Investigating Audio Data Visualization for Interactive Sound Recognition

Tatsuya Ishibashi
ishibashi.tatsuya@ist.osaka-u.ac.jp
Osaka University

Yuri Nakao
nakao.yuri@fujitsu.com
Fujitsu Limited

Yusuke Sugano
sugano@iis.u-tokyo.ac.jp
The University of Tokyo

ABSTRACT

Interactive machine learning techniques have a great potential to personalize media recognition models for each individual user by letting them browse and annotate a large amount of training data. However, graphical user interfaces (GUIs) for interactive machine learning have been mainly investigated in image and text recognition scenarios, not in other data modalities such as sound. In a scenario where users browse a large amount of audio files to search and annotate target samples corresponding to their own sound recognition classes, it is difficult for them to easily navigate through the overall sample structure due to the non-visual nature of audio data. In this work, we investigate the design issue for interactive sound recognition by comparing different visualization techniques ranging from audio spectrograms to deep learning-based audio-to-image retrieval. Based on an analysis of the user study, we clarify the advantages and disadvantages of audio visualization techniques, and provide design implications for interactive sound recognition GUIs using a massive amount of audio samples.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**.

KEYWORDS

Sound Recognition; Interactive Machine Learning; Visualization

ACM Reference Format:
Tatsuya Ishibashi, Yuri Nakao, and Yusuke Sugano. 2020. Investigating Audio Data Visualization for Interactive Sound Recognition. In *25th International Conference on Intelligent User Interfaces (IUI '20), March 17–20, 2020, Cagliari, Italy.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3377325.3377483

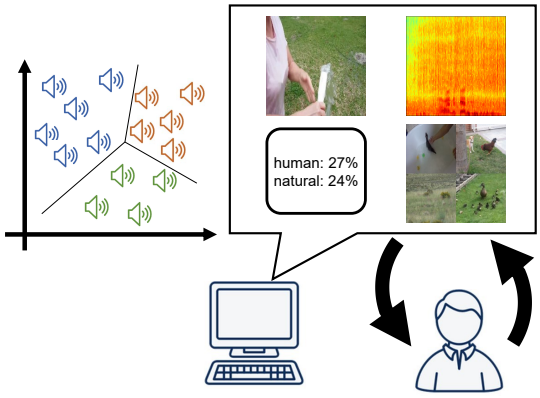

Figure 1: In this work, we comparatively investigate how to visualize audio data in an interactive sound recognition system that allows users to browse a large amount of unlabelled data.

1 INTRODUCTION

While recent advances in deep neural networks have dramatically improved the performance of media recognition tasks, it is still difficult to provide pre-trained models that meet diverse requirements of individual users. There is still a huge gap between actual and required recognition performances, and in most of the application scenarios, the recognition target often varies depending on the user. In order to address these limitations, there is a growing interest in interactive machine learning techniques [3]. In contrast to conventional machine learning techniques assuming the existence of labeled (or sometimes unlabeled) training data, users are directly involved in the incremental model update process of interactive machine learning by inspecting intermediate models and annotating training samples. In this way, interactive machine learning provides a way for individual users to examine the model training process, and adjust the behavior of the trained models so that they meet their own requirements. The key challenge is to design input and visualization frameworks where even novice users without machine learning expertise can interact with machine learning models in an intuitive manner.

However, one of the key limitations of existing studies is that most of them focused on image- or text-related tasks [17–19, 29, 38]. Images and texts are unique in the way that the meaning of individual samples are visually obvious, and even non-expert users can easily interpret the content by, e.g., looking at thumbnail images. This characteristic makes it easy to grasp the overview of a large amount of data, and the prior works proposed GUIs which allows users to browse candidate images and texts for annotating training data. In contrast, most of the other media such as sensor data and audio files are difficult to be visualized, and it is not a trivial task to interpret the meaning of individual data. Therefore, while sound recognition can be applied to a wide range of applications [13, 40, 47] and has a good potential for interactive customization, it is more challenging to design an interactive environment than image recognition scenarios. While there exists a spectrogram as a way of visualizing sound characteristics that has been commonly used by expert users [9, 16,

*Also with Fujitsu Laboratories Ltd..

IUI '20, March 17–20, 2020, Cagliari, Italy

ACM ISBN 978-1-4503-7118-6/20/03…$15.00
https://doi.org/10.1145/3377325.3377483

41], it is not clear whether it is beneficial in the context of interactive machine learning for novice users.

The goal of this work is to investigate the issue of sample visualization for designing interactive sound recognition GUIs. As shown in Fig. 1, we examine a scenario where a large amount of unlabelled audio samples are presented to users so that they can navigate through the samples and annotate samples according to their own target recognition categories. Audio samples are plotted in a two-dimensional space computed according to their feature similarities, and hierarchically clustered so that users can easily navigate through the samples. The most important design challenge here is how to make it possible for users to grasp a whole structure of the audio samples and to quickly interpret the semantic meanings of individual samples. To investigate the efficient visualization strategy, we compare several different visualization cues to represent samples and clusters in our GUI. They include audio spectrograms, thumbnail images obtained from reference videos, abstract audio class labels estimated from a pre-trained classifier, and cross-modal conversion using the state-of-the-art audio-to-image retrieval method [7]. Throughout an analysis of user study, we summarize the advantages and disadvantages of each cue and how they affect the user interaction.

The contribution of this work is summarized as follows. First, to the best of our knowledge, this is the first work to investigate the design of interactive machine learning system for sound classification. We propose and examine a novel interactive framework for training sound classification models from a large amount of unlabeled samples. Second, we provide a detailed analysis of user behaviors and questionnaires based on the user study, and clarify how novice users perform the interactive sound classification task and how each visualization influences the interactions. Finally, based on the analysis, we discuss design implications for interactive sound recognition GUIs.

2 RELATED WORK

Our work extends existing interactive machine learning literature through visualization scheme, and also related to sound annotation techniques for professional users.

2.1 Interactive Machine Learning

Interactive machine learning allows users to train their personal machine learning models through interaction and visual inspection [3, 18, 19]. The models are trained in an incremental manner through user interactions, and users can construct personalized models through trial and error. Interactive machine learning systems are expected to provide users simple and intuitive ways for examining the model output and adding training data. In prior work, there have been several attempts to present generic approaches for letting users directly interact with machine learning models regardless of the input modality via, e.g., confusion matrix [4, 30, 42]. However, most of the existing interactive machine learning techniques have been investigated in application-oriented ways, mainly in the fields of image- or text-related machine learning tasks.

Interactive machine learning has been often used to incorporate human knowledge into the process to compensate for the difficulty in the pre-defined recognition task. In this case, users are expected to improve recognition model performance by providing additional training data. In the image recognition domain, prior work allowed users to provide feedback on recognition results, and to add new sample annotations to the recognition model [15, 18]. Similarly, interactive machine learning has been also employed in text-related applications such as spam filtering and sentence proofing. Huang et al. proposed a system to support writing and reading online reviews [27]. Their system predicted and presented category score to the reviewers, while they could make a correction to wrong predictions. Abstrackr [43] was another online system for screening citations of systematic reviews. The system could improve its classification output of relevant citations based on the relevance labels provided by users. While in these approaches users could improve the performance of pre-defined recognition models, they did not discuss the interaction during the model definition phase and users were not allowed to define their own recognition models.

