



**Streamlining Production Workflows for  
Accessible Audio using AI Metadata  
Assignment**

**Kim Steele**

**University of York**

**Electronic Engineering**

MSc by Research

April 2023

# Abstract

Advancements in broadcast technology are granting new opportunities to improve listening experiences for consumers. One such advancement is object-based audio and the development of the accessible audio system termed Narrative Importance. This system provides users with an adjustable mix, that boosts sounds which are important to the story, such as dialogue, and attenuates non-essential background sounds, such as crowd chatter. This thesis will target one of the barriers to rolling out the Narrative Importance system - augmented production time due to implementation requirements. The specific focus will be furthering the investigation into whether machine learning can be trained to assign the requisite metadata for Narrative Importance.

A survey is deployed to collect label data for training the machine learning algorithm. Participants are asked to assign importance data to sounds in a mix for nine scenes. This data is then used to train a mixture model to categorise audio objects into 4 levels of importance. The results show that the method chosen here is not successful in its current form. Training with survey labels proves to be ineffective due to low levels of agreement amongst participants. Training with a single set of labels is shown to give better results.

The question of “What is object-based audio?” was also investigated, partially as a result of differing definitions in the existing literature. A survey of audio research professionals was undertaken. The results show that a robust definition for object-based audio does not exist in writing nor in practice. As a result, a definition is proposed in this thesis.

# Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

Signed :

A handwritten signature in black ink, appearing to read "King Steele", written in a cursive style. The signature is positioned above a horizontal line.

# Acknowledgements

Firstly, thanks must go to my supervisors, Lauren Ward and Gavin Kearney, who have always found time to support, reassure, and encourage me in my work (and side quests).

To all of my colleagues in the AudioLab, thank you for your support, participating in my surveys, and helping me feel like I have found my place in the world.

My gratitude to Matthew Paradis for his efforts in sourcing content and coding the online task interface. Thanks also go to Ben Shirley for providing the football content.

To my adopted supervisors, Michael McLoughlin, for his expertise on all things machine learning, and Alan Archer-Boyd, for being on “burn-out watch” and teaching me how to use quotation marks correctly.

Special mention goes to my family and friends, for always being willing to discuss my research, believing in my abilities more than I do, and making sure I don’t lose my social life to the quagmire of work.

Finally to BBC R&D Northlab, for being the best workplace I have had the pleasure of existing in. To everyone who has spent time listening to me figure out my own thoughts, helping me with code, and taking part in surveys, you are all wonderful humans.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivations . . . . .	1
1.2	Research Aim . . . . .	3
1.3	Key Contributions . . . . .	3
1.4	Thesis Structure . . . . .	3
<b>2</b>	<b>Background</b>	<b>5</b>
2.1	Introduction . . . . .	5
2.2	Why does Accessibility in Broadcasting Matter? . . . . .	5
2.2.1	Hearing Loss and Other Aural Diversities . . . . .	5
2.2.2	Progress So Far . . . . .	7
2.3	What is Narrative Importance? . . . . .	10
2.3.1	Narrative Importance . . . . .	10
2.3.2	The Importance of a Sound . . . . .	11
2.4	Why Machine Learning? . . . . .	13
2.4.1	Roll-out Barriers . . . . .	13
2.4.2	How Machines Learn . . . . .	13
2.4.3	State-of-the-Art . . . . .	14
2.4.4	Previous Work . . . . .	14
2.5	Conclusions . . . . .	15
<b>3</b>	<b>Data Collection Design</b>	<b>16</b>
3.1	Introduction . . . . .	16
3.2	Method . . . . .	16
3.2.1	Questionnaire . . . . .	17
3.2.2	Assignment Task . . . . .	18

3.2.3	Critical Reflection . . . . .	22
3.3	Conclusions . . . . .	23
<b>4</b>	<b>Content</b>	<b>24</b>
4.1	Introduction . . . . .	24
4.2	Object Categories . . . . .	24
4.3	Selected Content . . . . .	25
4.3.1	Protest . . . . .	25
4.3.2	The Vostok-K Incident . . . . .	26
4.3.3	Casualty Season 33 Episode 38 . . . . .	27
4.3.4	Penguins: Spy in the Huddle . . . . .	31
4.3.5	Football . . . . .	33
4.4	Processing of Scenes . . . . .	34
4.4.1	Processing Protest . . . . .	34
4.4.2	Processing The Vostok-K Incident . . . . .	35
4.4.3	Processing Casualty Season 33 Episode 38 . . . . .	35
4.4.4	Processing Penguins: Spy in the Huddle . . . . .	36
4.4.5	Processing Football . . . . .	37
4.5	Object Category Distribution . . . . .	37
4.6	Ground Truths . . . . .	38
4.6.1	Protest . . . . .	39
4.6.2	Casualty Season 33 Episode 38 . . . . .	39
4.7	Rejected Content . . . . .	42
4.7.1	European Athletics Championships - Women's Hurdles . . . . .	42
4.7.2	Radio Panel Shows . . . . .	43
4.7.3	The Watches Series . . . . .	43
4.8	Content Selection Problems . . . . .	43
4.9	Guidelines for Narrative Importance Content . . . . .	44
4.10	Conclusions . . . . .	46
<b>5</b>	<b>Data Collection Results</b>	<b>47</b>
5.1	Introduction . . . . .	47
5.2	Survey Demographics . . . . .	47
5.3	Participant Removal . . . . .	52

5.4	Fleiss Kappa . . . . .	54
5.5	The Assignments . . . . .	55
5.6	Results for Audio Professionals VS. Non-Professionals . . . . .	63
5.7	Feedback Quotes . . . . .	65
5.8	Conclusions . . . . .	69
<b>6</b>	<b>Machine Learning</b>	<b>70</b>
6.1	Introduction . . . . .	70
6.2	Method . . . . .	70
6.2.1	Overall Architecture . . . . .	70
6.2.2	Speech/Music/SFX Classification Model . . . . .	71
6.2.3	Speech/Music/SFX Classification Model Training . . . . .	73
6.2.4	Mixture Model . . . . .	75
6.2.5	Mixture Model Training . . . . .	76
6.2.6	Changes made to the Algorithm . . . . .	78
6.2.7	Critical analysis of Algorithm . . . . .	79
6.3	Results . . . . .	81
6.4	K-Nearest Neighbours Algorithms . . . . .	84
6.5	Training with Ground Truth Labels . . . . .	86
6.6	Conclusions . . . . .	88
<b>7</b>	<b>Further Understanding of Object-Based Audio</b>	<b>90</b>
7.1	Introduction . . . . .	90
7.2	The Problem/Motivation . . . . .	90
7.3	OBA in the Literature . . . . .	91
7.3.1	Discussion . . . . .	97
7.4	Survey . . . . .	97
7.4.1	Method . . . . .	98
7.4.2	Results . . . . .	100
7.4.3	Discussion . . . . .	106
7.5	Conclusions . . . . .	107
<b>8</b>	<b>Summary and Key Contributions</b>	<b>108</b>
8.1	Introduction . . . . .	108
8.2	Summary . . . . .	108

8.2.1	The Data Collection Survey . . . . .	108
8.2.2	Machine Learning . . . . .	109
8.2.3	Object-Based Audio . . . . .	109
8.3	Contributions to Knowledge . . . . .	110
8.3.1	Narrative Importance Content Guidelines . . . . .	110
8.3.2	Narrative Importance - What's Important? . . . . .	110
8.3.3	Machine Learning of Importance . . . . .	110
8.3.4	What is Object-Based Audio? . . . . .	111
8.4	Areas for Further work . . . . .	111
8.4.1	Narrative Importance - What's Important? . . . . .	111
8.4.2	Alternatives to Machine Learning . . . . .	111
8.4.3	Machine Learning Development . . . . .	112
8.4.4	Object-Based Audio Definitions . . . . .	112
8.4.5	The Recording Process . . . . .	113
8.5	Conclusions . . . . .	113
	<b>References</b>	<b>115</b>
	<b>Appendix A Object Tables</b>	<b>128</b>
	<b>Appendix B Object Creation Tables</b>	<b>132</b>
	<b>Appendix C Violin Plots</b>	<b>135</b>
	<b>Appendix D Speech/Music/SFX Classification Model Parameter Sweep</b>	<b>140</b>
	<b>Appendix E Mixture Model Distribution Plots</b>	<b>143</b>
	<b>Appendix F Mixture Model Results</b>	<b>151</b>
	<b>Appendix G 7NN Results</b>	<b>155</b>
	<b>Appendix H Mixture Model Training with Ground Truth Data</b>	<b>159</b>

# List of Figures

2.1	A graphical representation of the differences between OBA and traditional audio in broadcasting . . . . .	10
2.2	The gains applied to objects in each category . . . . .	11
3.1	A screenshot showing an example of the interface for the Protest content. Not all objects are shown. . . . .	21
4.1	The three screenshots provided to participants for Scene 4 from Casualty . . . .	29
4.2	The two screenshots provided to participants for Scene 46 from Casualty . . . .	30
4.3	The two screenshots provided to participants for Scene 49 from Casualty . . . .	31
4.4	The screenshot provided to participants . . . . .	32
4.5	The two screenshots provided to participants . . . . .	33
4.6	Pie charts showing the types of objects in each category for each scene . . . . .	38
4.7	A bar chart showing the ground truths of the objects in Protest. The Type bar at the top refers to the type of object as defined in Section 4.2. . . . .	39
4.8	A bar chart showing the ground truths of the objects in Scene 4 of Casualty. The Type bar at the top refers to the type of object as defined in Section 4.2. .	40
4.9	A bar chart showing the ground truths of the objects in Scene 46 of Casualty. The Type bar at the top refers to the type of object as defined in Section 4.2. .	41
4.10	A bar chart showing the ground truths of the objects in Scene 49 of Casualty. The Type bar at the top refers to the type of object as defined in Section 4.2. .	42
5.1	A bar chart showing the assignments for the Protest scene. Agreement refers to the value of agreement for each object as calculated in Eq. (5.1) . . . . .	56
5.2	A bar chart showing the assignments for Vostok-K extract 1. Agreement refers to the value of agreement for each object as calculated in Eq. (5.1) . . . . .	57

5.3	A bar chart showing the assignments for Vostok-K extract 2. Agreement refers to the value of agreement for each object as calculated in Eq. (5.1) . . . . .	58
5.4	A bar chart showing the assignments for Casualty Scene 4. Agreement refers to the value of agreement for each object as calculated in Eq. (5.1) . . . . .	58
5.5	A bar chart showing the assignments for Casualty Scene 46. Agreement refers to the value of agreement for each object as calculated in Eq. (5.1) . . . . .	59
5.6	A bar chart showing the assignments for Casualty Scene 49. Agreement refers to the value of agreement for each object as calculated in Eq. (5.1) . . . . .	60
5.7	A bar chart showing the assignments for Penguins Spy In The Huddle: Opening Credits. Agreement refers to the value of agreement for each object as calculated in Eq. (5.1) . . . . .	61
5.8	A bar chart showing the assignments for Penguins Spy In The Huddle: Cormorants and Seals. Agreement refers to the value of agreement for each object as calculated in Eq. (5.1) . . . . .	62
5.9	A bar chart showing the assignments for the Football scene. Agreement refers to the value of agreement for each object as calculated in Eq. (5.1) . . . . .	63
5.10	Three violin plots showing the distributions of the number of objects (normalised to between 0 and 1) each participant assigned to each importance level for Vostok-K S1. . . . .	64
5.11	A bar chart showing AP assignments for Vostok-K S1 . . . . .	65
5.12	A bar chart showing NAP assignments for Vostok-K S1 . . . . .	65
6.1	The structure of the legacy algorithm. Rectangular blocks represent the ML models. Shaded ellipses represent the databases used for training. . . . .	71
6.2	The final training with batch size = 16, LR = 0.01 . . . . .	74
6.3	Probability density functions for the four features when trained on all the data except for the Protest content . . . . .	78
6.4	Casualty Scene 4 results from training the mixture model with survey data . . .	82
6.5	Casualty Scene 46 results from training the mixture model with survey data . .	82
6.6	Casualty Scene 49 results from training the mixture model with survey data . .	82
6.7	Casualty Scene 4 results from training a 7NN model with survey data . . . . .	85
6.8	Casualty Scene 46 results from training a 7NN model with survey data . . . . .	85
6.9	Casualty Scene 49 results from training a 7NN model with survey data . . . . .	85
6.10	Casualty Scene 4 results from training the mixture model with GT data . . . . .	86

