2017-11-21

Perception of Auditory Objects in Complex Scenes: Factors and Applications

William Coleman
Technological University Dublin, d15126149@mytudublin.ie

Recommended Citation

Coleman, W., Adams, L., Cullen, C., & Yan, M. (2017). Perception of Auditory Objects in Complex Scenes: Factors and Applications. In Institute of Acoustics - *21st Century Developments in Musical Sound Production, Presentation and Reproduction (pp. 1–16). Nottingham, UK; November 21st*, 2017. doi.org/10.21427/dfvv-bb72

Perception of Auditory Objects in Complex Scenes: Factors and Applications

William Coleman[1], Linda Adams[1], Dr. Charlie Cullen[2], and Dr. Ming Yan[3]

[1]*School of Media, Dublin Institute of Technology*
[2]*University of the West of Scotland*
[3]*DTS Licensing Ltd.*

December 5, 2017

Abstract

Over the past twenty years, technological advances have driven the development of media consumption in both home and mobile contexts. While not ubiquitous, multi-channel audio home cinema systems have become more prevalent, as has the consumption of broadcast and gaming media on smartphone and tablet technology via mobile telecommunications networks. This has created new possibilities and poses new challenges for audio content delivery such as how the same content can be presented to greatest effect given that it may be consumed via either a surround-sound home entertainment system or in a mobile context using stereo headphones. This paper outlines research into the development of strategies to optimise audio delivery for broadcast, gaming and music content using audio-object theory informed by principles of Auditory Scene Analysis (ASA). The initial experiment in this project is a listening test that focuses on subject evaluation of audio objects isolated from context. This experiment will explore inherent inter-object hierarchies of importance using a foreground-background evaluation task. An overview of the experiment will be offered with a summary of initial findings. We envisage further experiments to investigate how factors such as expectation may influence music scene analysis and how this knowledge might be used in object-based delivery scenarios.

1 Introduction

Recent research in object-based broadcasting [1] and auditory object categorisation [2] has underlined a growing interest in this area. ASA involves a constant activity of sound categorisation which Bregman [3] outlines as both a conscious (schematic or "top-down") and unconscious (primitive or "bottom-up") process of soundscape perception. This can be further illustrated, see Figure 1, by considering ASA as a constant analysis of the surrounding sound scene which involves continual innate identification of interesting sounds which may then be consciously analysed for semantic information or further meaning. This unconscious process of background sound monitoring continues while conscious attention is focused on foreground sounds. When sounds deemed worthy of conscious attention are identified they cease to be part of the background sound scene and become part of a foreground sound scene. There is considerable sensory research regarding soundscapes and how such attentional processes affect our perception of the environment. However, there is little based specifically on the hierarchy of sound objects in complex auditory scenes and on the movement of sounds from unconscious, background attention to foreground, conscious attention. With the move towards object-based sound delivery in visual streaming scenarios, an understanding of how auditory objects are parsed and categorised from auditory scenes will be useful in the development of strategies for sound file delivery.

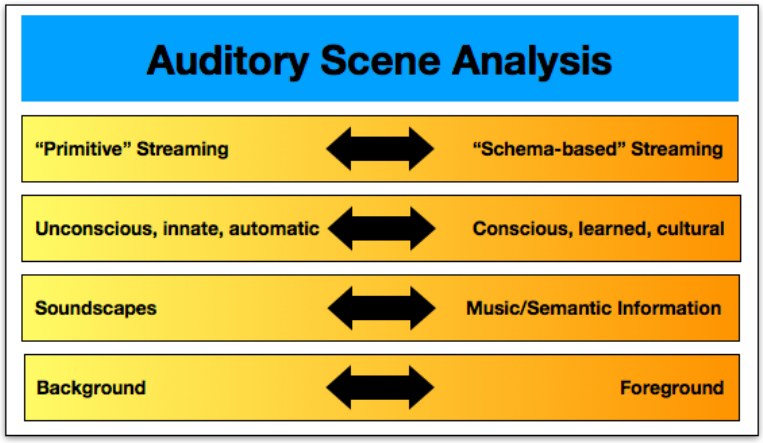

Figure 1: Auditory Scene Analysis

Many factors are known to influence the perception of sound. The physical properties of sounds [4], [5], the level of attention granted them by listeners [6], volume level [7], [8], proximity [9], sound event context [10], level of anticipation [11], prior training [12], [13] and experience [14], listening mode [10] and other senses (sight [15], [16], smell [17], touch [18]) are all known to affect our perception of sounds to some degree. However, the extent of the interaction of these factors, how they affect any inter-object hierarchy of importance and how this manifests in auditory scene perception is less well understood. Given that such factors are known to affect perception it is logical to assume that perception of such scenes may also differ based on the content type (e.g. music only versus broadcast and gaming content), or how it is being consumed (e.g. mobile consumption via stereo headphone versus home consumption via large screen devices with multi-channel audio). It is our hypothesis that perception of auditory scenes, and therefore any inter-object hierarchy, varies due to these factors.

In order to better understand the nature of these interactions a series of experiments are planned. Utilisation of stimuli analogous to visual streaming content is proposed as this is the most likely end-use of object-based audio in media consumption scenarios. As content type is one of the hypothesised parameters leading to change in perception, it is logical to isolate such a parameter for investigation where practicable.

In many forms of visual streaming content there is a direct linkage between visual and audio content. We see a referee blow their whistle on the screen and we hear it sounding from the television speakers. We see a movie character fire a gun and hear the gunshot. This is distinct from consumption of music content. Frequently, the only linkages between the music we hear and the visuals we see are imaginary, unless we happen to be watching a live musical performance where musician gesture is directly linked to the sounds perceived. However, much music may be heard without a direct visual analogy if listening on a bus, for instance, without a specifically designed visual accompaniment.

In light of the fact that vision is one of the factors known to affect perception of audio then it is logical to expect audio object perception of drama or sports content to differ from music perception, given that one contains direct linkages between visual and audio content and the other may not. We propose to make a distinction between visual streaming content which consists primarily of speech, sound effects and non-diegetic music (music whose source is not visible and is not implied to be present from the activity depicted on screen) and content which consists of music with and without accompanying visuals. For purposes of clarity we will refer to these divisions as the 'Speech/FX' and 'Music' research tracks respectively. This is illustrated in Figure 2.

For experimental purposes, it is considered that speech and effects predominant scenes which use primarily non-music stimuli will be analogous to visual streaming content such as drama or sports broadcasting and much computer game content. Stimuli selection will reflect this concern so as to maintain ecological validity of experiments.

