Görmeyi Uzaysal İşitme Duyusu ile İkame Eden Bir Sistemin Değerlendirilmesi
Assessment of a Visual to Spatial-Audio Sensory Substitution System

Ahmad Mhaish, Torkan Gholamalizadeh, Gökhan İnce, Damien Jade Duff
Department of Computer Engineering, Istanbul Technical University, Istanbul, Turkey
mhaish@itu.edu.tr, djduff@itu.edu.tr

***Özetçe* —Duyuların ikamesi, görme gibi tek bir kipteki duyusal bilginin işitme gibi bir başka kipteki bilgi ile bir birey tarafından özümsenmesi tekniğidir. Bu bildiri derinlik sensörü kullanılarak uzaklık verisinden oluşan bir diziyi konumsal ses bilgisine gerçek zamanlı olarak çevirerek, bir konumsal-ses duyu ikamesi sistemi yaratmaktadır. Deneylerde, katılımcılar duyu ikamesi sistemine ait sensörü gözlerinin önünde tutarak bir masaya oturmuş pozisyonda önce sistem ile eğitilmişler, daha sonra da gözleri bağlı olarak test edilmişlerdir. Deneylerde katılımcılardan hedef nesnenin masanın neresine konulduğunu iki adımda bulmaları istenmiştir: birinci olarak yönü, ikinci olarak da mesafeyi kestirerek. Katılımcılar önerilen sistemi kullanarak nesnenin yönünü bulmada yüksek bir doğruluk oranı (%90) ve mesafesini bulmada da %56'lık bir başarım göstermiştir.**

Anahtar Kelimeler*—*Yardımcı uygulamalar, binoral işaretler, nokta bulutları, insan bilgisayar etkileşimi, ses sentezi.

***Abstract*—Sensory substitution is a technique whereby sensory information in one modality such as vision can be assimilated by an individual in another modality such as hearing. This paper makes use of a depth sensor to provide a spatial-auditory sensory substitution system, which converts an array of range data to spatial auditory information in real-time. In experiments, participants were trained with the system then blindfolded, seated behind a table while equipped with the sensory substitution system while keeping the sensor in front of their eyes. In the experiments, participants had to localise a target on the table by reporting its direction and in its distance. Results showed that the using the proposed system participants achieved a high accuracy rate (90%) in detecting the direction of the object, and showed a performance of 56% for determining the object's distance.**

Keywords*—*Assistive applications, binaural cues, point clouds, human-computer interaction, sound synthesis.

I. Introduction

For many people, music can evoke a sense of space, of landscapes or objects. For some synaesthetic people, who experience one modality (such as sound) as if it were another (such as vision), this is even more true [1]. "Soundscapes" are musicians' attempt to invoke such a sense through engineering of layers of sound. The auditory sense is inherently spatial, which is why stereo and surround sound are important components of modern audio engineering. In our project we go a step further than soundscapes by attempting to use sound to invoke the real space, landscapes and objects that are captured by a depth camera attached to a user.

Such systems, that use one sense to present information to users extracted from another modality (sense), are called "sensory substitution" systems , and have been used to transmit information across many different modalities. In particular, sensory substitution systems that present visual or spatial information to users promise to provide perceptual support for blind people, giving them a sense of their immediate environment, objects that they might manipulate, or enabling them to navigate the space between obstacles. To-date this ambition has not been properly realised so our work can be considered a continuation of the exploration of the space of possible systems, inspired by the following principles, which also constitute the novelty of our proposal:

- Rather than transmitting visual information to a user, spatial information such as surfaces is transmitted. We hypothesise that it is more easily dealt with by the auditory modality and can be encoded using less sensory bandwidth.
- The human perceptual system is advanced and adaptive. Over-processing the information presented to the user, as is done in some systems that tell the user to head "right" or "straight", is avoided. Instead, we prefer to present the user with information that they can then make use of – trusting the adaptability of the user to different kinds of information. At the same time, the information needs to be presented in an intuitive form (analogous to the existing perceptual experience of users) so that users can quickly and more deeply master the system.
- Information will be delivered to the user in real-time so that they can explore their physical environment intuitively. We plan in later iterations of our system to give the user transparent control over what information is delivered to them, so that they can overcome the limits of the auditory modality, whose bandwidth is relatively limited.