6.11	Casualty Scene 46 results from training the mixture model with GT data . . . .	87
6.12	Casualty Scene 49 results from training the mixture model with GT data . . . .	87
7.1	Bar charts showing the responses to question 1 (which of the following best describes your work in audio?) and 2 (please give more detail) from the survey. No participants answered that they were a content creator for question 1. For question 2 ‘Other’ includes the software developer, mastering engineer, and the unspecified technical job . . . . .	101
7.2	A bar chart showing the responses to question 3 (how many years have you worked in audio?) of the survey. Participants’ answers were grouped into categories during analysis. The range was from 2 to 40 years. . . . .	102
7.3	A bar chart showing the responses to question 5 (how long have you been aware of/worked with object-based audio?) of the survey. Participants’ answers were grouped into categories during analysis. The range was from 8 months to 22 years.	102
C.1	Three violin plots showing the distributions of the number of objects (normalised to between 0 and 1) each participant assigned to each importance level for Protest. Protest contained 27 objects. . . . .	135
C.2	Three violin plots showing the distributions of the number of objects (normalised to between 0 and 1) each participant assigned to each importance level for Vostok-K Scene 2. Vostok-K Scene 2 contained 21 objects. . . . .	136
C.3	Three violin plots showing the distributions of the number of objects (normalised to between 0 and 1) each participant assigned to each importance level for Casualty Scene 4. Casualty Scene 4 contained 27 objects. . . . .	136
C.4	Three violin plots showing the distributions of the number of objects (normalised to between 0 and 1) each participant assigned to each importance level for Casualty Scene 46. Casualty Scene 46 contained 17 objects. . . . .	137
C.5	Three violin plots showing the distributions of the number of objects (normalised to between 0 and 1) each participant assigned to each importance level for Casualty Scene 49. Casualty Scene 49 contained 20 objects. . . . .	137
C.6	Three violin plots showing the distributions of the number of objects (normalised to between 0 and 1) each participant assigned to each importance level for Penguins Opening Credits. Penguins Opening Credits contained 23 objects. . . .	138

C.7	Three violin plots showing the distributions of the number of objects (normalised to between 0 and 1) each participant assigned to each importance level for Penguins Scene 1. Penguins Scene 1 contained 25 objects. . . . .	138
C.8	Three violin plots showing the distributions of the number of objects (normalised to between 0 and 1) each participant assigned to each importance level for Football. Football contained 11 objects. . . . .	139
D.1	Accuracy plots for the parameter sweep . . . . .	141
D.2	Loss plots for the parameter sweep . . . . .	142
E.1	Probability density functions when the mixture model is trained on all the data except for Vostok-K Scene 1 . . . . .	143
E.2	Probability density functions when the mixture model is trained on all the data except for Vostok-K Scene 2 . . . . .	144
E.3	Probability density functions when the mixture model is trained on all the data except for Casualty Scene 4 . . . . .	145
E.4	Probability density functions when the mixture model is trained on all the data except for Casualty Scene 46 . . . . .	146
E.5	Probability density functions when the mixture model is trained on all the data except for Casualty Scene 49 . . . . .	147
E.6	Probability density functions when the mixture model is trained on all the data except for Penguins Opening Credits . . . . .	148
E.7	Probability density functions when the mixture model is trained on all the data except for Penguins Scene 1 . . . . .	149
E.8	Probability density functions when the mixture model is trained on all the data except for Football . . . . .	150
F.1	A bar chart showing the mixture model’s assignments with Protest as the test data . . . . .	151
F.2	A bar chart showing the mixture model’s assignments with Vostok-K Scene 1 as the test data . . . . .	152
F.3	A bar chart showing the mixture model’s assignments with Vostok-K Scene 2 as the test data . . . . .	152
F.4	A bar chart showing the mixture model’s assignments with Penguins Opening Credits as the test data . . . . .	153



---

F.5	A bar chart showing the mixture model's assignments with Penguins Scene 1 as the test data . . . . .	153
F.6	A bar chart showing the mixture model's assignments with Football as the test data . . . . .	154
G.1	A bar chart showing the 7NN model's assignments with Protest as the test data	155
G.2	A bar chart showing the 7NN model's assignments with Vostok-K Scene 1 as the test data . . . . .	156
G.3	A bar chart showing the 7NN model's assignments with Vostok-K Scene 2 as the test data . . . . .	156
G.4	A bar chart showing the 7NN model's assignments with Penguins Opening Credits as the test data . . . . .	157
G.5	A bar chart showing the 7NN model's assignments with Penguins Scene 1 as the test data . . . . .	157
G.6	A bar chart showing the 7NN model's assignments with Football as the test data	158
H.1	A bar chart showing the mixture model's assignments when it was trained with the ground truth data, with Protest as the test data . . . . .	159

# List of Tables

3.1	A table showing N, the number of participants who completed assignments for each of the 9 the content pieces. . . . .	17
4.1	A table of categories which will be used to discuss the audio objects . . . . .	25
4.2	A table showing the 5 pieces of content and the 9 scenes that were extracted for use with the number of objects contained in each scene . . . . .	46
5.1	A table showing questions asked to all 50 participants with the number who selected each response in the <b>Counts</b> column . . . . .	48
5.2	A table showing questions asked to the 21 laypeople participants to gauge their audio and musical experience with the number of participants who selected each response in the <b>Counts</b> column . . . . .	48
5.3	A table showing the working background of the 29 audio professional participants. The number of participants who selected each response is in the <b>Counts</b> column. . . . .	50
5.4	A table showing the questions on OBA, NI, and AI asked to the 29 participants who worked in audio with the number of participants who selected each response in the <b>Counts</b> column . . . . .	51
5.5	A table showing N, the number of participants, and $\kappa$ , the value of Fleiss' Kappa, for each of the content pieces across three groups. The full dataset, the audio professionals, and laypeople. . . . .	55
6.1	Architecture of the VGGish based network. The top layer represents the input with the output at the bottom. . . . .	72
6.2	A table showing the precision, recall, F1 score, and accuracy figures after training the VGGish based model for 6 epochs. The 'Support' column refers to the number of samples in each class. . . . .	74

6.3	A table showing the precision, recall, F1 score, and accuracy for the model when tested with Scene 4 of Casualty. The ‘Support’ column refers to the number of samples in each class. . . . .	83
6.4	A table showing the precision, recall, F1 score, and accuracy for the model when tested with Scene 46 of Casualty. The ‘Support’ column refers to the number of samples in each class. . . . .	83
6.5	A table showing the precision, recall, F1 score, and accuracy for the model when tested with Scene 49 of Casualty. The ‘Support’ column refers to the number of samples in each class. . . . .	83
6.6	A table showing how many objects (left three columns) and how many Main Dialogue objects (right three columns) were assigned to each importance category by each $k$ NN . . . . .	84
6.7	A table showing the precision, recall, F1 score, and accuracy for the model, trained on the GT data, where Scene 4 of Casualty was the test data. The ‘Support’ column refers to the number of samples in each class. . . . .	87
6.8	A table showing the precision, recall, F1 score, and accuracy for the model, trained on the GT data, where Scene 46 of Casualty was the test data. The ‘Support’ column refers to the number of samples in each class. . . . .	88
6.9	A table showing the precision, recall, F1 score, and accuracy for the model when tested with Scene 49 of Casualty. The ‘Support’ column refers to the number of samples in each class. . . . .	88
7.1	A table of the 10 questions in the OBA survey . . . . .	99
A.1	A table showing all of the object names for Protest and the two Vostok-K scenes.	129
A.2	A table showing all of the object names in each of the Casualty scenes. . . . .	130
A.3	A table showing all of the object names in the two Penguins scenes and Football.	131
B.1	A table showing the creation of objects for the Protest scene. The first column shows the original track names, the second column shows the objects that came from the original tracks. Some objects were a down mix of several tracks, similiary some tracks were separated into multiple objects. The final column shows the ‘ground truth’. . . . .	133

- 
- B.2 A table showing the creation of objects for the two Vostok-K scenes. A blank cell indicates that the track wasn't included in that scene. Tracks that didn't feature in either scene have been omitted from the table. . . . . 134
- H.1 A table showing the precision, recall, F1 score, and accuracy when the model is trained with only ground truth data, where Protest was the test data. The 'Support' column refers to the number of samples in each class. . . . . 160

# Chapter 1

## Introduction

### 1.1 Motivations

Complaints about television audio continue to rise despite many intervention attempts [1]. In 2014 and 2016 respectively, BBC dramas “Jamaica Inn” and “Happy Valley” reportedly received hundreds of complaints of issues with dialogue comprehension [2], [3]. Traditional broadcasting systems tend to take a one-size fits all approach, providing a single stream of audio. This has been sufficient for the majority of the population so far. However broadcast audio often fails to meet the needs of people with hearing differences, at worst excluding them, at best providing sub-par experiences. Consumers with hearing loss often struggle to understand speech due to the level of sound effects (SFX) and/or music [4]. Alongside this, the progression of modern technology means a far larger variation of listening devices are being used to access content, and the single audio stream is more and more likely to fall short of expectations. The trend for television sets to be flat has imposed limitations on the size and direction of built-in loudspeakers [5]. Most modern televisions feature loudspeakers that don’t fire directly towards the viewer, e.g. downward firing speakers. This results in variable acoustic response depending on how the television is installed in the room, due to acoustic reflections from walls or surfaces such as a television stand [6], [7].

Several strategies have been proposed to improve this issue. Subtitles are perhaps the most widely available accessibility tool. The inclusion of subtitles in media has been widely campaigned for by the Royal National Institute for Deaf People (RNID), and recently they have reopened their survey on the matter [8], [9]. This campaign has resulted in a significant increase of Ofcom’s subtitling requirements for broadcasters [10]. Whilst this increase is certainly a pos-

itive step in the right direction, subtitles cannot solve accessibility for all people. The reading of subtitles requires a number of skills: good eyesight, sufficiently fast processing capabilities, and reading proficiency. These can all be impacted by various disabilities, which often coexist alongside hearing differences. The older population particularly are more likely to have comorbid disabilities, such as hearing loss and visual impairment [11], or hearing loss and cognitive decline [12].

Other approaches for improving the accessibility of audio have looked at exploiting 5.1 surround sound systems [13] but this solution neglects the financial and technical accessibility of such a system. Another approach proposed was clean audio, whereby a second audio track, which contained only speech was provided. This idea was tested in 2007 [14], however the additional cost required in production was too great for wider implementation. More recently, work has focused on dialogue enhancement techniques, where an alternative audio stream is provided with boosted dialogue stems, and Amazon have just launched this feature on their on-demand streaming service with a medium and high setting [15]. This is likely to improve the experiences of people who find it difficult to follow dialogue, such as those with a hearing impairment.

One of the latest advancements in audio technology that could improve the listening experience is object-based audio (OBA) [16]. Unlike with traditional channel-based audio, in OBA, metadata is utilised to facilitate some aspects of the mix being left to the user end of the production chain. The result of this is that user can be provided with a greater range of options for how they listen to content. OBA requires a set of rules to be developed that form the basis of the metadata. One such set of rules, in the area of accessibility, is the Narrative Importance (NI) system developed by Ward [17]. In the NI system, audio objects are assigned to one of four importance levels - Essential, High, Medium, and Low. From this, an adjustable mix where the less important sounds are attenuated and the more important ones are boosted can be derived. This system provides the best of both worlds, allowing a greater speech in noise ratio to be achieved whilst retaining sounds that aid understanding or immersion.

One of the barriers to rolling out the NI system is that it requires additional time within already strained production workflows to implement. Previous work done by Chourdakis and Ward [18] has shown potential in the area of machine learning (ML) to aid with this problem. In this work, an algorithm was trained on a single piece of content to categorise audio objects into importance levels. The training dataset was taken from a survey of audio production staff, where they were asked to assign importance ratings to objects in a mix. This thesis leads on directly from this work, aiming to extend it to other genres.

## 1.2 Research Aim

The advent of object-based media is affording opportunities for new, and better ways to provide accessible services through personalisable settings. The NI system devised by Ward is one way of providing a personalisable, accessible audio service [17]. A significant barrier to rolling out NI is the additional time taken to implement it during the production workflow.

