Journal of the Audio Engineering Society
Vol. 64, No. 6, June 2016 (© 2016)
DOI: http://dx.doi.org/10.17743/jaes.2016.0007

Categorization of Broadcast Audio Objects in Complex Auditory Scenes

JAMES WOODCOCK[1], WILLIAM J. DAVIES[1],
(j.s.woodcock@salford.ac.uk)

TREVOR J. COX,[1] *AES Member,* **AND FRANK MELCHIOR,[2]** *AES Member*

[1]*Acoustics Research Centre, University of Salford, Salford, M5 4WT, United Kingdom*
[2]*BBC R&D, Dock House, MediaCityUK, Salford, M50 2LH, United Kingdom*

This paper presents a series of experiments to determine a categorization framework for broadcast audio objects. Object-based audio is becoming an evermore important paradigm for the representation of complex sound scenes. However, there is a lack of knowledge regarding object level perception and cognitive processing of complex broadcast audio scenes. As categorization is a fundamental strategy in reducing cognitive load, knowledge of the categories utilized by listeners in the perception of complex scenes will be beneficial to the development of perceptually based representations and rendering strategies for object-based audio. In this study expert and non-expert listeners took part in a free card sorting task using audio objects from a variety of different types of program material. Hierarchical agglomerative clustering suggests that there are seven general categories, which relate to sounds indicating actions and movement, continuous and transient background sound, clear speech, non-diegetic music and effects, sounds indicating the presence of people, and prominent attention grabbing transient sounds. A three-dimensional perceptual space calculated via multidimensional scaling suggests that these categories vary along dimensions related to the semantic content of the objects, the temporal extent of the objects, and whether the object indicates the presence of people.

0 INTRODUCTION

The aim of the work presented in this paper is to determine a categorization framework for typical broadcast audio objects. The proliferation of new technologies with which broadcast audio is consumed has resulted in a need to shift from the channel-based paradigm traditionally adopted by broadcast media to a more format agnostic approach. In the transmission of broadcast audio, there are a variety of ways by which a virtual scene can be represented. Channel based representations are directly related to a specific loudspeaker layout such as stereo and 5.1 and are the most widely used method to represent virtual scenes. In this paper the term "sound scene" refers to a physical pressure field, the "auditory scene" refers to the listener's perception of the sound scene, and the "virtual scene" refers to some virtual representation of the scene that can be transmitted and reproduced. Transformation based representations such as ambisonics utilize spatially orthogonal basis functions, such as spherical harmonics, to represent the virtual scene as a set of transformation coefficients that are then decoded on the reproduction side. Object-based representations store and transmit different elements of the content along with metadata describing the position of each element in time and space for rendering at the reproduction end.

Object-based audio (OBA) is often considered to be the future of spatial audio transmission [1–4] and is the primary focus of this paper. The major advantage of OBA over traditional channel-based approaches is that, as the rendering is done at the receiver end, the virtual scene can be rendered in such a way as to optimize the reconstruction for the given reproduction device and listening environment [5]. The retention of audio objects through the transmission chain opens the potential for object level processing, such as specific rules for how to render different types of objects on different reproduction systems. In this paper the term "audio object" refers to an audio signal with associated metadata whereas the term "auditory object" refers to the perceptual construct. Although there have been several proposals of how to represent OBA (see, for example, Spatial Audio Object Coding [6], MPEG-4 AudioBIFS [7], and the Audio Scene Description Format [8]), knowledge regarding object level perception of complex broadcast audio scenes is rather limited.

In general, the aim of sound reproduction in the context of broadcast audio is to, as far as possible, faithfully recreate

what the content producer has experienced in the production environment. However, the content delivery chain can result in significant degradations of the auditory scene perceived by the listener. For the representation, transmission, and rendering of OBA it is therefore advantageous to understand how complex auditory scenes are cognitively processed by the listener.

Listeners make sense of their acoustic environment by parsing auditory input into auditory objects [9–11], and there is evidence to suggest that subsequent processing of these auditory events occurs at an object, rather than signal, level [12]. The formation of auditory objects (also often called auditory events or auditory streams in the literature) consists in assigning elements of the acoustic input to one or more sources. This process is termed auditory scene analysis and is driven by two processes: (1) a pre-attentive partitioning process based on Gestalt principles [9]; (2) a schema driven process that uses prior knowledge to extract meaning from the acoustic representation [13]. Object categorization is a fundamental process underlying human cognition [14]; without the cognitive process of categorization, people would not be able to cope with the volume of sensory information to which they are constantly subjected. Thus, an understanding of how listeners categorize and assign concepts to auditory objects is central to the understanding of the perception of complex auditory scenes.

Knowledge of general categories for broadcast audio objects will aid the translation of produced virtual sound scenes to the intended listener experience by: (1) providing a perceptually grounded framework for OBA representations, and (2) allowing the investigation of object category specific rendering rules (i.e., what to do when rendering an object of category X for system Y). Subsequent perceptual testing to optimize listener experience for different loudspeaker-based reproduction systems, as well as headphone and binaural reproduction, based on these categories will allow the development of intelligent rendering schemes that will maximize the quality of experience for a given listening situation. For example, experiments could be conducted to determine quantitative rules that can be used when rendering different categories of objects for different loudspeaker layouts. Expert listeners could be given control of a small number of parameters of an object based mix (examples of these parameters might include the level and position of objects of a certain category) and asked to vary these parameters for a variety of different speaker layouts. This knowledge could then be built into a rendering scheme in the form of an additional semantic layer to the signal level manipulations that are carried out to render object based audio content for different loudspeaker layouts; for example, it might be expected the signal level optimization for objects that carry dialogue might be different to diffuse background objects. What isn't currently known is how many categories listeners utilize and the nature of these categories.

Considering how fundamental categorization is to the human experience of the world, there is scant knowledge regarding the categorization of auditory objects. Soundscape research has suggested two generic cognitive categories for auditory events within urban soundscapes: "event sequences" where the source of the sound can easily be identified, and "amorphous sequences" where it cannot [15]. Based on this categorization, it appears that sounds are processed primarily as meaningful events and where source identification fails sounds are processed according to physical or low level perceptual parameters. This view is backed up by a number of neuro-cognitive studies, which have found that the processing of environmental sounds is dependent on the relationship the sound has to its referent. For example, in a behavioral study by Giordano et al. [16], it was found that the evaluation of sounds produced by non-living objects is biased towards low level acoustic features whereas the processing of sounds produced by living creatures is biased toward sound independent semantic information. There have been complementary findings in neuro-imaging studies, where object category specific temporal activations relating to non-living action/tool sounds, animal vocalizations, and human vocalizations have been observed (see, for example, Lewis et al. [17]).

Cognitive categories for environmental sounds have been explored by Gygi et al. [18] who found three distinct clusterings of sounds that related to harmonic sounds, discrete impact sounds, and continuous sounds. A study into auditory categories for environmental sounds in complex auditory scenes revealed two main categories relating to the presence or absence of human activity [19], and common categories for auditory events in soundscapes include "natural," "human," and "mechanical" [see Payne et al. [20] for a review]. It is important to note, however, that the cognitive categorization framework is contingent; this means that the categorization framework may change depending upon factors such as location and soundscape [21]. This has potentially important implications for broadcast audio, as the categorization framework will almost certainly change depending on the program material and factors such as the presence of a screen. This suggests that different rendering rules may be needed for different scene types; for example, material with accompanying visuals may require additional object categories to account for objects that appear on and off screen.

As well as investigations into the categorization of individual auditory objects, there has been some investigation into the categorization of complex auditory scenes. Rummukainen et al. [22] investigated the categorization of natural audiovisual scenes using a free card sorting paradigm. Based on the sorting experiment, five categories of scenes were identified that related to calm scenes, still scenes, noisy scenes, vivid scenes, and open scenes. A three dimensional multidimensional scaling solution was calculated (see Sec. 1.6.3), and the dimensions of the resulting perceptual space were found to relate to calmness, openness, and the presence of people.

From the literature detailed in this section, it can be seen that previous studies have focussed on the categorization of isolated auditory objects or the categorization of complex scenes as a whole. Object-based audio presents the opportunity to optimize the reproduction for each individual sound source in a produced scene. It would therefore be beneficial

to have knowledge of how listeners categorize audio objects in broadcast audio scenes. This would allow rendering schemes to have perceptually motivated, category specific optimization rules. Although previous work into categorization of soundscapes and environmental sounds provides insight into cognitive categories for everyday sounds in isolation, there is a lack of knowledge regarding how listeners categorize objects in complex auditory scenes.

Broadcast sound scenes differ from real world scenes because they have been produced; this implies that some structure has already been imposed on the scene by the content producer. What is not clear is how listeners perceive this structure, and the implications this has on the cognitive categorization scheme used by the listener. This paper reports on a series of exploratory experiments that were conducted with the primary aim of determining general cognitive categories for common broadcast audio objects. The experiments are in the form of case-studies that explore the categorization of audio objects for different types of program material produced in 5.0. The types of material explored are radio drama, live events, nature documentary, feature film, and naturalistic soundscape recordings. The nature documentary and feature film content also include video.

