2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
September 14-18, 2014. Chicago, IL, USA,

Visualization of auditory awareness based on sound source positions estimated by depth sensor and microphone array

Takahiro Iyama[1], Osamu Sugiyama[1], Takuma Otsuka[1], Katsutoshi Itoyama[1] and Hiroshi G. Okuno[1]

***Abstract*— We have developed a system for visualizing auditory awareness on the basis of sound source locations estimated using a depth sensor and microphone array. Previous studies on visualizing the acoustic environment viewed the level of sound pressures directly on the captured image, so the visualization was often based on a mixture of several sound sources. As a result, which targets to focus on was not intuitive. To help users selectively to find the targets and focus on the target analysis, we should extract the captured acoustic information and selectively propose it with the user demand. We have designed a three-layer visualization model for auditory awareness consisting of a sound source distribution layer, a sound location layer, and a sound saliency layer. The model extracts acoustic information by using the depth image and multi-directional sound sources captured with a depth sensor and microphone array. This model is used in the system we developed for visualizing auditory awareness.**

I. Introduction

Technology for observing and recognizing the ambient environment is essential for creating a safe and secure society. It would be particularly useful to the tele-operators of surveillance systems for recognizing from which direction sound is coming and what kind of sound it is. For instance, it would be useful for use in rescue robots searching for people buried in rubble and for use in obtaining situational awareness of abnormal situations in an area where people cannot enter [1][2]. Recent advances in microphone array technology based on multi-channel digital signal processing enables execution of multi-directional sound localization and separation in real time. These technologies can be introduced into surveillance systems, such as for visualizing them in RGB images. The question is how to visualize them.

When various kinds of sound information are being visualized on the basis of multi-channel digital signal processing, it is essential to make the users aware of the significant sounds, i.e., to provide auditory awareness of the environment. Auditory awareness includes not only the source positions and levels of sound pressure but also the positions of sound sources and changes in the states of the sound sources. For example, we should not only superimpose a two-dimensional level of sound pressure on RGB images but also provide depth information and visualize them on the basis of how far the sound sources are from the sensor. Even if the power of a sound source is low, if the sound source is far from the sensor, the system should represent it with higher power in consideration of the distance information.

*This work was supported by JSPS KAKENHI Grant Numbers 24220006 and 24700168.

[1]The authors are with the Graduate School of Informatics, Kyoto University, Sakyo, Kyoto, 606-8501, Japan. {tiayma, sugiyama, ohtsuka, itoyama, okuno}@kuis.kyoto-u.ac.jp

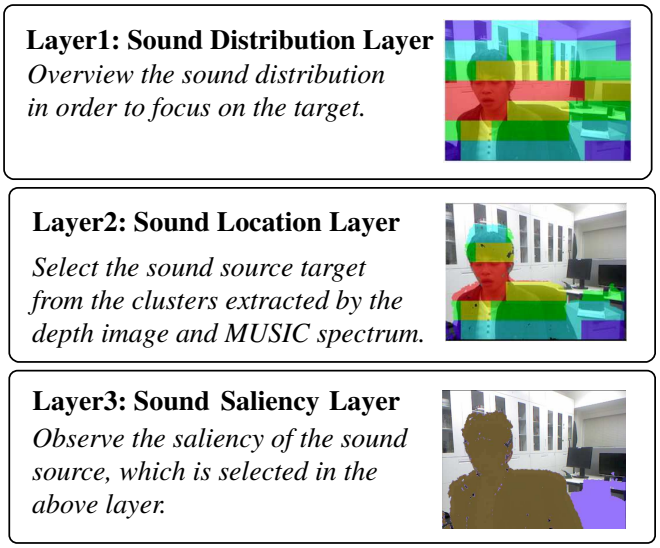


Fig. 1. Design of the three-layer visualization model for auditory awareness

The previous studies on visualizing the acoustic environment were conducted in various ways, such as superimposing the sound pressure on the RGB image [3], near-field acoustic holography [4], and displaying sound directional arrows on the RGB image [5]. There was also a study on designing an interface that visualizes the sound locations, the separated sounds, and the results of automatic speech recognition by using the visual information-seeking mantra proposed by Shneiderman [6][7]. This interface provides the features of *overview*, *zoom and filter*, and *details on demand* for the acoustic information so that users can flexibly focus on particular targets in the environment. These previous studies focused on providing acoustic information and enhancing the presence of the sound source, not on providing auditory awareness.