This thesis will investigate whether ML can assist with the generation of NI metadata for more genres. In other words, can a computer be trained to decide how important individual sounds are within a storyline?

A secondary research question will be addressed, “What is Object-Based Audio?”, in a chapter at the end of the work.

## 1.3 Key Contributions

This thesis makes several contributions to knowledge in the areas of NI, ML, and object-based audio (OBA), which are:

- The development of guidelines on NI content creation, storage, and sourcing (Chapter 4).
- Expansion of the investigation into how important specific sounds are within narrative content to include results from participants who don’t work in audio (Chapter 5).
- Advancement upon work done in the area of using ML for NI metadata creation (Chapter 6). This work improves upon the original algorithm methods and includes content from different genres.
- An investigation of how OBA is defined within the industry, resulting in a proposed definition with subsets of OBA classes (Chapter 7).

## 1.4 Thesis Structure

This thesis can be split into three main bodies of work. The first takes the form of a survey that will provide a database of training labels for the ML algorithm. The second is the ML itself. Finally an investigation into OBA and its definitions. The structure of the thesis will now be outlined.

In Chapter 2, the thesis will begin with an overview of audio accessibility in broadcasting and

the progress made so far in this area. This will be followed by an explanation of the NI system, and the various aspects that contribute to the importance of an individual sound within a mix. Finally, the chapter will end with an explanation of how ML might be helpful in this area, and the relevant work preceding this thesis.

Chapters 3 to 5 present the survey which aims to collect NI labels to inform the ML. In Chapter 3, the survey method will be explained. The survey is comprised of two parts, a questionnaire and a data collection task. The questionnaire portion will provide demographic information for participants. The data collection task facilitates the collection of NI labels, by asking participants to choose importance categories for objects in a mix. The results from the survey will be presented in Chapter 5.

Chapter 4 contains an in-depth discussion of the selection of broadcast content for use in the survey. Detailed descriptions will be provided for each of the scenes that will be used. Following this, an account of the work required to convert the audio to a suitable OBA format for the survey will be given. The content that was rejected will be discussed, along with justifications for why these pieces were deemed unsuitable. Finally, the chapter will finish with guidelines for the creation, selection, and storage of NI-friendly content.

Chapter 6 presents the method and results from the ML investigation. Two algorithms are trained on the survey data from Chapter 5. One algorithm is additionally trained on a set of “ground truth” labels.

In Chapter 7, a comprehensive exploration of the definition of OBA will take place. The chapter begins with a review of definitions found within standards and research literature. This is followed by a survey of audio professionals, focused on OBA definitions. The results from this survey are then presented and discussed.

Finally in Chapter 8, a summary of the body of work will be given. The key contributions and areas for further work will be identified. The chapter will finish with some concluding remarks.



# Chapter 2

## Background

### 2.1 Introduction

In this chapter, the importance of accessibility of broadcast audio is presented. Hearing loss will be the main focus, with considerations of other groups, such as neurodivergent people, and the complications of comorbidities of disability. This will lead into the work done so far to improve listening experiences and accessibility of content. An explanation of the Narrative Importance (NI) system precedes a discussion of the different functions of sounds within broadcasting. Finally, the previous work done on using ML to create NI metadata will be introduced.

### 2.2 Why does Accessibility in Broadcasting Matter?

#### 2.2.1 Hearing Loss and Other Aural Diversities

In the UK, the RNID reported that over 11 million people were affected by hearing loss in 2015 [19]. It is estimated that by 2035 that figure will be 15.6 million or 1 in 5 people in the UK due to the aging population. This predicted increase is reflected globally. Hearing loss can occur for a multitude of reasons. Age-related hearing loss (presbycusis) is the most common, with 71.1% of over 70s having some level of hearing loss [19]. These figures are reflected globally. In 2019 [20] estimated 1.57 billion people globally have hearing loss (which equates to 1 in 5 people). Data from 2001 to 2010 was used to estimate hearing loss in the US and found that a predicted 1 in 4 people (23%) over the age of 12 had hearing loss in at least one ear (unilateral loss) [21].

The World Health Organisation state that hearing loss is the fourth leading cause of disability

globally [22]. Hearing loss is associated with a lowered quality of life [23], and a higher risk of loneliness and social isolation [24], [25]. It has been observed as a potential risk factor for dementia, cognitive decline, and depression in later life [12], [26]. The links are not yet fully understood, though increased cognitive load, and social isolation are thought to be amongst the possible mediators [27].

Investigation on the impact of cognitive differences on listening is gaining traction. One study has shown that for older participants, speech in noise scores are lower than for younger participants, after screening all participants for normal hearing (bilateral audiometric thresholds  $\leq 20$  dB hearing loss at 0.125–6 kHz) [28]. This suggests that age related cognitive decline may affect people’s ability to process auditory information.

Similarly, research comparing autistic adults with neurotypical adults showed that speech in noise is more difficult to comprehend for autistic people [29]. Sturrock et al [30] conducted interviews with 9 autistic adults about their experiences of speech perception. Most of the participants reported significant difficulties when trying to focus on speech amongst competing sounds.

Television, radio and other forms of broadcast media play a large part in many older people’s lives, providing entertainment and a sense of companionship. The accessibility of broadcast media to those with hearing loss is vital to continue providing these benefits whilst decreasing additional strain on cognitive systems. Older people reportedly watch more television than the majority of the population. Nielsen [31] report that those 65 and over watch an average of 6 hours 39 minutes of live or time-shifted TV per day compared with 3 hours 41 minutes among all adults. An online survey was conducted in the United States in 2015, [4], consisting of 515 subjects with hearing loss, 260 of which used hearing aids (HA). 50% of HA users and 60% of non-HA users reported that the loudness of background music and sound effects impacted upon their ability to understand speech when watching television. The most common strategies participants reported using to improve their viewing experience were turning up the volume (80%) and using closed captions (45%). Notably, those with carpeted floors reported fewer difficulties, suggesting that depending upon a viewer’s acoustic environment their requirements for accessibility aids will vary.

### 2.2.2 Progress So Far

Accessibility of content for consumers with hearing loss has been extensively researched and campaigned for. In the last twenty years the vast development of technology has meant that many different approaches have been considered.

Good progress has been made in the area of subtitling for television with on-demand services slowly catching up. RNID have been running a campaign since 2015 entitled ‘Subtitle it!’ [9] and in recent years have been directing the focus of this campaign to on-demand services which currently have no legal requirements for subtitling of shows. Ofcom reported in 2021 that 71.4% of on-demand providers who opted to respond to their data request were offering subtitles [32]; 66.1% of the programming hours for these providers were subtitled. The figures on broadcast subtitles in 2021 were affected by an outage at Red Bee media which impacted several broadcasters, most notably Channel 4. The outage was caused by an acoustic shock wave resulting from a gas that was released upon the triggering of the fire suppression system at the Broadcast Centre in London. The shockwave damaged equipment, taking several broadcasters off-air [33]. Ofcom recently ruled that Channel 4 fell short of meeting their quota of subtitling, signing, and audio description services, as a result of ongoing disruption following the incident at Red Bee, which lasted nearly 2 months in 2021 [34]. Ofcom found that all other broadcasters met or exceeded their required quota for the year in spite of this issue.

Unfortunately, subtitling doesn’t provide a solution for all viewers since it relies on other abilities such as good eyesight and reading abilities, that will be impaired for some viewers. The older population are especially likely to have impaired vision, cognition and/or hearing [11].

Fastl and Zwicker in section 16.2.7 of [35] state that “For persons with hearing deficits, speech intelligibility deteriorates substantially in noisy environments.” This holds true for broadcast audio with the additional complication of lack of localisation cues compounding issues for some consumers. Localisation cues are limited when consuming broadcast audio due to all of the sound coming from the television or radio speakers, rather than being full 360 sound. Exploitation of localisation cues to improve audio for consumers with hearing loss has been explored in [13]. In this work the use of 5.1 systems to improve speech clarity was investigated and found to provide some benefit. Unfortunately 5.1 systems have not been widely adopted by home users. This is likely due to the cost and difficulty of installing and maintaining multiple speakers, with many technology sites recommending soundbars over surround sound systems these days [36], [37].

Armstrong gives a detailed review of the importance and issues of television audio in [14]. TV audio is vastly important as a narrative tool, but also as a function to keep the viewer's attention. TV audio contains diegetic and non-diegetic (visible and not visible) speech far more frequently than in film. This can be problematic for viewers whose ability to differentiate between voices is impacted by hearing loss. An example of this phenomena can be seen by programmes that contain narration such as many documentaries or reality programmes. At times, narration may play over other diegetic voices, which could be especially confusing for those who rely on lip-reading to boost their speech understanding. The practice of mixing voice-on-voice audio is sometimes known as 'ducking'. The 'ducking' practice of nine experts was compared to the preferred loudness difference (LD) of 12 non-expert listeners over the age of 57, [38]. 'Ducking' refers to the act of lowering the level of one audio track when another is present, for example a background track may be lowered in level when speech is present in a television show. The results showed a significant difference between professional practices and user preferences. Experts tended to LDs between 11.5-17 LU (loudness units) whereas non-experts chose a range of 20-30 LU. Since LU directly correspond to dB, 10 LU will be approximately equivalent to a perceived doubling, and therefore this difference is significant. This suggests that professional practice is not appropriate for some listeners. Another similar study looked at the 'ducking' between commentary and background sounds, and similarly found that non-expert listeners prefer LDs that are at least 4LU higher than those preferred by expert listeners [39].

One concept that has been explored in the past is that of providing a 'clean audio' track for viewers who find the background sounds interfere with the dialogue. A review of projects which used post-processing to achieve this showed little benefit in terms of speech intelligibility [40]. The alternative of creating a clean audio soundtrack during the production process was tested by the BBC in 2007 [14]. An episode of "The Nature of Britain: Secret Britain" was broadcast through the 'red button' service without any music. The main barrier to this system was the additional cost of production created. The episode required additional sound effects and narration to cover the gaps left by the removal of the music.

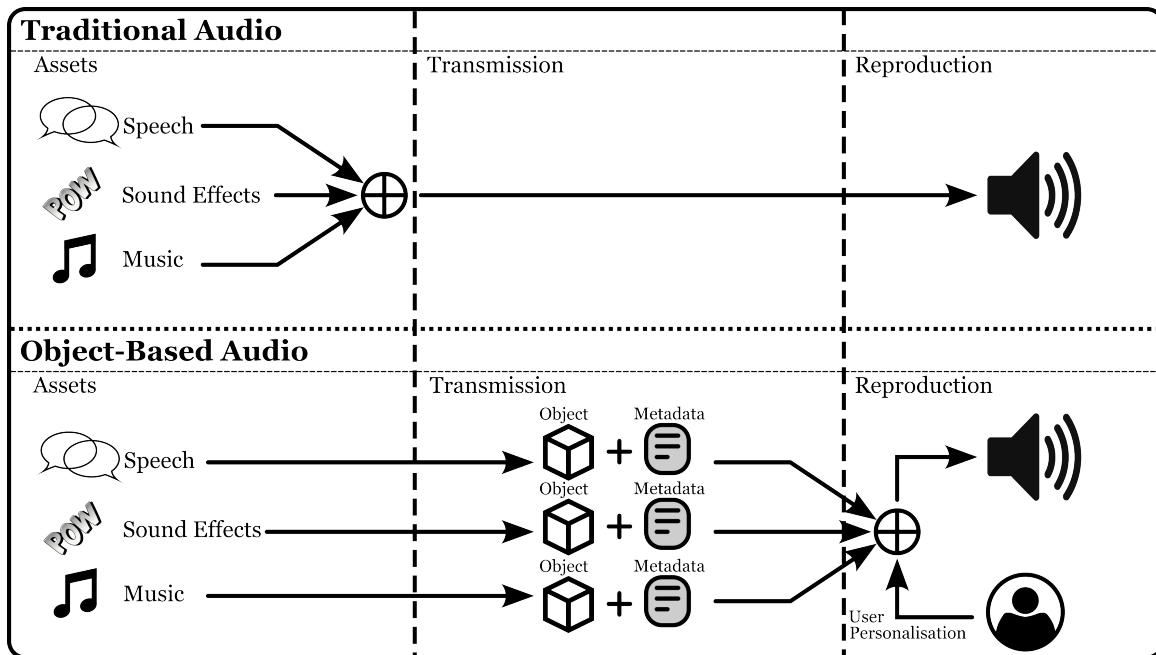
One of the user issues with a 'clean audio' approach is that, whilst it does enable speech to be better understood thanks to the removal of potential masking sounds, it also removes a lot of the context and emotive content from the show. Contextual sound effects have been shown to aid speech intelligibility scores [41]. The 'clean audio' concept also neglects the fact that TV viewing is often a social activity within family groups where not all viewers will have hearing

loss. For the viewers with normal hearing in a group, missing out entirely on the narrative enhancement provided by music and sound effects in order to accommodate a person with hearing loss's needs could be frustrating.