1 METHODS AND MATERIALS

1.1 Ethics

The experiments described in this paper were approved by the University of Salford ethics committee. Participants took part in the experiments voluntarily, and written consent was taken prior to the test session. Participants were told that they were free to withdraw from the experiment at any time without needing to give a reason to the researcher.

1.2 Participants

A total of 21 participants took part in the test. Ten of these participants had practical experience of audio engineering. The remaining 11 participants had neither experience of audio engineering nor formal training in acoustics or audio. Audiograms were not considered necessary as the aim of the experiments was to investigate the overall experience, rather than quantify the effects of lower level features. However, participants were asked if they had normal hearing prior to the experiment. Participants were recruited via an email invitation and through social media, and they were paid for their time.

1.3 Stimuli

In the experiments reported in this paper, five different types of program material were investigated:

1)Radio drama (BBC productions of the "Wizard of Oz" and "Hitchhiker's Guide to the Galaxy: Tertiary Phase")
2)Nature documentary (BBC production of "Life: Challenges of Life")
3)Live events (BBC productions of the last night of the proms, tennis at Wimbledon, and a soccer match)
4)Feature film (*Woman in Black*)
5)Naturalistic soundfield recordings of urban soundscapes around the city center of Manchester, UK

All of the broadcast program material was available in a 5.0 mix. A number of clips were selected from each of the content types for use in the test. The length of the clips ranged from 33 seconds to 4 minutes 32 seconds, and the clips were cut to be the length of a single scene. This was done so as to provide an ecologically valid set of stimuli; as the aim of the study is to understand how listeners categorize audio objects in typical broadcast audio scenes, it is important that listeners are able to understand the context of each object within the scene. The clips were selected so as to reflect a wide range of scene types from the different types of program material. Eleven clips were used for the radio drama content (15.5 minutes in total), 8 clips were used for the feature film content (14.5 minutes in total), 4 clips were used for the nature documentary content (13.5 minutes in total), 7 clips were used for the live event content (9.2 minutes in total), and 9 clips were used for the naturalistic recordings (11.2 minutes in total). It should be noted that categorization of complex stimuli can be influenced by the length of the stimulus [22]; however, in the case of the present study the length of the scene should not influence the categorization as it is the objects within the scene that are being categorized, not the scenes themselves.

The radio drama material, nature documentary material, and feature film material are all commercially available; the times of the clips used are detailed in Appendix A. The naturalistic soundfield recordings are available to download here http://dx.doi.org/10.17866/rd.salford2234293.

1.4 Reproduction

Audio was reproduced using Genelec 8030A active loudspeakers arranged in a 5.0 setup in accordance with ITU-R BS. 775 [23] in the University of Salford semi-anechoic chamber. The radius of the loudspeaker layout was 1.30 m and the listener was seated in the center of the array. The loudspeakers were adjusted to have equal gains by generating a full scale pink noise signal for each loudspeaker and adjusting the gain of the loudspeaker so that the sound pressure level in the center of the array was equal (85 dBA) for each loudspeaker. The program material was reproduced from 24-bit wav files sampled at 48 kHz via an RME UFX soundcard. The naturalistic soundfield recordings were decoded to 5.0 using the Soundfield Surround Zone VST plugin. The radio drama, feature film, live event, and nature documentary material were reproduced with no modifications to the gain of the original material and the naturalistic soundscape material was set to a comfortable listening level.

For the program material with associated video content (nature documentary and feature film), the video content was reproduced via a laptop with a 15.6" screen (1366 x 768 resolution). The laptop was positioned on a table in the test room and was approximately 0.8 m from the participant.

1.5 Procedure

Participants were required to complete a sorting task. A large number of variants of the sorting method exist, each of which results in different types of data. Details of the different methods can be found in Coxon [24]. The main differences between variants stem from whether the number of categories is determined by the researcher (fixed sorting) or the participant (free sorting) and whether the meaning of the categories is specified by the researcher (closed sorting) or the participant (open sorting). In the present study, as there were no *a priori* assumptions made regarding the number of categories or the meaning of the categories, a free and open sorting methodology was used.

For each type of program material, participants were given a set of cards. Each card was labelled with an object. Each set of cards contained all of the identifiable objects within the program material. Each card also contained an identifier to help the participant identify the clip in which the object occurred and the time of the first occurrence of the object in the clip. The objects printed on the cards were identified by a group of five expert listeners prior to the test. The aim of this exercise was to identify as many individual objects in the clips as possible. The expert listeners were given a list of objects for each of the clips, which had previously been identified by the main author of this paper; their task was then to identify any objects missing from the list, or to modify the description of any objects they disagreed with.

For the radio drama material there were 176 cards, for the feature film content there were 142 cards, for the nature documentary content there were 91 cards, for the live event content there were 105 cards, and for the naturalistic urban soundscape recordings there were 110 cards.

Participants were presented with an interface developed in Pure Data with which they could start, stop, rewind, fast forward, and switch between the different clips. The participants were asked to sort the cards into groups on the desk in front of them according to the following criteria: *"Please sort the cards into groups such that the sounds in each group serve a similar function or purpose in the composition of the scene."*

The participants were instructed that they were required to sort the objects according to their function in the scene and not necessarily according to the similarity of the sounds themselves. If the participants asked for an example they were told the following: "Consider that you were asked to sort the instruments in an orchestra so that the instruments in each group serve a similar function or purpose in the orchestra. You may decide to make a *percussion* group which contains the timpani, triangle, and snare drum. Although these instruments all have a different sound, they each serve a similar purpose in the orchestra."

Participants were instructed that they could form as many or as few categories as they wished and that the relative positions of the categories on the desk was unimportant. They were asked to use all of the cards for the given type of program material, such that at the end of the test all of the cards from all of the scenes for that type of material had been sorted into categories. This procedure is often referred to in the literature as a free sorting task [24].

Once the participant was happy with their grouping, they were asked to give a short label to each of the categories they had formed, and also to give a rating from 0 to 10 of the importance that category of audio objects had in their overall experience of the scene. Note that, as participants were required to sort all of the objects and some participants were unable to identify some of the objects in the clips, this procedure resulted in a small number of the participants forming a category of sounds they could not identify.

This procedure was carried out for each of the types of program material; therefore, each participant completed five separate card sorts. The participants were told that they were free to make new categories for the different content types. The order in which the different types of program material were presented was randomized for each participant.

After the participant had completed the card sort for each type of content, they were presented with the all of the category labels they had generated throughout the entire procedure. The participant was asked to sort the categories into groups that represented the same concept. The aim of this was to investigate commonalities and differences between the categorization structure for the different types of program material.

Participants were allowed 3.5 hours to complete the test, and the participants were given the opportunity of 2 short comfort breaks throughout the test. Due to this time restriction, 4 of the participants did not manage to complete card sorts for all 5 types of content. The data for the tests they did complete are used in the subsequent analyses reported in this paper. It is interesting to note that despite the length of the test, most of the participants stated that they found the process enjoyable and not overly fatiguing.

1.6 Analysis

1.6.1 Data Preparation

For each type of program material, a categorization matrix was formed that took on a value of 1 if an object had been grouped in a given category and a 0 otherwise. A categorization matrix encompassing all of the program material types was formed in the same way based on each participants' sorting of their category labels.

A co-occurrence matrix was generated for each participant for all of the audio objects over all of the different types of program material. This matrix was constructed from pairwise similarities of the objects by assigning pairs of objects that had been grouped in the same category a 1 and pairs of objects that were not grouped in the same category a 0. The individual similarity matrices were averaged over the participant group to generate an average similarity matrix.

A graphical representation of the construction of these matrices and the subsequent analysis is shown in Fig. 1.