An additional use of sensory substitution systems is exploration of human cross-modal perception. When considering cross-modal perception, such questions arise as: How independent of modality are human perceptual processes, or how potentially independent of modality could they be given plasticity or training? Are there characteristics of different

978-1-5090-1679-2/16/$31.00 ©2016 IEEE

modalities that can be illuminated by sensory substitution systems that use them? Could it be that the auditory pathway is more adaptable to spatial information about spatially located surfaces, since information about spatially located surfaces is more directly analogous to information that the auditory modality processes normally - for example, the sound that is generated by spatially located excitement of surfaces. As the proposed system is designed to generate sound from surfaces and in a way roughly analogous to the physical synthesis of sound, our work is an early step in the exploration of this hypothesis.

In the present paper we restrict ourselves to tabletop scenarios with only one object at a time. This allows us to explore the capabilities of our proof-of-concept system with respect to the tasks of object localisation.

II. Literature Review

Sensory substitution as a subject of earnest investigation can be traced back to at least 1957 with the beginning of work on the use of active touch effectors as a means of communication, and the late 1960s with attempts to provide dense visual information to users (including sight-impaired users) in the form of vibrating tactile arrays [2].

With early impressive results, many users of touch and audio sensory substitution systems report experiencing spatial rather than simply sensory experiences, and are able to somewhat master the spatial world, recognizing objects, locating and tracking objects, detecting obstacles, and even catching falling objects [3,4]. Despite these early advances, these systems have not been used much. They can be used in everyday life only in specific situations in which the user is required to put a lot of effort into understanding the object or scene being examined, and often are considered novelties or curiosities. The main reasons for this are twofold:

- There is an acuity trade-off between temporal, spatial and amplitude discrimination meaning that the user often cannot access the full richness of perceptual information achievable by vision.
- The signal being received by the user is made up of a lot of irrelevant visual information, such that, in real-world situations, the signal of interest is crowded out by irrelevant visual information.

Although the first of these drawbacks might be considered a fundamental limit on the bandwidth of information transfer, we believe that control by the user over the sensory substitution system can be used to somewhat mitigate this limit. In our system we work on making a system that is as transparent as possible to the user. Ultimately we plan to make our system allow users to control which part of a visual scene is processed and how. The second drawback is addressed in our system by focusing on surface information rather than visual, providing spatial information directly to a largely spatial modality, namely audition.

One the most well-known sensory substitution systems is the vOICe [4], which transmits a full grey-scale image by scanning an image from left to right. The amplitudes of pixels high in the image are converted to frequencies in the high part of the audio spectrum and the amplitudes of pixels low in the image are converted to frequencies in the low part of the audio spectrum. It does this without 3D spatial audio. The system has been shown to enable the recognition and localisation of objects in high contrast environments, however the scanning process causes a delay (approximately 1 second) which makes it easy for users to get lost and indeed reduces the sense of spatial embeddedness.

Among tactile sensory substitution systems Bach-y-Rita's tongue system presents visual information, typically in the form of grey-scale images, as tactile or electrical stimulation to the tongue [3]. This kind of responsive sensory substitution systems that produce tactile stimulation allow users to respond to external stimuli, such as ball-catching, in real-time, but only under appropriate controlled high-contrast conditions.

The most similar previous work to the proposed project is that of Dunai et al. [5], that use stereo image processing to extract the horizontal location and height of a moving object in a scene and then transform that location and height into a point in the visual field (distance from the user being represented as amplitude), with object speed represented by pitch. The aim in the project proposed in the current application is similar, except that it is based around attempting to communicate a dense field of range information to the user in real-time, rather than a single object. In its current iteration, our system is only tested on single objects but still maintains the distinction of providing some object discrimination capabilities and finer location information. The use of time-of-flight technology and point cloud processing is a novelty that will also ultimately allow an exploration of characteristics of arbitrary surface sections, such as orientation and curvature rather than full objects, giving the user information more focused on the potential navigation and manipulation of objects.

