Sonification of 3D Object Shape for Sensory Substitution: An Empirical Exploration

Torkan Gholamalizadeh, Hossein Pourghaemi, Ahmad Mhaish, Gökhan İnce and Damien Jade Duff
Faculty of Computer & Informatics Engineering
Istanbul Technical University
Email: djduff@itu.edu.tr,gokhan.ince@itu.edu.tr

***Abstract*—Different approaches to sonification of 3D objects as part of a sensory substitution system are experimentally investigated. The sensory substitution system takes 3D point clouds of objects obtained from a depth camera and presents them to a user as spatial audio. Two approaches to shape sonification are presented and their characteristics investigated. The first approach directly encodes the contours belonging to the object in the image as sound waveforms. The second approach categorizes the object according to its 3D surface properties as encapsulated in the rotation invariant Fast Point Feature Histogram (FPFH), and each category is represented by a different synthesized musical instrument. Object identification experiments are done with human users to evaluate the ability of each encoding to transmit object identity to a user. Each of these approaches has its disadvantages. Although the FPFH approach is more invariant to object pose and contains more information about the object, it lacks generality because of the intermediate recognition step. On the other hand, since contour-based approach has no information about depth and curvature of objects, it fails in identifying different objects with similar silhouettes. On the task of distinguishing between 10 different 3D shapes, the FPFH approach produced more accurate responses. However, the fact that it is a direct encoding means that the contour-based approach is more likely to scale up to a wider variety of shapes.**

Keywords–Sensory substitution; sensory augmentation; point clouds; depth cameras; sound synthesis.

I. INTRODUCTION

Sensory substitution is the use of technology to replace one sensory modality with another. In visual-to-audio sensory substitution, visual information captured by a camera is presented to users as sound. Such systems promise help for the sight-impaired: imagine users navigating using space/obstacle information, grasping novel objects, eating meals with utensils, and so forth. By not falling into the trap of many artificial intelligence-based assistive systems of aggressively abstracting the data provided to users, user agency is preserved and the user's own advanced cognitive data processing capabilities are leveraged. Sensory substitution systems also provide interesting platforms for exploring synaesthesia and cross-modal sensory processing [1].

Recent work in utilizing depth cameras for sensory substitution promises to increase the usefulness of such visual-to-audio sensory substitution systems [2][3]. Mhaish et al.'s system [2] uses a 3D depth camera to create point clouds characterizing the surfaces of objects in a scene and presents those surfaces to a user using spatial audio. See Figure 1 for a summary of the information flow in that approach. The present work extends that system, offering an investigation of different ways of encoding 3D spatial surfaces as audio, an area ripe for exploration in the context of sensory substitution systems.

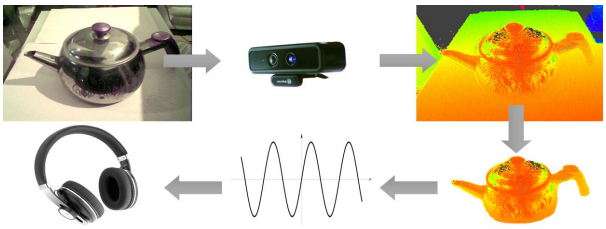

Figure 1. The flow of data in the full sensory substitution system, from real-world objects, via the depth camera, to a point cloud, and a segmented tracked object, and finally a sound waveform played to a user.

Figure 2. Left: the system being used "in the wild". Right: physical set-up of the experiments described in this paper.

Broadly speaking, the process of encoding information as (non-speech) sounds is called sonification. Sonification is used in applications such as medical imaging where it is used for example to differentiate a healthy brain from an unhealthy one, geological activity detection, and so forth. In the present paper, we present two approaches to sonification of object shape. Object shape is particularly important in providing functional information about objects, particularly for blind users who may wish to perceive objects in their environment in order to recognize them, avoid them, or manipulate them. The first approach to shape sonification described in the present paper is encoding based on 2D object contours and the second is based on a 3D object recognition descriptor called the Fast Point Feature Histogram.

The rest of this work will be presented in four sections. In the next section, a short summary of related work is presented. In Section 3, an overview of the sensory substitution system is given and two different sound generation approaches are explained in detail. Then in Section 4, details of experiments and results are shown. Finally, results are summarized and further work discussed in Section 5.

 ISBN: 978-1-61208-538-8

II. Related Work

Sensory substitution systems are systems that map visual information to audio in an attempt to create an effect like vision but channeled through a different sense. More broadly, systems that map any kind of information (including visual and graphical) to audio are called sonification systems. In general, there are two kinds of sonification systems, high-level and low-level sonification systems, where the high-level approaches are designed to convert information to speech. A significant subset of these systems are text-to-speech applications, widely used for visually impaired people. Examples include VoiceOver and JAWS [1]. In addition to text-to-speech applications there are some other high-level sonification systems which are more complex and can detect objects and identify them and return their names in real-time, like LookTel [4] and Microsoft Seeing AI project [5] that can read texts, describe people and identify their emotions. These high-level sonification systems are easy for users and do not require training, but they can fail in sonifying complex environment or shapes for which the system has not been adapted.