1.6.2 Agglomerative Hierarchical Clustering

The categorization matrices were analyzed using agglomerative hierarchical clustering. Hierarchical clustering is a technique that produces a nested sequence of partitions

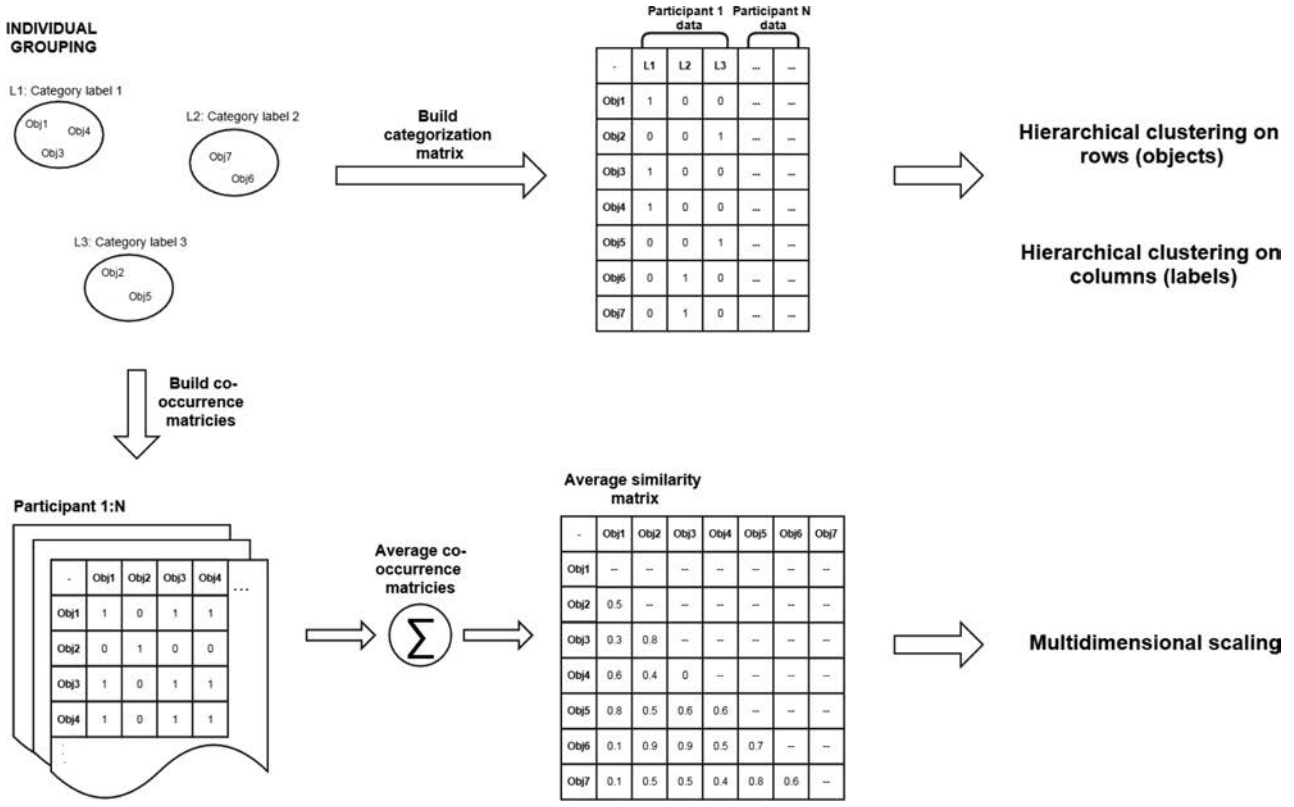

Fig. 1. Graphical representation of the construction of the data matrices and subsequent analysis for the sorting task for different types of program material. The same procedure was used in the sorting of the category labels generated across the program items.

of a dataset, with a top level cluster that encompasses all objects and a bottom level consisting of each object as an individual. Intermediate levels show the merging of two clusters from the lower level. The results of hierarchical clustering are most often displayed as dendrograms that graphically represent this merging process. In agglomerative clustering, the merging process starts at the bottom level, with all objects as individual clusters. At each subsequent stage, the closest pair of clusters are merged. The dendrograms produced by this analysis can then be cut at different levels to examine the structure of the data; this cut is often made to give the average number of groups formed by the participants [22].

The clustering using the Ward's minimum variance method, which aims to minimize the total within cluster variance defined by the sum of the squared Euclidean distances within each cluster [25], was conducted both row-wise and column-wise on the categorization matrices, thus resulting in two clustering solutions for each of the types of program material; one solution relating to the clustering of the audio objects and the other solution relating to the clustering of the descriptive labels participants attributed to their groups.

1.6.3 Multidimensional Scaling

The co-occurrence similarity matrix (generated by averaging the individual similarity matrices over the participant group) was analyzed using non-metric multidimensional scaling [26]. Multidimensional scaling is an exploratory data analysis technique the aim of which is to determine a configuration of a group of objects in a low dimensional multidimensional space. The resulting configuration provides a visual representation of pairwise distances or (dis)similarities between objects in the group. This low dimensional representation is assumed to represent a latent perceptual space, with the dimensions representing salient orthogonal perceptual features. Multidimensional scaling has been used extensively in sensory sciences, and for sound perception in areas such as the perception of musical timbre [27, 28], the perception of concert hall quality [29], and product sound quality [30].

2 RESULTS

The following sections show the results of the hierarchical cluster analysis for the different types of program material investigated in this paper. Due to the number of labels in the clustering solution, it is not possible to reproduce the full clustering solutions in this paper. By way of example, Fig. 2 and Fig. 3 show a truncated version of the full clustering solution (the first cluster of category labels generated by the participants and the first cluster of audio objects for the radio drama program material). In the figures that accompany these results, labels have been assigned summarizing each of the clusters that are formed when cutting the clustering of category labels at a level that results in a number of clusters equal to the median number of clusters formed across participants for that type of program

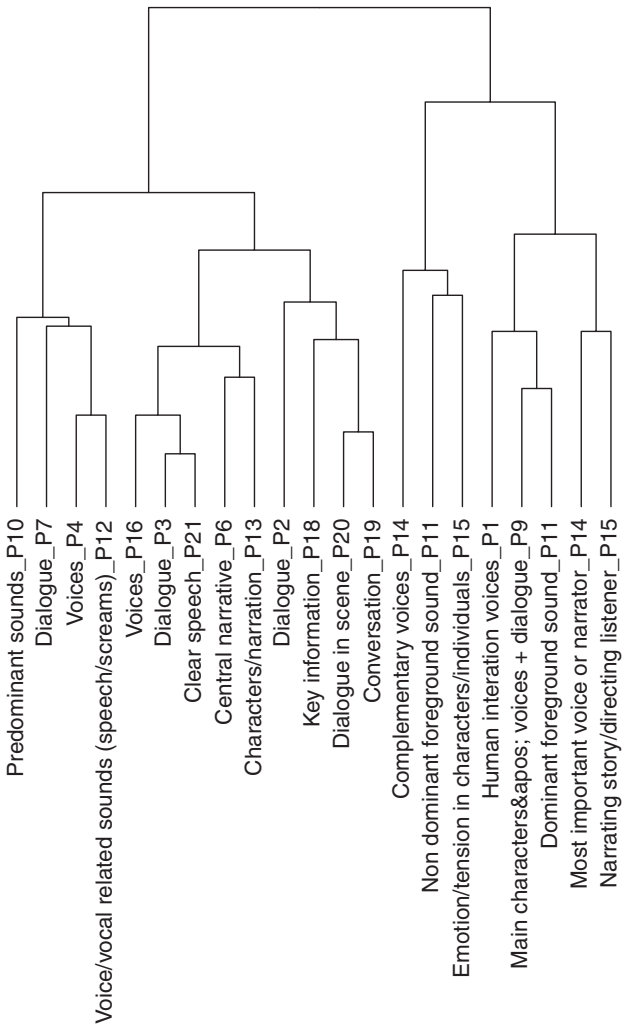

Fig. 2. First cluster of category labels for the radio drama program material. PN indicates that the category label was produced by participant *N*.

material. The median number of clusters was taken as the starting point to interpret the clustering solutions. Any further interpretable subclusters are also discussed. In the case of the cluster of category labels shown in Fig. 1, this cluster was summarized by the researcher as "Clear speech." The full clustering of audio objects and categories can be found at http://dx.doi.org/10.17866/rd.salford2234293.

2.1 Radio Drama

Fig. 4 shows the results of the cluster analysis with respect to the category labels for the radio drama program material. The median number of categories formed for this program material was 5, and as such the dendrogram shown in Fig. 4 has been cut so as to show 5 clusters.

From the clustering of the category labels (from left to right), the first cluster relates to clear speech; participants' category labels for this cluster include "Dialogue in scene (P20[1])," "Clear speech (P21)," and "Dialogue (P2, P3, and P7)" and objects in this group include the main character voices. The second cluster of category labels relates to sounds that coincide with actions or movement; participants' category labels for this cluster include "Sound of movement (P19)," "Activity sounds (P1)," and "Plot forwarding/vital sounds (P9)," and related objects include footsteps, opening of doors, and clinking of glasses. The third cluster of category labels relates to non-diegetic (here, "non-diegetic" refers to whether the audio object is implied to be present in the scene): music and effects; participants' category labels for this cluster include "Musical sounds (P1)," "Music and SFX (not part of scene) (P3)," and "Music (outside scene) (P20)," and related objects include musical instruments along with low frequency rumbling and whooshing sounds. The fourth cluster of category labels relates to both localizable and continuous background sounds; participants' category labels that appear in this cluster include "Background effects (P2)," "Ambient sound (P4)," and "Background noise. Set a location (P13)," and related objects include rain sounds, wind sounds, and birds tweeting. The interpretation of the fifth cluster of category labels in less clear and seems to encompass a number of different less well defined categories including vocalizations, attention grabbing impact sounds, diffuse atmospheric sounds, and diegetic music. Vocalizations appeared as a well defined cluster of audio objects.

2.2 Feature Film

Fig. 5 shows the results of the cluster analysis with respect to the category labels for the feature film program material. The median number of categories formed for this program material was 5, and as such the dendrogram shown in Fig. 5 has been cut so as to show 5 clusters.