A more recent area of interest has been that of dialogue enhancement. In recent years Source Separation algorithms have greatly improved in their capabilities. As a result, the use of them to extract dialogue from broadcast audio in order to provide an enhanced dialogue track has been explored [42], [43]. Deployment of an enhanced dialogue stream in Sweden was found to reduce audio complaints [44]. In April 2023, Amazon Prime added a Dialogue Boost option on Amazon Original content worldwide [15]. The feature has a medium and high setting.

In [45] reproduced environmental noise was found to influence preference for foreground-background balance. They found two clusters of participants, one group who increased the level of the background sounds in noise, and another group who increased the foreground noise or kept the balance the same. In [46] a model was developed to predict the benefits of providing dialogue enhancement technologies. This model predicts a substantial reduction in listening effort for a variety of mix ratios ranging from 4.5 to 10 dB.

Object-based audio (OBA) provides new opportunity for personalisable accessible experiences. The literature on definitions of OBA will be discussed in more depth in Chapter 7. For the purposes of this thesis, prior to that point OBA will be defined as an audio production technique where ascribed metadata allows aspects of the mix to be rendered at the user end rather than at the production stage. A requirement of this is that sounds are not downmixed into beds, rather left as distinct “objects”, i.e. audio events with associated metadata. A graphical representation of this difference can be seen in Fig. 2.1.



**Figure 2.1:** A graphical representation of the differences between OBA and traditional audio in broadcasting

Interest in OBA and its potential applications has grown significantly over the last decade or so. Since OBA leaves some of the audio rendering until the user end it has the capability to give users a personalisable audio experience. OBA has the potential to enable a wealth of creativity within content production. Four key areas of development have been identified for OBA [47]. These are interactivity, immersion, personalisation and accessibility. Examples of this include audio device orchestration [48], (where users can create a surround sound system with smartphones, tablets, etc), provision of alternative languages [49], and spatial audio for interactive broadcasting [50].

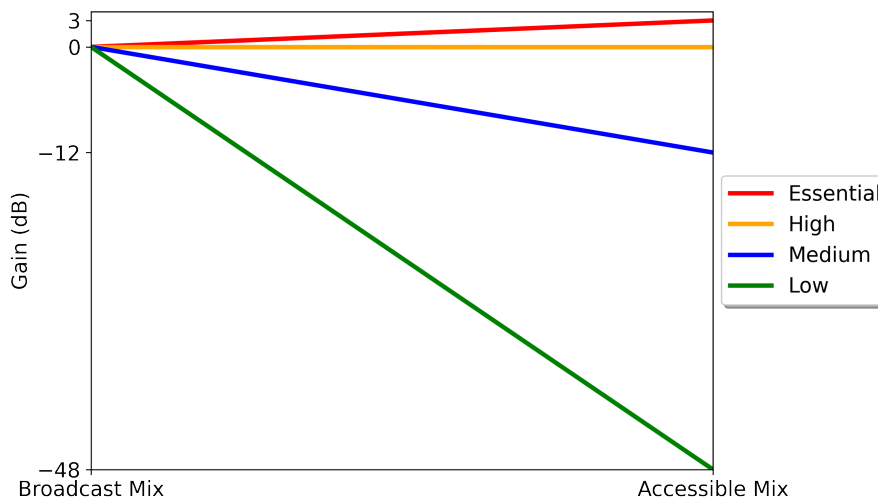
## 2.3 What is Narrative Importance?

### 2.3.1 Narrative Importance

The Narrative Importance (NI) concept was developed in [17]. The system was designed to improve accessibility of broadcast content for those with hearing loss and has been shown to be successful in a national public trial [51]. It has since been shown that it may have benefits for other groups of the population such as neurodivergent people [52], or people accessing content not in their native language.

The NI system hinges on personalisable audio, where sounds levels can be adjusted by the user

depending on their importance to the storyline of the content. There are four categories of importance: Essential, High, Medium, and Low. Each category has a fixed associated gain. Metadata is assigned to each sound object in the mix to ‘place’ them into a category. The end user is presented with an ‘Accessible Mix’ control which looks like a standard volume slider, but adjusts the gains of the objects according to the gains shown in Fig. 2.2. At the far end of the scale (labelled Accessible Mix in Fig. 2.2) the gains applied are +3dB for Essential objects, 0dB for High objects, -12dB for Medium objects, and -48dB for Low objects.



**Figure 2.2:** The gains applied to objects in each category

### 2.3.2 The Importance of a Sound

One of the key aspects of the NI system is that it requires all audio to be assigned to one of the four importance levels. The decision about how a sound’s importance is often fairly complex and context dependent. In an interview with a Foley artist in [53], the reasoning behind the additions of certain sounds is discussed. One example is the use of a pair of clogs to create the sense of a ‘body’ in a radio drama; the clogs are recorded shuffling slightly to portray the presence of a character.

Previous work has shown that for normal hearing listeners the inclusion of contextual sound effects can increase word recognition by 69.5% [41]. The results for listeners with hearing loss proved more complex [54]. For mild to moderate loss, findings supported the idea that semantically relevant sound effects can aid comprehension.

Consideration of enjoyment and immersion should also be made. The removal or reduction of all sound effects and music can significantly change the emotional response to a piece of content. [55] undertook an analysis of the effect of film soundtracks on IMBD ratings. Their

findings indicated that the soundtrack of a film is influential on ratings, even when other factors are controlled for. Similarly, [56] found that musical genre can significantly affect a character's likeability. Findings in [57] indicate that music and/or SFX significantly increase the immersion and suspense for a viewer.

Music in TV audio is often non-diegetic and can serve many different functions, which have varying degrees of importance. Some of the purposes of music as given in [14] are listed:

- Masking - covering gaps and unwanted noise.
- Continuity - musical themes can signal continuity or discontinuity between scenes.
- Directing attention - music can be used to direct the viewers attention to particular on screen actions.
- Mood - affecting the viewers feelings about a scene.
- Communication of meaning - Clarifying ambiguous scenes. Can be used to signify location
- Memory cue - Links can be formed between visual images and music to reinforce recognition of a scene.
- Arousal and focal attention - music can increase the stimulation of viewers by activating different parts of the brain.
- Musical aesthetics - association of the pleasure of music with the programme. It can also work in the reverse.

The above functions can also be fulfilled by sound effects to varying degrees. Diegetic sound effects are very important to the congruence of a scene. For example, imagine watching someone slam a door silently. Equally non-diegetic sound effects can be crucial to the storyline by informing the audience of off-screen events. This is often done with events such as crashes or explosions to save on the production costs of creating the visual of an explosion.

Work predating the NI concept attempted to come up with ways of categorising sounds in broadcast audio by undertaking a survey of 21 participants [58]. They found that participants identified at least 7 categories of sound within broadcast audio: “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.”



## 2.4 Why Machine Learning?

### 2.4.1 Roll-out Barriers

One of the main barriers to the roll-out of the Narrative Importance system is the additional time needed within a production workflow. Most production timelines are already tight with many teams producing multiple episodes a week. In practice, implementation of NI requires assigning each audio object to an importance category. Each importance category is in effect a bus - an additional track which individual audio tracks are routed to so they can be manipulated as a group. This can be quite a time consuming process, especially at first introduction to the system. This is where ML may be of help to the NI workflow. If a ML algorithm can be trained to generate importance metadata then a plug-in based on this could be implemented to save producers' time.

### 2.4.2 How Machines Learn

Broadly speaking ML is the field of development of computer systems which learn and adapt with being given explicit rules or instructions. ML begins with data, from this data, patterns can be detected, and inferences are drawn. There are a wide variety of algorithms and models which each have their unique systems for 'learning'. In this section, some general definitions will be given to give context to the main methods used in this thesis.

Firstly, supervised learning is a method where the dataset being used has labels [59]. These labels are used to train the algorithm and then verify its performance. This is usually done by splitting the dataset into training and testing sets (sometimes a third split is made, to additionally give a validation set, though this is not done in this thesis). The training set is used to train the model, and the testing set is left as 'unseen' data. After training, the testing data is fed to the trained model, and its performance can be assessed by comparing the model's labels with the true labels.

Another method of learning is unsupervised learning, where a machine is trained to cluster data into groups based on features of the data. An example of this could be grouping music into similar genres. Features are the input variables of ML models. In the context of audio, features are usually calculated from the audio itself. According to [60], some examples of common audio features are zero-crossing rate, log attack time, short time energy, root mean square (RMS), peak frequency, and short-time Fourier transform.

A third type of ML is semi-supervised learning. As the name suggests, this falls somewhere between supervised and unsupervised, where the dataset might not all have labels, or the labels could be considered ‘weak’ [59]. This is the situation that will be encountered in this thesis. Some of the dataset will have ‘ground truth’ values to compare to, but these are subjective and should be considered flexibly.

### 2.4.3 State-of-the-Art

Recent developments have been made in the area of machine learning for automatic mixing. A large proportion of the work has been focused on intelligent music production, rather than broadcast audio. Venkatesh et al. explored the idea of using word embeddings as the input layer for a neural network (NN) [61]. The NN generated EQ settings from the semantic embeddings. This technique generated an improvement in performance over models without the semantic layer. [62] used ‘out-of-domain’ data such as wet or processed multi-track music recordings (as opposed to clean, dry tracks) to train supervised deep learning models. The model successfully achieves automatic loudness, EQ, DRC, panning, and reverberation music mixing. [63] developed an algorithm, termed Wave-U-Net, to carry out intelligent drum mixing. User preference ratings of the Wave-U-Net mix were found not to be significantly different from a human-made mix.

[64] developed an algorithm called “You Only Hear Once” (YOHO) which carries out audio segmentation and sound event detection. YOHO was shown to be six times faster than segmentation by classification. Segmentation has applications in audio content analysis, speech recognition, audio-indexing and music information retrieval.

Another area of interest is that of emotional classification of music. As discussed, the emotional relevance of music can impact its importance within the context of a narrative, [65] developed a linear regression model for emotionally rating music based on features that map to arousal and valence. Another work presents an application for classifying emotions in film music [66]. Nine emotional states are modelled, and colours are assigned to each state using colour theory. Subjective testing is utilised to validate the emotional model.

### 2.4.4 Previous Work

The use of ML to assign the Narrative Importance metadata has already been explored in [18]. This algorithm forms the starting point for the work in this thesis and will be discussed in more

depth in Chapter 6. The work used the data collected in chapter 10 of [17]. The data was collected by sending out a survey and task to 33 audio professionals. The task was to assign the importance levels for an excerpt of the object-based radio drama, ‘The Turning Forest’ which was produced as part of the S3A project [67]. Despite the results from the task having very low agreement amongst participants the results from the ML carried out in [18] showed promise. The approach taken utilised transfer learning to calculate ratios of speech, music and sound effects in each object. Then other features were calculated for each object and used along with the classification data to model the importances using a mixture model based on Stochastic Variational Inference [68].

The transfer learning done was based on a CNN termed VGGish [69]. The work used a dataset of YouTube videos [70] to test the viability of using CNNs on audio; previously CNNs had mainly been used to successfully classify images. From this work the VGGish algorithm was trained, which was based on the VGG algorithm for image recognition [71]. Other models have since been trained on similarly large datasets to classify audio [72]. The scope of this thesis did not investigate changing the transfer learning algorithm, however future work could focus on this.

## 2.5 Conclusions

This chapter has outlined the area of broadcast audio accessibility. It is clear that a single solution to the problem does not exist but the area of audio personalisation will at least provide users some control over their listening experience. The NI system is one such method which has been shown to be effective and desirable in nationwide public trials [51]. Rolling out the NI system requires development of tools to assist production staff and subsequently reduce additional strain to workflows. This thesis will continue work done to this end, with the implementation of a ML algorithm to generate importance metadata for audio objects.