On the other hand, low-level sonification systems generate sound directly based on visual information. The main difference between these systems and high-level sonification systems is that users need to be trained before using these systems to be able to understand the relation between the generated sounds and properties of observed objects. Though these kinds of systems can seem difficult to use, they can be more flexible for new environments and undefined objects because they produce sounds based on characteristics directly calculated from input data [6]. One of the most well-known systems of this group is the sensory substitution system The vOICe [1], which uses the gray-level image of the scene and scans the image from left to right and generates and sums audible frequencies based on pixels' location with amplitude based on pixels' intensity. The disadvantages of the vOICe system are that it requires 1 second to scan the image. Further, the image-sound mapping is somewhat abstract if used with depth images without adaptation and does not explore physical or metaphorical synergies with shape in particular. However, our proposed system is conceived as a system for generating spatial audio generated based on surface and shape information for helping users to localize objects and identify them in real-time.

Systems closest to our own include the electro-tactile stereo-based navigation system ENVS of [7] with ten channels of depth information calculated from stereo transmitted to ten fingers, which focuses on navigation but not shape understanding and uses the tactile pathway, and the depth-camera visual-to-audio based sensory substitution system See ColOr of [3] which, though using depth-cameras, concentrates on bringing color (and not space or shape) to blind users by mapping different intervals of hue and value to instruments like violin, trumpet, piano etc. Conversely, finding a proper method for mapping shape information to audio is a vital step in many low-level sonification systems. In this area, the work of Shelley et el. [8] is close to the proposed system, focusing on sonification of shape and curvature of 3D objects in an augmented reality environment as part of the SATIN project, where the user of the system is able to touch and alter the 3D objects using the visual-haptic interface of the system. In that article, object cross sections (and associated curvature) are used to modulate the frequency of a carrier signal or the parameters of physical sound generation [8].

As discussed above choosing a good approach to sonification plays an important role in achieving good performance of low-level sonification systems. Therefore in the current work, two different sound generation methods are provided for Mhaish et al.'s [2] system and their accuracies are measured on the task of synthetic 3D object identification. The idea of using synthetic objects instead of real objects is to evaluate the performance of different sound generation methods isolated from the performance of other components and environmental noise. In future work, the best approach or mix of these approaches will be applied in the identification of real objects.

III. Technical Details

Output from a head-mounted depth camera (DepthSense 325 or ASUS Xtion) is converted to a head-centred point cloud, which is segmented by curvature and point-distance in real-time [9] into surface primitives. These surface primitives are tracked using simple data association, selected using size and closeness criteria, and presented to the user as spatially-located audio (played using a wrapper around the spatial audio library OpenAL [10], the wrapper taking care of time tracking and interpolation).

Heavy use of the Point Cloud Library [11] is made in the point-cloud processing steps and particular care is made to keep processing of point clouds at 15+ frames per second so as to provide responsive sensory feedback to user probing motions. An illustration of the system being tested can be found in the left picture of Figure 2.

Note that this system is designed to segment surface primitives rather than objects. Although for some applications, such as tabletop object manipulation, short-cuts can be taken to extract separate objects, general object segmentation is an unsolved problem. Since the current paper is focused on sonification (making sounds to represent data), the focus here is on the sonification of whole but mostly simple objects.

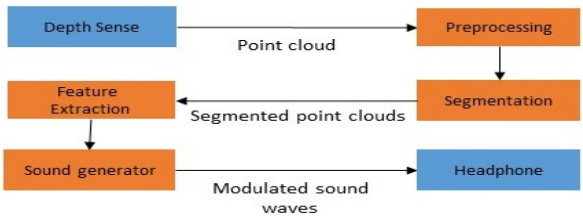

Figure 3. Software level of data flow diagram.

Before going into detail about the approaches used to sonify shape, the processing steps used by the system to extract visual information and process it to audio will be explained. Figure 3 shows the data flow architecture of the system. The steps in the architecture are further explained as follow:

1) Preprocessing: RGB and depth information produced by the time of flight or structured light camera is passed to a preprocessing step in the form of a point cloud and in this step normals are calculated from the point cloud "organized" in a 2D array of points, using a real time integral image algorithm supplied by Holzer et al [12].

 ISBN: 978-1-61208-538-8

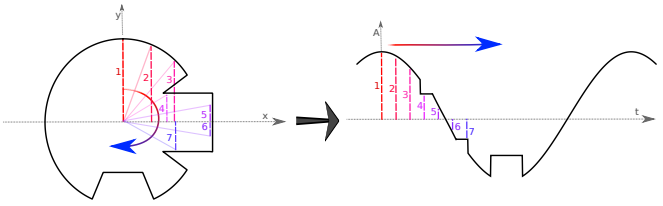

Figure 4. Rotary-contour-based encoding. **Left:** Original object contour in x-y image space. **Right:** Resulting waveform as a plot of amplitude(A) against time(t).