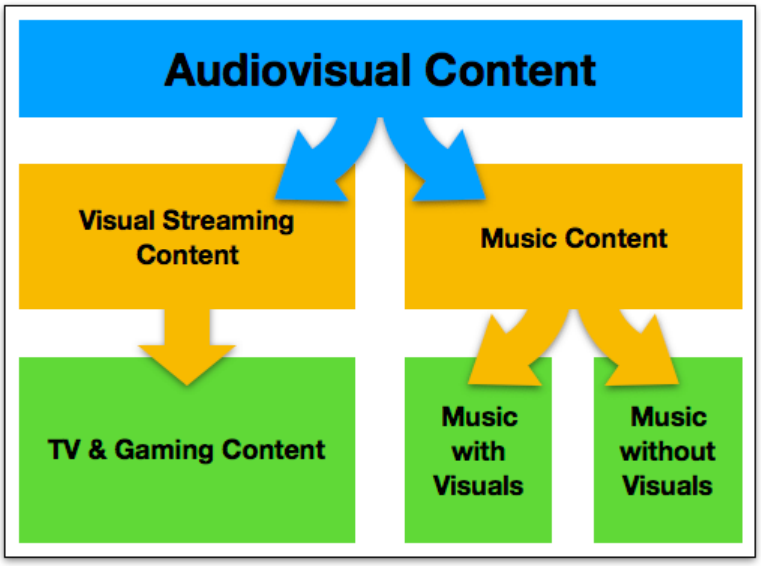

Figure 2: Illustration of research tracks

The following sections contain a brief summary of progress and current state of research for both the Speech/FX and Music tracks. This will entail a review of relevant prior art covering Speech/FX concerns and a brief description of an initial background versus foreground categorisation experiment. Summary findings will also be offered. This will be followed by a literature review of relevant research for the Music research path with details on a proposed experiment. Finally, future research intentions will be outlined.

2 Auditory Scene Analysis

A comprehensive literature review of existing sound taxonomies and soundscape research was undertaken to establish current practice in complex auditory scene perceptual research and ascertain what principles could be applied to audio object perception. A general summary of this review will be useful to frame discussion of salient points.

Bregman [3] has described ASA as the process by which auditory scenes are parsed into individual sounds which we are referring to as auditory 'objects'. This is a complex task because sounds are interleaved and overlap in both temporal and frequency domains, and the human auditory system only has access to an amalgam of all sounds that are presented to the ear at the same time. Bregman describes processes of sequential and simultaneous grouping where perception is governed by low-level primitive and high-level schematic structures that parse the sound scene presented to the ear for individual objects.

Sequential grouping occurs when similarities in sounds from one moment to the next result in them being grouped to form a 'stream'. This is demonstrable via variations in tempo, frequency, timbre, spatial direction and duration of exposure (what Bregman describes as 'cumulative effects' [19]). Simultaneous grouping occurs when properties of the sound scene match patterns that tend to be true when components of sound come from the same source. If a subset of frequencies are detected that are all multiples of a common fundamental, this suggests that the subset is from a common source. Sounds which have a different fundamental frequency tend to be segregated and be considered separate sounds. Periodic sounds, such as the human voice and many musical instruments, are an example of this phenomenon. Other factors known to aid sequential grouping are sound onset/offset synchrony, frequency components that come from the same spatial location, components which have the same pattern of fluctuation and also, those that are close together in frequency.

Both forms of grouping are functions of primitive and knowledge-based processes (see [20] & [21]).

We have referred to primitive processes as ”bottom-up”, unconscious processes which are thought to be innate, have been found in non-human animals [22], in the perception of speech [23] and of music [3]. We have referred to knowledge-based processes as schematic, or ”top-down” processes which involve conscious attention or past experience [24].

These functions in and of themselves do not approach the question of foreground/background categorisation which is central to our research. To this end, an overview of sound categorisation examples will be illustrative of how sound objects can be organised and how salient such organisations may be to foreground/background allocation. Central to this is the validity of arbitrary sound categorisation as such structuring will be intrinsic to planned experiments on inter-object hierarchies.

2.1 Sound Taxonomies

Gaver [25] outlines the taxonomy reproduced in Figure 3, which he presents as a simple map of sound events distinguished by classes of materials and by interactions which may cause them to sound. Gaver further suggests the thesis, supported by [26], that everyday listening, or “the experience of listening to events rather than sounds” (pg. 2), focuses on acoustic factors most useful for source identification, as distinct from musical listening, where the “perceptual dimensions and attributes of concern have to do with the sound itself” (pg. 1). This separation also supports the division between perception of musical and non-musical scenes proposed in the introduction. It is interesting to note that this taxonomy is outined according to qualities of the sounds themselves rather than the objects which produce the sounds, a facet which is prevalent in more recent similar taxonomies.

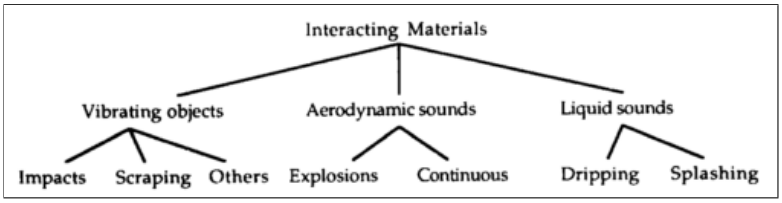

Figure 3: Gaver’s simple taxonomy of sound events, reproduced from [25]

R. M. Shafer outlines an extensive catalogue of sound types as used in the World Soundscape Project in [27]. The organisation used in the catalogue is arbitrary, but also comprehensive, having been built up over a period of years, and is empirically derived. Regarding the bias inherent in any such organisation of objects, Shafer makes the point that ‘the only framework inclusive enough to embrace all man’s undertakings with equal objectivity is the garbage dump’ [ibid., pg. 137]. An illustration of the broadest categories of sounds is offered in Figure 4.

Natural Sounds: Bird, chicken, rain, sea shore	Human Sounds: Laugh, whisper, shouts, talk, cough	Sounds & Society: Party, concert, grocery store	Mechanical Sounds: Engine, cars, air conditioner	Quiet & Silence: Wild space, silent forest	Sounds as Indicators: Clock, doorbell, siren

Figure 4: Categories of sounds used for the World Soundscape Project

A more recent example of such organisations is offered by Gemmeke et al. [28] which consists of a dataset of sounds[1] manually curated from over 2 million YouTube[2] videos. These events are organised using a hierarchically structured ontology of 632 audio classes the top-level structure of which is outlined in Figure 5.

[1] https://research.google.com/audioset/ontology/index.html
[2] http://www.youtube.com