In this study, we designed a three-layer visualization model for auditory awareness with multimedia sensor input obtained from a microphone array, a Kinect depth sensor, and a camera. We developed an auditory awareness visualization system based on this model. The model consists of three layers: sound distribution layer, sound location layer, and sound saliency layer (Fig. 1), which follows the design of visual information-seeking mantra. The sound distribution layer provides users with features for overviewing the sound distribution in the environment (*overview first*). With this layer, users can overview the sound distribution and select a sound source target, for which they want to observe the

sound information. The sound location layer provides users with the features for spatially extracting sound information from the environment (*zoom and filter*). With this layer, users can distinguish the sound source cluster, which is clustered using the depth sensor image and MUSIC spectrum, which is the signal autocorrelation matrix created by the MUSIC (Multiple Signal Classification) algorithm and represents the source position and the level of its sound pressure [8]. Then they can select the target on which to focus for their preferred depth range. Lastly, the sound saliency layer provides users with features for extracting sound information from the time sequence domain. In other words, the system can notify the users of changes in sound "saliency," such as the appearance of a new target or a significant change in the power of a source (*details on demand*). With those layer features, users can selectively choose and observe their targets of interest.

We also developed an acoustic visualization system with a microphone array and depth sensor based on our three-layer model. The design of the software is based on the MVC (Model-View-Controller) framework, with which the users can flexibly switch, combine, and adjust the three-layer model features and parameters as desired. Thus, the system interface facilitates user awareness of sound source saliency in the environment. Usage examples of the interface are presented in this paper.

The rest of the paper is organized as follows: Section II describes our three-layer visualization model for auditory awareness. Section III describes the development of the auditory awareness visualization system based on the three-layer model. Section IV presents usage examples of the developed system. Section V concludes the paper with a summary of the key points.

II. Three-Layer Visualization Model for Auditory Awareness

The three-layer visualization model was designed to visualize auditory awareness. The model consists of a sound distribution layer, a sound location layer, a sound saliency layer. By switching or combining the features from each layer, the user can extract sound information from three different perspectives; overview, zoom and filter, and details on demand. Each layer consists of two sub-processes: back-end and front-end processing. The back-end processing handles the data processing required in the next higher layer, while the front-end processing handles the generation of the sound images, which can be controlled by the user and rendered in the display. The following sections describe the design of each layer in detail.

A. Sound Distribution Layer (layer 1)

The sound distribution layer provides users with features for overviewing the sound distribution in the environment. With this layer, users can view the combined image of the power distribution of the MUSIC spectrum and RGB image (Fig. 2, right side) and roughly identify the location of the sound sources. When users want to analyze the environment, they first look over the environment with the colored image