III. System description

A. Hardware

At present we make use of the following hardware (see Figure 1 for an image of the system during training). We developed the system and tested it using SoftKinetic DepthSense 325 [6] in our experiments as it has a smaller minimum range than competing depth sensors that can be bought off the shelf. As a portable processing station we used a laptop with Core i5 processor, 8 GB RAM and SSD hard drive, running Kubuntu 14.04, the Point Cloud Library [7], relevant sensory drivers, and the OpenAL [8] library as the base framework for generating sounds. We also utilised headphones to play created sounds to the participants.

B. Software

Figure 2 shows an overview of the proposed software architecture. Data is acquired in the form of point clouds and segmented according to surface normal direction (obtaining contiguous and similar surfaces). Salient surface segments (in practice only one) are extracted and tracked, and the surface is presented as an audio signal located spatially as per the centroid of the points in the surface. We manage to do this at 30Hz - as fast as data comes from the sensor. More discussion of each component is below.

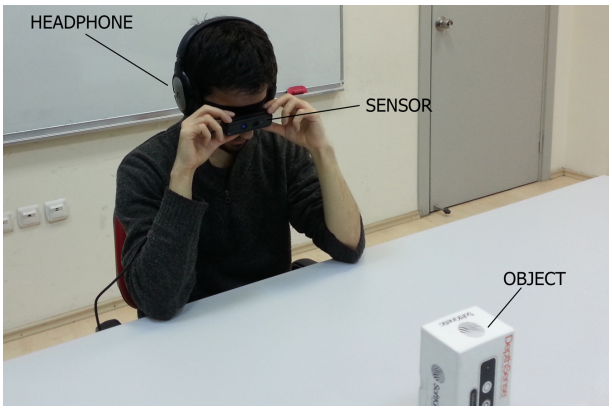

Figure 1. Participant interacting with the system.

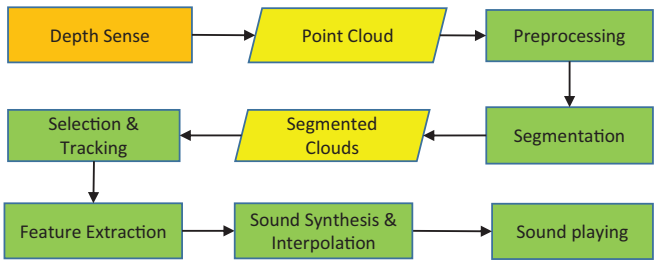

Figure 2. Software level process flow diagram.

1) Preprocessing: The most expensive part of the processing pipeline is surface normal extraction, needed for curvature-based segmentation. For real-time performance, the 2D "organised" property of point clouds from range cameras is exploited using the integral image algorithm supplied by Holzer et al. [9]. We used a maximum depth change factor of 0.5 to ensure sensible surfaces from which to calculate normals and and normal smoothing size of 70 points. For our tabletop experiments we also throw away horizontally oriented surface points so that we can ignore the tabletop on which our target object is supported, improving subsequent segmentation results.

2) Segmentation: Segmentation proceeds according to the method of Trevor et al. [10], exploiting the fact that the depth data is stored in a dense 2D array. Each point's neighbours are accessible in constant time. In two passes a set of segments is found such that neighbouring points within each segment possess the property that their depth ratio is under some threshold (0.08 in our case) and the cosine of the angle between their surface normals is above a separate threshold (0.997). By adding the latter criterion, surfaces from a tabletop can be effectively segmented without needing to explicitly calculate and extract the tabletop. This algorithm is linear time in the number of points, which allows our system to work in real-time, and yet makes it configurable enough (as long as we can express our requirements for segmentation as pairwise constraints) for future extension. Horizontally oriented segments are thrown away by getting the average of normals in the y direction and comparing it with a threshold (0.6 in our case). Also, to accommodate noise, any small segments are removed using a threshold for segment size (50 points in our case). An example segmentation can be seen in Figure 3.