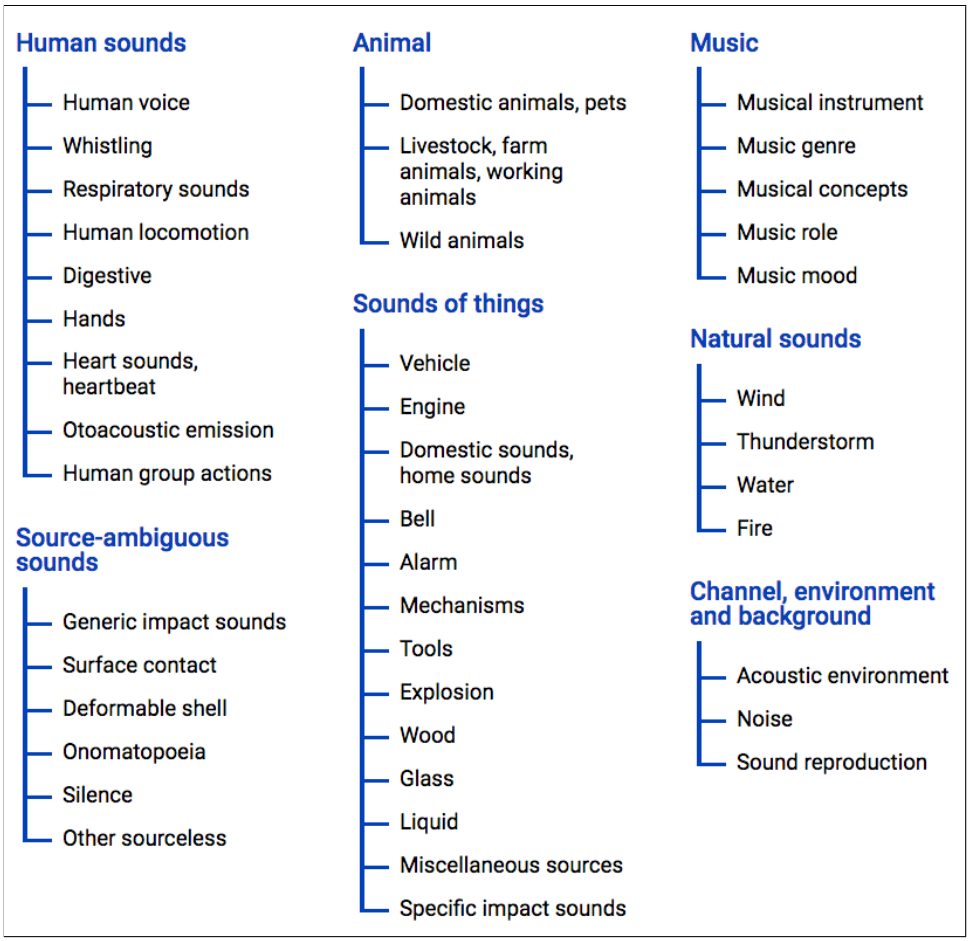

Figure 5: Top-level organisation of the Audioset dataset

These taxonomies reflect the arbitrary nature of the sound categorisation task but are indicative of the general principles used in the research and reflect much of the subsequent literature. They illustrate that multiple approaches are possible, though consistency can be observed. Note the presence of 'human sound' and 'natural sound' categories in both the taxonomy of Shafer and the ontology from Gemmeke et al. We envisage structures similar to these being useful in content organisation for subsequent experiments.

2.2 Sound Categorisation & Listening Modes

Concerning the validity of such structures, Thorogood et al. [29] examine the consistency of an arbitrary background/foreground categorisation of sounds drawn from the World Soundscape Project Tape Library (WSPTL) [30]. Subjects' were asked whether they agreed or disagreed with the categorisation provided by the WSTPL. Strong levels of consensus were observed between study participants and the arbitrary tagging of the WSPTL on what constitutes a foreground sample (80%), background sample (92%) and background with foreground samples (75%).

This supports the view that foreground/background categorisation of a sound can be established with a reasonable degree of confidence, with the caveat that this could not be considered a universal, unchanging categorisation and that caution should be exercised. We could further extrapolate from this data that a consensus on what constitutes a background sound is easier to arrive at than a consensus on what constitutes other categories of sound. We will see subsequently in Section 3 that this is not necessarily always the case. Foreground/background categorisation, in other words, retains a somewhat subjective nature, dependant on other factors.

Thorogood et al. make several further points about the nature of the foreground/background categorisation task which are worthy of mention:

- Dependant on context, sounds can be classed either as either foreground or background, an observation consistent with [31]. A drop of water in a bath tub could be a foreground sound, for instance, whereas a drop of water in the ocean is more likely to be a background sound.
- Attention is a significant factor in the foreground/background classification task. Sound from the TV is foreground when the user is paying attention to the program, but shifts to background if the focus becomes a conversation with a person in the next room.
- Background sounds either seem like they are further away than foreground sounds or are unchanging to the extent that they blend into the rest of background noise.
- Ubiquitous sound can be thought of as the background quality of a soundscape. As summarised by [32], sound can seem to come from everywhere and nowhere, from a single source and from many sources. Urban drones and the sound of insects are two examples of such ubiquitous sounds.
- Foreground sounds can be said to stand out clearly from the background.
- Listening, as outlined by Truax [10], Chion [33] and Wolvin & Coakley [34] is a dynamic process of numerous listening modes, which can treat a sound as background or foreground depending on the amount of attention being paid to the sound. A useful summary of various listening modes outlined in the literature is adapted from [35] and mapped against a background to foreground scale in Figure 6.

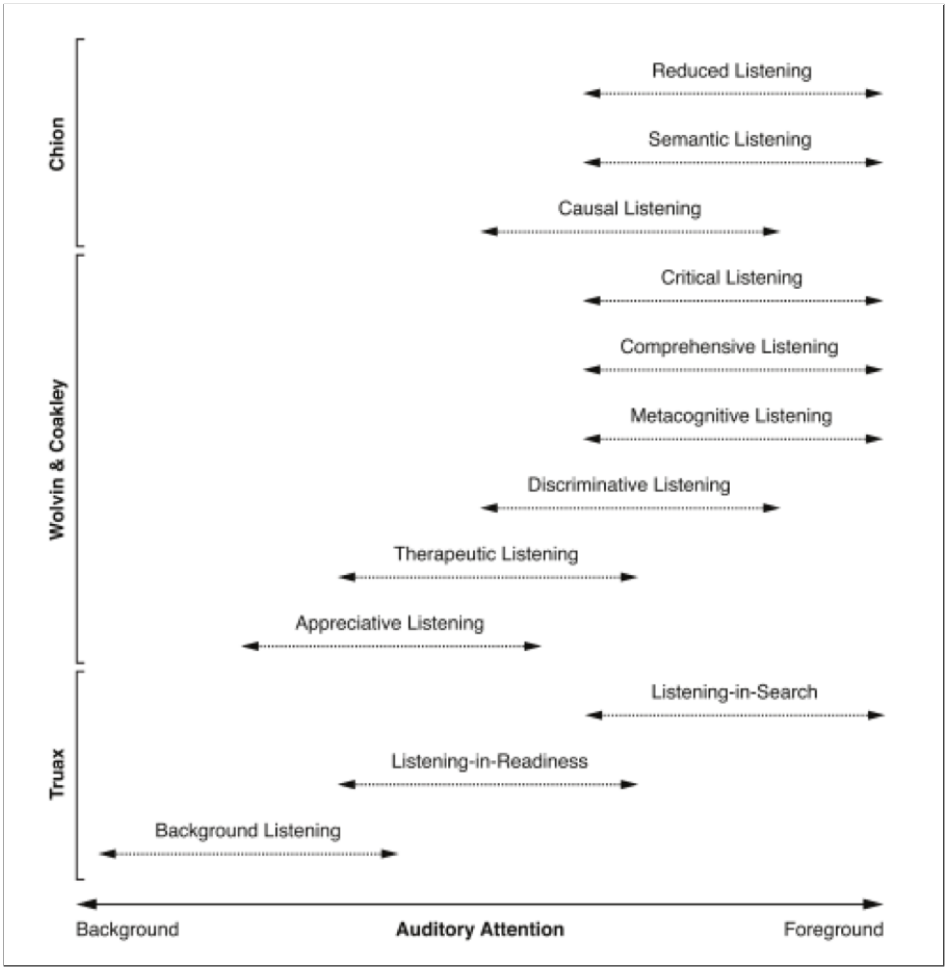

Figure 6: Summary of listening modes as outlined in [35]

Truax's [10] listening modes provide a useful framework of different levels of auditory perception. While subliminal auditory perception is acknowledged as controversial, Kotzé & Moller [36] note significant Galvanic Skin Response (GSR) to subliminal auditory stimuli. Dupoux et al. [37] suggest that conscious and unconscious processing are distinguished by 'high-level perceptual streaming factors' rather than stimulus energy and duration. This is distinct from Truax's description of 'background listening', which he posits as a level of sounds which we are aware of, though not actively listening for. Referring to the Gestalt example of figure/ground perception, sounds that we are actively paying attention to can be thought of as 'figure' sounds, while others form the 'ground'. The 'cocktail party effect' [38] highlights the ability of the auditory system to pull different auditory objects in and out of focus as required. This frames background listening as a complex process of constant evaluation and re-evaluation of the auditory scene, where objects are continually evaluated as to whether or not they are worthy of greater attention. 'Listening-in-readiness' is described as being an intermediate mode of listening where familiar sounds, such as the sound of our own name, are continually monitored while primary attention is focused elsewhere. Truax highlights the classic example of the parent capable of sleeping through traffic noise who wakes at the sound of their baby crying. 'Listening-in-search' is when listening is most analytical, where the sound itself is searched for meaning. This is illustrated by the cocktail party effect, where a conversation within one group can be focused on to the exclusion of the conversations of others.