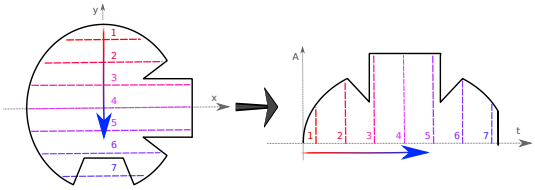

Figure 5. Vertical-contour-based encoding. **Left:** original object contour. **Right:** the resulting waveform as amplitude (A) against time (t).

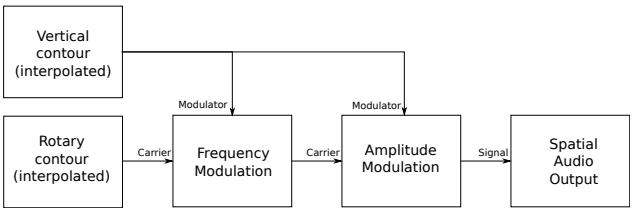

Figure 6. A simplified diagram of the direct encoding of an object contour as sound.

2) Segmentation: The 2D organized point cloud is then segmented by the method of Trevor et al [9] and segments are obtained characterized by slowly changing surface normal vectors and no intervening gaps. Further processing can be applied to find and remove tabletop surfaces for tabletop scenarios.

3) Feature Extraction: Information to characterize the acquired segments is extracted. In the current work, contour-based and FPFH-based approaches are explored.

4) Sound Generation: In the system presented by Mhaish et al. [2], a simple sonification approach was proposed based on a conversion of principle object dimensions to frequencies. A circular buffer is used to create, update and interpolate sound waves and their envelopes and the rate at which frames are arriving is estimated in order to send the appropriate number of samples to the OpenAL spatial sound system. In the current system, FPFH signatures are converted via a recognition step to different instruments from the STK simulation toolkit [13] and object contours are converted via interpolation and modulation data processing steps to sound waves.

The present paper focuses on the feature extraction and sound generation steps, proposing the contour- and FPFH-based approaches, explained in the next sections.

The sound is played using the OpenAl library which is provided as many samples as necessary from the filled circular buffer and generates spatial-audio based on binaural cues or, alternatively, Head Related Transform Functions (HRTFs).

A. *Contour-based sonification*

In the contour-based encoding, object contours are translated directly into auditory waveforms, and frequency and amplitude modulations of waveforms.

In the variation on this idea tested in this paper, the rotary-contour is extracted from the object and used to generate a carrier signal. In the rotary-contour-based encoding, a path is traced out around the contour of the object and the distance of each contour point from the object's horizontal axis (defined by the centroid of the points in the object) becomes an instantaneous amplitude in the sound waveform (normalized to fit within the range of acceptable sample amplitudes). Spherical or circular objects thus translate perfectly into sinusoidal waveforms. For instance, the object on the left side of Figure 4 becomes the waveform plot (amplitude vs time) on the right hand side, with radial distances converted into instantaneous amplitudes which are then potentially interpolated.

Because the signal waveform depends on the object contour, some timbre properties also depend on the object contour. The carrier signal is then modulated at a slower (consciously perceivable) time-scale using frequency and amplitude modulation by another time-varying function which we call here vertical-contour-based encoding. In this kind of encoding, which is illustrated in Figure 5, the top to bottom scanned width of the object silhouette is converted to the amplitude of a modulating signal which is then applied to the carrier signal as frequency and amplitude modulation. Thus, multiple perceptual channels are used to transfer information to the user. For a sketch of the signal processing flow used to generate the resulting waveform, see Figure 6.

The contour-based approach is motivated both by the conceptual clarity of the mapping, but also by the fact that sounds already arise as vibrations in objects and spaces, and travel through the objects, reflecting in the resulting waveforms the shape and size of these spaces; thus, the method, depending on the exact encoding used, has an analogue in the physics of real sound generation and consequently natural synergies with perception.

B. *FPFH-based sonification*

The FPFH is a feature extracted from point clouds or point cloud parts, designed for representing information about the shape of the cloud that is invariant to rotation. It is comparable to a histogram of curvatures measured in different ways across the object surface.

FPFH is a 33-bin histogram extracted from the points and normals in the point cloud. This histogram counts 3 different curvature measures with 11 bins for each measure. The relative position and surface-normal vector of each point is processed and the bin into which the point falls for each of the 3 dimensions incremented [14]. Note: the FPFH is not originally designed as a full object descriptor but it has proved sufficient for current purposes: other more or less viewpoint-invariant or object-global descriptors can also be easily adapted to this purpose. Examples of FPFH descriptors extracted from point clouds used in experiments in this paper can be found in Figure 7. As can be seen in the figure, different shapes generally correspond to different histograms and different sizes

