mean(x), and std(x) denote the absolute value, the mean value, and the standard deviation of x, respectively.

Table 1. Clustering performance comparison.

	VRC	Number of exemplars	Running time [s]
Original AP method	32.93	4.7	0.0078
Adaptive AP method	35.99	5.5	0.2146
Proposed IAP method	38.47	9.6	0.1013

Further, $\boldsymbol{X}$ denotes all data in a cluster consisting of data XK". The parameters α and β are positive constants greater and less than one, respectively. Therefore, the AP algorithm proposed in this study is implemented by adding the original rules (1)-(4) to rules (5) and (6).

1.2. Improved Affinity Propagation (IAP) Performance

In this subsection, the proposed IAP method is compared with the original and adaptive AP methods, using the 2D random data points $\boldsymbol{x} = [x1, x2]$ generated with $\boldsymbol{x} \in [0,1]$. The data point number is $N = 30$. The original and adaptive AP methods are implemented using the MATLAB program obtained from [19]. The performances of the three algorithms are evaluated using the Calinski-Harabasz criterion [20], which is the ratio of the between-cluster variance to the total within-cluster variance, defined as

$$\mathrm{VRC} = \left(\mathrm{SS_B}/\mathrm{SS_W}\right) \times \left(\left(N-k\right)/\left(k-1\right)\right). \tag{7}$$

Here, k denotes the number of clusters and $\mathrm{SS_B}$ is the overall between-cluster variance, which is essentially the variance of all the cluster centroids from the grand centroid in the dataset, defined as

$$\mathrm{SS_B} = \sum_{i=1:k} n_i \left\| \boldsymbol{x}_i - \boldsymbol{m}_x \right\|^2. \tag{8}$$

Here, n_i indicates the number of elements per cluster, $\boldsymbol{x}_i$ is the centroid of cluster i, $\boldsymbol{m}_x$ is the overall mean of the dataset, and $\left\|*\right\|^2$ is the L2 norm of *. Further, $\mathrm{SS_W}$ is the overall within-cluster variance, defined as

$$\mathrm{SS_W} = \sum_{i=1:k} \sum_{x' \in c_i} \left\| \boldsymbol{x}' - \boldsymbol{m}_i \right\|^2, \tag{9}$$

where $\boldsymbol{x}'$ is a data point, c_i is the ith cluster, and $\boldsymbol{m}_i$ is the centroid of c_i. A large positive value of VRC indicates that the clustering performance is superior. In this study, α and β in (5) and (6) were set to 1.5 and 0.9, respectively.

Table 1 shows the performance results, which were averaged over the results of 30 trials. It is apparent that the proposed IAP method has a longer computational time and more exemplars than the original AP method. However, the former exhibits superior clustering performance, compared with the two conventional methods. Note that the running time was calculated using a PC (CPU: i7-4770@3.4 GHz, RAM: 8.0 GB). Hence, it can be concluded that rules (5) and (6) work effectively in the original AP method.