2.3 The Foreground/Background Categorisation Task

Framing our investigation of the foreground/background categorisation task through the listening modes of Truax, this positions the categorisation of auditory objects as fluctuating due to perceived importance relative to activity in the observed scene. Existing studies of sound categorisation have been reviewed to establish what consistencies may be observed in subject approach to such a task. Dimensions of such a categorisation-space will be useful in the formulation of any rule-set to predict sound object foreground/background ranking.

Lewis et al. [39] present a study where subjects were asked to rank sounds as either 'object-like' or 'scene-like'. In general, mechanical sounds tend to be ranked as more 'object-like' than environmental sounds and vice versa. Additionally, 'scene-like' sounds tend to have a more gradual change characteristic, differentiating continuous sounds from those with more abrupt change characteristics. In a study investigating the categorisation of broadcast audio objects [40], Woodcock et al. identified three dimensions in sound object categorisation using multidimensional scaling (MDS). One of these dimensions ranged between continuous and discrete impact sounds. Another was proposed to be related to the presence of absence of humans. A third dimension progresses from continuous background sounds to clear speech. The authors maintain that this dimension is related to whether the sound carries semantic meaning or not, which is mirrored in neuro-cognitive studies such as [41] & [39]. Interestingly, subjects' perceived importance of sound objects correlated with this dimension, suggesting that sound objects which carry semantic information are more important than those which do not. Collett et al. [42] found that musical and vocal stimuli were easier to categorise than environmental sounds which, supported by [25], [26] & [43] suggests that sound categorisation is easier when more semantic information is discernible from the sound. Additionally, Guastavino [44] suggests that people organise sounds and soundscapes in terms of the meaning attached to a sound as a semantic clue to source identification as opposed to any abstract physical property of the sound.

Gygi & Shafiro [45] demonstrate what they term an incongruency advantage by showing that sounds perceived as out of place in an auditory scene are more likely to be noticed. This is supported by Sussman-Fort & Sussman [46], who suggest that the auditory system maintains a representation of the environment that is only updated when new information indicates that reanalysing the scene is necessary. This is consistent with Rummukainen et al. [24] who find that humans are attentive to perceived movement, noisiness and eventfulness when analysing real-life urban environments. They note that arousal can affect selective attention, increasing focus on certain sounds to the detriment of attention paid to others.

Salamon et al. [47] present a taxonomy of urban sounds which they have labelled with a saliency

characteristic, which indicates a subjective labelling of the sound on a foreground/background scale. A subsequent categorisation experiment found that background sounds were significantly more difficult to identify than foreground sounds with only one exception – a siren noise. This suggests that subjective labelling, if done with care, is a robust mechanism for sound categorisation.

Guastavino [20] suggests that sounds are either classified into taxonomic categories ('car', 'truck', 'street', 'acceleration') according to low-level features or into script categories ('doing the groceries', 'taking a walk', 'having a drink') according to high-level features concerned with the situation of use or the end-use purpose of the object. Raimbault & Dubois [48] support the idea that certain noises are identified in terms of the semantic content the sound suggests, and also outline research that suggests that psychological and sociological factors can affect sound scene perception. They suggest that street scenes which look 'pretty' will generally be thought to sound more pleasant than those which do not.

2.4 A Framework for Future Investigation

From similarities in groupings observed in these sources, the author has derived a series of axes which outline relationships between sounds and define the dimensions of a categorisation space for object classification. These may be of use in investigating the fluctuation of relative importance between sounds as a function of time and are outlined in Figure 7.

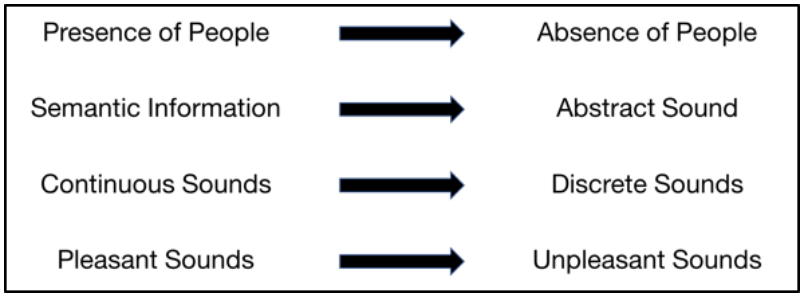

Figure 7: Sound object relationship axes to guide future study

The first of these axes reflects dimensions of sounds which range from those that suggest the presence of humans, to those that do not. The second axis outlines the difference between sounds which carry a high degree of semantic information about the object, action or event that caused their creation, and sounds that do not. This axis could also be referred to as 'sounds that are often described by the event that caused them', and thus easier to identify, versus 'sounds that are often described using some abstract quality of the sound itself', which are more difficult to identify. The next axis concerns continuous sounds (more likely to be background and harder to identify) versus discrete sounds (connected to an object or event, easier to identify). The final axis outlines pleasant (people, nature, music, harmonic, lively ambiances) versus unpleasant (traffic, alerts, inharmonic, alert) sounds.

For the initial experiment of the current research it was decided to investigate sound objects in isolation to determine what hierarchy, if any, exists in this state. The complex nature of auditory scenes means that many factors have an influence on perception of a scene and the categorisation of objects within it. The foreground/background nature of isolated sounds can then be used to inform a test set of sounds to be used in further experiments investigating these factors. The following sections will give an overview of the experimental design of the initial experiment and will then briefly outline initial results.

3 An Initial Experiment Using Non-Music Stimuli

Listening tests generally focus on one of two broad areas of research. The first of these, which we will refer to as categorisation experiments, is broad auditory scene analysis – how do we perceive and

parse sound scenes? The second, which we will term evaluation experiments, generally investigate the perceived basic audio quality (BAQ) of system components. Compression codecs, loudspeakers and microphones have all been the subject of such evaluation research.

Methods differ somewhat between these two purposes. The first is generally related to the process of our perception of sound and has given rise to a variation of experimental design approaches. The most prominent proponent of this research is Albert Bregman [19], who investigated our perception of auditory 'streams' using a series of experiments that often made use of synthetic tones to establish the basic principles of auditory scene analysis. See [49] for a review of recent research in the area. In a similar vein, soundscape research investigates human perception of complex sound scenes to evaluate how such scenes, and the audio objects that comprise them, are perceived and categorised by listeners.

Evaluative research, often based around ITU test methods and standards (e.g. BS.1116-3 [50] & BS.1534-3 (MuSHRA) [51]), is concerned with evaluation of some element of a sound delivery system. Such tests are used to rate factors such as headphones, loudspeakers or audio compression codecs and are generally interested in forensically parsing audio stimuli to detect fractional differences between the factors under investigation to determine which is superior. The stimuli used in such experiments generally reflect the intended end use of the factor under investigation, so a listening test comparing headphones, for example, will often use popular music for stimuli.