# Chapter 3

## Data Collection Design

### 3.1 Introduction

This chapter will discuss the design and implementation of the data collection survey. The aim of the survey was to provide a database of Narrative Importance (NI) metadata from audio professionals and members of the public. This database would provide the labels for the training of the machine learning algorithm. The survey consisted of a questionnaire and an assignment task, which are both presented.

### 3.2 Method

The survey and task design lead directly from the work done in Chapter 10 of [17]. In that body of work a similar experiment was run, initially to gauge production staffs' opinions of NI and to see how assignments varied across different areas of the discipline. The format of that was a questionnaire followed by an assignment task. The assignment task presented participants with an interface that resembled a digital audio workstation (DAW) and asked them to choose the importance of each object in an object-based mix. The task used a 100 second segment of an open source radio drama called 'The Turning Forest' [67].

Using that design as a starting point the questions were broadened to include members of the public. The wording of the introduction and instructions were changed to make them more accessible to people without knowledge of audio terminology. The task was largely coded by Dr. Matthew Paradis (BBC R&D) since he had coded the original interface used in [17].

The assignment task was modified to present each participant with three pieces of content rather

than just one. In total 9 scenes were selected from 5 different content pieces and participants were each presented with 3 of these nine. The content will be discussed in more detail in Chapter 4. The 3 were selected in a pseudo-random manner whereby the number of times each content piece had been presented was logged and the site would attempt to choose the scenes with the least presentations. This system was decided upon in an attempt to keep the spread between scenes even. Unfortunately this system was not foolproof since the count was logged when participants loaded the page, rather than after they had completed the tasks. Some erroneous counts were logged, despite the author checking and correcting this count daily, resulting in uneven distribution over the content pieces. The number of participants who completed each scene ranged between 15 and 19. The final counts for each of the nine content pieces can be seen in Table 3.1. Details of the content pieces will be discussed in Chapter 4.

Content	N
Protest	19
Vostok-K S1	19
Vostok-K S2	16
Casualty S4	15
Casualty S46	15
Casualty S49	18
Penguins Credits	16
Penguins Seal	15
Football	16

**Table 3.1:** A table showing N, the number of participants who completed assignments for each of the 9 the content pieces.

The questionnaire was hosted on qualtrics.com and the assignment task was hosted on the University of York’s AudioLab web server. In order to link the two parts of the study together participants were guided to create an anonymous unique ID consisting of the last three digits of their postcode and phone number.

### 3.2.1 Questionnaire

The first section of the questionnaire gave participants some background information on the study and explained briefly what they would be asked to do. They were then presented with a consent form that explained how their data would be used and stored. If participants declined consent they were directed to the final page of the survey thanking them for their time.

After reading the information and giving consent, the questionnaire began by asking parti-

Participants were asked some demographics questions. All demographics questions were asked with the aim of comparing the assignments from different groups of participants. They were asked if they identified as neurodivergent, what their first language was and if they had any known hearing loss. If participants answered yes to having hearing loss they were then asked if they had hearing aids and if they were wearing them to complete the task. Then participants were asked if they currently or had previously worked in audio. From this answer participants were sent down one of two routes. If participants answered no, they were asked two questions: whether they were familiar with digital audio workstations (DAWs) and whether they had any musical training.

Participants who answered yes to the audio profession question were asked 10 multiple choice questions with some opportunities to provide further information in text boxes. The closed-ended questions began by asking some data questions about how long they had worked in audio, the country they most often worked in, the medium and genre they worked in most often, and their job role. They were then asked a series of questions which were more specific to the task. This began with asking how familiar they were with object-based audio (OBA). They were then asked if they consider the sounds' importance to the narrative when mixing. They were asked to rate how comfortable they would feel with the audience being able to control the levels in a mix. Then a few questions were asked regarding the use of AI. Firstly, whether they had already used AI in their workflow, then how comfortable they would feel using an AI plug-in.

Finally all participants were advised to complete the task using headphones but in the interest of making the task as accessible as possible, were told they could complete it on any device. They were then asked what type of device they were using with 4 choices: headphones, earbuds, inbuilt speakers or loudspeakers. They were asked to input the make and model of the device if they knew it.

### 3.2.2 Assignment Task


The assignment task began by asking participants their memorable ID and then asking them to listen to the first content piece in full. The content pieces will be discussed in more detail in Chapter 4. Written descriptions of the scenes (and screenshots for scenes that had accompanying video) were presented at this point. These descriptions and screenshots can be seen in Chapter 4. Once the participants had listened through to the whole piece they were then able to click start to navigate to the task page. The following are the instructions that the participants were given for carrying out the assignment task. These instructions were available to participants throughout the task.

### Instructions

We are asking you to categorise sounds from the scene you just listened to based on how important the sound is to understanding the story.

For each sound you should select one of the following categories:

- Essential (**red**),
- High Importance (**yellow**),
- Medium Importance (**blue**),
- Low Importance (**green**),

based on how important you feel it is to following the content's narrative. Select the categories using these buttons  below.

**You must assign at least one sound in each importance category in order to submit your results.**

Each object has a bar showing the periods when the object is active. You can navigate the track by clicking on these bars. Each object has a preview button below its name (left hand side) allowing you to listen to a small clip of the sound (this will pause the overall mix). You can start and pause the track using the Play/Pause button.

---

### Personalised Playback

There are three different playback options demonstrating a personalised audio implementation which would allow the user to adjust the mix based on their hearing needs or listening scenario.

The three playback options alter the sounds' reproduction levels based on the category you have assigned them to. The options are:

- **Original Mix:** This is the original broadcast mix. All objects are reproduced without level alteration.
- **Moderate Mix:** The sound elements most vital (category Essential) to un-

derstanding the narrative are increased in volume, the level of category High Importance remain unchanged, whilst less important objects (category Medium and Low) are slightly attenuated. This mix is designed for listeners with mild hearing loss or listeners in a noisy environment.

- **Accessible Mix:** The most vital sound elements to the narrative are further increased in volume whilst low importance objects are heavily attenuated. This mix is designed for listeners with moderate to severe hearing loss or listeners in very noisy conditions.

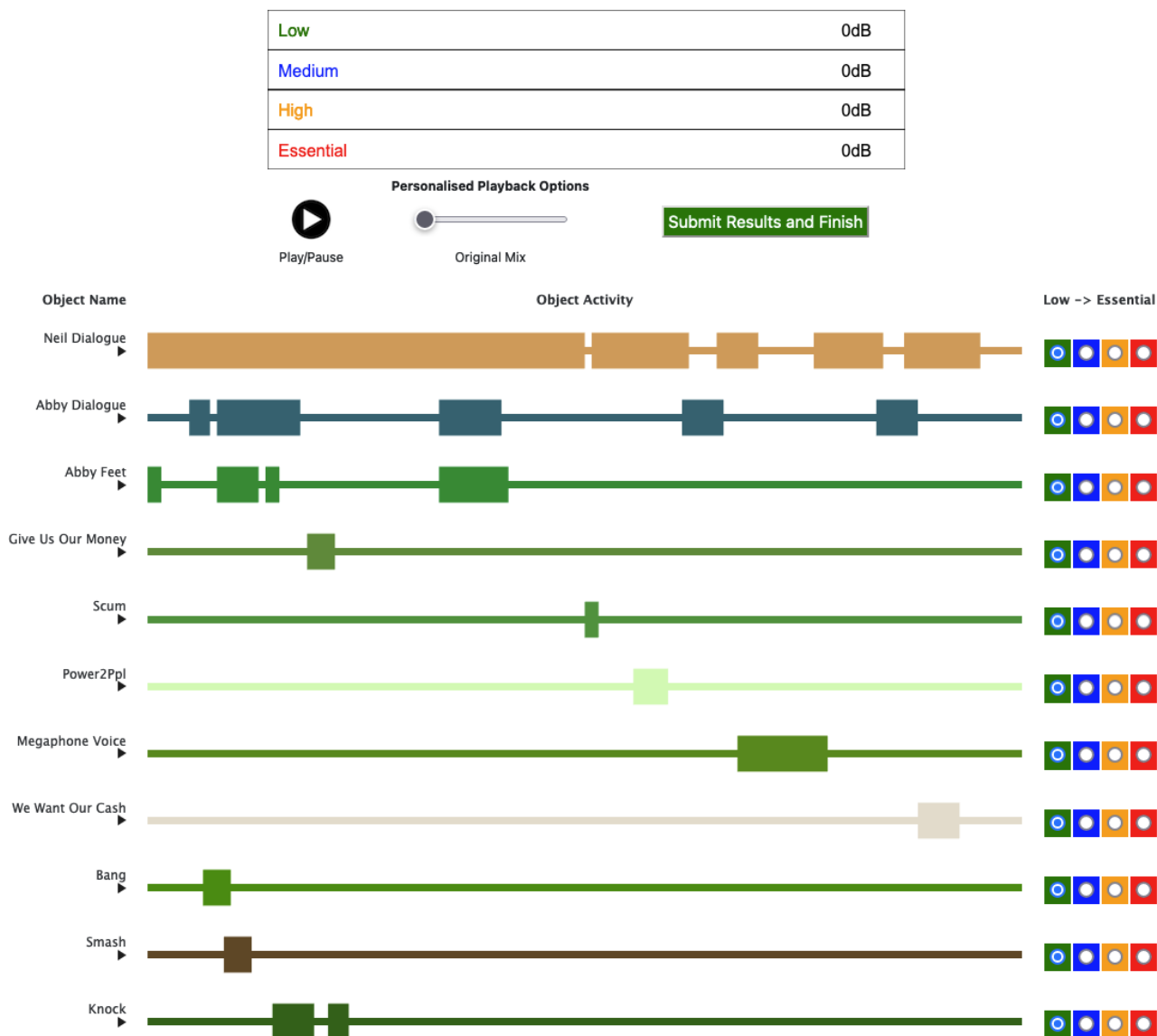
The control is to the right of the Play/Pause button. The default position is the Original Mix. The sounds all start in the Low Importance category. This means that if you move the playback slider to the Moderate or Accessible Mix position then you will hear the entire mix get quieter. We suggest trying this and then moving one object to higher importance categories to hear how the assignment affects the volume of the sounds.

As you categorise the sounds in the mix we suggest using the playback control to compare the original mix with the moderate and accessible mixes. As you move the playback control you will see the dB values in the grid change according to the control position.

---

Fig. 3.1 is a screenshot demonstrating the DAW-like interface the was used for the task. Five trial participants tested the interface before it was released. Feedback from these trials was used to improve the instructions and make them clearer. Trial participants who had little audio knowledge or listening test experience provided the most feedback on the instructions and interface, as would be expected.





**Figure 3.1:** A screenshot showing an example of the interface for the Protest content. Not all objects are shown.

The images used to represent the audio were generated using a code written in MATLAB. The code plotted 0.5 second blocks whenever the root mean square (RMS) was greater than  $5 \times 10^{-5}$ . For stereo tracks the code calculated the maximum RMS value for the two channels. It was decided to visualise stereo tracks as mono to avoid confusing participants without any audio experience and to keep the interface as simple as possible.

Participants had an unlimited amount of time to complete the task.

It was decided to enforce assigning at least one object to all 4 levels of importance. The previous study (in chapter nine of [17]) that asked production staff to complete this task had not enforced the use of all four levels. The result of this was that a number of participants only used three

of the importance levels and two participants didn't assign any objects to the essential and low categories. In order for the NI concept to fulfil its true purpose the full scale of the importance levels will ideally be used (content depending). This discrepancy was believed to contribute to the very low level of agreement between participants (Fleiss' Kappa = 0.11) for that study. It was hypothesised that enforcing the use of all four levels might help to combat this.

### 3.2.3 Critical Reflection

If the test were run again then several changes would be made based on the results. Firstly, more tracking of metrics could prove lucrative. The time each participant spent on each of the scenes was tracked, namely the time between them clicking the start button and the submit button. Other than this, no metrics tracking how they interacted with the interface were recorded. It would be useful to know whether participants changed the setting of the 'Personalised Playback Options' since this allowed them to hear how their assignments changed the mix. If participants didn't move this slider then the mix they heard was always the original broadcast mix.

A way of determining whether participants listened through the full scene after assigning all of the objects would be helpful. The number of times they used the 'Play/Pause' button could indicate this, especially if combined with timestamps and information about their assignment activity. Tracking assignment activity might provide information about which objects participants found more difficult to decide. The use of the preview buttons would show whether participants listened to each individual sound or relied on the track names and what they could hear in the mix. It's possible that this information could indicate more troublesome objects, as listening to an object multiple times implies difficulty in deciding its importance.