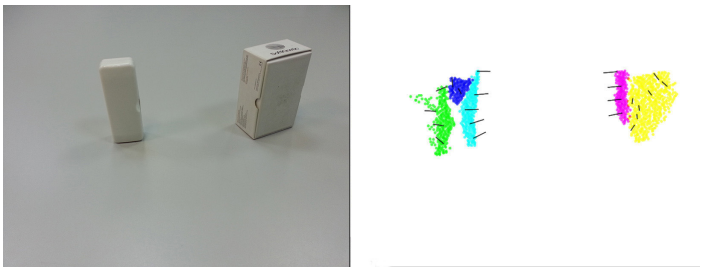

Figure 3. Screenshots form the recorded video (left panel) and segmentation process (right panel). The segmentation output also shows a selection of surface normals.

3) Selection & Tracking: Next the centroid of each segment obtained from the preceding step and the distance to it are calculated, segments are associated over time and segments chosen to be further processed - at this stage we would like to give the user more control over the segments to be processed, enable multiple segments, and use head-tracking capabilities to add stability; but the existing system uses just the single nearest large segment.

4) Feature extraction: After selecting the target segment we have the point cloud of the segment from which we can extract many spatial properties, such as size and shape, but at present we make use only the distance to the segment. This value is smoothed with a low-pass filter for stability.

5) Sound Synthesis and Interpolation: The distance to the target segment is turned into a vibration in a space proportional to the distance to the object, and appropriate frequencies introduced at the centroid of the segment using spatial audio. The system keeps an updated estimate of the amount of time between frames and attempts to construct a sound on the basis of the currently extracted segment features. The sound is constructed so that it will likely last until approximately the next frame. A circular buffer of samples is filled but envelopes are applied so that from frame to frame sounds transition smoothly. The sound synthesis component acts as a producer for the sound player consumer.

6) Sound player: The sound playing component reads as many samples as necessary from the circular buffer and acts as a consumer of information placed in the circular sound buffer, using OpenAL to play spatially located sound.

The OpenAL audio system [8] is an audio playback system capable of taking sample arrays and playing them spatially for the user using either basic binaural cues or Head-Related Transfer Functions (HRTFs). It is this out-of-the-box functionality that attracts us to OpenAL but its disadvantage is that fine temporal control is poor – we cannot for example, learn the progress of OpenAL in playing a sample to a finer resolution than about 40Hz. It does not keep an accurate record of which buffers that we have provided to it have or have not played at a finer resolution. Since our system functions at 30Hz this is insufficient. Instead we must estimate the inter-frame duration and maintain a small lag in order to prevent clicks when the interframe duration is unexpectedly long. We also try to incorporate what information OpenAL provides us about its progress in the resulting control system.

IV. Training and Experiment

The performance of the current proposed system was measured with a tabletop localisation experiment. Before conducting experiments, a training session took place. During the training session participants were seated behind a labelled table (see Figure 4 for the layout) and were given a short verbal explanation about how the system works and the meaning of the table labelling. The experimenter explained the relation between the location of the object and the frequency and apparent spatial location of sounds produced by the system. During this session participants were told about the rules of the experiment and they were given the opportunity to familiarise themselves with the system for 5 minutes while their eyes were open. Both in training and trial sessions participants held the camera in front of their eyes so that the camera touched their forehead; they were not allowed to move the camera freely, but they could observe the object by moving their head in different directions.

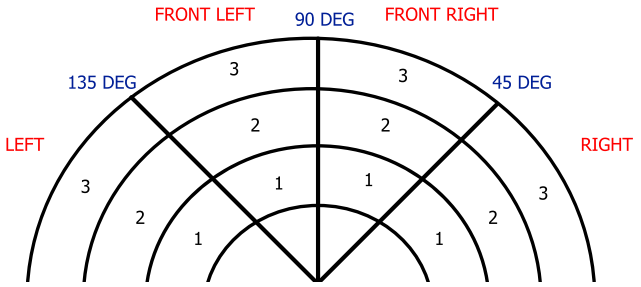

Figure 4. The labelling of the target sections.