For the initial experiment, it was decided to use a foreground/background categorisation task with non-musical, isolated sound objects as stimuli. We hope to gain an insight into what, if any, inherent hierarchy exists between sounds when they are removed from the context of an auditory scene. As the forensic level of detail afforded by the BS.1116-3 and MuSHRA standards was deemed inappropriate in this instance the experiment was conducted online using a purpose-built website where all subjects were asked to complete the test using headphones in a quiet environment. It has been found that there is minimal difference between laboratory and online experiments for comparable tests ([52] & [53], for example). Subjects were required to submit basic demographic information and then rate 40 sounds in a background – neutral – foreground evaluation task. If unsure as to whether a stimulus was background or foreground subjects were advised to mark the sound as neutral. Stimuli were sourced from the ESC-50 [54] sound set and presentation was randomised so as to minimise presentation order effects. A total of 110 complete test results were collected.

Subject scores were collated for each stimulus. It was found that there were several sounds which subjects deemed strongly foreground or background when isolated from context. Most sounds however exhibited no clear consensus as to their position on this axis, possessing a slight majority for one position with a significant minority for the opposite position. Figure 5 shows the proportion of scores for each stimulus. Four sounds could be said to be strongly foreground; 'Clock Alarm', 'Glass Breaking', 'Baby Crying' and 'Door Knock'. The consensus is not so strong for background sounds, but the 'Crickets', 'Clock Tick', 'Keyboard Tapping', 'Fire' and 'Birds' sounds have the highest background scores.

This suggests that, while some sounds have an inherent property which ranks them firmly either foreground or background, many sounds are capable of being ranked between either extreme, dependant on other factors. It is expected that by manipulating such factors a map can be created that illustrates how such foreground to background fluidity may function.

3.1 Summary

The previous sections have outlined general considerations for ASA as they pertain to soundscape research and sound object categorisation and have detailed an experiment into the foreground background categorisation task for isolated sound objects. A series of factors which affect auditory perception were outlined and relevant research reviewed. These factors include loudness, context, attention, prior experience, training, other senses, expectation and others. The influence and interaction of these factors on the perception of auditory scenes is complex, as is evident from a consideration of how anticipation and expectation may play a role in our perception of musical stimuli, which will be addressed

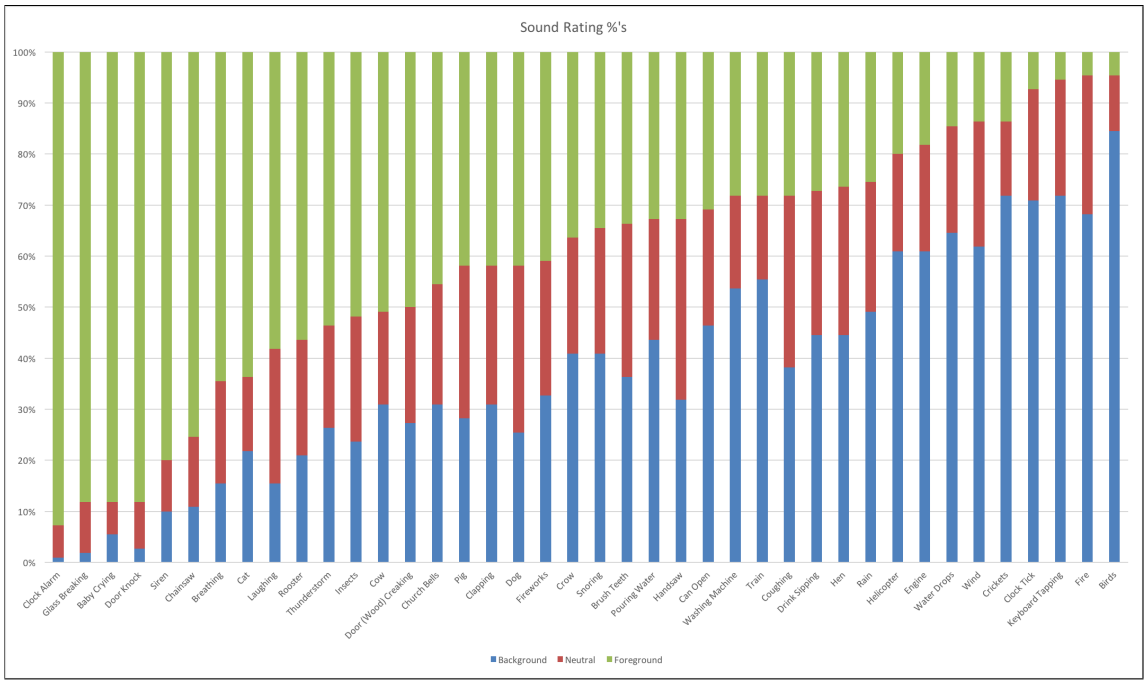

Figure 8: Foreground - Neutral - Background ratings of audio objects

in the next section. An overview will now be offered of considerations relevant to a planned experiment using musical stimuli to investigate how factors such as attention and expectation can affect our perception of such scenes. Factors relevant to how inter-object importance may change in a musical context will be reviewed.

4 ASA and Music

In his book 'Auditory Scene Analysis', Albert Bregman distinguishes between two distinct forms of Auditory Scene Analysis (ASA): automatic streaming and schema-based streaming [3]. The former encompasses processes that are universal and innate, and that occur without conscious effort. The latter describes processes that are the result of learning, or concentrated attention. Both are involved in the cognition of music [55], affecting perception of melody, rhythm and emergent properties of these such as tonality and harmony. Automatic ASA may affect perception of melody via Gestalt principles of continuity and proximity. Several studies have found that melodic coherence is dependent on the frequency proximity of melodic tones [56], and how this frequency proximity competes with contour [57], tempo, and rhythm [58]. These Gestalt principles create innate and automatic expectations for melodic continuity within particular frequency ranges, or along established trajectories. Rhythm is affected by automatic streaming in that rhythms tend to emerge within auditory streams, and listeners have difficulty comprehending rhythms across streams [59]. Listeners may use rhythmic cues in a top-down manner to differentiate interleaved melodies in order to facilitate streaming [60].

The perception of low-level elements of music may be governed by innate automatic processes, but comprehension of music as an emergent property of combinations of partials and their temporal organisation involves schema-driven processes, namely the input of knowledge and expectancies acquired through exposure to a particular musical culture. This is demonstrated by studies that have shown that listeners' musical expectancies are primarily derived from the culture with which they are familiar [61] [62], and evidence from research into musical development showing that although infants demonstrate melodic and rhythmic expectancies based on Gestalt principles [63], knowledge of harmony and tonality do not emerge until there has been some exposure to a musical culture [64]. The following section will discuss a planned experiment exploring how musical analysis of tonality and harmony is performed

within this schema-driven ASA process, and how it may affect the foreground/background positioning of music in an auditory scene.

4.1 Schema-driven ASA and Expectation

Schema-driven, "top-down" attention affects both the streaming of auditory objects [65] and the allocation of objects into foreground/background categorisations [66]. It has been found that the integration/segregation of sequential sounds into streams can be affected by how the listener directs their attention [59] [67]. These findings are corroborated by neuro-physiological studies that have found ERP enhancement when subjects attend to auditory objects [68] and enhanced neural representation of intended targets when attention is focused on them [69]. Dowling has found that the focus of attention on musical stimuli may create "expectancy windows" during which musical processing is enhanced [70]. Woodward et al. and Schröger have found that unexpected auditory objects divert attentional resources in a "bottom-up" manner, increasing reaction times and lowering performance levels on concurrent tasks [71] [72].