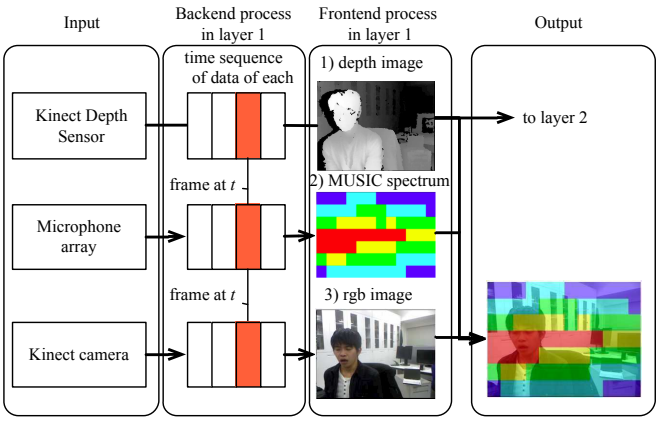


Fig. 2. Integration of MUSIC spectrum and RGB image

of the MUSIC spectrum superimposed on the RGB image. Then they select the targets to be extracted. This layer provides features to overview the sound distribution. Basically what the users do in this layer is adjust the transparency rate of the colored image of the MUSIC spectrum and the threshold power to make the distribution visible. By adjusting the transparency rate and the threshold, users can focus on the changes in the RGB images as well as the power changes on the MUSIC spectrum images.

Figure 2 shows the data flow and processing in the sound distribution layer. The input is a time series of RGB frames from a Kinect camera, depth images from a Kinect depth sensor, and RGB frames of the MUSIC spectrum from a microphone array. The layer creates the colored image of MUSIC spectrum by assigning a color for each sound pressure level and superimposing it on the RGB images for the same frame time t.

B. Sound Location Layer (layer 2)

The sound location layer provides users with features for extracting the sound source locations in the environment. With this layer, users can view the image of the MUSIC spectrum overwrapped on the clusters in the RGB image (Fig. 3, right side). Clusters are a sets of points that are close together and at the same depth level. Users can set the depth range and extract the clusters in that range that they want to observe. These features are useful for narrowing down the targets on which to focus. For instance, suppose there are people near the sensor and the user wants to focus on a background object behind them. By adjusting the depth range, the user can highlight the background object.

Each cluster consists of a set of points, $p_{u,v}$, which are clustered with a K-means++ algorithm [9]. The clustering is done by first extracting the depth image for the user-defined depth range. Pixels $d'_{u,v}$ in the depth image are given by

$$d'_{u,v} = \begin{cases} d_{u,v} & (d_{\min} \leq d_{u,v} \leq d_{\max}), \\ 0 & (\text{otherwise}) \end{cases} \quad (1)$$

where u is the index of the width in the depth image, v is the index of the height in the depth image, and $d_{\min}$ and

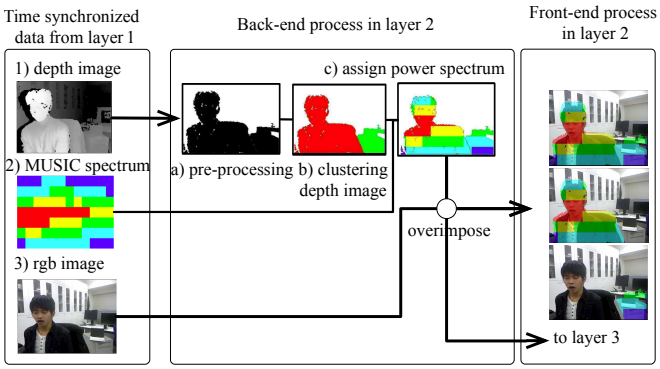


Fig. 3. Sound source location visualization

$d_{\max}$ are the user-defined depth ranges. The distance map given by equation 1 is converted into points $p_{u,v}$ in the depth sensor coordinates with constraints (here, the Kinect sensor constraints). Point $p_{u,v}$ is given by

$$p_{u,v} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \left(u - \frac{w}{2}\right) d'_{u,v} \frac{\tan(a/2)}{h/2} \\ \left(v - \frac{h}{2}\right) d'_{u,v} \frac{\tan(e/2)}{w/2} \\ d'_{u,v} \end{pmatrix}, \quad (2)$$

where w and h are the width and height of the depth image and a and e are the vertical and horizontal fields of view of the depth sensor (here, $w = 640, h = 480$, $a = 59°, e = 53°$). To reduce the load of K-means++ clustering, the system compresses the point array by 75%. Point $p'_{u,v}$ in the compressed array is given by