Model definition is one of the most essential parts in machine learning, and interactive machine learning systems also tried to provide users a way to freely explore unlabeled samples to define their own class labels. Enders et al. proposed a design for visual analytic interaction, where a large document dataset was displayed in a 2D layout and users could perform data analysis on these distributed samples [17]. The system updated the corresponding parameters of the analytic reasoning, and the visualization was also updated according to the parameters. Some prior work also tried to design interactions for defining user-specific image recognition models [5, 19, 23, 26, 36]. Cueflik [19] was an interface which allows users to create personal rules for image search, and the authors showed that users can perform the model definition task more efficiently when the best and worst matching samples are displayed. Similarly, in order to let users easily find similar groups of training samples, there have been some visualization approaches using two-dimensional feature embeddings [26, 36]. In Sharkzor [36], unlabeled images were embedded into a two-dimensional space by using the t-SNE algorithm [34], and users could perform rearrangement and grouping operations directly on the embedded space.

While our work has a similar motivation to these prior work, we explore the potential of interactive machine learning in the sound recognition scenario. As summarized above, most prior works relied on the visualization of training samples and recognition results, which was relatively straightforward for texts and images. The key research question in this work is whether it is possible to extend the same idea to the audio data domain.

2.2 Sound Recognition and Annotation

Sound recognition has been applied to many applications including intelligent noise-canceling systems [20, 46] and assistive devices for deaf and hard-of-hearing people [10, 35]. While recent advances in deep neural networks have also significantly improved the state-of-the-art performance of generic sound recognition [21, 25], it still has difficulties in accuracy and task diversity, and user adaptation is often required to meet real-world demands. Therefore, there is a great potential for interactive sound recognition systems. However, it is more difficult to browse and annotate a large amount of audio data than texts and images. Users cannot easily grasp the audio contents at a glance, and need to listen to individual audio samples.

Spectrogram has been commonly used for visualization in the context of annotation systems for professional purposes such as bioacoustic recordings [9, 16, 41]. It has been reported that expert annotators do not even have to listen to audio signals, and perform the annotation task only with spectrograms [41]. However, despite the efficiency on identifying changes in the audio event, past studies have also suggested that it requires some experiences to recognize and interpret the visual patterns of spectrogram representation [12]. In the context of interactive machine learning for novice users, it is not clear whether spectrograms can be used to inspect audio contents and whether the spectrogram representation is the best way for information visualization.

Cross-modal conversion or retrieval techniques have been also actively investigated to visualize sounds by converting them into images or texts with the similar semantic meanings [2, 32, 37, 45]. In recent years, deep learning-based approaches have been attracting more attention and conversion/retrieval performance has been greatly improved. Prior researches proposed to generate images from sounds using Generative Adversarial Networks [14, 44], or learned a deep feature representation which can be used for cross-modal retrieval [8]. Arandjelović and Zisserman further proposed to learn such a feature representation from a large collection of video clips without human-annotated labels [6, 7]. However, the effectiveness of these methods was evaluated only in an experimental laboratory setup, and the usability of these state-of-the-art cross-modal conversion techniques has never been investigated in the context of interactive machine learning.

There have been a few attempts for enabling efficient annotation of a large amount of audio samples [22, 31, 39]. Shuyang et al. proposed an idea to use clustering results based on auditory characteristics for fast annotations [39]. When a user assigned a class label on the representative sample of a cluster, all samples in the same cluster were labeled as the same class. Some researches have also reduced the annotation cost by using semi-automated approaches [22, 31]. In these methods, samples similar to the user-labeled ones and important samples around the decision boundary were presented to the user. While these studies have only discussed the visual layout of audio samples, we extensively investigate the interpretability issue of auditory signals in interactive machine learning including the sample browsing and model definition phase.

3 INTERACTIVE SOUND RECOGNITION

Figure 2 illustrates the basic overview of our GUI for interactive sound recognition. The interface consists of three components. The left panel (Fig. 2 (a)) shows frames corresponding to user-defined classes and annotated samples. The center panel shows the sample and class embedding (Fig. 2 (b)), and the right panel shows details of the selected sample (Fig. 2 (c)). The goal of our GUI is to let users easily browse a large amount of audio files for labeling, and train their personal sound recognition models.

All audio samples are embedded into two-dimensional space according to their similarities in feature space, and hierarchically clustered for a better overview. While this 2D hierarchical representation shows the relative relationships between samples, the key difficulty is how to visualize the actual contents of each sample and cluster. Although we tentatively set speaker icons to show audio

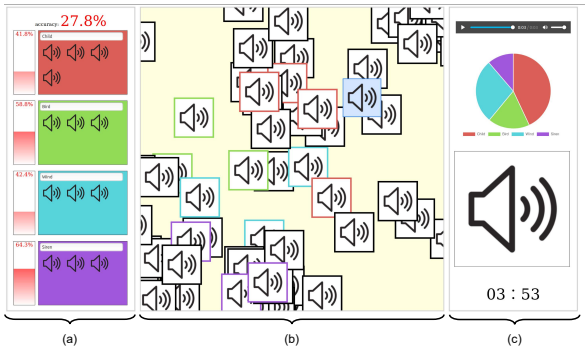

Figure 2: Overview of the GUI design for interactive sound recognition. (a) The left panel shows frames corresponding to user-defined classes and annotated samples. (b) The center panel shows the sample and class embedding, and (c) the right panel shows details of the selected sample. Audio samples are illustrated with speaker icons, but in this work we examine different visualization techniques using this base GUI.

samples in feature space in Fig. 2, they are obviously not informative enough. The main purpose of this work is to investigate how to visualize these audio samples in an intuitive manner for users. In this section, we first describe the basic interaction flow of our GUI. We then describe technical details of our visualization techniques, followed by implementation details of our machine learning backend.

3.1 User Interaction

Figure 3 illustrates the flow of our proposed interaction. Users first define target classes in the left panel. Users can add boxes corresponding to a new sound class, and define the class name in the text boxes (Fig. 3 (1)). For each class, users can optionally set reference validation samples to quantitatively evaluate the classification accuracy during interaction.

Users then add training data to these user-defined classes from the set of unlabeled samples displayed in the center panel. All samples are displayed in the form of hierarchical clusters, and users can both move and zoom in/out through the sample visualization (Fig. 3 (2)). In our implementation, we use the t-SNE algorithm [34] to compute the sample embedding in the center panel. Samples with similar audio features become close to each other in the 2D embedding space. Clustering is directly performed on the embedded 2D coordinates, and the hierarchical structure is obtained by applying k-means clustering multiple times to the cluster centroids. The embedded features are further clustered using the k-means algorithm [33]. Users can move the sample cluster horizontally and vertically by drag operation, and zoom in (out) the sample hierarchy by rotating the mouse wheel up (down). The panel shows the corresponding sample clusters according to the zoom level. The panel shows sample clusters with their representative samples at higher levels of the sample hierarchy, and higher hierarchy shows fewer, more high-level clusters. At the lowest level, the panel shows individual audio samples.

When users double-click individual audio samples, they can play the audio file and check details of each data at the right panel (Fig.