Several studies have explored this phenomenon within music scene analysis, and have found increased reaction times and attentional diversion for deviant harmonic and melodic stimuli within music. Reactions to deviant musical stimuli are related to musical expectation, which are themselves dependent on the listener's familiarity with a particular musical culture. Trainor and Trehub have found that although very young infants do not differentiate between tonal and non-tonal melodic violations [63], by age 4, children have expectations for melodies to continue within a fixed tonality [73]. By the age of 7, children can detect melodic violations that imply deviant harmony, suggesting that they have expectations for particular harmonic patterns within a given diatonic context [64]. In a study by Schellenberg et al., children's reaction times in timbre judgement tasks were slower when associated chords violated Western harmonic conventions [74]. Adults with no musical training have been found to have considerable implicit knowledge of the rules of Western tonality and harmony, and strong expectations that music will follow these rules. Several studies have found behavioural and neurophysiological reactions to violations of tonal and harmonic rules in adults with no explicit knowledge of music theory [75] [76] [13]. These reactions to harmonic deviations demonstrate how important expectation is in our understanding of music. Meyer and others have theorised that contextual violations of expectation may even be linked to aesthetic reactions to music [77] [11].

If violations of musical expectation are indeed related to aesthetic reactions to music, then it could be assumed that aesthetic enjoyment of a piece would be reduced on repeated exposure, given that we would be aware of upcoming deviations in advance. However, the opposite appears to be case; enjoyment of a piece of music tends to increase with repeated exposure [78] [79]. Paradoxically, musical expectations have been found to be consistently violated even within musical pieces that a listener may be familiar with [80]. Research has found that repeated exposure to deceptive harmonic cadences does not significantly diminish the increase in reaction time that is typically found for such deviations, suggesting that the attentional resources required to process the deviation are not influenced by previous exposure [81] [82]. This may be because the veridical expectations associated with specific pieces of music are processed separately to schematic expectations arising from knowledge of a musical framework.

The experiment in preparation will use a cross-modal paradigm to explore the extent to which attention is affected by previously untested harmonic deviations commonly found in contemporary music, such as deceptive resolution of secondary dominants, dominant modulations and modal interchange. Unexpected harmonic changes are dependent on a listeners' (implicit) schematic knowledge of contemporary Western harmony, while reactions to unexpected changes in loudness, timbre and tempo derive from more innate, automatic processes. Therefore, reactions to violations of loudness, timbre and tempo will be examined and compared to reactions to harmonic violations. The results of these tests may indicate which factors influence foreground/background categorisation of music in an auditory scene. Since musicians' schematic musical knowledge is likely to be more comprehensive than non-musicians', these two groups will be compared to determine if this has any effect on attention.

The effects of veridical expectations will be studied to determine if repeated exposure affects reaction time to harmonic violations in comparison to tempo/timbre/loudness violations in musicians and non-musicians.

The results of this experiment may yield information on how attentional resources are dynamically allocated within a musical auditory scene, and may provide initial guidelines for how bandwidth should be assigned to music within a broader auditory scene in an object-based audio context.

5 Conclusions

This paper has considered factors relevant to the functioning of auditory objects in auditory scenes comprising of both non-music and music-only elements. A research review has identified existing methods and practices which are relevant to currently planned research and which have informed the design of an initial experiment on the non-music research track. It is intended to use these findings to further investigate the interaction of the identified factors in complex auditory scenes comprising both non-music and music content.

6 Future Work

This research will inform the development of a set of rules which will describe how inter-object hierarchies of importance within auditory scenes change over time. This rule set will be used to formulate codecs for use in the generation of audio content for different media forms and for differing consumption paradigms. With regard to predominantly Speech & FX scenes, it is envisaged that audio objects tagged with appropriate metadata can be used to vary the delivery of audio over time as is deemed optimal depending on content type (broadcast, game or music audio), end-user configurations (stereo, headphones or multi-channel) and other factors (varying bandwidth capacities, individual preferences and differing environments). A test codec will then be validated using an environment which simulates the consumption of different media forms and delivery modalities.

An experiment investigating background/foreground categorisation of isolated objects has been briefly described. To further development of a codec for object-based delivery of audio content we envisage a series of experiments to investigate how audio object importance can change over time. This experiment series will consider music content separately from visual streaming (predominantly speech & FX) content.

Our goal with the initial experiment in the Speech/FX research track was to establish what, if any, inherent foreground/background ranking is evidenced by single sound objects in isolation. Potential factors for future investigation include the physical properties of sounds, attention, volume, proximity, context, anticipation, prior training & experience, & other senses (sight, smell & touch). A similar progression is envisaged for the Music research track, taking cognisance of the differing factors at play in the perception of music. The influence of anticipation as mediated by schematic and veridical expectations is only one possible route of enquiry. The insights thus gained will be critical in developing delivery strategies for broadcast, game, music and other forms of audio content.

Subsequent experiments will seek to build an understanding of these interactions which will inform development of a matrix that outlines how inter-object importance changes over time. The validity of this matrix will then be tested by applying it in an environment which investigates subjects' perception of auditory scenes which have been manipulated subject to the parameters of the matrix.

Acknowledgements

This work was supported by the Irish Research Council and DTS Licensing Ltd. under project code EBPPG/2016/339. We would also like to gratefully acknowledge the input of our supervisors in the preparation of this paper.