 ISBN: 978-1-61208-538-8

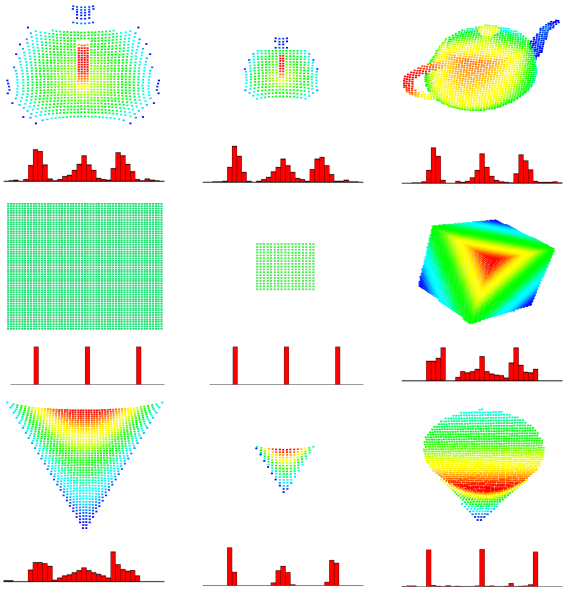

Figure 7. Sample point cloud views with normalized histogram shapes (FPFH). **Top row:** teapot. **Middle row:** cube. **Bottom row:** cone. **Left:** baseline view. **Middle:** a different object size. **Right:** a different view direction.

TABLE I. THE OBJECT-INSTRUMENT MAPPING IN THE FPFH-BASED APPROACH.

Object	Instruments
Teapot (Tp)	Shakers
Cube (Cb)	Struck Bow
Cuboid (Cd)	Drawn Bow
Cylinder (Cl)	High Flute
Cone (Cn)	Plucked String
Elipsoid (El)	Hammond-style Organ
Icosahedron (Ic)	Saxophone
Stretched Cylinder (SC)	Low Flute
Sphere (Sp)	Clarinet
Torus (Ts)	Sitar

and scales generally do not affect the histograms radically. However, the external contours do not always affect the result, as can be seen by comparing the bottom view of the cone and the side view of the cube.

After an object is encoded using FPFH, a database of existing FPFH descriptors is searched (using an indexing KD-tree) for the closest descriptor and the resulting object label retrieved. A mapping (Table I) is provided from object label to instrument type and the relevant instrument is synthesized using the Synthesis ToolKit (STK) [13].

3D object recognition techniques are attractive for the current application since the field of robotic vision has well-established approaches, and many descriptors are available for representing shape, having rotational invariance built in for example [14]. Moreover, synthetic instrument models provide highly discriminable sounds, which can support a sound-object mapping approach to the task under consideration.

IV. Training and Experiment

To evaluate the ability of the encodings discussed above to transmit shape as sound, the ability of users to identify objects under changing conditions was investigated. For a

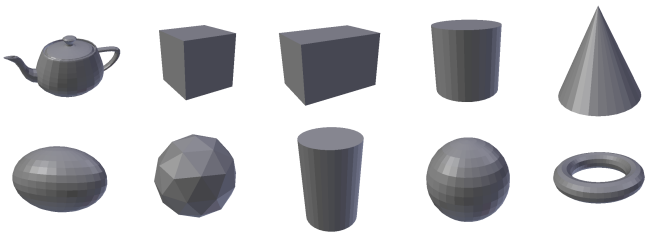

Figure 8. The set of objects used in experiments. Top: Teapot (Tp), Cube (Cb), Cuboid (Cd), Cylinder (Cl), Cone (Cn), Bottom: Elipsoid (El), Icosahedron (Ic), Stretched Cylinder (SC), Sphere (Sp), Torus (Ts).

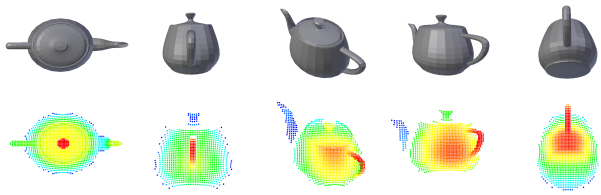

Figure 9. The set of poses used in the pose-varying experiment. **Top**: the object model as seen from different viewpoints. **Bottom**: the point cloud resulting from each viewpoint. Colour in the point cloud represents normalized distance from the camera of the points.

clear evaluation of the relationship between shape and sound, point clouds presented to the user were based on point-cloud samplings of views of the ten object meshes shown in Figure 8. Performance of the proposed encodings was measured by conducting two experiments. In the first experiment, the location from which objects are viewed was varied among five different equi-distant viewpoints, illustrated in Figure 9, and in the second experiment, five different scales of objects were presented, scale here stands for either size or viewing distance but only in the context of the encodings used in the present paper - not all point cloud encodings will confuse size and distance. The five sizes used are shown in Figure 10.

The main idea of choosing these two experiments is that these parameters are the most changing parameters in wild. Other possible parameters include lighting conditions, but our cameras use active lighting, or material properties, but these depend on the particular choice of depth-sensing device, to which our approach is designed to be mostly agnostic.