From the clustering of category labels (left to right), the first cluster relates to sounds relating to actions and movement; participants' category labels for this cluster include "Dominant/meaningful event sound (P11)," "Single event sounds (P9)," and "Sounds resulting from human activities (P21)," and objects related to this category include footsteps, impacts of objects on tables, and doors opening. The second cluster of category labels relate to clear speech and dialogue; participants' category labels for this cluster include "Human voice (P14)," "Dialogue (P2)," and "Key information (P18)," and objects related to this category include the main character voices along with vocalization such as screaming. The third cluster of category labels couldn't be clearly interpreted as a whole; it did however encompass a clear cluster of prominent, attention grabbing sounds that occur off-screen; participants' category labels for this cluster include "Off-screen but significant (P7)," "Things happening out of the scene (P13)," and "Impact sound, loud, distinct (P14)," and related objects include impact sounds from upstairs (off-screen), clattering of cart wheels, and a glass smashing. The fourth cluster of category labels relate to non-diegetic music and effects; participants' category labels for this cluster include "Music (P2)," "Mood defining (usually music) (P6)," and "Music and sound effects (not part of scene) (P3)." The related objects for this cluster could be seen to clearly cluster into non-diegetic music (i.e., strings and synth pads) and effects (i.e., low frequency rumbling and high frequency whispering). The

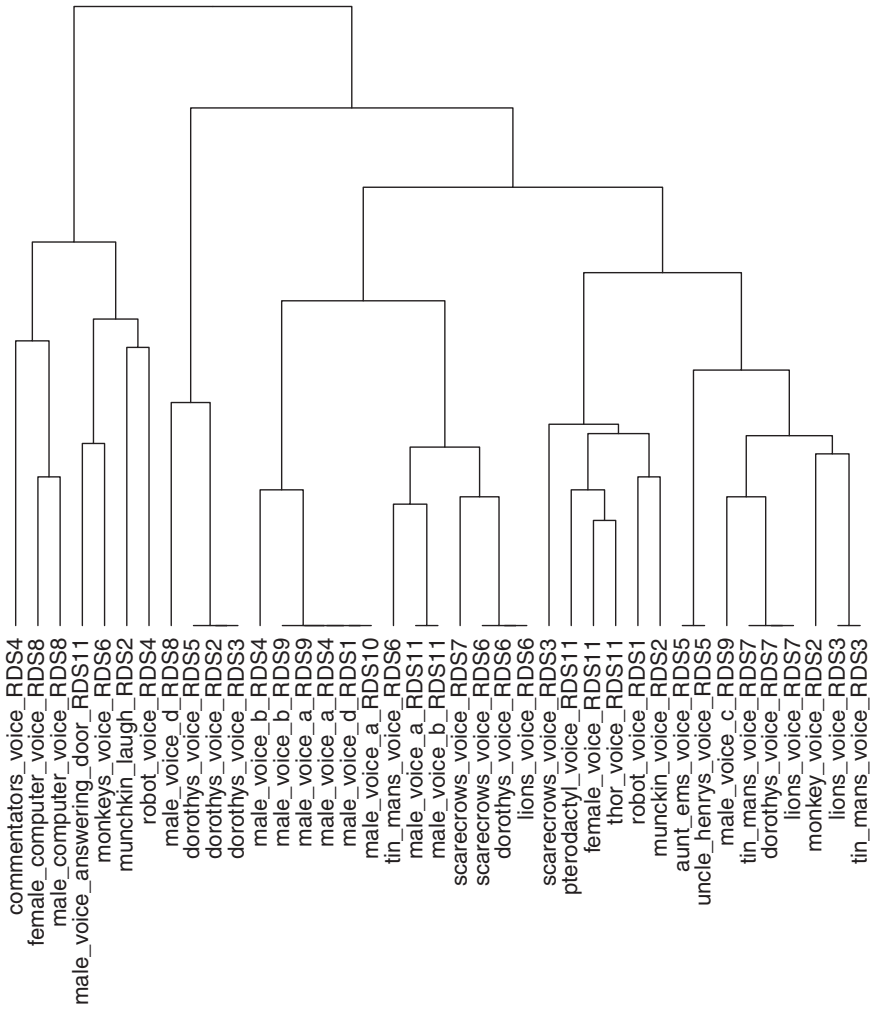

Fig. 3. First cluster of audio objects for the radio drama program material.

fifth cluster of category labels relate to both diffuse and localizable background sounds; participants' category labels for this cluster include "Background (P18)," "Diffuse atmos (P20)," and "Scene setting (P5)," and objects related to this cluster include birdsong, wind whistling, and crowd babble. Further inspection of this cluster of audio objects reveals a clear clustering of continuous (i.e., wind and crowd babble) and transient background sounds (i.e., birdsong and horses hooves).

2.3 Nature Documentary

Fig. 6 shows the results of the cluster analysis with respect to the category labels for the nature documentary material. The median number of categories formed for this program material was 5, and as such the dendrogram shown in Fig. 6 has been cut so as to show 5 clusters.

From the clustering of category labels (left to right), the first cluster relates to sounds relating to actions or movement; participants' category labels for this cluster include "Dominant event sound (related with the video) (P11)," "Sounds resulting from animals movements/actions (P21)," and "Sounds directly relating to actions on-screen (P12)," and the audio objects related to this category include animal footsteps, splash of animals entering water, and the crunch of a venus fly trap closing. Within the cluster of category labels, two clear subcategories can be seen relate to sounds coinciding with on-screen action and sounds that don't have a visual counterpart. The second cluster of category labels relates to non-diegetic music and effects; participants' category labels for this cluster include "Musical instruments (P16)," "Music/SFX (P7)," and "Non-diegetic music (P9)," and the audio objects related to this cluster include musical instruments and synthesized effects. The third cluster of category labels relates to localizable and diffuse background sounds; participants' category labels for this cluster include "Quieter sounds (P10)," "Envelopment/scene setting (P6)," and "Non-dominant event sound (P11)," and the audio objects related to this category include bird calls, the sound of rustling grass, and the sound of splashing water. The fourth cluster of category labels relates to the narration; participants' category labels for this cluster include "Narrator/Narration (many participants)," "Key information (P18)," and "Dialogue outside scene (P20)." The fifth cluster of category labels presents no clear grouping, but contains a subgroup of prominent animal vocalizations; participants' category labels for this sub-cluster include "Animal noises observable (P3)," "Prominent animal vocalizations (normally on screen) (P20)," and "Important sounds

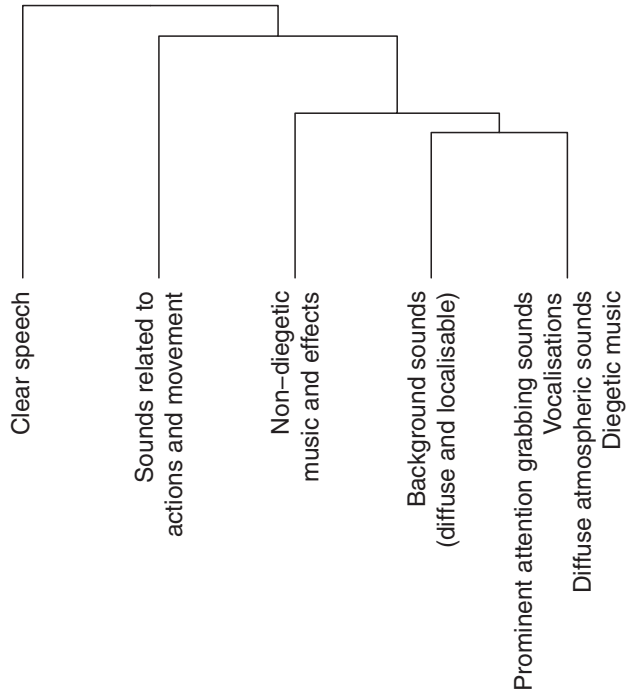

Fig. 4. Dendrogram showing hierarchical agglomerative clustering of category labels for the radio drama program material.

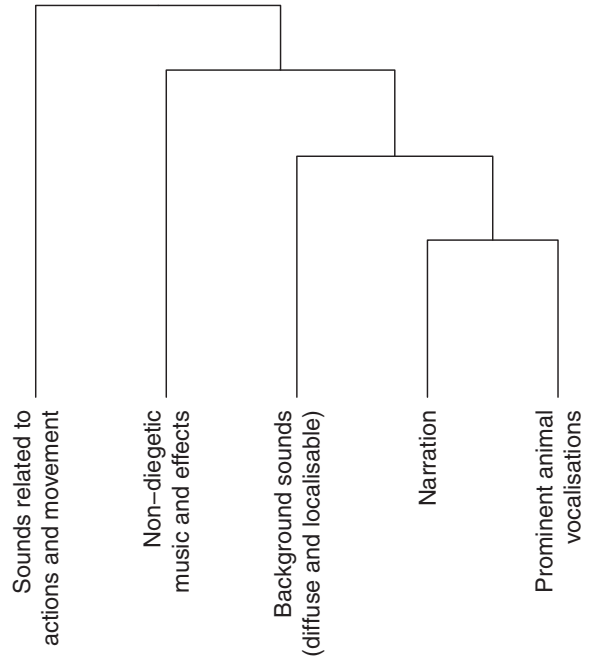

Fig. 6. Dendrogram showing hierarchical agglomerative clustering of category labels for the nature documentary program material.