$$p'_{u,v} = G(\{p^i\}_{i \in S}) = \frac{1}{N} \sum_{i=1}^{N} p^i, \quad (3)$$

where S is the point set corresponding to one pixel of the compressed array in the original point array, and $G(\{p^i\}_{i \in S})$ is the function used to calculate the centroid in the point set. The compressed point array is then given to the K-means++ clustering process, and the system receives the clustered point set. Evaluation function ϕ in K-means++ clustering is given by

$$\phi = \sum_{p_j \in X} \min_{i \in k} \|p_j - c_i\|^2, \quad (4)$$

where X represents the points in the compressed point array, p_j is a point in the array, and c_i is the centroid of the cluster. With these processes, the sound location layer creates the clustered point set and generates the cluster image, which is superimposed on the RGB image (Fig. 3).

C. Sound Saliency Layer (layer 3)

The sound saliency layer provides users with features for extracting the time-sequential differences, in other word, saliency, in the environment. With this layer, users can obtain awareness of the saliency, which is expressed as a colored cluster in the RGB images (Fig. 4). The saliency represents the changes in the positions or distribution of the clusters described in sub-section B. When there is a change in the power of a sound source target or a change of its position, the target gains saliency. On the other hand, if a sound source target has no changes, it loses saliency. The sound saliency is expressed on the image as the transparency rate of the cluster. If the target gains saliency, the rate decreases. If the target loses saliency, the rate increases. The user can adjust the weight of the MUSIC spectrum or the depth when the system detects saliency. The user can also adjust how long the indication lasts. With these features and the features of layers 1 and 2, the system can focus on the sound targets of interest and detect time-sequential differences by saliency mapping.

The saliency of a sound source target is calculated by first calculating the differences in depth cluster l_d and MUSIC spectrum cluster l_m between frames t and $t-1$ on the basis of Kullback-Leibler divergence [10], which is a non-symmetric measure of the difference between two probability distributions. The l_d and l_m are given by

$$\begin{cases} l_d = \frac{1}{2}\left[\log \frac{|\Sigma_{d2}|}{|\Sigma_{d1}|} + tr\{\Sigma_{d2}^{-1}\Sigma_{d1}\} \right. \\ \qquad \left. + (\mu_{d1} - \mu_{d2})^T \Sigma_{d2}^{-1} (\mu_{d1} - \mu_{d2}) - 3\right], \\ l_m = \frac{1}{2}\left[\log \frac{\sigma_{m2}^2}{\sigma_{m1}^2} + \frac{\sigma_{m1}^2}{\sigma_{m2}^2} + \frac{(\mu_{m1} - \mu_{m2})^2}{\sigma_{m2}^2} - 1\right] \end{cases} \quad (5)$$

where Σ_{d1} and Σ_{d2} are covariance matrixes of the depth clusters at t and $t-1$, μ_{d1} and μ_{d2} are the centroids (means) of the depth clusters at t and $t-1$, μ_{m1} and μ_{m2} are the means of the MUSIC spectrum clusters at t and $t-1$, and σ_{m1} and σ_{m2} are the variances of the MUSIC spectrum clusters at t and $t-1$.

Then, with the Kullback-Leibler divergences of the depth and MUSIC spectrum clusters, distance d_c is calculated using a weighted average,

$$d_c = \alpha \cdot l_d + (1.0 - \alpha) \cdot l_m, \quad (6)$$

where α is the weight. Sound saliency $s(t)$ is calculated using d_c:

$$s(t) = \begin{cases} s(t-1) - s_d & (d_c \leq \psi), \\ 1.0 & (d_c > \psi) \end{cases} \quad (7)$$

where s_d is the rate of saliency decrease when there are no changes in the divergence, and ψ is the threshold value used to identify whether the divergence is significant.