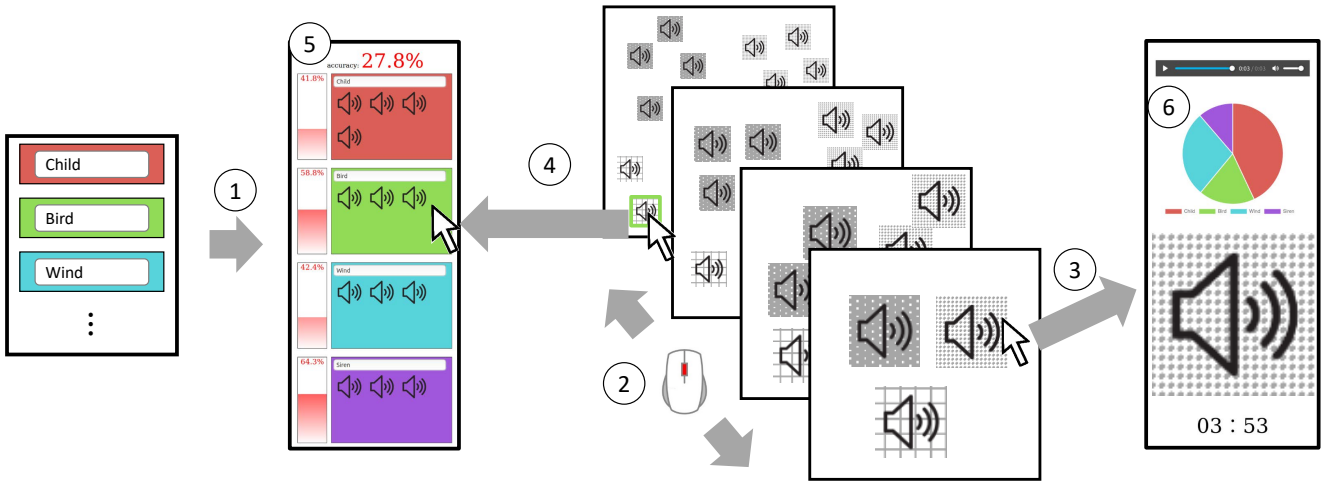

Figure 3: The overall usage flow of our proposed system. 1) Users first define target classes in the left panel. 2) They then browse unlabeled samples through the visualization, and 3) check individual samples in detail. 4) They can select and add positive samples to the user-defined classes. 5) Every training process, the left panel shows the accuracy of each class and the overall accuracy for the validation data, and 6) This further shows current estimation results on individual samples. Users repeat this labeling and training process to update the model until it achieves the desired performance.

3.2 (3)). The right panel shows an audio player to playback the data, together with the enlarged view of the visualization cue. In order to add positive samples to the user-defined classes, users first select target samples by ctrl-clicking individual samples or right-dragging the target area. These selected samples can be added to one of the user-defined classes by clicking the target class frame in the left panel (Fig. 3 (4)). Similarly, users can also select the lowest cluster and add all the samples belonging to the cluster. Then the colors of the samples border are assigned with the same color as the user-defined classes on the left panel. Users can also remove the samples from classes by ctrl-click.

When users add samples to the user-defined classes, the classification model is re-trained using the current training data. Once the user-defined classification model is updated, the left panel shows the accuracy of each class and the overall accuracy for the validation data (Fig. 3 (5)). In addition, the right panel also shows the class probability scores (Fig. 3 (6)). Users repeat this labeling and training process to update the model until it achieves the desired performance.

3.2 Sample and Cluster Visualization

Based on the interaction design described above, we examine four visualization cues listed in Fig. 4 for samples and clusters.

Video Thumbnail. If the audio sample is taken from a video clip, it is possible to extract a representative frame as a thumbnail image. Although this visualization is not always applicable to audio samples, we can at least insert a few video data with thumbnail images as references. In our implementation, the middle frame of the original video (if available) is extracted as the thumbnail image representing the audio data.

Spectrogram. As discussed earlier, spectrogram has been commonly used to visualize audio characteristics. Audio spectrograms can be obtained by applying a short-time Fourier transformation to the waveforms of audio samples. We adopt hanning window as window function. The number of data points used in each block for the Fourier transformation is 256 and the number of points of overlap between blocks is 128. The power scale is expressed in dB.

Semantic Representation. Another potential approach is to present textual descriptions of each audio data. Since describing diverse audio data is still a difficult task and it is a chicken-and-egg problem to provide descriptions exactly matching the user demand, we use classification results with highly abstract sound categories. In our implementation, we pre-trained a generic sound classifier using the AudioSet dataset [21]. In order to provide high-level abstract understandings of audio characteristics, we took seven semantic categories ("Human sounds", "Animal", "Music", "Sounds of things", "Natural sounds", "Source-ambiguous sounds" and "Channel, environment and background") from the highest level of the AudioSet ontology. Once trained, this classifier can be used to estimate semantic class probabilities of arbitrary audio samples. We use estimated class probabilities to represent individual samples and clusters. The probability scores of clusters are computed from the averages of probability scores of all samples belonging to the cluster.

Audio-to-Image Retrieval. While thumbnail images are only available for audio samples taken from videos, we can also convert arbitrary audio samples to images by using cross-modal retrieval techniques. In this work, we use the state-of-the-art audio-to-image retrieval method [7] to obtain images representing the semantic meaning of audio data. This method learns two convolutional neural networks that embed both audio and image data into a unified 128-dimensional feature representation so that their Euclidean distance

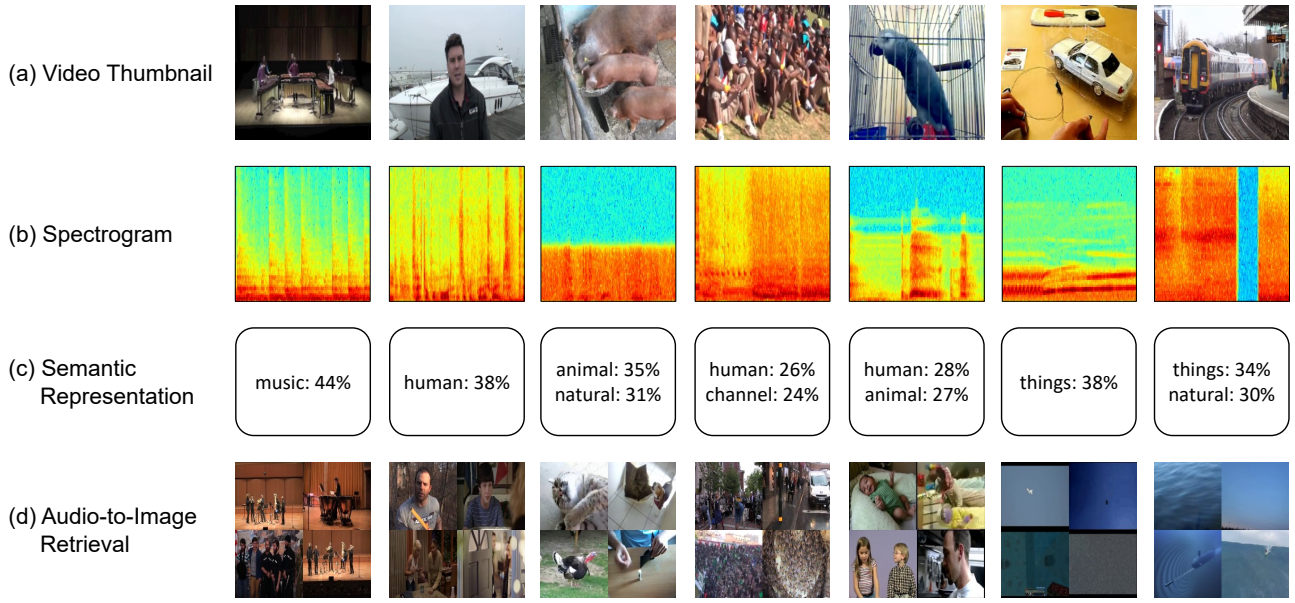

Figure 4: Examples of visualization cues used in this work. Each column shows visualization results of the same audio sample using the technique denoted on the leftmost column.

can be used to discriminate whether two embedded features are taken from the same video source. In other words, this method obtains the cross-modal feature representation in a self-supervised manner, by training the network to classify matching audio and image pairs from randomly shuffled ones. As a result, this feature representation can be used for audio-to-image retrieval; we compute the 128D feature from individual audio samples, and retrieve images whose 128D features are closest to the audio feature. Each sample cluster is represented using images retrieved from the most audio sample closest to the centroid.