Participants were advised to 'set a comfortable listening level' and asked not to change this throughout the task during the questionnaire part of the task and reminded of this upon entering the task. The participants were asked to set their listening volume during the first example, meaning that each participant had a different content piece to set their volume to. Whilst each scene was normalised to -23 LUFS it may be argued to have been better to have a short clip that was the same for all participants to set their volume to.

One suggestion that arose during trials but was too lengthy to implement at that stage was a training video and/or mini task. This would have been particularly helpful for participants with no audio background, since the interface would have been very unfamiliar for them. A video embedded in the main task page demonstrating each of the controls and how they should

be used would have allowed participants to refer back to it as they completed the task. Equally a mini task that walked them through the controls by highlighting them and requiring them to interact with them before moving on would have been useful, though this may have frustrated some participants with high levels of audio knowledge and experience using DAWs.

### **3.3 Conclusions**

This chapter has outlined the method used to run the data collection survey. The survey began with a questionnaire which led participants to complete an assignment task. Participants were asked to assign each object in a mix to one of four importance levels based on how important they felt to object was to the overall narrative of the scene. Each participant completed this task for three scenes. Finally, the chapter finished with a critical reflection of the work. A number of points for improvement were identified. The content used will be discussed in Chapter 4, and the results of the survey will be reported in Chapter 5.

# Chapter 4

## Content

### 4.1 Introduction

This chapter will discuss the content used in the data collection study. Five suitable pieces of content were sourced, with 9 excerpts being selected from them for use in the data collection task. All content was either open-source or provided by the BBC, bar one piece which was sourced from a contact of the investigators. Significant processing was undertaken to ensure the content was object-based, in line with the definition given in Section 2.2.2, and compatible with the Narrative Importance system. The chapter will begin by defining the categories of sounds within the content. Next will be a description of each of the 9 scenes. The processing of each piece of content will then be outlined. This will be followed by a brief discussion of content that was provided but deemed unsuitable. The end of the chapter will outline guidance for the creation and storing of NI suitable content. The number of objects in each category is discussed for each content piece below. A graphical representation of the breakdown of all scenes can be seen in Section 4.5.

### 4.2 Object Categories

Before discussing the selected content it is worth defining some terms that will be used to simplify the discussion of the sounds. Across the 9 selected scenes there are a total of 198 distinct sound objects. Generally within broadcast audio, sounds are defined as being one of three types: speech, music, or SFX. It is possible to divide these categories further in many different ways. Often sounds are further categorised by whether they are diegetic or non-diegetic

Type	Term	Description
Speech	Main Dialogue	Dialogue spoken by named characters including narration and commentary
	Extra Dialogue	Dialogue spoken by extras
	Vocal Atmos	Background chatter, crowd noise
SFX	Acute SFX	Distinctive (usually short) SFX
	Atmos SFX	Background, longer SFX tracks
Music	Music	Any non-diegetic music

**Table 4.1:** A table of categories which will be used to discuss the audio objects

(heard by the characters or not). For the purposes of this thesis these terms were avoided, since they aren't applicable within all genres (e.g. sports, documentary). Instead a set of terms which appropriately describe the sounds contained within the 9 scenes are defined in Table 4.1. These categories will be used throughout the following work to discuss the content.

## 4.3 Selected Content

### 4.3.1 Protest

'Protest' is a radio-drama that was produced as part of the S3A project [73]. The three radio dramas produced in this work were '*designed to demonstrate different features of 3D audio.*' 'Protest' is described in [67] as depicting a protest '*being staged outside a bank. The scene begins inside the bank and evolves from a front dominant image to full 3D sound as the action moves from indoors to the protest outside. The scene demonstrates immersive crowd atmos, individually identifiable voices popping out of the atmos, and moving localizable sources at different heights.*'

In full 'Protest' is 2 minutes 38 seconds in duration. For the assignment task a 1 minute 2 second segment was extracted. The participants for the task were presented with the following description of the scene before undertaking the task:

This piece is a segment from a radio drama called Protest. Protest begins with two members of staff, Neil (male) and Abby (female), inside a bank whilst a protest is happening outside. Neil and Abby leave the bank and enter the crowd of the protest. The selected excerpt focuses on the section where the staff decide to leave the bank and enter the protest.

The segment was chosen to ensure the inclusion of specific sound objects. There were 27 objects

in this scene. Tables detailing the names of all of the objects for each scene can be found in Appendix A. Before the staff leave the bank the protest can be heard happening outside. One point of interest is when children are heard shouting ‘power to the people’ outside and Neil’s next line is ‘who brings a child to a protest?’ Another notable moment is the sound of a window being smashed, which causes Abby to gasp. Once Neil and Abby leave the bank there are many sounds to consider as they travel through the crowd. As well as recordings of crowds chanting there are individual voices amongst the crowd. It was hypothesised that not all of these vocalisations would be classed as ‘essential’ by participants and would test the algorithm’s ability to discern between important dialogue and less important speech. The scene did not contain any music tracks, however some of the crowd ambience tracks contain drumming. The scene contained 27 objects in total. There were 2 main dialogue, 5 extra dialogue, 10 vocal atmos, 6 acute SFX and 5 atmos SFX objects.

### 4.3.2 The Vostok-K Incident

‘The Vostok-K Incident’ is a 13 minute science fiction radio drama produced to investigate orchestration of personal devices [48]. A trial was run where participants could use personal devices such as phones and tablets as additional speakers to create a surround sound system. ‘The Vostok-K Incident’ is described in [48] as: *‘The drama is set during the Cold War, and takes place in the cockpit of a fighter aircraft. The pilot receives a radio message and is redirected to investigate a mysterious spacecraft. Whilst the action unfolds, a taped conversation between a general and a cosmonaut (occurring 20 years after the events of the drama) helps to explain aspects of the story.’*

Two segments were extracted for use in the assignment task. They will be referred to as extract 1 and 2 respectively. Both scenes have dense soundscapes, including music, internal and external engine sounds, gunfire, thunder, rain, and explosions.

#### Extract 1

The first extract was 1 minute in length. The participants were asked to read the following description whilst listening to the segment before undertaking the task:

This piece is a segment from a science fiction radio drama called The Vostok-K Incident. The Vostok-K Incident is set in the cockpit of a Cold War fighter aircraft. Joe (male), the pilot, receives a radio message from Sam (male), his correspondent,

instructing him to investigate a mysterious spacecraft. The drama also features an interview between a general (male) and a cosmonaut, Tatiana (female), which provides some explanation of what is happening in the story. The selected scene features the attack on Joe starting and Tatiana trying to warn him not to retaliate.

Several considerations informed the selection of the first segment. Firstly a section which included music was chosen. The music in the segment builds tension in the scene as the attack intensifies. The segment also features radio transmissions which are created using two objects, one which is the dialogue and another which is the static of the radio. There are 27 objects in total. 5 of these were classed as Main Dialogue, 17 as Acute SFX, 4 as Atmos SFX and 1 as music.

### **Extract 2**

The second segment was 45 seconds long. The following is the description provided to participants before embarking on the task:

This piece is a segment from a science fiction radio drama called The Vostok-K Incident. The Vostok-K Incident is set in the cockpit of a Cold War fighter aircraft. The drama also features an interview between a general (male) and a cosmonaut, Tatiana (female), which provides some explanation of what is happening in the story. The selected scene features Joe, under attack, realising that the mysterious aircraft he is fighting with is his own fighter jet.

This segment was selected as it features an alarm and verbal warning which directly precede an explosion. It also contained music which builds throughout the scene and adds emphasis to the explosion. The segment contained 21 objects in total. These were split into 2 Main Dialogue, 14 Acute SFX, 4 Atmos SFX and 1 Music object.

### **4.3.3 Casualty Season 33 Episode 38**

Three scenes were extracted from an episode of the BBC drama 'Casualty'. 'Casualty' is the longest running medical drama in the world and is set in the Accident and Emergency (A&E) department of an NHS hospital. The entire episode had a runtime of 49 minutes 25 seconds. There are multiple plot lines running throughout the 53 scenes making up the episode. A large number of the scenes in this episode were unsuitable for use due to the inclusion of a deaf character who speaks sign language. The limitation of using only audio in the trial meant that

these scenes would be incomprehensible to participants. In the interest of brevity, only the scenes selected for use in the study will be summarised with relevant context given. A full breakdown of the plot and scenes can be found in chapter 11 of [17].

#### **Scene 4**

The first excerpt selected for the study was a 40 second section of the fourth scene in the episode. This scene centres on the aftermath of a moped accident which occurred in the second scene. The participants were asked to read the following description, whilst listening to the excerpt. They were also provided with the three screenshots in Fig. 4.1.

This piece is a scene from an episode of *Casualty*, a drama set in the Accident and Emergency department of a hospital. In this scene Dani (female), a character previously seen in other episodes, pretends to be a paramedic and attempts to provide medical assistance to Barbara (female). Barbara was hit by a moped in the opening scene and is lying on the floor struggling to breathe. Dani takes a penknife out of her bag intending to perform a tracheotomy, a procedure where a tube is inserted into the windpipe through an incision in the throat. Ruby (female) and Iain (male), genuine paramedics who know Dani arrive before an incision is made. Upon Ruby and Iain's arrival Dani runs away, dropping the penknife. A group of people who saw the accident happen are gathered around Barbara throughout the scene.





(a) A frame showing Dani checking Barbara's pulse



(b) A frame showing Ruby and Iain arriving



(c) A frame showing Dani placing the penknife on Barbara's neck

**Figure 4.1:** The three screenshots provided to participants for Scene 4 from Casualty

The scene was chosen as it was one of the most sonically diverse scenes in the episode. It contained 27 objects in total. The notable sound objects were the sounds associated with the penknife, the ambulance siren and engine sound, the laboured breathing of Barbara, and the voices of extras expressing confusion about who Dani was when she runs away. The objects were categorised as 3 Main Dialogue, 4 Extra Dialogue, 2 Vocal Atmos, 11 Acute SFX and 7 Atmos SFX. One of the Vocal Atmos objects was the breathing object, whilst this is not strictly a vocalisation it was felt that this should be classed differently from the other SFX tracks.

### Scene 46

The second excerpt from Casualty was 1 minute 27 seconds in length, making it the longest excerpt presented to participants. The scene shows a conversation between Ruby and Dani. It was chosen due to its inclusion of emotive music at the end of the scene. The participants were given the following description of the scene alongside the frames shown in Fig. 4.2.

This piece is a scene from an episode of Casualty, a drama set in the Accident and Emergency department of a hospital. In this scene Ruby (female), a paramedic,

talks to Dani (female), a character Ruby befriended after Dani's mother died. Ruby has previously encouraged Dani to pursue a career as a paramedic. Now Ruby tells Dani they cannot be friends because Dani endangered an injured woman earlier in the episode by attempting to treat her. Dani is lying on a hospital bed in a busy ward and Ruby is stood at the foot of the bed.



(a) A frame showing Dani in the bed



(b) A frame showing Ruby standing at the foot of the bed

**Figure 4.2:** The two screenshots provided to participants for Scene 46 from *Casualty*

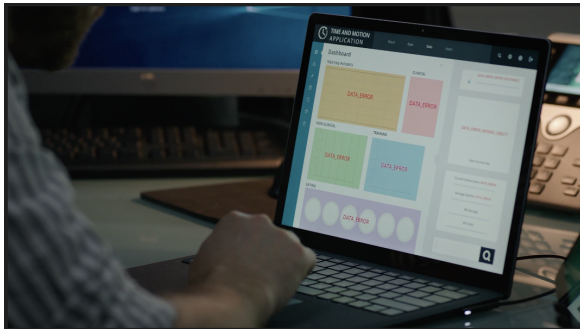
This excerpt contained 17 objects despite being the longest excerpt chosen. It was decided that the length was acceptable given the low number of objects. Cutting the scene further would have eliminated too many objects and resulted in the scene feeling incomplete for participants. This scene was included as it was one of the only scenes in the episode that contained non-diegetic music. The majority of the objects aside from the music and dialogue were atmospheric, creating the sound of a hospital ward. The objects were categorised as 2 Main Dialogue, 8 Acute SFX, 6 Atmos SFX and 1 Music object.