With these processes, the sound saliency layer shows

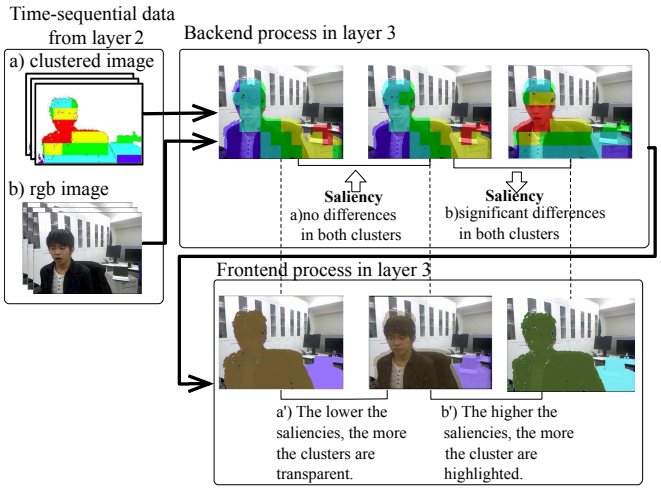


Fig. 4. Saliency visualization

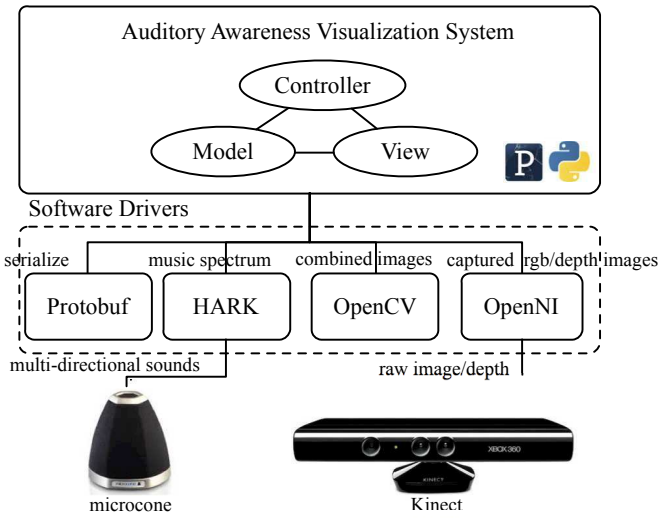


Fig. 5. Auditory awareness visualization system configuration

whether the tracking clusters have saliency and generates a dynamic cluster image, which is superimposed on the RGB image (Fig. 4).

III. Auditory Awareness Visualization System

Using our three-layer visualization model for auditory awareness, we developed an auditory awareness visualization system. As shown in Fig. 5, the input is from a Kinect depth sensor and a microcone, which is a microphone array consisting of six microphones hexagonally located on the side surface and one microphone located on the top surface. The raw RGB image and depth image are captured with an OpenNI library and passed to the system. The multi-directional sounds are captured with a HARK (Honda Research Institute Japan Audition for Robots with Kyoto University) [11] and passed to the system. The system uses an OpenCV for rendering various images on its GUI and uses a protocol buffers library to serialize and deserialize the processed data. The system was coded using python and Processing 2.0 and is designed with the MVC framework, which enables users to flexibly combine and adjust the features of the three-layer visualization model. The system is explained in detail in the following sections.

A. Software Design

We designed the auditory awareness visualization system by using the MVC framework. Figure 6 shows the MVC design . The model part consists of sound distribution extraction, sound location extraction, and sound saliency extraction modules, which correspond to our three-layer model and execute the data processing in each layer. The view part consists of an image integrator, a time-sequence controller, and a GUI renderer. The controller part (Fig. 6, top) consists of event function and feature controller modules. With the input from the GUI, the system can control the kind of extraction module to be activated each moment as well as which combination of images to render in the view part. The following sub-section describes the detailed data processing

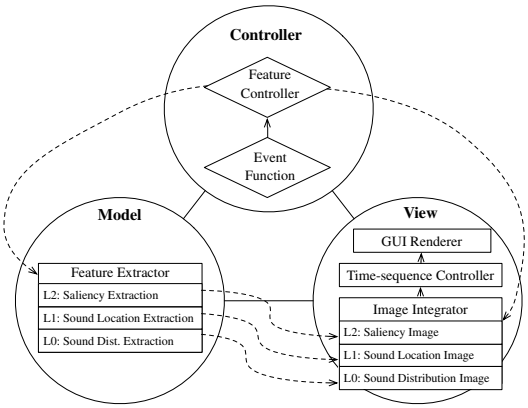


Fig. 6. MVC design of auditory awareness visualization system

in the model part and the user interface design in the view part.