3.3 Feature Extraction and Classification

The audio features for sound recognition are extracted using the VGGish convolutional neural network [25] pre-trained on the AudioSet dataset [21]. Audio spectrograms are obtained by applying a short-time Fourier transformation to the waveforms, and the spectrograms are then integrated into 64 mel-spaced frequency bins and transformed into mel-spectrograms. The mel-spectrograms are fed into the VGGish network, and audio features are calculated for every second. The output feature vector from the VGGish network is 1024-dimensional, and it is then reduced to 128-dimensional representation by applying PCA [28] and whitening.

In order to train the sound classifier, we chose the Random Forest algorithm [11] considering the ease of hyperparameter tuning and computational cost. Given the user-defined classes, the multi-class Random Forest classifier is trained on the 128-dimensional feature vectors. The number of trees in a random forest is set to be 100, and each tree is expanded until all leaves are less than two samples.

4 EXPERIMENTS

We conduct a user study to verify the effectiveness of our interaction design for sound recognition and examine how each visualization influences the interactive machine learning process. We asked participants to search and annotate training samples for classifying multiple sound categories, and compare the usability scores, classification performances, and user interaction logs under different visualization.

Throughout the user study, we used audio samples from the AudioSet dataset [21]. As the classification target, we selected four sound categories: *child*, *bird*, *wind*, and *siren*. These four classes were selected from the AudioSet ontology as reasonably abstract audio categories which are not too similar to each other. We manually selected 2,000 audio samples from the *unbalanced train* subset of the AudioSet dataset, 500 clips per category, as validation data. We further picked three sets of 10,000 audio samples from the AudioSet as unlabeled data. They were randomly selected from the *balanced train* and *evaluation* subsets of the AudioSet dataset, while balanced to contain equal amounts of samples corresponding to the four target classes. We randomly trimmed the original data to three seconds. The task for participants is to classify the four target classes by using the interactive system to search and annotate corresponding samples from the 10,000 unlabeled samples.

4.1 Visualization Approaches

During experiments, we compare three approaches illustrated in Fig. 5. Each 10,000 samples were clustered into four hierarchies in our visualization, and Fig. 5 illustrates samples and clusters visualized in the center panel (Fig. 2). The raw embeddings of 10,000 samples were first clustered into 1,000 clusters, and then their cluster

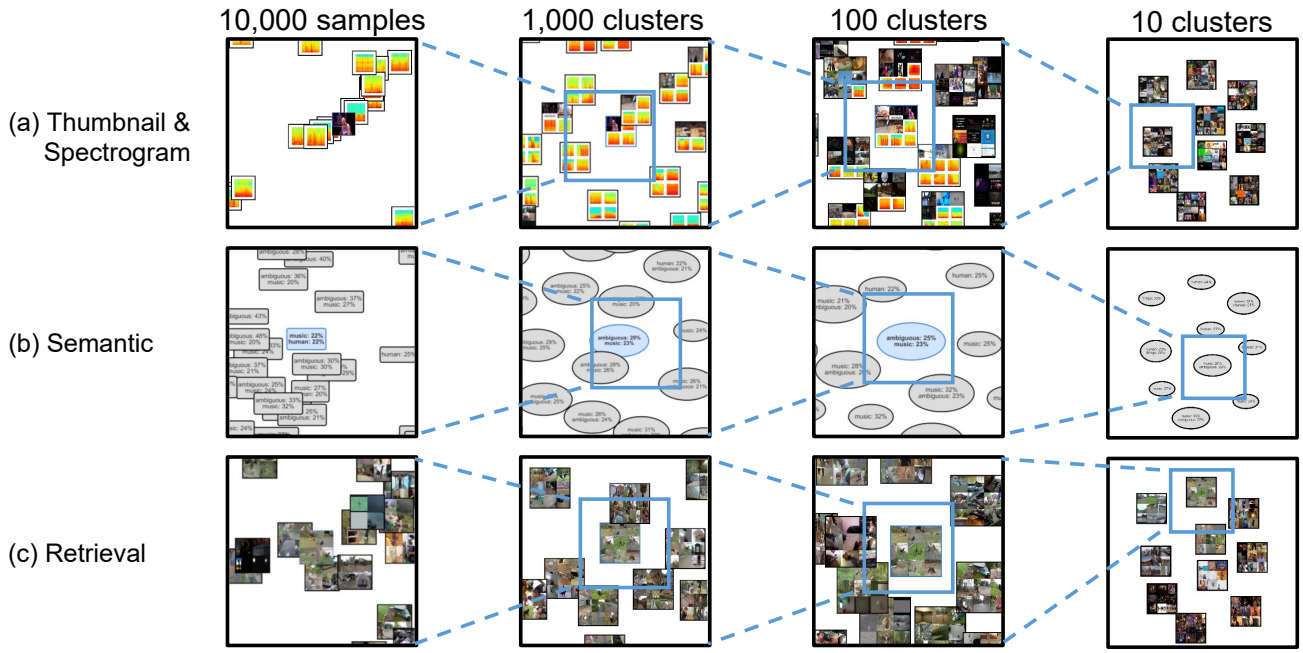

Figure 5: Candidate visualization approaches in our experiments. Each column shows the samples or clusters visualized using each approach.

centroid locations were clustered into 100 clusters, and then their cluster centroid locations were further clustered into 10 clusters. The rightmost column in Fig. 5 is the highest hierarchy where users can see the entire sample distribution. As the user goes to the left (lower) hierarchies, users can browse more fine-grained clusters and individual samples at the lowest hierarchy.

Thumbnail & Spectrogram Since thumbnail images are not always available and spectrograms alone are too difficult for novice users to interpret, we combine these two visualization cues as the first approach (Fig. 5 (a)). To obtain thumbnail images, we use videos corresponding to the AudioSet samples from the YouTube-8M dataset [1]. In our experiments, we assume 1,000 out of the 10,000 samples have the reference videos. These 1,000 samples are represented using the corresponding thumbnail images, and the remaining 9,000 images are represented using spectrograms. The clustered representation uses samples near the centroid as representative samples, but thumbnail-associated samples are prioritized over spectrogram-associated samples.

Semantic As the second approach, we visualize all samples and clusters with the semantic class probabilities (Fig. 5 (b)). The generic sound classifier is trained using 1-second audio data from the *unbalanced train* subset of AudioSet, which is not overlapping with the unlabelled samples in our experiments. We extract audio features and train the Random Forest classifier in the same manner as described in Section 3.3. The estimated class labels and their probabilities are used for visualization, and each sample and cluster shows class labels with probabilities more than 20%.