A. Training session

Each experiment comprises two conditions, the FPFH condition and the contour-based condition, presented to the individuals in a random order. For each condition, an independent training session was conducted. A single training session takes 15 minutes, including 2-3 minutes for describing the principles of the system followed by a free experimentation period. During the training sessions participants were given the ability to play sounds for all five different viewpoints of

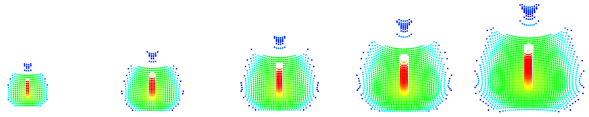

Figure 10. The set of object sizes used in the size-varying experiment.

 ISBN: 978-1-61208-538-8

TABLE II. THE VALUES OF RELEVANT PARAMETERS USED IN THESE EXPERIMENTS.

General parameters	
Input point cloud size (width × height)	320 × 240
Object distances to camera (metres)	3
Sound generation: Sample rate (Hz)	44100
Sound generation: Bits per sample	16
Contour-based approach	
Pixel-sample ratio: rotary-contour carrier	1:1
Pixel-sample ratio: vertical-contour modulator	600:1
Peak frequency deviation proportion: modulator	1.0
Peak amplitude deviation proportion: modulator	1.0
FPFH-based approach	
FPFH database size (num object views)	100
FPFH database size (num objects)	10

all ten objects for the viewpoint experiment and for all five scales for the scale/size experiment. Users were allowed to play with the system, choosing objects and viewpoints/sizes from the training set and playing the sounds as well as viewing a 3D visual representation from that viewpoint/size and they were asked to remember the sounds related to each object.

B. Experimental session

Experiments were conducted with 16 non-disabled participants with an average age of 24 years, divided into two groups with each group containing both male and female participants. Participants of each group performed only either the viewpoint or the scale experiment. Both sound generation approaches were tested with each participant. In experimental sessions, sounds of randomly selected objects with randomly selected viewpoints/sizes from the training set were presented and participants were asked to identify the objects. In these sessions, for each approach 30 trials were conducted with each participant and the participant was informed after each trial whether the answer was correct, and when the answer was wrong, the experimenter informed the participant the actual object identity. Answers were recorded in confusion matrices. In these experiments, users were not supposed to guess the viewpoints or scales; they were asked only to identify the objects based on the sound that they were hearing. The physical set-up of the experiment can be found on the right of Figure 2. Parameters of the system used in experiments are shown in Table II.

C. Results

The complementary properties of the two methods tested can be observed in the confusion matrices in Tables IV and III. In these matrices, numbers in cells represent the number of times the object in the row header was identified by participants as the object in the column header. Zero values are left blank. The inability of the contour-based approach to take account of the depth information in the interior of an object leading it to confuse objects with similar silhouettes, as it can be seen in the tables.

With an overall accuracy of 60% and 57% on the two experiments vs. 36% and 42% for the contour-based method, the FPFH approach performed better (verified with χ^2 tests, which are applicable when class sizes are balanced, 1 D.O.F., $p = 0.01$). However, the FPFH-based approach was still unable to account for the contour on the silhouette of an object, leading it to confuse objects with similar visible curvatures.

TABLE III. RESULTS OF SIZE-VARYING EXPERIMENT. (BOLD NUMBERS SHOWS THE MOST CHOSEN OBJECT(S) BY PARTICIPANTS, WHEN THE OBJECT IN THE ROW HEADER WAS PRESENTED).

Contour-based

Actual \ Response	Tp	Cb	Cd	Cl	Cn	El	Ic	SC	Sp	Ts
TP	**33.3**				11.1	11.1	11.1	11.1		22.2
Cb		**33.3**	25.0	8.3	8.3			8.3		16.6
Cd		16.6	**50.0**							16.6
Cl	6.6	13.3	13.3	**33.3**	6.6			13.3	13.3	
Cn				10.0	**70.0**	10.0			10.0	
El	11.1		11.1	11.1		11.1	**33.3**		11.1	11.1
Ic	15.3			15.3	**23.0**		7.6	7.6	**23.0**	7.6
SC					11.1	11.1	22.2	**55.5**		
Sp	25.0							25.0	**50.0**	
Ts						33.3				**66.6**

FPFH-based

Actual \ Response	Tp	Cb	Cd	Cl	Cn	El	Ic	SC	Sp	Ts
TP	**100.0**									
Cb		**76.9**		7.6					15.3	
Cd		21.4	**64.2**	14.2						
Cl		16.6		33.3			**50.0**			
Cn		11.1	5.5	5.5	**55.5**		5.5	5.5	11.1	
El	7.6			7.6		15.3	7.6	15.3	**46.1**	
Ic	6.6			**33.3**	6.6		26.6	6.6	13.3	
SC				37.5			6.2	**56.2**		6.6
Sp		16.6		33.3					**50.0**	
Ts							11.1			**88.8**

TABLE IV. RESULTS OF VIEWPOINT-VARYING EXPERIMENT. (BOLD NUMBERS SHOWS THE MOST CHOSEN OBJECT(S) BY PARTICIPANTS, WHEN THE OBJECT IN THE ROW HEADER WAS PRESENTED).