### Scene 49

The final scene from *Casualty* was a 52 second section of scene 49. The scene features three members of staff having a conversation in one of the main ward areas. The participants were provided with the screenshots in Fig. 4.3 and the following description to read:

This piece is a scene from an episode of *Casualty*, a drama set in the Accident and Emergency department of a hospital. In this scene, Ciaran (male) discovers that the data from the new initiative has been deleted from his computer. The initiative monitored staffs' activity throughout the working day and was explained in a scene near the beginning of the episode. The staff generally objected to the

initiative. The scene starts with Ciaran in his office looking at his computer, he then exits his office to discuss the missing data with two staff members, Archie (female) and Will (male), in the main ward area.



(a) A frame showing the computer data error



(b) A frame showing the three main characters of the scene

**Figure 4.3:** The two screenshots provided to participants for Scene 49 from Casualty

This scene contained 20 objects in total. It begins with a series of clicks and beeps, recognisable as a computer error. Amongst the atmospheric objects are the sound of a lift, including the spoken ‘lift going down’ message, and a phone ringing. The objects were placed in the categories: 3 Main Dialogue, 10 Acute SFX, and 7 Atmos SFX.

#### 4.3.4 Penguins: Spy in the Huddle

‘Penguins: Spy in the Huddle’ is a 3-part documentary which used over 50 spycams to film 3 species of penguins over the course of a year. Two excerpts were chosen from episode 3 of the series.

##### Extract 1 - Opening Credits

The opening credits of the episode were chosen as the first extract. This was 47 seconds long and the following description was provided to participants, along with the screenshot below:

This piece is the opening credits from an episode of a nature documentary called Penguins: Spy in the Huddle. The documentary was filmed over a year using 50 spycams disguised as life-size penguins to capture footage of penguins. The series details three different species of penguins, emperor penguins in Antarctica, rockhopper penguins on the Falkland Islands, and Humboldt penguins in Peru’s

Atacama Desert. The opening credits feature shots of the three penguins (and other wildlife) filmed by the spycams.



**Figure 4.4:** The screenshot provided to participants

This scene was chosen to be included in the study as most programming contains some kind of opening credits. The scene contained 23 objects, in the categories: 1 Main Dialogue, 7 Acute SFX, 14 Atmos SFX, and 1 Music object. The music is jovial and whilst it doesn't add to the narrative could be considered as important to signify the beginning of the programme. The credits also have a voiceover which introduces the programme. The video shows various clips of penguins from the show so the accompanying sounds are heard.

### **Extract 2 - Cormorant and Seal Scene**

The next scene was 55 seconds in length. It features cormorants, seals and penguins. The following description was provided to the participants along with the two screenshots below:

This piece is a scene from an episode of a nature documentary called Penguins: Spy in the Huddle. The documentary was filmed over a year using 50 spycams disguised as life-size penguins to capture footage of penguins. In this scene seals are in the sea, preventing the Humboldt penguins and cormorants from fishing. The cormorants attack the seals, moving them on to the shore, allowing the penguins and cormorants to enter the water to fish.





(a) A frame showing the beach full of penguins



(b) A frame showing the penguins in the water

**Figure 4.5:** The two screenshots provided to participants

This scene was chosen as it contained a variety of different sound objects, 25 in total. As well as calls of the three animals and a narrator, there were many tracks of water and wind sounds to build atmosphere. The number of objects in each category were: 1 Main dialogue, 16 Acute SFX, 7 Atmos SFX and 1 Music.

### 4.3.5 Football

The final piece of content was an enhanced clip of a football match. In total, 3 minutes 2 seconds of a match was provided. The selected excerpt was 20 seconds in length. Audio objects from a broadcast of a football match were combined with enhanced audio to create enough objects for the task. The participants were given the following description:

This piece is an excerpt from a broadcast of a football match. The match was broadcast with additional audio such as a ‘kick enhancement’ which emphasises the sound of the players kicking the ball. In the clip after a couple of passes a player takes a shot on goal but misses.

The scene contained only 11 objects, making it the least complex scene included in the study. This was thought to be an accurate portrayal of most sports broadcasts. Commentary and crowd noise made up the majority of the objects, there was also the kick enhancement sound and a player shouting on the pitch. The sounds were categorised as 1 Main Dialogue, 1 Extra Dialogue (the player shouting), 6 Vocal Atmos, and 3 Acute SFX.

## 4.4 Processing of Scenes

This section will discuss the selection of the scenes and the processing that was required to make each scene suitable for use in the survey. Some of the processing would be needed to use NI metadata regardless of how it was being assigned. Other processing was done to make the content more accessible to participants, for example track names were changed from names like FX, foley, dial to be specific descriptions of the track content. The detailed discussion of processing here pertain to the specific excerpts selected for the data collection task. The full sessions for each content piece would require more processing for use with NI. Tables for the two open-source pieces, ‘Protest’ and ‘Vostok-K’, detailing the conversion from the original tracks into the objects used can be found in Appendix B.

### 4.4.1 Processing Protest

As mentioned previously ‘Protest’ was created for the S3A project which was an exploration of spatial audio. The original session was made up of 64 mono tracks, 12 of which were left and right channels of stereo tracks. Once the tracks were downmixed to create 6 stereo tracks, the session contained 58 tracks.

Sixteen of the tracks in the session were reverb tracks created using the SPAT software from IRCAM. These tracks were removed as they were not required for this work. In fact they would actively counteract the NI metadata since they were effectively downmixes of the entire session. This would result in most sounds being heard as part of two separate objects, potentially with different importance assignments. For example, if the reverb tracks were assigned to the High category, some Low Importance objects may be heard on the reverb tracks at a much higher level than intended.

Six of the tracks contained dialogue. The first two were the dialogue of the two main characters, Neil and Abby. These two tracks were left unchanged but were renamed to ‘Neil Dialogue’ and ‘Abby dialogue’. The next 4 tracks were made up of various distinct voices coming out of the crowd. ‘Dialogue 3’ also contained the sound of Abby’s feet. These four tracks were split into 6 objects for the section used in this test. The names of the objects were ‘Abby Feet’, ‘Give Us Our Money’, ‘Scum’, ‘Power2Ppl’, ‘Megaphone Voice’, and ‘We Want Our Cash’.

The next set of tracks were five stereo FX tracks. These contained multiple objects including the bank door opening, knocking, a window smashing, drums, horses hooves, and crowd noise. These were split into the objects: ‘Bang’, ‘Smash’, ‘Knock’, ‘Bank Door Opens’, ‘Horse Hooves’,

‘Crowd Transition’, ‘Crowd + Cars’, ‘Crowd Outside’, and ‘Drums’.

The next set of tracks were 4 crowd tracks. The tracks were all split at the point that the bank door opens. The crowd sound before that point is heard as though from inside the bank, so is muffled and distant. It was thought that this sound might be rated differently to the full crowd sound after the two main characters move outside. Since participants only have limited control during the task the split had to be done prior to the task for them to be able to rate this differently. The resulting object names from these four tracks were ‘Scum Crowd 1’, ‘Scum Crowd 2’, ‘Crowd inside’, and ‘Crowd + Drums’.

There was a stereo music track which contained the intro and outro music for the piece, neither of which were included in the chosen excerpt. This track did contain the sound which became the object ‘Lowfi Boom’.

Finally there were 8 Atmos, 9 Crowd FG, 9 Crowd BG tracks. These were downmixed to stereo. Then the crowd tracks were split at the door opening section as done previously. The atmos track only began after the door opened so did not require splitting. They became the objects ‘Crowd inside Background’, ‘Crowd inside Foreground’, ‘Crowd outside Background’, ‘Crowd outside Foreground’, and ‘External Atmos’.

#### 4.4.2 Processing The Vostok-K Incident

The Vostok-K Incident was the most NI compatible of the assets before processing. It was originally intended for use with spatial audio. Two sets of stems were provided, the delivered session and the raw set. A large number of objects had been downmixed in the delivered session to a single track. This made these stems unusable for NI. Fortunately the raw stems were still in an unmixed format and needed little processing, compared to other assets. Both excerpts were treated similarly. The music spanned several tracks, so was downmixed to a single stereo track. The largest chunk of processing was the creation of the ‘Bullet Ricochets’ objects which spanned multiple tracks. The separate objects were created based on the time at which the different shot sounds happened. A table (Table B.2) of the tracks and objects can be found in the appendix.

#### 4.4.3 Processing Casualty Season 33 Episode 38

This episode of Casualty was used for a nationwide trial of Narrative Importance. Despite this, the tracks required significant processing to use them in the study. The entire episode was

provided which had a full run time of 49 mins 25secs. There were 91 tracks in total. The first step taken to process the episode into usable assets was to tag each of the 53 scenes in the episode. From here the author listened and annotated scenes which were sonically interesting and suitable.

Once a selection of suitable scenes had been chosen, the objects contained in them were inspected more closely. Scenes which contained similar objects to each other were grouped together, to avoid too much repetition.

Tracks without any audio for the chosen scenes were removed. The tracks often contained several distinct sound events and were organised by the recording method and the NI level. For example there were 12 Stereo FX tracks split into 5 High, 5 Medium and 2 Low.

Tracks were renamed and divided into distinct objects. This was partially to make them more accessible to general public members. Track names such as dial, sync, and ADR are unlikely to mean anything to people outside of broadcasting.

Some of the processing was required due to the recording techniques used on the set of Casualty. Casualty is filmed with a multiple camera setup to speed up production by minimising the number of takes. To avoid any mics being in shot lapel mics are used instead of boom mics. Unfortunately the result of this is that a lot of motion sounds get captured especially for paramedic characters as their outfits are made from a noisy waterproof material. Where possible the noise was removed from the dialogue objects and became a separate ‘movement’ object. This was not possible for all instances of noise, as sometimes the noise was too intertwined with dialogue to be separated cleanly.

Dialogue was often split across multiple tracks, even for a single character. In some instances this appeared to be due to bleed between the lapel mics when actors were in close proximity to each other. There was one instance where a single syllable from a sentence was on a separate track to the rest of the sentence. The decision was taken to ‘collect’ all of the dialogue for each character and downmix them to a single object. This was done largely to avoid confusing participants with too many objects. It is also believed that the machine learning will not handle individual words/syllables well.

#### 4.4.4 Processing Penguins: Spy in the Huddle

The stems provided for ‘Penguins: Spy in the Huddle’ were in a 5.1 format. Since the task would only support stereo format the stems required downmixing. Only three of the 6 channels



in the 5.1 stems contained audio. Unfortunately using standard 5.1 downmixing coefficients yielded a mix where the centre channel was almost inaudible compared to the left and right channels. To combat this the left and right channels were attenuated by 6dB, as well as the centre channel being attenuated by the customary 3dB. This improved the levels but still required some tweaking of the mix by the author.

#### 4.4.5 Processing Football

The football audio that was provided included tracks that were recorded using an ambisonics mic. These tracks were converted to stereo using a binaural decoder [74].