Retrieval As the last approach, we visualize samples and clusters using audio-to-image retrieval results (Fig. 5 (c)). For training the cross-modal feature embedding network, we use 500,000 1-second video clips from the YouTube-8M dataset, which are part of the *unbalanced train* subset of AudioSet and not overlapping with the unlabelled samples. The network architecture and the training procedure follows the original paper [7], but the audio and image sub-networks are each initialized weights by training auto-encoders in each modality. We use the audio spectrogram computed from the middle one second of the input audio sample, and candidate images for image retrieval are taken from the same data used to train the feature embedding network. Image retrieval during experiments is done from the other two unlabelled sample subsets, i.e., 10,000 audio samples in each sample set are visualized with similar images taken from the videos corresponding to the other 20,000 samples. Each sample is represented using the four most similar images, and each cluster is represented using the nine most similar images.

4.2 Procedure

We recruited eighteen (nine female) participants ranging from 18 to 25 (M = 22.06, SD = 2.09) years old from university mailing lists. None of them had previous experience in either machine learning or sound recognition technologies.

The study was conducted using a desktop computer with a 27" 4K monitor, and the participants used a headphone to preview audio clips. The participants were first given a detailed explanation about the task and the interactive machine learning system. They were also encouraged to perform some dummy classification tasks to

familiarize with the interfaces. All validation data were presented to the participants before starting the task. After inspecting the validation data, the participants freely used the system to annotate unlabeled audio samples to accurately classify these test samples. The participants could also check these validation samples during experiments. In order to set a high standard on the task performance and to encourage active engagement from participants, we instructed them to perform the annotation task aiming at 70% classification accuracy on validation data[1]. Regardless of whether or not this goal was achieved, all sessions were forced to be 30 minutes. The above process was repeated using each of the three visualization approaches. The order of using visualization approaches and the combination between visualizations and unlabelled sample sets was counterbalanced among participants.

We recorded all possible user operation logs such as mouse click and scrolling, and labeling results (selected samples for each target class) together with timestamps. After each session, the participants were asked to rate the perceived workload of the annotation task using a questionnaire based on NASA-TLX [24]. We asked the participants to rate 5 out of the NASA-TLX evaluation items, *Mental Demand*, *Temporal Demand*, *Effort*, *Performance*, and *Frustration*, in 10 levels. After all sessions, the participants were further asked to rate each interface using subjective preference scores in 5-point Likert scale. As the post-experiment questionnaire, we also asked their subjective feedback on positive and negative aspects of each interface in a free-form survey.

4.3 Results

In this section, we first summarize quantitative data obtained from the interactive classification task performance, interaction logs, and workload rating results. We then describe the results of preference rating from participants, together with a summary of their subjective feedback.

4.3.1 Quantitative Analysis. Figure 6 shows the transition of accuracy and number of annotation operations during experimental sessions. Horizontal axis corresponds to time in each experimental session, and all plots show the mean values across all participants at every 1 minute. Left vertical axis (bold solid line) corresponds to the mean accuracy of the trained classifier, which is calculated using the 2,000 (500×4 categories) validation data presented to the participants. Following the evaluation protocol in prior work [25], estimated class probabilities for each second are averaged over each 3-second audio clip and the class with the highest probability is treated as the output for each audio clip. Right vertical axis (dashed line) corresponds to the mean number of annotation operations by participants, including both addition and removal to all of the target classes. At the end of experiments, the *Thumbnail & Spectrogram* approach shows the best accuracy with a significant difference from the *Semantic* method ($p < 0.05$, Wilcoxon signed-rank test). The *Thumbnail & Spectrogram* approach also achieves the largest number of annotation operations, with a significant difference from the *Retrieval* method ($p < 0.01$, Wilcoxon signed-rank test). The instructed performance target of 70% classification accuracy was achieved at 10 out of 18×3 sessions.

[1] If the classifier is trained with labels provided by the AudioSet, i.e., if the users can find all "ground-truth" samples belonging to each class, the mean accuracy is 80.8%.

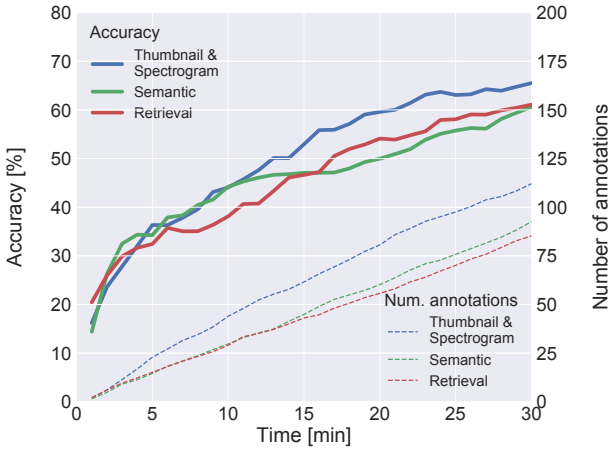

Figure 6: Transition of classifier accuracy (left, bold lines) and number of annotation operations (right, dashed lines) during experimental sessions. Horizontal axis corresponds to time in each experimental session, and all plots show the mean values across all participants at every 1 minute. Accuracy is calculated using the validation data presented to the participants. Number of annotation operations include both addition and removal to all of the target classes.

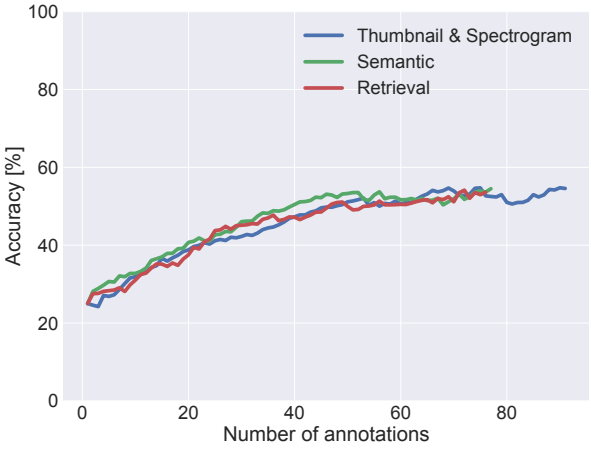

Figure 7: Mean classification accuracy with respect to the number of annotations. Each plot shows the mean accuracy at the time with the same number of annotation operations given by (at least nine) participants.

To provide another perspective on Fig. 6, Fig. 7 further shows the mean accuracy with respect to the number of annotations. While the vertical axis corresponds to the accuracy calculated in the same way as Fig. 6, these plots show the mean accuracy at the time with the same number of annotation operations given by participants. We plot each line up to the point with more than nine participants, and thus *Thumbnail & Spectrogram* has a longer plot than others. As can be

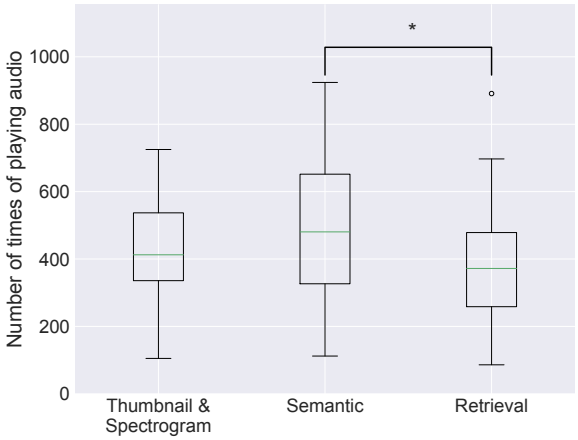

Figure 8: Number of times participants played audio samples during experiments. Each box plot corresponds to each visualization approach, and the number of playing operations includes both individual samples and clusters.