B. Layered Data Processing

In our auditory awareness visualization system, the data is simultaneously processed for the different time scales and passed from lower layer to higher layer. This section describes the data flow of the model part in detail. Figure 7 shows an overview of the data processing. In each layer, there are two main processes, the back-end process and the front-end process to handle the multimedia data. The back-end process mainly processes the data and the front-end process creates the images for the visualization. The following paragraph describes in detail features of each layer.

The back-end process in the sound distribution layer is time synchronization of the MUSIC spectrum, depth, and RGB images captured by the Kinect depth sensor and microcone. The data for each are captured at different frame rates, and the layer module synchronizes these timings. The front-end module generates a color-map image corresponding to the power in the MUSIC spectrum. The synchronized data are sent to the sound location layer while the generated image

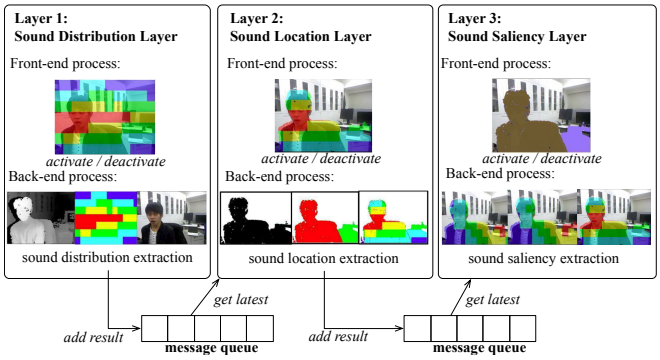


Fig. 7. Layered data processing

is sent to the sound distribution image generator in the view part.

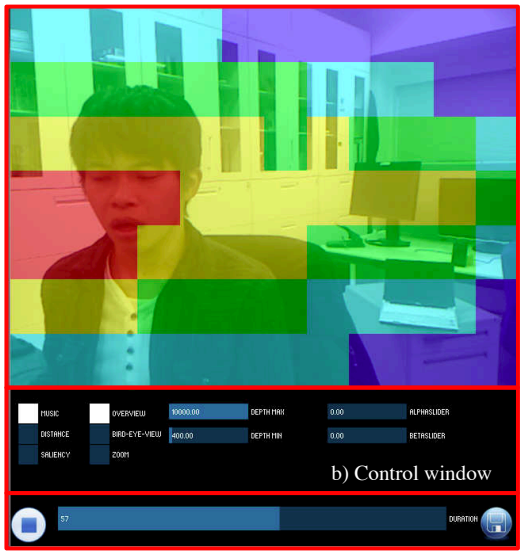
Next the back-end process in the sound location layer clusters the depth image and MUSIC spectrum given by layer 1. The cluster information is sent to the sound saliency extraction module as well as to the front-end module in layer 2. The front-end module in layer 2 generates a cluster image on the basis of the cluster information and MUSIC spectrum. Then the module passes the generated image to the sound location image generation module in the view part. Finally, the back-end process in the sound saliency extraction module tracks the cluster received from layer 2 and calculates the saliency. The front-end module of sound saliency layer generates a cluster image on the basis of the calculated saliency. Activation of the front-end process in each layer is controlled from the controller part, so the layered module in the model part can properly handle the data processing by using a combination of the features of the three-layer model.

The processing speeds of the layers is differ due to differences in process heaviness. Thus, each layer is not piped but linked with the message queues, and each module acquires the latest results from the message queue at each calculation timing.