Contour-based

Actual \ Response	Tp	Cb	Cd	Cl	Cn	El	Ic	SC	Sp	Ts
TP	**56.2**			6.2	25.0		6.2			6.2
Cb		**29.4**	17.6	17.6			7.6			7.6
Cd			**69.2**	7.6	7.6		7.6			7.6
Cl	9.2	**28.5**	14.2	14.2	4.7	9.2		9.2	4.7	4.7
Cn	9.0	9.0	4.5	4.5	**36.3**	9.0	4.5		18.1	4.5
El		8.3		8.3	8.3	**41.6**		8.3		25.0
Ic	6.6			**20.0**	**20.0**	**20.0**	**20.0**		13.3	
SC			**50.0**			25.0		25.0		
Sp	5.2	5.2		5.2	15.7	10.4	10.4		**36.8**	10.4
Ts			11.1			22.2	11.1		22.2	**33.3**

FPFH-based

Actual \ Response	Tp	Cb	Cd	Cl	Cn	El	Ic	SC	Sp	Ts
TP	**100.0**									
Cb		**37.5**	**37.5**					25.0		
Cd		23.5	**52.9**		11.7			11.7		
Cl		15.3		**53.8**		7.6	15.3	7.6		
Cn		**36.3**		9.0	**36.3**		9.0		8.0	
El		12.5		12.5		**50.0**	12.5		12.5	
Ic			28.5		14.2		**57.1**			
SC		5.5	11.1	11.1				**66.6**		5.5
Sp		9.0		9.0	9.0	18.1			**54.5**	
Ts					7.1					**92.8**

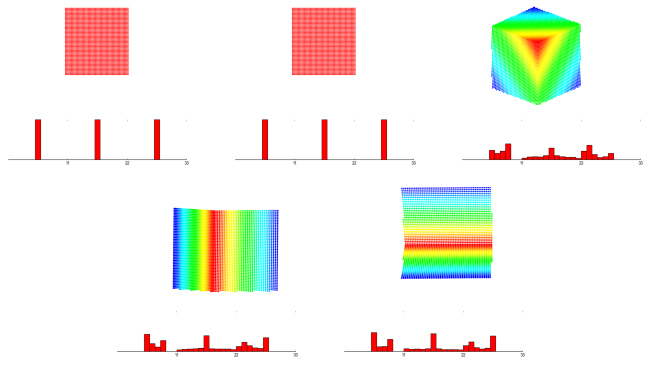

Figure 11. Five different view points of cube and their histograms.**Top:** Left: Top view. Middle: Frontal view. Right: Right front top corner view. **Bottom:** Left: Right Front edge view. Right: Front bottom edge view.

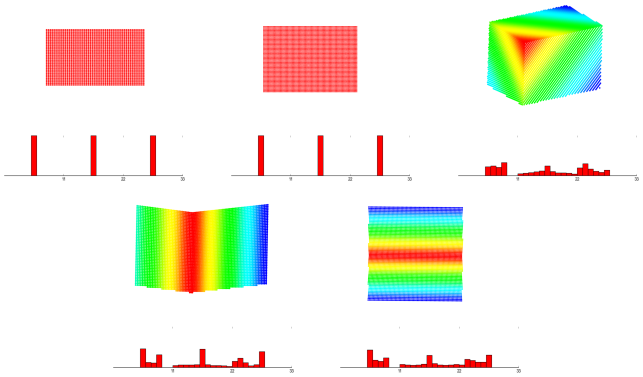

Figure 12. Five different view points of cuboid and their histograms.**Top:** Left: Top view. Middle: Frontal view. Right: Right front top corner view. **Bottom:** Left: Right Front edge view. Right: Front bottom edge view.

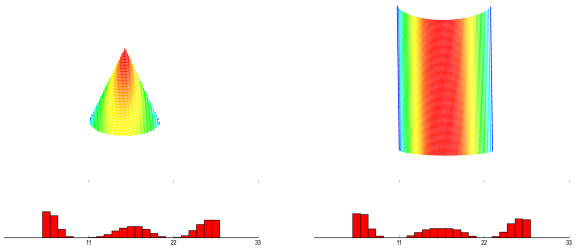

Figure 13. Similarity in histograms (FPFH) causes the system to mis-classify the objects. **Left:** FPFH for third view of cone. **Right:** FPFH of largest stretched cylinder(with second viewpoint)

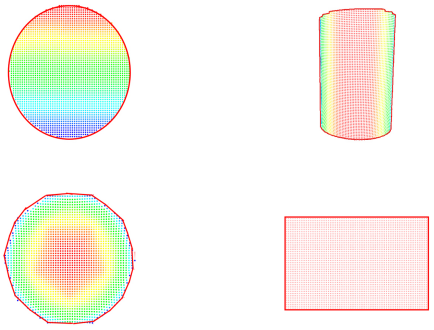

Figure 14. Example of objects contours for which the proposed contour based approach can not generate sufficiently distinguishable sounds(red lines around the objects represent objects contours). **Left column:** Top: sphere contour. Bottom: icosahedron contour. **Right column:** Top: cylinder(front view). Bottom: rectangle(front view)