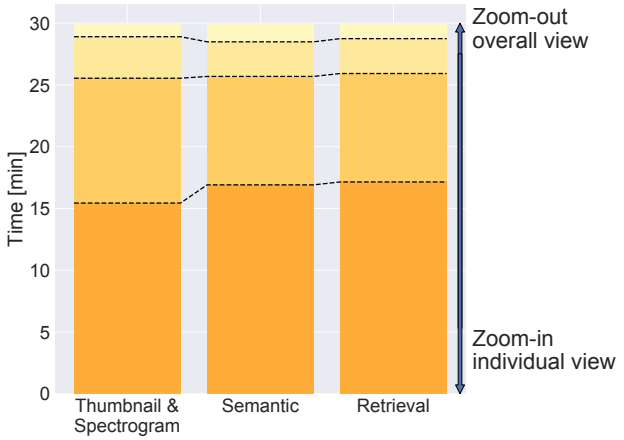

Figure 9: Total time spent in each of the four hierarchies of sample clusters. Each stacked bar plot shows the mean total time the participants spent in each of the four hierarchies, from bottom (lowest) to top (highest hierarchy).

seen, there is almost no difference between the three approaches with the same number of annotations. This indicates that the superiority of the *Thumbnail & Spectrogram* approach in classification accuracy is brought mainly by the highest number of annotations.

To illustrate the differences in user behavior, Fig. 8 shows the number of times participants played audio samples during experiments. Each box plot corresponds to each visualization approach, and the number of playing operations includes both individual samples and clusters. It can be seen that participants listened to more audio samples while using the *Semantic* approach. In particular, there is a significant difference from the *Retrival* approach ($p < 0.05$, Wilcoxon signed-rank test). While both *Semantic* and *Retrival* approaches synthesize the representations without relying on auxiliary video data, participants tend to annotate samples without listening to them while using the *Retrival* approach.

Figure 9 visualizes the total time spent in each of the four hierarchies of sample clusters. Each stacked bar plot shows the mean total time the participants spent in each of the four hierarchies. Users can see the overview of the top 10 clusters at the highest hierarchy (top in Fig. 9), while they can zoom into individual samples at the lowest hierarchy (bottom in Fig. 9). Although the difference is not significant, participants tend to spend more time in the lower hierarchy while using the *Retrieval* approach.

Figure 10 shows the summary of workload assessments by participants. Each bar shows the mean rating of all participants, and the error bars show their standard deviations. Overall, the *Thumbnail & Spectrogram* approach achieves the best rating while the *Semantic* approach achieves the worst. Especially, the *Thumbnail & Spectrogram* approach shows significantly better rating on performance than the *Retrieval* and *Semantic* approaches ($p < 0.05, 0.01$, Wilcoxon signed-rank test). It also shows significantly better rating on frustration than the *Semantic* approach ($p < 0.05$, Wilcoxon signed-rank test).

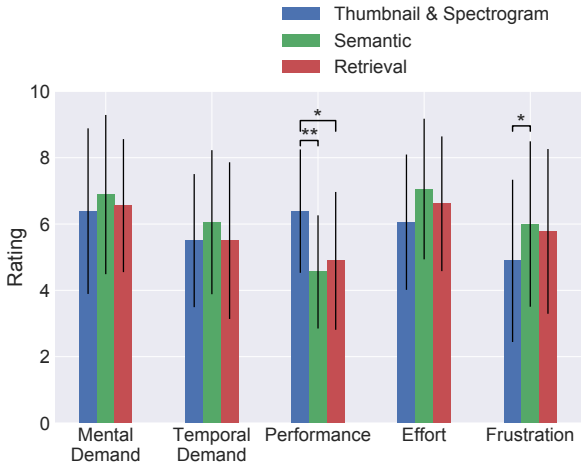

Figure 10: Summary of workload assessments by participants. Each bar shows the mean rating of all participants, and the error bars show their standard deviations.

4.3.2 Subjective Feedback. Figure 11 shows 5-point Likert scale preference scores for each approach from participants. Each bar shows the mean score of all participants, and the error bars show their standard deviations. The *Thumbnail & Spectrogram* approach achieves the highest ratings among them, with the significant difference from the *Semantic* method ($p < 0.01$, Wilcoxon signed-rank test). When we asked all participants which interface is the most preferred, 11 out of 18 answered that the *Thumbnail & Spectrogram* interface.

Most participants relied on visual inspection to browse unlabelled samples, and thus they preferred visualization approaches using images. Eight participants mentioned that they could imagine actual

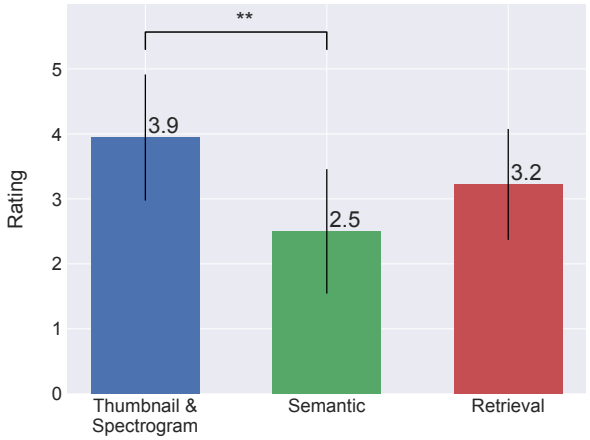

Figure 11: 5-point Likert scale preference scores from participants. Each bar shows the mean score of all participants, and the error bars show their standard deviations.

sounds from thumbnail images. However, at the same time, there were three participants who commented that the thumbnail images did not match the actual sound contents. Similarly, there was some positive feedback on the audio-to-image retrieval results for sample browsing. Eleven participants mentioned that they could imagine the actual sounds from the visualization using audio-to-image retrieval. Eight participants also mentioned that, from audio-to-image retrieval results, they could guess where their target audio samples are in the overall distribution. One participant commented: "*It was easy to comprehend the visual commonality between audio samples, because each audio was represented by multiple images.*" However, some participants also mentioned negative aspects of the audio-to-image retrieval, mainly in terms of correspondence with actual sound contents. Fifteen participants commented that they were confused because the audio-to-image retrieval results did not match the audio samples. One participant said: "*The actual sounds were often different from the sounds I imagined from the visualization. The matching accuracy in my impression was about 50%. During the last half (of the experimental session), I didn't refer to the visualization much.*" There were also two participants who mentioned about the information overload with too many images: "*There was too much information on the screen. The lower (sample cluster) hierarchy I went, the more the visualization got crowded and difficult to grasp. I was a little tired of browsing.*"

Eight participants mentioned that they could also guess where their target audio samples are from the semantic representation of sample clusters. One participant stated: "*If I just want to grasp the overview, I could quickly recognize from the text description what kind of sound data is distributed there,*" and another said: "*Even though the information is rough, text has relatively stronger power of expression (than image representation).*" However, seven participants commented that they were confused because the class probabilities often did not match the audio contents. Four participants also mentioned that it was difficult to imagine sounds from non-visual representation.