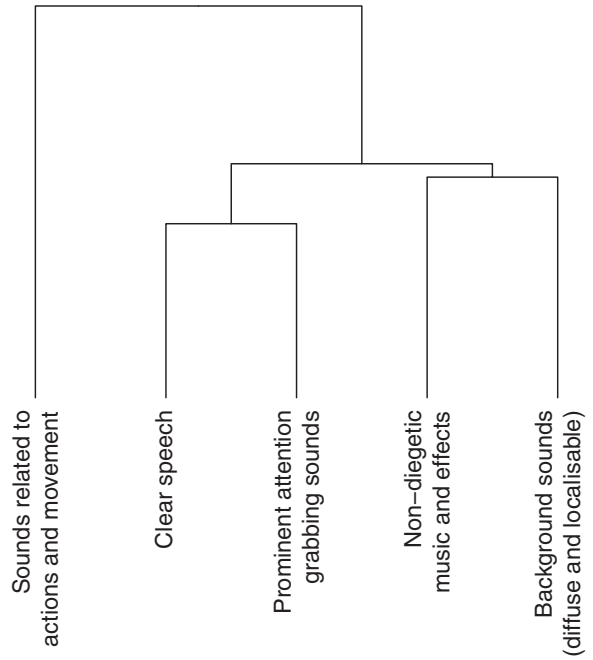

Fig. 5. Dendrogram showing hierarchical agglomerative clustering of category labels for the feature film program material.

(P16)," and the audio objects related to this cluster include ostrich vocalizations, seal vocalizations, and the sound of a whale blowing.

2.4 Live Events

Fig. 7 shows the results of the cluster analysis with respect to the category labels for the feature live event material. The median number of categories formed for this program material was 5, and as such the dendrogram shown in Fig. 7 has been cut so as to show 5 clusters.

From the clustering of category labels (left to right), the first cluster relates to commentary and clear speech; participants' category labels related to this cluster include "Commentary (P4)," "Key information/narrative (P18)," and "Verbal description/direction (P18)," and audio objects related to this category include commentators' voices, tennis umpires' voices, and stadium announcements. The second cluster of category labels relates to primary event sounds; participants' category labels related to this cluster include "Primary event sounds (P7)," "Primary (P17)," and "Target music from the stage/field (P21)." This category can split into two further categories relating to music where the focus of the live event is music (related audio objects include individual musical instruments) and event sounds for sporting events (related audio objects include ball kicks, referee's whistle, and the impact of a tennis ball on a racket). The interpretation of the third category is less clear, it does however contain a subcategory of impact sounds; participants' category labels related to this cluster include "Movement/impact (P15)," "Sounds related to actions (P20)," and "Foreground sound effects (P2)." The fourth cluster of category labels is related to the reaction of the crowd to events; participants' category labels related to this cluster include "Crowd noise (P1)," "Crowd reaction (P4)," and "Collective sounds/vocalizations (P9)," and audio objects related to this category include applause, crowd cheering, and laughter. The fifth cluster of category labels is related to localizable background sounds; participants' category labels related to this cluster include "Non-dominant event sound

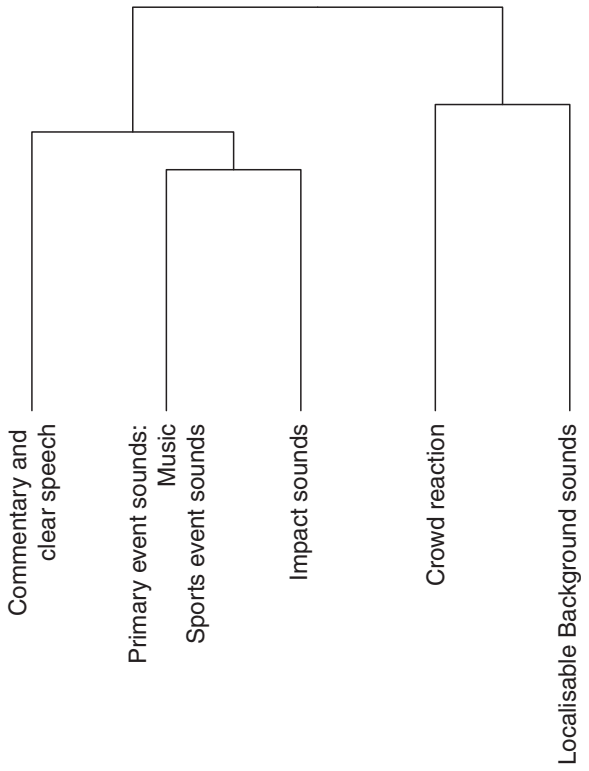

Fig. 7. Dendrogram showing hierarchical agglomerative clustering of category labels for the live event program material.

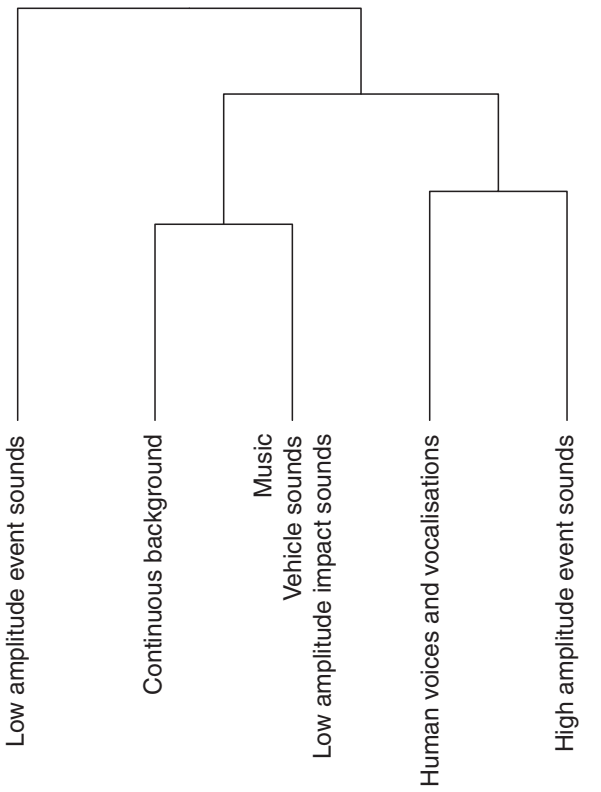

Fig. 8. Dendrogram showing hierarchical agglomerative clustering of category labels for the naturalistic urban soundscape recordings.

(P11)," "Individually identifiable sounds (P3)," and "Quiet sounds which are identifiable from the background noise (P10)." Closer inspection of this cluster of audio objects reveals a clear grouping of living (i.e., coughs and crowd whistling) and non-living sounds (i.e., clicks, party poppers, and fireworks).

2.5 Naturalistic Recordings

Fig. 8 shows the results of the cluster analysis with respect to the category labels for the naturalistic recordings of urban soundscapes. The median number of categories formed for this program material was 5, and as such the dendrogram shown in Fig. 8 has been cut so as to show 5 clusters.

From the clustering of category labels (left to right), the first cluster relates to low amplitude localizable event sounds; participants' category labels related to this cluster include "Non-dominant event sound (P11)," "Low level event sound (P20)," and "Sounds resulting from human activities (P21)," and audio objects related to this category include the rustling of paper, jangling of coins, and various impact sounds. The second cluster of category labels is related to continuous background sounds; participants' category labels related to this cluster include "Ambient sounds (P9)," "Background filler/bed (P2)," and "Background sounds which indicate the scene (P21)," and audio objects related to this category include unintelligible voices, distant traffic noise, and air conditioning sounds. The interpretation of the third cluster of category labels is less clear. Within this cluster there are a number of clear subgroups; the first is related to music (related audio objects include music in shops), the second is related to vehicle sounds (related audio objects include vehicle acceleration sounds, cars starting, and the clunk of a vehicle passing over a manhole cover), and the third is related to low level impact sounds (related audio objects include various unidentifiable impacts). The fourth cluster of category labels relates to human voice and vocalizations; participants' category labels related to this cluster include "Human voice (P1)," "Presence of people (P15)," and "Human generated sounds/noises/vocalizations (P9)," and audio objects related to this category include voices, laughter, and coughing sounds. The fifth cluster of category labels relates to high amplitude localizable event sounds; participants' category labels related to this cluster include "Dominant and meaningful event sound (P11)," "Louder sounds (P10)," and "High level foreground event sounds (P20)," and audio objects related to this category include doors closing, mobile phone notifications, and the sound a chair scraping against the floor.