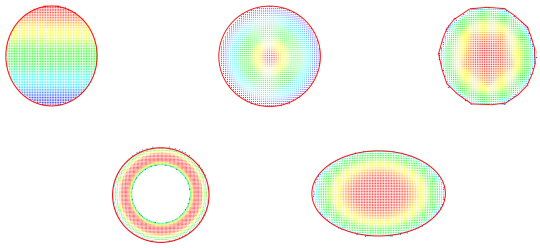

Figure 15. Similar contours of multiple objects used in experiment(red lines around the objects represent objects contours). **Top:** Left: sphere, Middle: cone, Right: icosahedron. **Bottom:** Left:torus Right: ellipsoid

For the FPFH-based approach, the natural user strategy to the identification problem is to learn a sound-object mapping. This worked as long as the system could find the correct mapping, but the system itself did not always use all available information. For instance, the flat bottom of a cone and cube or cuboid produce the same FPFH signature - see Figure 7. The same confusion occurred between cube and cuboid. Participants using the system frequently misclassified cube as cuboid in all the of 5 viewpoints of cube, as is shown in Figure 11 and Figure 12. The FPFH of the cube is so similar to cuboid as to cause the system to mis-classify the cube. This is sufficient to explain why in Table IV the cube is classified as a cuboid almost as much as it is a cube. The lack of distinguishability of FPFH signatures between the top view of the cylinder, the stretched cylinder and the cube is apparent because they all have a single flat surface visible. There are also some unexpected confusions such as the recognition system itself wrongly identifying third view of cone as largest scale of stretched-cylinder (see Figure 13). Since the frequency of occurrence of this confusion was low, participants were able to hear the related sound for the cone more frequently, so it did not affect their performance and they could treat the second sound as noise.

There were some objects that the system did not have any difficulty in identifying, such as the icosahedron, ellipsoid, sphere, teapot and torus. However, their classification accuracy varies from one-in-two to near-perfect. For example, ellipsoid, sphere and icosahedron are correctly identified in 50.0%, 54.5% and 57.1% of trials, while teapot and torus were identified perfectly (100% for teapot and 92.8% for torus-see Table IV). This high difference in confusion rates is due to the choice of instrument corresponding to each object, a fact that was mentioned by most of the participants during the experimental session. They believed that identifying the teapot and torus was easy because their sounds (shakers and sitar) are more distinct than the others. Hence, putting similar sound for shapes that are geometrically similar to each other may not be a good idea or work should be done to ensure that instruments are more distinguishable from each other. However, using dissimilar instruments for similar shapes can defeat any attempt to sonify subtle differences in shape.

In the contour-based approach, it was also observed that some participants preferred the abstract learning strategy of learning identity-sound associations rather than understanding the sound-shape mapping representation as well. For this approach, participants reported that some important object properties were not available to them, leading them to confuse objects like the sphere with the icosahedron or the front view of the cuboid with the same view of the cylinder (see Figure 14). For these two pairs of objects, the output of the system does not produce exactly the same result but similar results which makes it hard for users to distinguish them from each other and they need to put in more effort to understand the differences. However, it should be noted at this point that visually similar objects should be expected in any successful system to pose a larger challenge. Moreover, as discussed before, this approach is viewpoint-variant and for some viewpoints of different objects which have similar contours, it generates identical or

 ISBN: 978-1-61208-538-8

too-similar sounds which causes the user to choose the wrong object. For instance, as shown in Figure 15, the top view of cylinder, cone, torus and sphere all have a circular contour which makes their sounds exactly the same.

V. Conclusion and Future Work

Two different encodings of 3D shape into sound were presented, i.e., *contour*-based and *FPFH*-based. The contour-based approach presented maps directly from shape to sound. This is an advantage in that any new object can be represented in sound, and that similarly shaped objects produce similar sounds. However, the encoding attempted here only transmits the image-contour of the object and is not robust to viewpoint. Some participants also preferred to learn the abstract object mapping, suggesting that work is needed on making this approach more intuitive when it comes to the relationship between shape and sound.

The FPFH-based approach solves these problems by using data about the full 3D object shape and by representing features that are somewhat invariant to viewpoint (though only to the extent that surfaces are visible). The FPFH-based approach also has the advantage when creating distinguishable sounds of using a mature sound-synthesis system with highly recognizable objects. However, again, the use of discrete instruments reduces flexibility in encoding different shape properties. In order to make the system work, object exemplars must be paired with sounds, restricting the generalizability of the system to new objects and abstracting some of the user's agency, not fully utilizing their cognitive capacity.

The next step in this work is to extend these approaches to reduce the above-mentioned limitations. In the case of the contour-based approach, a more sophisticated encoding is needed, that takes into account 3D aspects of the object. Adding some viewpoint invariance may be desirable, but it would be a subject of empirical investigation as to whether this viewpoint invariance would actually be helpful when considering other tasks that users might want to do with objects, such as manipulation, in which users need to perceive also the orientation of the object. In the case of the FPFH-based approach, a way is needed of generalizing from the exemplars in an appropriate way, for example by using machine learning techniques in conjunction with user input. Other point cloud features with different properties also should be systematically investigated.