Interestingly, ten participants suggested that the spectrograms worked complementary to thumbnails and improved the understanding of visualization. They mentioned that the sounds could be roughly guessed with thumbnails, and the spectrograms can be used to understand the characteristics of the audio samples. One participant said: "*As I kept looking at the spectrograms and listening to the actual sounds, I could gradually understand how to narrow down the search to the target samples. For example, I could guess samples with discrete patterns appearing in the spectrograms are not what I want.*" Similarly, another participant also said: "*I could roughly grasp the actual contents of audio samples from thumbnail images, and the magnitude of the sound from spectrograms. (The spectrograms were) Especially helpful when I was searching for sounds of* ." However, this seems to depend on the user characteristics, and three participants commented that it was difficult to imagine sounds from the spectrograms. One participant stated: "*I could not imagine (actual sounds) by only looking at spectrograms, and I couldn't help searching for thumbnails.*"

Overall, eight participants provided positive feedback on the base visualization approach to map audio samples into the two-dimensional space using t-SNE. One participant stated: "*Similar samples were arranged close to each other, and I can select and annotate multiple samples together.*" Two participants also positively mentioned the cluster structure and their visualization. They commented that such a hierarchical structure made it easy to grasp the characteristics of a group of audio samples and helped them decide where to explore.

One participant also said: "*I was enjoying that the system displays the current model accuracy immediately, and I feel it provided good feedback.*" However, there was some negative feedback regarding the accuracy display and model behavior. Three participants, including the one gave the positive comment above, mentioned that they felt frustrated when the classification accuracy dropped down even with "correct" annotations according to their standards. One of them stated: "*Annotations for child and bird classes were particularly difficult. In the former case, I felt that the accuracy had dropped when I added voices of crying babies and talking infants. In the latter case, the same thing happened when I added the sound of chirping birds and noisy crowing.*"

5 DISCUSSIONS

Throughout the user study, we found some key characteristics of the visualization cues we examined. These findings also lead to some design implications for the future development of interactive sound recognition systems.

5.1 Key Findings

According to the qualitative performance and user feedback, *Thumbnail & Spectrogram* approach was the best among our candidate designs. We assume that this is because the combination of thumbnail and spectrogram has produced a synergistic effect when jointly used with the similarity-based sample embedding. While users can understand the acoustic characteristics (high or low frequency, discrete or continuous pattern) from spectrograms, it is difficult for novice users to guess their semantic meanings. In contrast, thumbnail images can serve as a rough indicator of the semantic meanings,

while the details of the actual sound contents are completely omitted. In this way, users could grasp the overview of sample distribution from thumbnail images, and at the same time rely on spectrograms to infer the characteristics of individual audio samples. However, as discussed earlier, thumbnail representation requires that the audio is taken from a video clip, and it is not always possible to obtain the corresponding thumbnail for arbitrary audio data. It is still an open question to design an interface for large-scale audio data annotation without any reference visual information.

Both semantic class probabilities and audio-to-image retrieval results have an advantage that they can efficiently guide users to browse a large amount of audio data and find the target samples. This can be seen from the user feedback on *Semantic* and *Retrieval* approaches, and from less time user spent in higher sample hierarchy (Fig. 9). Compared to the fully data-driven video thumbnail representation, these learning-based methods provide stronger abstractions of sound contents and represent high-level overviews of the semantic meanings. However, while users could roughly guess the sound category from semantic class probabilities, it was far more difficult to imagine the actual details of the audio data than visual representation. Therefore, as shown in Fig. 8, users played and listened to many audio samples while using the *Semantic* approach. This could have influenced the lower ratings with NASA-TLX as shown in Fig. 10, and indicates a strong disadvantage of textual representation especially when applied to scenarios users cannot listen to audio samples.

In contrast, according to the user feedback, the *Retrieval* approach was more beneficial for grasping the overview of the sample distribution, and users did not have to listen to many samples with the audio-to-image retrieval. Participants spent more time in lower hierarchies (Fig. 9), and eight participants commented that they could imagine where the target samples are located in the distribution. This indicates that the cluster structure was visualized in a more organized way, and users could browse the sample distribution more quickly. One of the potential reasons is that the visualization in the *Retrieval* approach is more unified and continuous than in the *Thumbnail & Spectrogram* approach, while providing more concrete information than the *Semantic* approach. The audio-to-image retrieval method is built upon a deep feature representation based on both visual and acoustic similarities, and visually similar images also tend to be close to each other in the cross-modal feature space. This improves the similarities between images inside the 2D audio embedding, and make the visualization more continuous than with thumbnail images which fully depends on individual video data. However, the obvious disadvantage of audio-to-image retrieval is that the technique itself is addressing a quite challenging task, and visualization results are not satisfactory enough even with the state-of-the-art method. The actual sound is often different from the one imagined from visualization, which gave negative impressions to users as can be seen from NASA-TLX rating and subjective feedback.

5.2 Design Implications and Future Work

Our study indicates that each visualization has both advantages and disadvantages, and it is important to combine different visualization techniques for better visualization. First, spectrograms are beneficial for the purpose of understanding details of the sound characteristics even for novice users. While spectrograms alone cannot tell the meaning of sound, it helps the visualization when jointly used with other visualization techniques to describe individual samples.

Second, visual representations like thumbnail and audio-to-image retrieval work complementary to spectrograms by conveying the semantic meaning of sound to users in more intuitive ways. Audio-to-image retrieval can provide a more unified overview of the sample distribution because of its stronger abstraction power and continuous nature in the cross-modal feature space. Such a learning-based visualization has the potential to make the annotation process more efficient by appropriating the overall sample structure in the 2D embedding. Thumbnail images are, if available, also effective to give more concrete pictures of audio data especially at lower levels of the sample hierarchy, and can compensate for the lack of accuracy in audio-to-image retrieval.

Third, on the other hand, it is possible that text representation of semantic sound categories makes the process of annotation more efficient as a supplementary information. While user-defined target classes are expected to be different from them, information about pre-defined class probabilities can be helpful to understand the semantic meaning of sound data more quickly than visual representations. It can be, for example, also used to visualize representative regions for each pre-defined sound category in the background of the sample distribution, rather than representing individual samples and clusters.

It is an important future work to investigate further possibilities to jointly use different visualization techniques. Learning-based visualization techniques also have a large design space and room for technical improvement, and various audio-to-image conversion and textual sound description techniques in general should be examined too. The optimal design can also vary according to the target application scenarios and user characteristics, and there will be required more studies focusing on specific task domains.

One of the most important application domains is accessibility, and there is a good potential of interactive, personalized sound recognition for deaf and hard-of-hearing (DHH) users. While we believe this study provides some insights such as the limitation of text representation, the optimal design for DHH users could be further different, also depending on whether the target user has congenital or acquired hearing loss. Since even in our experiments we observed some user-specific tendencies, it will be also important to investigate approaches for user adaptation.

6 CONCLUSION

In this work, we conducted a comparative study of audio visualization techniques for interactive sound recognition. Our proposed GUI visualizes the distribution of audio samples using two-dimensional feature embedding and hierarchical clustering, while each sample or cluster is represented by different visualization cues. Through the user study, we clarified the advantages and disadvantages of each visualization technique, and draw some design implications for efficient audio data visualization in interactive sound classification.

ACKNOWLEDGMENTS

This work was supported by JST CREST Grant Number JPMJCR1781, Japan.

REFERENCES

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* (2016).