2.6 All Material

From the free sort of category labels across all of the program material types, participants formed a median of seven groups. The first of these clusters consists of sounds related to actions and movement. The second and third clusters relate to background sounds, with the second cluster mainly relating to transient background sounds and the

third cluster mainly relating to continuous ambient sounds, crowd reaction, and sounds indicating the presence of people. Some overlap was observed in the category labels for these two groups, with, for example, the category "secondary action sounds" being in the same cluster as diffuse and ambient categories. The fourth of these clusters relates to clear speech and dialogue. The fifth cluster relates to non-diegetic music and effects. The interpretation of the sixth cluster was less clear, but contained a cluster of clear speech that is outside of the scene, music that occurs within the scene, and human vocalizations. The seventh cluster related to prominent transient sounds.

From the sorting of the category labels, a similarity matrix was built using the method described in Sec. 1.6. The data was subject to non-metric multidimensional scaling (MDS). Whereas the hierarchical clustering solutions presented in the preceding section gave a hierarchical view of the categorization structure, MDS provides a different way of interpreting the data by allowing the investigation of the independent perceptual dimensions along which the objects and categories vary. To determine an optimum dimensionality of the scaling, solutions were calculated in 2 to 9 dimensions and the stress was inspected. A three-dimensional solution gives a non-metric stress of 0.12, which suggests a fair fit with the original data [31]. The Pearson's correlation between the original and fitted distances for a three dimensional solution is 0.89 ($p < 0.001$).

Fig. 9 shows the configuration of audio objects in the three dimensional multidimensional scaling solution. The points in this figure relate to individual audio objects. The groupings have been formed by a hierarchical agglomerative clustering of the dissimilarity matrix; this resulted in slightly different grouping than the cluster analysis that was conducted on the co-occurrence matrix, with crowd reactions and sounds indicating the presence of people emerging as a cluster and prominent transient sounds being grouped with sounds relating to actions.

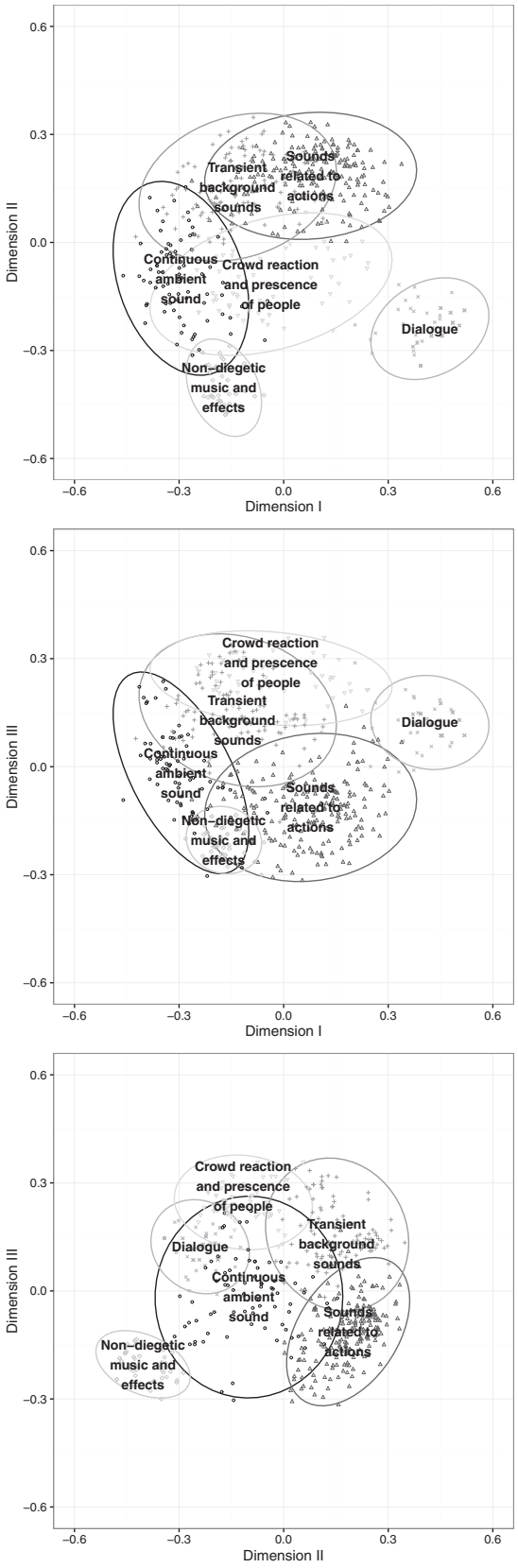

Fig. 9. Configuration of audio objects in a three dimensional non-metric multidimensional scaling solution. Ellipses show the clusterings identified in Sec. 2.6. For clarity, objects that fell into the unclear cluster have not been plotted. A color version of this figure is available at http://dx.doi.org/10.17866/rd.salford2234293.

3 DISCUSSION

3.1 Interpretation of Perceptual Dimensions

From the ordering of objects along the dimensions of the multidimensional scaling configuration shown in Fig. 9, some interpretation can be made of the meaning of these dimensions. The first dimension appears to be related to the relationship the object has to it's referent; that is to say, whether the object carries semantic information such as clear speech or is related to an action. This can be seen in the progression along the first dimension of object categories from continuous background objects (exemplified by sounds such as low frequency rumbling, birdsong, and distant traffic noise) through to short localizable background sounds, action sounds, vocalizations, and finally dialogue and clear speech. This progression of object categories along the first perceptual dimension parallels findings from neuro-cognive studies [17, 16] where differences have been found in the processing of non-living action/tool sounds, animal vocalizations, and human vocalizations. The sec-

ond dimension appears to be related to the temporal extent of the audio objects. This can be seen in the progression of object categories along this dimension, from music and dialogue at one extreme and transient sounds at the other. This supports the findings of Gygi et al. [18], who derived a perceptual space for the categorization of environmental sounds and found that the second perceptual dimension differentiated between continuous sounds and impact sounds. The interpretation of the third perceptual dimension is less clear, but it appears to relate to the presence of people with non-diegetic music and effects at one extreme of the dimension, and dialogue and crowd reactions appearing at the other extreme. This is consistent with findings in soundscapes research [19] and research into the perception of complex audiovisual scenes [22].

3.2 Differences between Naive and Expert Listeners

Research into audio quality and the perception of urban soundscapes has revealed differences between listeners who have training in acoustics or audio engineering (so called "expert listeners") and those who don't (so called "naive listeners"). For example, Guastavino [32] [described in Guastavino and Katz [33]] found that the preferred audio reproduction method for urban soundscapes varied depending on whether the listener is a sound engineer, acoustician, or non-expert. Sound engineers were found to give greater precedence to localization and precision of sources, whereas non-expert listeners and acousticians gave greater precedence to presence and spatial distribution of sound. Perceptions of audio quality can change depending on the experience and role of the listener. For example, Rumsey et al. [34] have investigated the relationships between experienced listener ratings of multichannel audio quality and naive listener preferences. It was found that timbral fidelity, frontal spatial fidelity, and surround spatial fidelity contributed to expert listeners' ratings of basic audio quality, however only timbral fidelity and surround spatial fidelity contributed significantly to naive listeners' ratings of preference.

To explore if there were any differences in the categorization strategy for expert and naive listeners, data from each of the types of program material were split into two subsets. The first subset contained data from those participants who stated that they had previous practical experience of audio engineering and the second subset contained data from those participants who stated that they had no previous practical experience of audio engineering. Hierarchical agglomerative clustering (Ward method) was conducted for the audio objects on each of these subsets of data for each type of program material. The similarity of the clustering solutions for the two different groups of listeners was then assessed using the Rand Index [35], which is a measure of the agreement between two clustering solutions. The measure takes into account true positive decisions where two objects have been classified in the same cluster and true negative decisions where two objects have been classified in different clusters. The Rand Index is then expressed as a percentage indicating the sum of the true positives and true negatives over the the number of all possible pairs of objects in the clustering solution.

Based on the calculated Rand Index between the expert and non-expert clustering solutions, for the radio drama program material 79% of pairs of objects were categorized in the same way, for the feature film material 75%, for the nature documentary material 86%, for the live events material 87%, and for the naturalistic recordings 78%. Faye et al. [36] suggest that the free sorting with naive participants leads to similar results as descriptive analysis by an expert panel, and the similarity between the clustering for expert and non-expert groups appears to support this claim. Differences between the two listener groups included a tendency for the expert listener group to use more technical language such as *foley* and *diegetic*. Further, the categorization structure was found to be more homogeneous across the expert listener groups, with the non-expert listener group creating more unique categories.

Table 1. Results of a linear regression model relating object position in a 3 dimensional MDS solution to mean object importance. MDS1 is the position of the object on the first perceptual dimension, MDS2 is the position of the object on the second perceptual dimension, and MDS3 is the position of the object on the third perceptual dimension. Numbers in brackets are standard errors.