References

[1] Tony Churnside. *Object-Based Broadcasting*. 2013. (Visited on 10/27/2017).
[2] J Woodcock, W J Davies, and T J Cox. "A Cognitive Framework for the Categorisation of Auditory Objects in Urban Soundscapes". In: *Applied Acoustics* 121.2017 (2017), pp. 56–64.
[3] Albert S Bregman. *Auditory Scene Analysis: The Perceptual Organisation of Sound*. Cambridge, MA: The MIT Press, 1990.
[4] Stephen Handel. *Listening: An Introduction to the Perception of Auditory Events*. Cambridge, MA, USA: The MIT Press, 1989.
[5] Stephen Handel. "Timbre Perception and Auditory Object Identification". In: *Listening*. Ed. by Brian C. J. Moore. London, UK: Academic Press, 1995. Chap. 12.
[6] Shihab Shamma et al. "Temporal Coherence and the Streaming of Complex Sounds". In: *Basic Aspects of Hearing: Physiology and Perception*. Ed. by Brian C. J. Moore et al. Vol. 787. 765. New York, USA: Springer-Verlag, 2013. Chap. 59, pp. 109–118.
[7] J. C. Webster and P. O. Thompson. "Responding to Both of Two Overlapping Messages". In: *The Journal of the Acoustical Society of America* 26.3 (1954), p. 396.
[8] Irwin Pollack and J.M. Pickett. "Cocktail Party Effect". In: *Journal of the Acoustical Society of America* 29.11 (1957), pp. 1262–1262.
[9] Catherine Lavandier, Catherine Lavandier, and Boris Defréville. "The Contribution of Sound Source Characteristics in the Assessment of Urban Soundscapes". In: *Acta Acustica united with Acustica* 92 (2006), pp. 912–921.
[10] Barry Truax. *Acoustic Communication*. Ed. by Melvin J. Voigt. 1st. Norwood, NJ, USA: Ablex Publishing Corporation, 1984, xxi, 244 p.
[11] David Huron. *Sweet Anticipation: Music and the Psychology of Expectation*. Cambridge, MA, USA: The MIT Press, 2006.
[12] Søren Bech. "Selection and Training of Subjects for Listening Tests on Sound-reproducing Equipment". In: *Journal Audio Eng. Soc.* 40.7/8 (1992), pp. 590–610.
[13] E Bigand and B Poulin-Charronnat. "Are We "Experienced Listeners"? A Review of the Musical Capacities that do not Depend on Formal Musical Training". In: *Cognition* 100.2006 (2006), pp. 100–130.
[14] Stephen McAdams. "Recognition of Sound Sources and Events". In: *Thinking in Sound: The Cognitive Psychology of Human Audition*. Ed. by Stephen McAdams and Emmanuel Bigand. Oxford, UK: Clarendon Press, 1993. Chap. 6, pp. 146–198.
[15] Jesper Udesen, Tobias Piechowiak, and Fredrik Gran. "Vision Affects Sound Externalization". In: *Proceedings of the 55th International Conference of the Audio Engineering Society*. Helsinki, Finland; August 27-29, 2014, pp. 1–4.
[16] Kurtis G Gruters et al. "The Eardrum Moves when the Eyes Move: A Multisensory Effect on the Mechanics of Hearing". In: *bioRxiv* 156570 (2017).
[17] Jin Yong Jeon et al. "Non-auditory Factors Affecting Urban Soundscape Evaluation". In: *The Journal of the Acoustical Society of America* 130.6 (2011), pp. 3761–3770.
[18] Jochen Steffens, Daniel Steele, and Catherine Guastavino. "Situational and Person-related Factors Influencing Momentary and Retrospective Soundscape Evaluations in Day-to-day Life". In: *The Journal of the Acoustical Society of America* 141.3 (2017), pp. 1414–1425.
[19] Albert S Bregman. *Auditory Scene Analysis*. 2004.
[20] Catherine Guastavino. "Categorization of Environmental Sounds". In: *Canadian Journal of Experimental Psychology* 61.1 (2007), pp. 54–63.

[21] Claude Alain, S R Arnott, and T W Picton. “Bottom-up and Top-down Influences on Auditory Scene Analysis: Evidence from Event-Related Brain Potentials”. In: *Journal of Experimental Psychology: Human Perception and Performance* 27.5 (2001), pp. 1072–1089.

[22] Amy B. Wisniewski and Stewart H. Hulse. “Auditory Scene Analysis in European Starlings (Sturnus Vulgaris): Discrimination of Song Segments, their Segregation from Multiple and Reversed Conspecific Songs, and Evidence for Conspecific Song Categorisation.” In: *Journal of Comparative Psychology* 111.4 (1997), pp. 337–350.

[23] Chris J. Darwin and R. P. Carlyon. “Auditory Grouping”. In: *Handbook of perception and cognition: Hearing.* Ed. by B. C. J. Moore. 2nd. London, UK: Academic Press, 1995. Chap. 11, pp. 387–424.

[24] Olli Rummukainen et al. “Categorization of Natural Dynamic Audiovisual Scenes”. In: *PLoS ONE* 9.5 (2014), p. 14.

[25] William W. Gaver. “What in the World do we Hear?: An Ecological Approach to Auditory Event Perception”. In: *Ecological Psychology* 5.1 (1993), pp. 1–29.

[26] Brian Gygi, Gary R. Kidd, and Charles S. Watson. “Similarity and Categorization of Environmental Sounds”. In: *Perception & Psychophysics* 69.6 (2007), pp. 839–855.

[27] Raymond Murray. Schafer. *The Soundscape: Our Sonic Environment and the Tuning of the World*. Rochester, Vermont: Destiny Books, 1994.

[28] Jort F Gemmeke et al. “Audio Set: An Ontology and Human-Labeled Dataset for Audio Events”. In: *Proc. IEEE ICASSP 2017, New Orleans, LA (to appear)*. New Orleans, LA, USA; March 5-9, 2017.

[29] Miles Thorogood, Jianyu Fan, and Philippe Pasquier. “Soundscape Audio Signal Classification and Segmentation Using Listener’s Perception of Background and Foreground Sound”. In: *PAPERS Journal of the Audio Engineering Society* 64.7/8 (2016), pp. 484–492.

[30] Barry Truax. *World Soundscape Project Tape Library*. 2015. (Visited on 03/07/2017).

[31] A. L. Brown, Jian Kang, and Truls Gjestland. “Towards Standardization in Soundscape Preference Assessment”. In: *Applied Acoustics* 72.6 (May 2011), pp. 387–392.

[32] Jean-Francois Augoyard and Henri Torgue. *Sonic Experience: A Guide to Everyday Sounds*. London, UK: McGill-Queen’s University Press, 2005.

[33] Michel Chion. *Audio-Vision: Sound on Screen*. Ed. by Claudia Gorbman. New York, NY, USA: Columbia University Press, 1994, pp. 25–34.

[34] Andrew D. Wolvin and Carolynn Gwynn Coakley. “A Listening Taxonomy”. In: *Perspectives on Listening*. Ed. by A. D. Wolvin and C. G. Coakley. Norwood, NJ, USA: Ablex Publishing Corporation, 1993, pp. 15–22.

[35] John Mcgee and Charlie Cullen. “The Role of Semantic Processing in the Allocation of Auditory Attention in Competitive Acoustic Scenarios”. In: *13th Annual Auditory Perception, Cognition and Action Meeting (APCAM 2014)*. Long Beach, CA, USA; November 20th, 2014.

[36] H.F. Kotzé and A.T. Möller. “Effect of Auditory Subliminal Stimulation on GSR.” In: *Psychological Reports* 67 (1990), pp. 931–934.

[37] Emmanuel Dupoux, Vincent de Gardelle, and Sid Kouider. “Subliminal Speech Perception and Auditory Streaming”. In: *Cognition* 109.2 (2008), pp. 267–273.

[38] Colin E. Cherry. “Some Experiments on the Recognition of Speech, with One and with Two Ears”. In: *Journal of the Acoustical Society of America* 25.5 (1953), pp. 975–979.

[39] James W Lewis et al. “Distinct Cortical Pathways for Processing Tool versus Animal Sounds”. In: *Journal of Neuroscience* 25.21 (2005), pp. 5148–5158.

[40] James Woodcock et al. “Categorization of Broadcast Audio Objects in Complex Auditory Scenes”. In: *Journal of the Audio Engineering Society* 64.6 (2016).

[41] Bruno L. Giordano, John McDonnell, and Stephen McAdams. “Hearing Living Symbols and Nonliving Icons: Category Specificities in the Cognitive Processing of Environmental Sounds”. In: *Brain and Cognition* 73.1 (2010), pp. 7–19.

[42] E. Collett et al. “Categorization of Common Sounds by Cochlear Implanted and Normal Hearing Adults”. In: *Hearing Research* 335 (2016), pp. 207–219.

[43] James W. Lewis et al. “Auditory Object Salience: Human Cortical Processing of Non-Biological Action Sounds and their Acoustic Signal Attributes”. In: *Frontiers in Systems Neuroscience* 6.May (2012), pp. 1–15.

[44] Catherine Guastavino. “The Ideal Urban Soundscape: Investigating the Sound Quality of French Cities”. In: *Acta Acustica united with Acustica* 92.2006 (2006), pp. 945–951.