C. Interface Design

The interface of the developed system was designed so that users can intuitively extract and observe sound source information in the environment. Figure 8 shows the design of the GUI, which consists of a rendering window, a control window, and a time-sequence window. In the rendering window, the system renders the images, which are combinations of images generated by the three-layer model. The user can change the rendered images by changing the parameters used in the three-layer model in the control window. Table I shows the set of parameters that are used in the control window. The parameters are arranged by layer, and users can flexibly combine the feature parameters in the three-layer model and freely change the output image rendered in the rendering window. The time-sequence window consists of a play/stop button, a time-sequence bar, and a save button. Users can jump to a section of interest and repeatedly view the image in the captured data. The save button is used to save the captured data.

a) Rendering window



c) Time sequence window

Fig. 8. Interface design

TABLE I
INTERFACE PARAMETERS

Layer	Parameters	Description	range
1	transparency, t	transparency of music power spectrum image	$0.0 < t < 1.0$
	power threshold, p_{thr}	power threshold used to make lower power region transparent	$0 < p_{thr}$
2	distance min value, d_{min}	min value of depth range	$0 < d_{min}$
	distance max value, d_{max}	max value of depth range	$d_{min} < d_{max}$
3	weight, α	weighting parameter	$0 < \alpha < 1.0$

IV. Usage of Developed System

This section describes trial usage of the developed system. The trial data was captured in an environment with a speaker close to the sensor and a laptop playing music in the middle distance from the sensor. Suppose the user would like to analyze the sound changes for either the speaker or the laptop.

A. Overview acoustic environment (layer 1)

From the perspective of sound distribution, the user can overview the sound distribution in the environment, as shown at the top in Fig. 9. The user can observe that sound is coming from the region near the speaker, the laptop, and the displays (top level of Fig. 9, a). By adjusting the transparency rate of the superimposed image, the user can make the RGB image clearer to facilitate searching for the sound source target (top level, b). By adjusting the threshold to make the sound visible, the user can highlight the region where the level of the MUSIC spectrum is higher than the set threshold (top level, c).

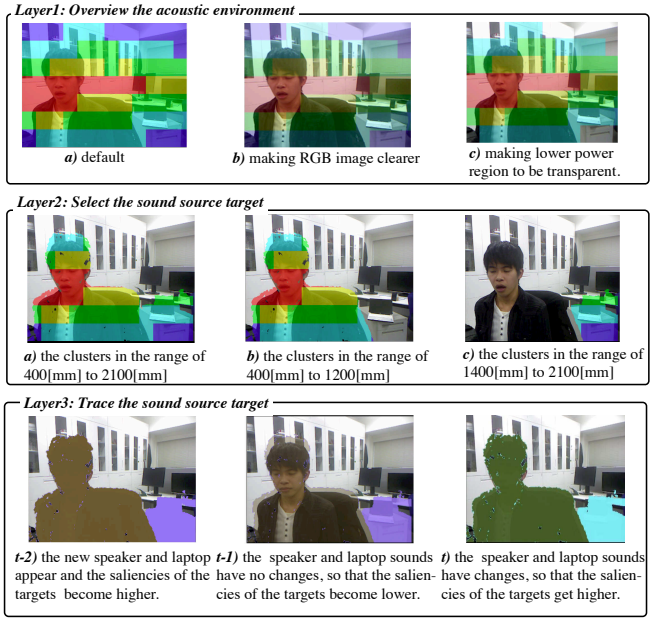


Fig. 9. Usage example of developed system

B. Select sound source target (layer 2)

From the perspective of sound location, the sound source can be extracted for the depth range, as shown in the middle level of Fig. 9, which shows several examples of highlighting the sound source clusters for a certain depth range. The left figure, a), shows the clusters in the range of 400 to 2100 [mm]. The middle figure, b), shows the clusters in the range of 400 to 1200 [mm]. And the left figure, c), shows the ones in the range of 1400 to 2100 [mm]. The first example shows tracing of the sound changes for both the speaker and laptop while the second and third examples show tracing of the sound changes for either the speaker or the laptop. Once the user decides which sound source target to observe, the user switches the features of the sound saliency layer.