	Dependent variable: Importance
MDS1	3.42***
	(0.103)
MDS2	−1.19***
	(0.110)
MDS3	−1.35***
	(0.139)
Constant	6.60***
	(0.023)
Observations	624
R^2	0.679
Adjusted R^2	0.677
Residual Std. Error	0.585 (df = 620)
F Statistic	435.568*** (df = 3; 620)

Note: *p<0.1; **p<0.05; ***p<0.01

3.3 Importance of Groups

For each object, a mean importance rating was calculated by assigning each object the importance rating given by the participant of the group in which it was included and taking an average of these ratings across participants. A multiple linear regression model was calculated with the positions of the sounds on the three axes calculated in the multidimensional scaling analysis as independent variables and the mean importance of each object as the dependent variable. The results of this model are shown in Table 1. The model was found to be a significant fit and accounted for 68% of the variance in the importance scores ($R^2_{adj} = 0.68$, $p < 0.001$). A forward-backward stepwise regression resulted in no dropping of variables in the model. This suggests that each of the dimensions are significantly

related to the perceived importance of each of the object categories.

Taking the model coefficient for the first perceptual dimensions as an example, the interpretation of Table 1 is such that a 3.41 increase in an object's position on the first perceptual dimension corresponds to a unit increase in the object's perceived importance. The sign of the regression coefficients suggest that perceived importance increases as sounds progress along Dimension I and decreases as sounds progress along Dimension II and Dimension III. The first perceptual dimension was found to be related to the semantic information carried by the object. The coefficient for the first perceptual dimension in the model shown in Table 1 therefore suggests that objects carrying semantic information have the greatest weighting on the perceived importance of an object to a scene.

3.4 Consequences for Object Based Audio

The results presented in this paper provide a framework for the categorization of broadcast audio objects in complex auditory scenes. Considering the median number of clusters produced for each type of program material, the results presented in Sec. 2 suggest that listeners utilize around five categories for each of the types of program material. Overall, there appear to be at least seven unique categories across the program material, suggesting that the categorization structure is somewhat contingent on the type of material.

Object-based audio opens up the possibility of object level manipulation of audio content, where different categories of object can be subject to different rules and manipulations. This would allow the signal level manipulation used in the rendering of spatial audio to be optimized on a category by category basis. The results presented in this paper provide a perceptual basis for such a categorization framework, ensuring that the categories used are relevant to how listeners parse complex auditory scenes.

Knowledge of the categorization structure will allow the investigation of high level semantic rules that can be used to optimize the rendering of spatial audio material. For example, sounds relating to actions or movements may be treated differently to continuous background sounds when rendered to different loudspeaker layouts.

Finally, the categories presented in this paper provide a perceptual basis for future metadata specifications for object-based audio and could provide the basis for future high level languages for the description of the rendering of spatial audio. In terms of object based workflows, this may take the form of a metadata field that allows content producers to tag and group different objects in the production according to the categories presented in this paper.

4 CONCLUSIONS

This paper has presented a series of experiments conducted to determine categories for auditory objects in complex broadcast audio scenes. Twenty-one participants completed free sorting tasks for five types of program material. Hierarchical agglomerative cluster analysis revealed at least seven categories across the different types of program material. These categories relate to sounds indicating actions and movement, continuous and transient background sound, clear speech, non-diegetic music and effects, sounds indicating the presence of people, and prominent attention grabbing transient sounds. A three-dimensional perceptual space calculated via multidimensional scaling suggests that these categories vary along the dimensions of semantic content, continuous-transient, and presence-absence of people. The position of an audio object on the dimensions of the perceptual space were found to be related to the perceived importance of the object. These results are well supported by findings in environmental psychology, soundscape research, and neuro-cognitive studies, and have applications in psychological research into complex auditory scene perception, multimedia quality-of-experience testing, and the development of object based audio processing.

5 ACKNOWLEDGMENTS

This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and the BBC as part of the BBC Audio Research Partnership. The author would like to thank Chris Pike and Steve Marsh from BBC R&D for their help in sourcing the material for the tests. Finally, the author would like to thank the participants of the listening tests for their time. The experimental data underlying the findings are fully available without restriction, details are available from http://dx.doi.org/10.17866/rd.salford2234293. Due to copyright restrictions, the radio drama, live events, nature documentary, and feature film program material used in the listening experiments is not available from this link. A metadata record of these data can be found at http://dx.doi.org/10.17866/rd.salford2234413.

6 REFERENCES

[1] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties "MPEG-H Audio—The New Standard for Universal Spatial/3D Audio Coding," *J. Audio Eng. Soc.*, vol. 62, pp. 821–830 (2014 Dec.). http://dx.doi.org/10.17743/jaes.2014.0049

[2] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter "Spatial Sound with Loudspeakers and its Perception: A Review of the Current State," *Proceedings of the IEEE*, vol. 101, pp. 1920–1938 (2013 Jul.). http://dx.doi.org/10.1109/JPROC.2013.2264784

[3] R. Oldfield, B. Shirley, and J. Spille "An Object-Based Audio System for Interactive Broadcasting," presented at the 137th Convention of the Audio Engineering Society (2014 Oct.), convention paper 9148.

[4] C. Kim, "Object-Based Spatial Audio: Concept, Advantages, and Challenges," in *3D Future Internet Media* (Springer–Verlag, New York, 2014), pp. 79–84.

[5] B. Shirley, R. Oldfield, F. Melchior, and J.-M. Batke, "Platform Independent Audio," in *Media Production, Delivery and Interaction for Platform Independent Systems: Format-Agnostic Media*, John Wiley & Sons, Chichester, 2013), pp. 130–165.

[6] J. Herre, H. Purnhagen, J. Koppens, O. Hellmuth, J. Engdegård, J. Hilper, L. Villemoes, L. Terentiv, C. Falch, A. Hölzer, et al., "MPEG Spatial Audio Object Coding—The ISO/MPEG Standard for Efficient Coding of Interactive Audio Scenes," *J. Audio Eng. Soc.*, vol. 60, pp. 655–673 (2012 Oct.).

[7] E. D. Scheirer, R. Vaananen, and J. Huopaniemi "Audiobifs: Describing Audio Scenes with the MPEG–4 Multimedia Standard," *IEEE Multimedia*, vol. 1, pp. 237–250 (1999 Sept.). http://dx.doi.org/10.1109/6046.784463

[8] M. Geier, J. Ahrens, and S. Spors "Object-Based Audio Reproduction and the Audio Scene Description Format," *Organ. Sound*, vol. 15, pp. 219–227 (2010 Dec.). http://dx.doi.org/10.1017/S1355771810000324

[9] A. S. Bregman *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, 1994), pp. 1–45.

[10] T. D. Griffiths and J. D. Warren "What Is an Auditory Object?" *Nat. Rev. Neurosci.*, vol. 5, pp. 887–892 (2004 Nov.). http://dx.doi.org/10.1038/nrn1538

[11] J. K. Bizley and Y. E. Cohen "The What, Where and How of Auditory-Object Perception," *Nat. Rev. Neurosci.*, vol. 14, pp. 693–707 (2013 Sept.) http://dx.doi.org/10.1038/nrn3565

[12] N. Ding and J. Z. Simon "Emergence of Neural Encoding of Auditory Objects While Listening to Competing Speakers," *P. Natl. Acad. Sci. USA*, vol. 109, pp. 11854–11859 (2012 Jul.). http://dx.doi.org/10.1073/pnas.1205381109

[13] S. A. Shamma, M. Elhilali, and C. Micheyl "Temporal Coherence and Attention in Auditory Scene Analysis," *Trends Neurosci.*, vol. 34, pp. 114–123 (2011 Mar.). http://dx.doi.org/10.1016/j.tins.2010.11.002

[14] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem "Basic Objects in Natural Categories," *Cognitive Psychol.*, vol. 8, pp. 382–439 (1976 Jul.). http://dx.doi.org/10.1016/0010-0285(76)90013-X

[15] D. Dubois, C. Guastavino, and M. Raimbault, "A Cognitive Approach to Urban Soundscapes: Using Verbal Data to Access Everyday Life Auditory Categories, " *Acta Acust. United Ac.*, vol. 92, pp. 865–874 (2006 Nov.).

[16] B. L. Giordano, J. McDonnell, and S. McAdams "Hearing Living Symbols and Nonliving Icons: Category Specificities in the Cognitive Processing of Environmental Sounds," *Brain Cognition*, vol. 73, pp. 7–19 (2010 Jun.). http://dx.doi.org/10.1016/j.bandc.2010.01.005

[17] J. W. Lewis, J. A. Brefczynski, R. E. Phinney, J. J. Janik, and E. A. DeYoe "Distinct Cortical Pathways for Processing Tool versus Animal Sounds," *J. Neurosci.*, vol. 25, pp. 5148–5158 (2005 May). http://dx.doi.org/10.1523/JNEUROSCI.0419-05.2005

[18] B. Gygi, G. R. Kidd, and C. S. Watson "Similarity and Categorization of Environmental Sounds," *Percept. Psychophys.*, vol. 69, pp. 839–855 (2007 Aug.). http://dx.doi.org/10.3758/BF03193921

[19] C. Guastavino "Categorization of Environmental Sounds," *Can. J. Exp. Psychol.*, vol. 61, pp. 54–63 (2007 Mar.). http://dx.doi.org/10.1037/cjep2007006

[20] S. Payne, W. Davies, and M. Adams *Research into the Practical and Policy Applications of Soundscape Concepts and Techniques in Urban Areas* (Department of Environment, Food and Rural Affairs, London, 2009). pp. 30–35.