[45] Brian Gygi and Valeriy Shafiro. “The Incongruency Advantage for Sounds in Natural Scenes”. In: *Proceedings of the 125th Convention of the Audio Engineering Society*. San Francisco, CA, USA; October 2-5, 2008, p. 6.
[46] Jonathan Sussman-Fort and Elyse Sussman. “The Effect of Stimulus Context on the Buildup to Stream Segregation”. In: *Frontiers in Neuroscience* 8.8 APR (2014), pp. 1–8.
[47] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. “A Dataset and Taxonomy for Urban Sound Research”. In: *Proceedings of the ACM International Conference on Multimedia - MM ’14*. Orlando, Florida, USA; November 3-7, 2014, pp. 1041–1044.
[48] Manon Raimbault and Daniele Dubois. “Urban Soundscapes: Experiences and Knowledge”. In: *Cities* 22.5 (2005), pp. 339–350.
[49] Susann Deike, Susan L. Denham, and Elyse Sussman. “Probing Auditory Scene Analysis”. In: *Frontiers in Neuroscience* 8.September (2014), p. 293.
[50] International Telecommunication Union. “ITU-R BS.1116-3, Methods for the Subjective Assessment of Small Impairments in Audio Systems”. In: *ITU-R Recommendation* 1116.3 (2015).
[51] International Telecommunication Union. “ITU-R BS.1534-3, Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems”. In: *ITU-R Recommendation* 1534-3 (2015).
[52] Alastair C Disley, David M Howard, and Andy D Hunt. “Timbral Description of Musical Instruments”. In: *9th International Conference on Music Perception and Cognition*. Bologna, Italy; August 22-26, 2006.
[53] K O McGraw, M D Tew, and J E Williams. “The Integrity of Web-delivered Experiments: Can You Trust the Data?” In: *Psychological Science : A Journal of the American Psychological Society / APS* 11.6 (2000), pp. 502–506.
[54] Karol J. Piczak. *ESC: Dataset for Environmental Sound Classification*. Brisbane, Australia; October 26-28, 2015.
[55] Diana Deutsch. *Grouping Mechanisms in Music*. 2013, pp. 183–248.
[56] W J Dowling. “The Perception of Interleaved Melodies”. In: *COGNITIVE PSYCHOLOGY* 5 (1973), pp. 322–337.
[57] Yves Tougas and Albert S. Bregman. “Crossing of auditory streams.” In: *Journal of Experimental Psychology: Human Perception and Performance* 11.6 (1985), pp. 788–798.
[58] Albert S. Bregman and Jeffrey Campbell. “Primary auditory stream segregation and perception of order in rapid sequences of tones.” In: *Journal of Experimental Psychology* 89.2 (1971), pp. 244–249.
[59] L. P. A. S Van Noorden. “TEMPORAL COHERENCE IN THE PERCEPTION OF TONE SEQUENCES (Unpublished Doctoral Dissertation)”. In: Technische (1975).
[60] Orsolya Szalárdy et al. “The effects of rhythm and melody on auditory stream segregation”. In: *The Journal of the Acoustical Society of America* 135.3 (Mar. 2014), pp. 1392–1405.
[61] Carol L. Krumhansl et al. “Melodic Expectation in Finnish Spiritual Folk Hymns: Convergence of Statistical, Behavioral, and Computational Approaches”. In: *Music Perception: An Interdisciplinary Journal* 17.2 (Dec. 1999), pp. 151–195.
[62] C Krumhansl. “Cross-cultural music cognition: cognitive methodology applied to North Sami yoiks”. In: *Cognition* 76.1 (July 2000), pp. 13–58.
[63] Laurel J. Trainor and Sandra E. Trehub. “A comparison of infants’ and adults’ sensitivity to Western musical structure.” In: *Journal of Experimental Psychology: Human Perception and Performance* 18.2 (1992), pp. 394–402.
[64] Laurel J Trainor and Sandra E Trehub. “Key membership and implied harmony in Western tonal music: Developmental perspectives”. In: *Perception & Psychophysics Deutsch & Feroe* 56.2 (1994).
[65] Diana Deutsch. *Memory and Attention in Music*. December 1977. 1977, p. 33.
[66] Jonathan B Fritz et al. “Auditory attention — focusing the searchlight on sound Introduction and overview”. In: *Current Opinion in Neurobiology* 17 (2007), p. 12.
[67] Robert P. Carlyon et al. “Effects of attention and unilateral neglect on auditory stream segregation.” In: *Journal of Experimental Psychology: Human Perception and Performance* 27.1 (2001), pp. 115–127.
[68] Joel S Snyder, Claude Alain, and Terence W Picton. “Effects of Attention on Neuroelectric Correlates of Auditory Stream Segregation”. In: *Journal of Cognitive Neuroscience* 18.1 (Jan. 2006), pp. 1–13.

[69] Mounya Elhilali et al. “Interaction between Attention and Bottom-Up Saliency Mediates the Representation of Foreground and Background in an Auditory Scene”. In: *PLoS Biology* 7.6 (June 2009). Ed. by Timothy D. Griffiths, e1000129.
[70] W. Jay Dowling, Kitty Mei-Tak Lung, and Susan Herrbold. “Aiming attention in pitch and time in the perception of interleaved melodies”. In: *Perception & Psychophysics* 41.6 (Nov. 1987), pp. 642–656.
[71] Steven; Woodward et al. “Probing the Time-Course of the Auditory Oddball P3 with Secondary Reaction Time”. In: *Psychophysiology* 28.6 (1991), pp. 609–618.
[72] Erich Schröger. “A Neural Mechanism for Involuntary Attention Shifts to Changes in Auditory Stimulation”. In: *Journal of Cognitive Neuroscience* 8.6 (1996), p. 11.
[73] Sandra E. Trehub et al. “Development of the perception of musical relations: Semitone and diatonic structure.” In: *Journal of Experimental Psychology: Human Perception and Performance* 12.3 (1986), pp. 295–301.
[74] E. Glenn Schellenberg et al. “Children’s implicit knowledge of harmony in Western music”. In: *Developmental Science* 8.6 (Nov. 2005), pp. 551–566.
[75] Carol L. Krumhansl. *Cognitive foundations of musical pitch*. Oxford University Press, 1990, p. 307.
[76] Shuang Guo and Stefan Koelsch. “Effects of veridical expectations on syntax processing in music: Event- related potential evidence”. In: *Nature Publishing Group* (2015), p. 9.
[77] Leonard B. Meyer. *Emotion and meaning in music*. University of Chicago Press, 1956, p. 307.
[78] I Peretz, D Gaudreau, and A M Bonnel. “Exposure effects on music preference and recognition.” In: *Memory & cognition* 26.5 (Sept. 1998), pp. 884–902.
[79] Carlos Silva Pereira et al. “Music and emotions in the brain: familiarity matters.” In: *PloS one* 6.11 (2011), e27241.
[80] W. Jay. Dowling and Dane L. Harwood. *Music cognition*. Academic Press, 1986, p. 258.
[81] T C Justus and Jamshed J Bharucha. “Modularity in musical processing: the automaticity of harmonic priming.” In: *Journal of Experimental Psychology: Human Perception and Performance* 27.4 (2001), p. 11.
[82] Barbara Tillmann and Emmanuel Bigand. “Musical structure processing after repeated listening: Schematic expectationsresist veridical expectations”. In: *Musicae Scientiae* Special Issue (2010), pp. 33–47.