[2] Philippe Aigrain, HongJiang Zhang, and Dragutin Petkovic. 1996. Content-based representation and retrieval of visual media: A state-of-the-art review. *Multimedia tools and applications* 3, 3 (1996), 179–202.

[3] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120.

[4] Saleema Amershi, Max Chickering, Steven M Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. Modeltracker: Redesigning performance analysis tools for machine learning. In *Proc. CHI*. ACM, 337–346.

[5] Saleema Amershi, Bongshin Lee, Ashish Kapoor, Ratul Mahajan, and Blaine Christian. 2011. CueT: human-guided fast and accurate network alarm triage. In *Proc. CHI*. 157–166.

[6] Relja Arandjelovic and Andrew Zisserman. 2017. Look, listen and learn. In *Proc. ICCV*. 609–617.

[7] Relja Arandjelovic and Andrew Zisserman. 2018. Objects that sound. In *Proc. ECCV*. 435–451.

[8] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2017. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932* (2017).

[9] Rolf Bardeli, Daniel Wolff, Frank Kurth, Martina Koch, K-H Tauchert, and K-H Frommolt. 2010. Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. *Pattern Recognition Letters* 31, 12 (2010), 1524–1534.

[10] Danielle Bragg, Nicholas Huynh, and Richard E Ladner. 2016. A personalizable mobile sound detector app design for deaf and hard-of-hearing users. In *Proc. ASSETS*. 3–13.

[11] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[12] Mark Cartwright, Ayanna Seals, Justin Salamon, Alex Williams, Stefanie Mikloska, Duncan MacConnell, Edith Law, Juan P Bello, and Oded Nov. 2017. Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 29.

[13] Daniel Chamberlain, Rahul Kodgule, Daniela Ganelin, Vivek Miglani, and Richard Ribón Fletcher. 2016. Application of semi-supervised deep learning to lung sound analysis. In *Proc. EMBC*. IEEE, 804–807.

[14] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu. 2017. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*. ACM, 349–357.

[15] Yin Cui, Feng Zhou, Yuanqing Lin, and Serge Belongie. 2016. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In *Proc. CVPR*. 1153–1162.

[16] Allan G de Oliveira, Thiago M Ventura, Todor D Ganchev, Josiel M de Figueiredo, Olaf Jahn, Marinez I Marques, and Karl-L Schuchmann. 2015. Bird acoustic activity detection based on morphological filtering of the spectrogram. *Applied Acoustics* 98 (2015), 34–42.

[17] Alex Endert, Patrick Fiaux, and Chris North. 2012. Semantic interaction for visual text analytics. In *Proc. CHI*. ACM, 473–482.

[18] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proc. IUI*. 39–45.

[19] James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. 2008. CueFlik: interactive concept learning in image search. In *Proc. CHI*. 29–38.

[20] Kanika Garg and Goonjan Jain. 2016. A comparative study of noise reduction techniques for automatic speech recognition systems. In *Proc. ICACCI*. IEEE, 2098–2103.

[21] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. ICASSP*. 776–780.

[22] Sébastien Gulluni, Slim Essid, Olivier Buisson, and Gaël Richard. 2011. An Interactive System for Electro-Acoustic Music Analysis. In *Proc. ISMIR*. 145–150.

[23] Martin Halvey, David Vallet, David Hannah, and Joemon M Jose. 2009. ViGOR: a grouping oriented interface for search and retrieval in video libraries. In *Proc. JCDL*. 87–96.

[24] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[25] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *Proc. ICASSP*. 131–135.

[26] Nathan Oken Hodas and Alex Endert. 2016. Adding semantic information into data models by learning domain expertise from user interaction. *arXiv preprint arXiv:1604.02935* (2016).

[27] Shih-Wen Huang, Pei-Fen Tu, Wai-Tat Fu, and Mohammad Amanzadeh. 2013. Leveraging the crowd to improve feature-sentiment analysis of user reviews. In *Proc. IUI*. ACM, 3–14.

[28] Ian Jolliffe. 2011. *Principal component analysis*. Springer.

[29] Mayank Kabra, Alice A Robie, Marta Rivera-Alba, Steven Branson, and Kristin Branson. 2013. JAABA: interactive machine learning for automatic annotation of animal behavior. *Nature methods* 10, 1 (2013), 64.

[30] Ashish Kapoor, Bongshin Lee, Desney Tan, and Eric Horvitz. 2010. Interactive optimization for steering machine classification. In *Proc. CHI*. 1343–1352.

[31] Bongjun Kim and Bryan Pardo. 2017. I-SED: An interactive sound event detector. In *Proc. IUI*. 553–557.

[32] Dongge Li, Nevenka Dimitrova, Mingkun Li, and Ishwar K Sethi. 2003. Multimedia content processing through cross-modal association. In *Proceedings of the eleventh ACM international conference on Multimedia*. ACM, 604–611.

[33] Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE transactions on information theory* 28, 2 (1982), 129–137.

[34] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[35] Matthias Mielke and Rainer Brueck. 2015. Design and evaluation of a smartphone application for non-speech sound awareness for people with hearing loss. In *Proc. EMBC*. 5008–5011.

[36] Meg Pirrung, Nathan Hilliard, Artëm Yankov, Nancy O'Brien, Paul Weidert, Courtney D Corley, and Nathan O Hodas. 2018. Sharkzor: Interactive Deep Learning for Image Triage, Sort and Summary. *arXiv preprint arXiv:1802.05316* (2018).

[37] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 251–260.

[38] Ehsan Sherkat, Seyednaser Nourashrafeddin, Evangelos E Milios, and Rosane Minghim. 2018. Interactive Document Clustering Revisited: A Visual Analytics Approach. In *Proc. IUI*. ACM, 281–292.

[39] Zhao Shuyang, Toni Heittola, and Tuomas Virtanen. 2017. Active learning for sound event classification by clustering unlabeled data. In *Proc. ICASSP*. 751–755.

[40] Jason Smith, Dillon Weeks, Mikhail Jacob, Jason Freeman, and Brian Magerko. 2019. Towards a Hybrid Recommendation System for a Sound Library.. In *IUI Workshops*.

[41] Kyle A Swiston and Daniel J Mennill. 2009. Comparison of manual and automated methods for identifying target sounds in audio recordings of Pileated, Pale-billed, and putative Ivory-billed woodpeckers. *Journal of Field Ornithology* 80, 1 (2009), 42–50.

[42] Justin Talbot, Bongshin Lee, Ashish Kapoor, and Desney S Tan. 2009. EnsembleMatrix: interactive visualization to support machine learning with multiple classifiers. In *Proc. CHI*. 1283–1292.

[43] Byron C Wallace, Kevin Small, Carla E Brodley, Joseph Lau, and Thomas A Trikalinos. 2012. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. In *proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. ACM, 819–824.

[44] Chia-Hung Wan, Shun-Po Chuang, and Hung-Yi Lee. 2019. Towards Audio to Scene Image Synthesis Using Generative Adversarial Network. In *Proc. ICASSP*. IEEE, 496–500.

[45] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. 2016. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215* (2016).

[46] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. 2015. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 1 (2015), 7–19.

[47] Kateřina Žmolíková, Marc Delcroix, Keisuke Kinoshita, Takuya Higuchi, Atsunori Ogawa, and Tomohiro Nakatani. 2017. Speaker-Aware Neural Network Based Beamformer for Speaker Extraction in Speech Mixtures. In *Proc. Interspeech 2017*. 2655–2659.