[21] W. J. Davies, M. D. Adams, N. S. Bruce, R. Cain, A. Carlyle, P. Cusack, D. A. Hall, K. I. Hume, A. Irwin, P. Jennings, et al., "Perception of Soundscapes: An Interdisciplinary Approach," *Appl. Acoust.*, vol 74, pp. 224–231 (2013 Feb.). http://dx.doi.org/10.1016/j.apacoust.2012.05.010

[22] O. Rummukainen, J. Radun, T. Virtanen, and V. Pulkki "Categorization of Natural Dynamic Audiovisual Scenes," *PloS One*, vol. 9, e95848 (2014 May). http://dx.doi.org/10.1371/journal.pone.0095848

[23] International Telecommunication Union, *"ITU-R BS.775-2, Multichannel Stereophonic Sound System with and without Accompanying Picture,"* (International Telecommunication Union, Geneva, 2006).

[24] A. P. M. Coxon *"Sorting Data: Collection and Analysis"* (Sage Publications, Thousand Oaks, 1999), pp. 1–104.

[25] J. H. Ward "Hierarchical Grouping to Optimize an Objective Function," *J. Am. Stat. Assoc.*, vol. 58 pp. 236–244 (1963). http://dx.doi.org/10.1080/01621459.1963.10500845

[26] I. Borg and P. J. Groenen *"Modern Multidimensional Scaling: Theory and Applications"* (Springer–Verlag, New York, 2005), pp. 3–14.

[27] J. M. Grey "Multidimensional Perceptual Scaling of Musical Timbres," *J. Acoust. Soc. Am.*, vol. 61, pp. 1270–1277 (1977 May). http://dx.doi.org/10.1121/1.381428

[28] S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, and J. Krimphoff "Perceptual Scaling of Synthesized Musical Timbres: Common Dimensions, Specificities, and Latent Subject Classes," *Psychol. Res.*, vol. 58, pp. 177–192 (1995 Dec.). http://dx.doi.org/10.1007/BF00419633

[29] M. R. Schroeder, D. Gottlob, and K. Siebrasse "Comparative Study of European Concert Halls: Correlation of Subjective Preference with Geometric and Acoustic Parameters," *J. Acoust. Soc. Am.*, vol. 56, pp. 1195–1201 (1974 Oct). http://dx.doi.org/10.1121/1.1903408

[30] E. Parizet, E. Guyader, and V. Nosulenko "Analysis of Car Door Closing Sound Quality," *Appl. Acoust.*, vol. 69, pp. 12–22 (2008 Jan.). http://dx.doi.org/10.1016/j.apacoust.2006.09.004

[31] J. B. Kruskal "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis," *Psychometrika*, vol. 29, pp. 1–27 (1964 Mar.). http://dx.doi.org/10.1007/BF02289565

[32] C. Guastavino *"Etude sémantique et acoustique de la perception des basses fréquences dans l'environnement sonore urbain,"* Ph.D. thesis, Université Paris 6 (2003).

[33] C. Guastavino and B. F. Katz "Perceptual Evaluation of Multi-Dimensional Spatial Audio Reproduction," *J. Acoust. Soc. Am.*, vol. 116, pp. 1105–1115 (2004 Aug.). http://dx.doi.org/10.1121/1.1763973

[34] F. Rumsey, S. Zielinski, R. Kassier, and S. Bech "Relationships between Experienced Listener Ratings of Multichannel Audio Quality and Naïve Listener Preferences," *J. Acoust. Soc. Am.*, vol. 117, pp. 3832–3840 (2005 Jun.). http://dx.doi.org/10.1121/1.1904305

[35] W. M. Rand "Objective Criteria for the Evaluation of Clustering Methods," *J. Am. Stat. Assoc.*, vol. 66, pp. 846–850 (1971). http://dx.doi.org/10.1080/01621459.1971.10482356

[36] P. Faye, D. Brémaud, M. D. Daubin, P. Courcoux, A. Giboreau, and H. Nicod "Perceptive Free Sorting and Verbalization Tasks with Naive Subjects: An Alternative to Descriptive Mappings," *Food Qual. Prefer.*, vol. 15, pp. 781–791 (2004). http://dx.doi.org/10.1016/j.foodqual.2004.04.009

APPENDIX

The clips for the radio drama material were taken from "The Hitchhiker's Guide to the Galaxy: Tertiary Phase. BBC, 2004" and "The Wonderful Wizard of Oz. BBC, 2009". The clips from "The Hitchhiker's Guide to the Galaxy: Tertiary Phase" occurred at approximately (min:sec) 05:00–06:30 (Episode 1), 14:40–17:40, 18:20–19:40 (Episode 2), 10:30–13:05, 18:20–20:20, 30:20–33:20 (Episode 3), 20:00 (Episode 4), and 24:50 (Episode 5). The clips from "The Wonderful Wizard of Oz" occurred at approximately 00:00–02:30, 04:15–5:37, 20:00–22:40, and 22:50–25:40.

The clips for the nature documentary material were taken from "Life. Episode 1, Challenges of Life. BBC, 2009." The clips used occurred at approximately 04:30–06:00, 11:19–15:30, 21:53–23:40, and 27:20–29:32.

The clips for the feature film material were taken from "The Woman in Black, 2012." The clips used occurred at approximately 04:40–6:31, 10:49–12:15, 14:35–16:36, 16:36–18:05, 19:51–23:38, 23:38–25:10, 45:43–46:57 , 53:09–54:19.

THE AUTHORS


James Woodcock


William J Davies


Frank Melchior


Trevor J Cox

James Woodcock is a research fellow at the University of Salford. His primary area of research is the perception and cognition of complex sound and vibration. James holds a B.Sc. in audio technology, a M.Sc. by research in product sound quality, and a Ph.D. in the human response to whole body vibration, all from the University of Salford. James is currently working on the EPSRC funded S3A project. His work on this project mainly focuses on the perception of auditory objects in complex scenes, the listener experience of spatial audio, and intelligent rendering for object-based audio.

•

Bill Davies is professor of acoustics and perception at the University of Salford. He researches human response to complex sound fields in areas such as room acoustics, spatial audio, and urban soundscapes. He led the Positive Soundscape Project, an interdisciplinary effort to develop new ways of evaluating the urban sound environment. Bill also leads work on perception of complex auditory scenes on the S3A project. He edited a special edition of Applied Acoustics on soundscapes, and sits on ISO TC43/SC1/WG54 producing standards on soundscape assessment. He is also an Associate Dean in the School of Computing, Science and Engineering at Salford, and Vice-President of the Institute of Acoustics (the UK professional body). Bill holds a B.Sc. in Electroacoustics and a Ph.D. in auditorium acoustics, both from Salford. He is the author of 80 academic publications in journals, conference proceedings, and books.

•

Frank Melchior received the Dipl.-Ing. degree in media technology from the Ilmenau University of Technology, Germany, in 2003 and the Dr.-Ing. degree from Delft University of Technology, The Netherlands, in 2011. Since 2012 he is leading the audio research group and the BBC Audio Research Partnership at BBC Research and Development. From 2009 to 2012 he was the Chief Technical Officer and Director Research and Development at IOSONO GmbH, Germany. From 2003 to 2009 he worked as a researcher at the Fraunhofer Institute Digital Media Technology, Germany. He holds several patents and has authored and co-authored a number of papers in international journals and conference proceedings. His research is currently focused on next generation audio for broadcast and interdisciplinary innovations for new audience experiences in an IP based broadcast world of the future. Dr. Melchior is member of the Audio Engineering Society, the German Acoustical Society, and represents the BBC in the International Telecommunication Union and the European Broadcasting Union.

•

Trevor Cox is Professor of Acoustic Engineering at the University of Salford and a past president of the UK's Institute of Acoustics (IOA). Trevor's diffuser designs can be found in rooms around the world. He is co-author of *Acoustic Absorbers and Diffusers* (3rd edition 8/16). He was awarded the IOA's Tyndall Medal in 2004. He is currently working on two major audio projects. www.goodrecording.net combines perceptual testing and blind signal processing to detect recording errors in user generated content. S3A is investigating future technologies for spatial audio in the home. Trevor was given the IOA award for promoting acoustics to the public in 2009. He has presented shows to 15,000 pupils including performing at the Royal Albert Hall. Trevor has presented 24 documentaries for BBC radio including "The Physicist's Guide to the Orchestra." For his popular science book *Sonic Wonderland* (in USA: *The Sound Book*), he won an ASA Science Writing Award in 2015. @trevor_cox