In the trial, a box (size $15\times10\times7$ cm) was put on the table at three different distances from the user: 80 cm (range 3), 60 cm (range 2), and 40 cm (range 1), and in 4 different sections $0^\circ-45^\circ$ (right), $45^\circ-90^\circ$ (front right), $90^\circ-135^\circ$ (front left) and $135^\circ-180^\circ$ (left), so that we could measure both range and angular error. We put the object in all possible twelve regions in a random order and participants were allowed to interact with the object using our sensory substitution system for a maximum of 20 seconds, and then were asked to determine the current object direction and its distance. Answers of participants were recorded in a confusion matrix.

V. Results

The localisation experiment was done on 9 non-disabled participants with an average age of 25 years. The confusion matrices for angle and range can be found in Table I and Table II respectively.

Participants were able to choose the correct direction (angle) in 89.8% of runs as shown in Table I. The factor causing most of the mistakes (seeing the object at left front or right front when it was ahead) occurred because the system would sometimes detect the edge of the table as an object.

Range errors (Table II) were much more common. Participants only chose the right distance 56.5% of the time. The most confusion between related distances occurred as the difference in frequency was hard to detect, and because users had trouble orienting themselves with respect to the table, particularly when the object was at the front of the table since then the object often became closer than the minimum effective range of the sensor (15cm).

Table I. Direction confusion matrix

Real vs. Estimated Direction	**Right**	**Front Right**	**Front Left**	**Left**
Right	**27**	0	0	0
Front Right	1	**25**	1	0
Front Left	2	0	**25**	0
Left	2	2	3	**20**

Table II. Range confusion matrix.

Real vs. Estimated Range	**Range 1**	**Range 2**	**Range 3**
Range 1	**20**	14	12
Range 2	12	**17**	7
Range 3	4	8	**24**

VI. Conclusion

We presented a spatial-audio sensory substitution system, which is unique in that it segments objects or surfaces from point-clouds and presents them in real-time using spatial audio. The architecture of the system allows for real-time performance and tracking of perceptual objects by users. The provided system showed sufficient performance for locating an object but was only tested in the tabletop scenario.

We plan to extend the system to other scenarios, deal with multiple surface primitives, encode more information about surfaces in a scene, and increase stability and user-system interaction by giving the user more control over selection, including the ability to focus on specific parts of a scene.

Reference List

[1] S. Day, "Synaesthesia and Synaesthetic Metaphors," *Psyche*, vol. 2, no. 32, 1996.

[2] B. W. White, F. A. Saunders, L. Scadden, P. Bach-Y-Rita, and C. C. Collins, "Seeing with the skin," *Perception & Psychophysics*, vol. 7, no. 1, pp. 23–27, Jan. 1970.

[3] P. Bach-y Rita and S. W. Kercel, "Sensory substitution and the human–machine interface," *Trends in Cognitive Sciences*, vol. 7, no. 12, pp. 541–546, 2003.

[4] 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, pp. 416–430, 2007.

[5] L. Dunai, G. Fajarnes, V. Praderas, B. Garcia, and I. Lengua, "Real-time assistance prototype - A new navigation aid for blind people," in *IECON*, Glendale, Arizona, USA, Nov. 2010, pp. 1173–1178.

[6] "DepthSense Cameras." [Online]. Available: http://www.softkinetic.com/en-us/products/depthsensecameras.aspx

[7] R. B. Rusu and S. Cousins, "3D is here: Point cloud library (PCL)," in *ICRA*, Shanghai, China, 2011, pp. 1–4.

[8] "OpenAL library." [Online]. Available: https://www.openal.org/

[9] 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 *IROS*, Vilamoura, Algarve, 2012.

[10] A. J. Trevor, S. Gedikli, R. B. Rusu, and H. I. Christensen, "Efficient organized point cloud segmentation with connected components," in *ICRA Workshop on Semantic Perception (SPME)*, Karlsruhe, Germany, 2013, pp. 1363–1370.

This work was supported by TÜBİTAK (The Scientific Research Council of Turkey) Project No. 114E443.