C. Trace saliency of sound source target (layer 3)

Finally, the sound saliency layer is used to track how the saliencies of the sound source targets change. The bottom level of Fig. 9 shows the time-sequential changes in saliency for the speaker and the laptop. Between the left and middle figures in the bottom level, the distribution of sound on the speaker has not changed. The transparency rate of the speaker drops as the saliency of the speaker decrease. On the other hand, between the middle and right figures, the distribution of sound on the laptop has changed significantly. Therefore, the laptop gets highlighted to emphasize the changes in the sound distribution.

With those manipulations of each layer's features, users can flexibly select the sound source target on which to focus and observe the target behavior. Thus, our developed auditory awareness visualization system improves the auditory awareness of users when observing a sound environment.

V. Conclusion

Our three-layer visualization model for auditory awareness has a sound distribution layer, a sound location layer, and a sound saliency layer. It was implemented in an auditory awareness visualization system using a Kinect and a microcone. The sound distribution layer provides features for overviewing the sound distribution in the environment. The sound location layer provides features for spatially extracting the sound information. The sound saliency layer provides features for extracting the sound information form the time sequence domain. The system thus visualizes sound saliency, such as the appearance of a new target or a significant change in the power of a source. With those features in the three-layer model, the user can selectively choose a target of interest, and the system interface facilitates analysis of sound saliency in the environment.

References

[1] N. Keiji, H. Ishida, S. Yamanaka, and Y. Tanaka, "Three-dimensional localization and mapping for mobile robot in disaster environments," in *Intelligent Robots and Systems, 2003. (IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on*, vol. 4, Oct 2003, pp. 3112–3117 vol.3.

[2] H. Sun, P. Yang, L. Zu, and Q. Xu, "A Far Field Sound Source Localization System for Rescue Robot," in *Control, Automation and Systems Engineering (CASE), 2011 International Conference on*, July 2011, pp. 1–4.

[3] J. Naoshi, Y. Oikawa, and Y. Yamasaki, "Visualization of sound environment using multi channel acoustic measurement system," in *Acoustic Society Symposium,2008*, Sep 2008, pp. 1509–1510.

[4] J. Even, N. Kallakuri, Y. Morales, C. Ishi, and N. Hagita, "Creation of radiated sound intensity maps using multi-modal measurements onboard an autonomous mobile platform," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, Nov 2013, pp. 3433–3438.

[5] T. Mizumoto, K. Nakadai, T. Yoshida, R. Takeda, T. Otsuka, T. Takahashi, and H. Okuno, "Design and implementation of selectable sound separation on the Texai telepresence system using HARK," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, May 2011, pp. 2130–2137.

[6] Y. Kubota, M. Yoshida, K. Komatani, T. Ogata, and H. Okuno, "Design and Implementation of 3D Auditory Scene Visualizer towards Auditory Awareness with Face Tracking," in *Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on*, Dec 2008, pp. 468–476.

[7] B. Shneiderman, C. Plaisant, M. Cohen, and S. Jacobs, *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, 5th ed. USA: Addison-Wesley Publishing Company, 2009.

[8] K. Nakamura, K. Nakadai, and H. G. Okuno, "A real-time super-resolution robot audition system that improves the robustness of simultaneous speech recognition," *Advanced Robotics*, vol. 27, no. 12, pp. 933–945, 2013.

[9] D. Arthur and S. Vassilvitskii, "K-means++: The Advantages of Careful Seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '07. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[10] J. Hershey and P. Olsen, "Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, April 2007, pp. IV–317–IV–320.

[11] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design and Implementation of Robot Audition System "HARK" – Open Source Software for Listening to Three Simultaneous Speakers," *Advanced Robotics*, vol. 24, no. 5-6, pp. 739–761, 2010.