Further work also involves testing these sonifications "in the wild" and with multiple objects, which will require work on more aggressive noise elimination and object (or surface primitive) tracking. In both approaches, it is important to exploit and extend the intuitive mappings from shape to sound whose exploration was begun here, for quick learning and application of the system, as well as for recruiting the advanced cognitive capabilities of users.

Acknowledgments

This work was supported by the Scientific and Technological Research Council of Turkey (TÜBITAK), Project No 114E443.

References

[1] M. Auvray, S. Hanneton, and J. K. O'Regan, "Learning to perceive with a visuo-auditory substitution system: Localisation and object recognition with 'The vOICe'," Perception, vol. 36, no. 3, 2007, pp. 416 – 430, URL:http://www.perceptionweb.com/abstract.cgi?id=p5631 [retrieved: 2017-02-04].

[2] A. Mhaish, T. Gholamalizadeh, G. İnce, and D. Duff, "Assessment of a visual-to-spatial audio sensory substitution system," in Signal Processing and Communications Applications(SIU). Zonguldak, Turkey: IEEE, May 2016 24th, pp. 245–248, URL:http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7495723 [retrieved:2017-02-04].

[3] J. D. Gomez Valencia, "A computer-vision based sensory substitution device for the visually impaired (See ColOr)," Ph.D. dissertation, University of Geneva, 2014, URL:http://archive-ouverte.unige.ch/unige:34568 [retrieved:2017-02-04].

[4] J. Sudol, O. Dialameh, C. Blanchard, and T. Dorcey, "Looktel—A comprehensive platform for computer-aided visual assistance," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2010, pp. 73–80, URL:http://ieeexplore.ieee.org/abstract/document/5543725/ [retrieved:2017-02-01].

[5] "Microsoft," 2016, URL: https://news.microsoft.com/videos/microsoft-cognitive-services-introducing-the-seeing-ai-app/#sm.00001gckhn7usbfpbyeceij35wz8i#570rbWhUTwooz7X5.97 [accessed: 2017-02-02].

[6] T. Yoshida, K. M. Kitani, H. Koike, S. Belongie, and K. Schlei, "EdgeSonic: image feature sonification for the visually impaired," in Proceedings of the 2nd Augmented Human International Conference. ACM, 2011, p. 11, URL:http://dl.acm.org/citation.cfm?id=1959837 [retrieved:2017-02-01].

[7] S. Meers and K. Ward, "A vision system for providing 3d perception of the environment via transcutaneous electro-neural stimulation," in International Conference on Information Visualisation. London, UK: IEEE, Jul. 2004, pp. 546–552, URL:http://ieeexplore.ieee.org/abstract/document/1320198/ [retrieved:2017-02-02].

[8] S. Shelley, M. Alonso, J. Hollowood, M. Pettitt, S. Sharples, D. Hermes, and A. Kohlrausch, "Interactive Sonification of Curve Shape and Curvature Data," in Haptic and Audio Interaction Design. Springer, Berlin, Heidelberg, Sep. 2009, pp. 51–60, DOI: 10.1007/978-3-642-04076-4_6, URL:http://link.springer.com/chapter/10.1007/978-3-642-04076-4_6 [retrieved:2017-02-01].

[9] A. J. Trevor, S. Gedikli, R. B. Rusu, and H. I. Christensen, "Efficient organized point cloud segmentation with connected components," Semantic Perception Mapping and Exploration (SPME), 2013, URL:https://pdfs.semanticscholar.org/c96b/ad70db489701ade007b365fe215478303003.pdf [retrieved: 2017-02-04].

[10] G. Hiebert, "Openal 1.1 specification and reference," 2005, URL:http://www.citeulike.org/group/12573/article/6483595 [retrieved: 2017-02-04].

[11] R. B. Rusu and S. Cousins, "3d is here: Point cloud library (PCL)," in IEEE Intl. Conf. on Robotics & Automation. Shanghai, China: IEEE, 2011, pp. 1–4, URL:http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5980567 [retrieved:2017-02-04].

[12] S. Holzer, R. B. Rusu, M. Dixon, S. Gedikli, and N. Navab, "Adaptive neighborhood selection for real-time surface normal estimation from organized point cloud data using integral images," in Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on. IEEE, 2012, pp. 2684–2689, URL:http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6385999 [retrieved:2017-01-20].

[13] G. P. Scavone and P. R. Cook, "RtMidi, RtAudio, and a synthesis toolkit (STK) update," Synthesis, 2004, URL:http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.143.997&rep=rep1&type=pdf [retrieved: 2017-02-04].

[14] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (FPFH) for 3d registration," in IEEE Intl. Conf. on Robotics & Automation. Kobe, Japan: IEEE, 2009, URL:http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5152473 [retrieved: 2017-02-04].

 ISBN: 978-1-61208-538-8