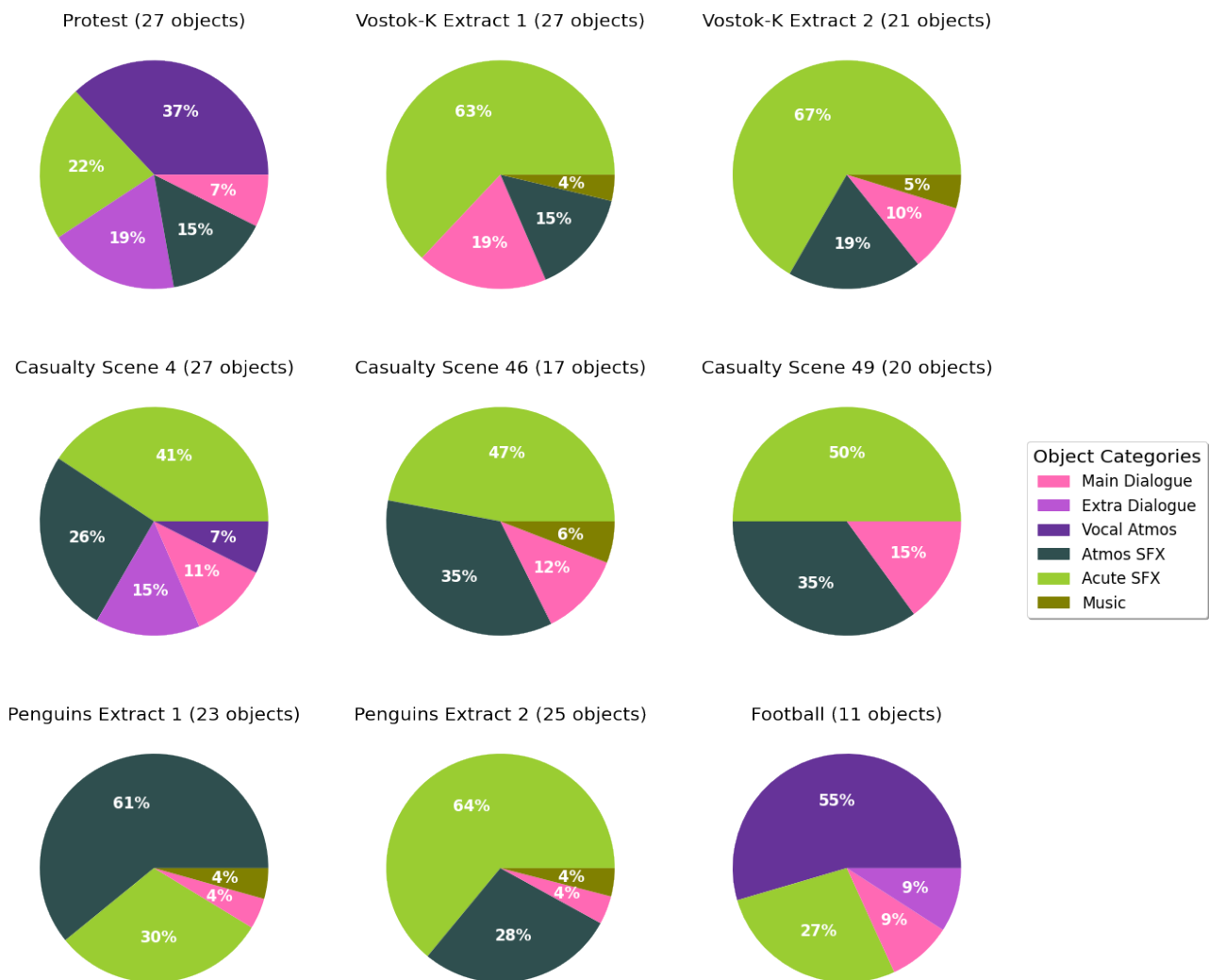
The crowd tracks were separated into objects where the crowd noise was ambient and reactionary. That is to say that in this 20 second segment the crowd sound before the near miss on goal was separated from the crowds reaction to the shot. This was done to give participants as much control as possible over how they decided to assign the importance.

The commentary track contained a lot of crowd sound. It was decided to separate as much of the crowd sound as possible. The sections where the commentator was speaking still contained a lot of crowd sound. This is often an issue in sports content, and whilst post-processing can combat this somewhat, it is an issue for accessibility that should be looked at in more detail.

There were three kick enhancement sounds on a single track. These were separated into separate objects, again to give as many options for participants.

### 4.5 Object Category Distribution

The distribution of objects for each of the scenes can be seen in Fig. 4.6. The most common object type was Acute SFX, which dominated 6 of the 9 scenes. The Protest and Football scenes had a majority of Vocal Atmos objects. Penguins Extract 1 was mainly Atmos SFX, this scene is the opening credits of the show, which is perhaps why the sounds are more building an atmosphere, rather than specific sound events relating to the narrative. All scenes contained at least one Main Dialogue object as this was a requirement of the selection process. No scene contained all 6 categories of objects, the maximum is 5 (Protest and Casualty Scene 4) and the minimum is 3 (Casualty Scene 49). 5 of the scenes contained music, and no scene contained more than one music object.



**Figure 4.6:** Pie charts showing the types of objects in each category for each scene

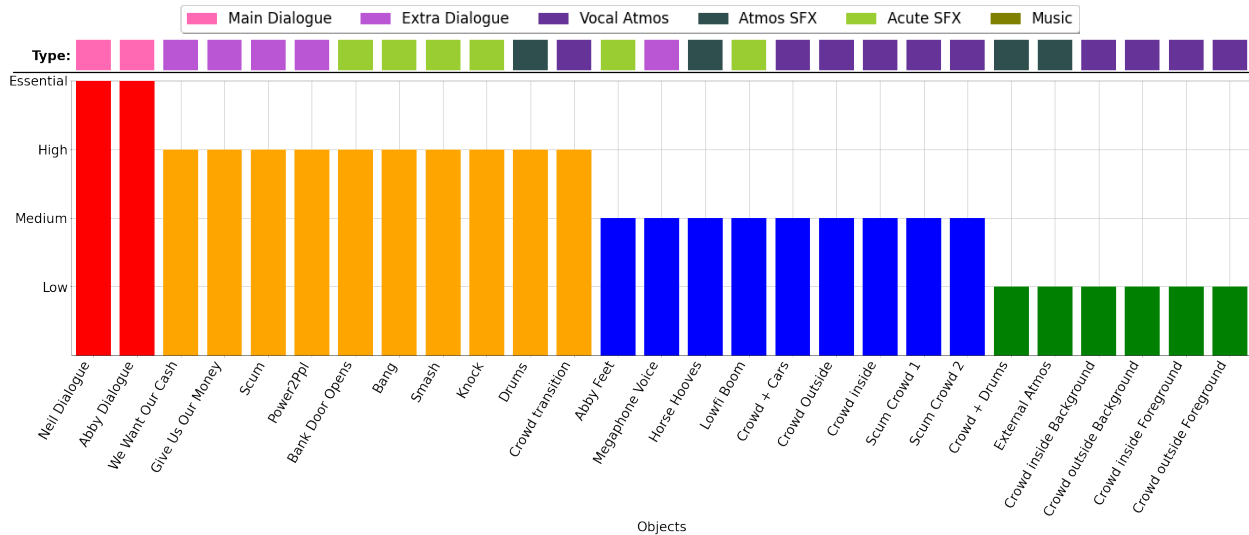
## 4.6 Ground Truths

Of the 5 pieces of content two were considered to have a ‘ground truth’ originating from work previously done in [17]. These were ‘Protest’ and ‘Casualty’. In machine learning the term ground truth is used to refer to the correct answer or label for an asset. For example if the task were to identify animals in pictures then the ground truths would be the types of animals in each picture. In this work the ground truth is somewhat unusual since Narrative Importance is a subjective concept. What is considered important may vary greatly from person to person. This section will discuss these ‘ground truths’.

### 4.6.1 Protest

Protest was one of the pieces of content which could be considered to have a ‘ground truth’. The ‘ground truth’ was taken as the work done in [17] where the author spent two days with a producer working on assigning NI metadata to ‘Protest’ and ‘The Turning Forest’.

The ground truth assignments for the segment considered here are shown in Fig. 4.7. The object categories are indicated above the bars using the same colour systems as in Fig. 4.6.



**Figure 4.7:** A bar chart showing the ground truths of the objects in Protest. The Type bar at the top refers to the type of object as defined in Section 4.2.

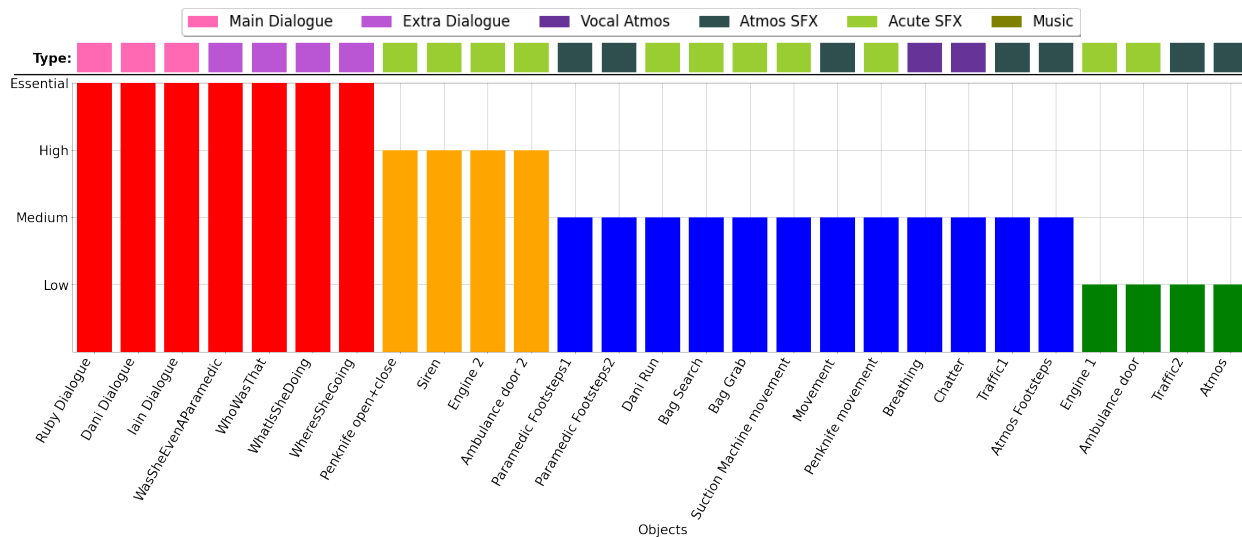
Three objects were comprised of objects that originally were assigned to different importance levels. These are ‘Crowd Inside Background’, ‘Crowd Inside Foreground’, and ‘Crowd Inside’. All three contained both low and medium objects. ‘Crowd Inside’ contained two medium and one low object, so its ground truth is taken to be medium. The other two objects contained 1 medium and 2 low, so they are taken as low. The ‘Crowd Transition’ object came from cutting the ‘Drums’ object so its importance should be viewed flexibly. The rearrangement of these objects can be seen in more detail in Table B.1.

### 4.6.2 Casualty Season 33 Episode 38

This episode of Casualty was used for a large-scale public trial of the narrative importance concept in [51]. Meaning that this was another asset with a ‘ground truth’ reference, as it was broadcast with the narrative importance control. It is notable that this ground truth was created with the visual of the scene in mind and that the work that follows this relied purely on

the audio. This may lead to significant differences in importance due to diegetic/non-diegetic considerations.

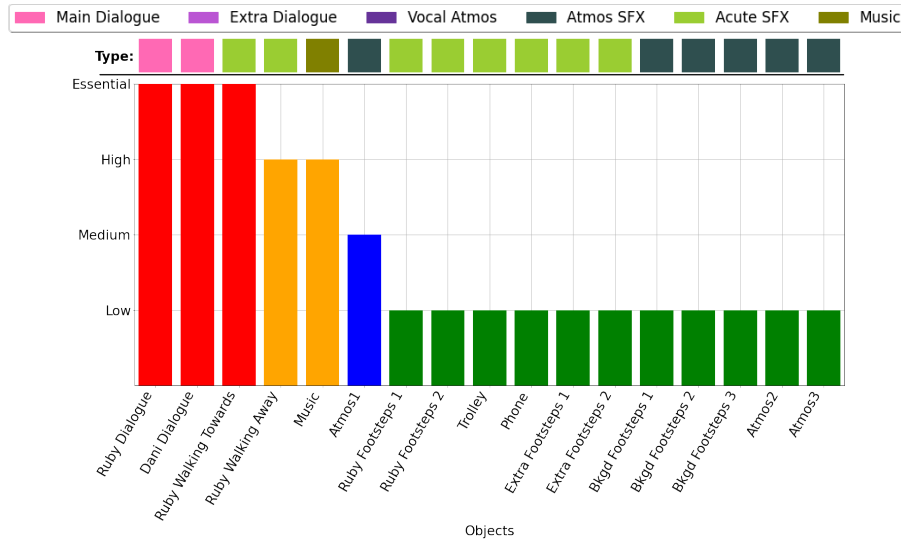
## Scene 4



**Figure 4.8:** A bar chart showing the ground truths of the objects in Scene 4 of Casualty. The Type bar at the top refers to the type of object as defined in Section 4.2.

The ‘Breathing’ object had originally been separated by the producer so that the first couple of breaths were in the high category. The remainder of the breathing was placed into medium so that it didn’t distract from the main dialogue. It was decided to make this into one object for two reasons. Firstly it was thought that the separation of the two breaths might be leading for participants. Secondly, the decision to separate a single object into two objects, to emphasize the importance of that sound without interfering with the dialogue is well beyond the capabilities of the machine learning algorithm in this work. Decisions such as this would have to be made by producers.

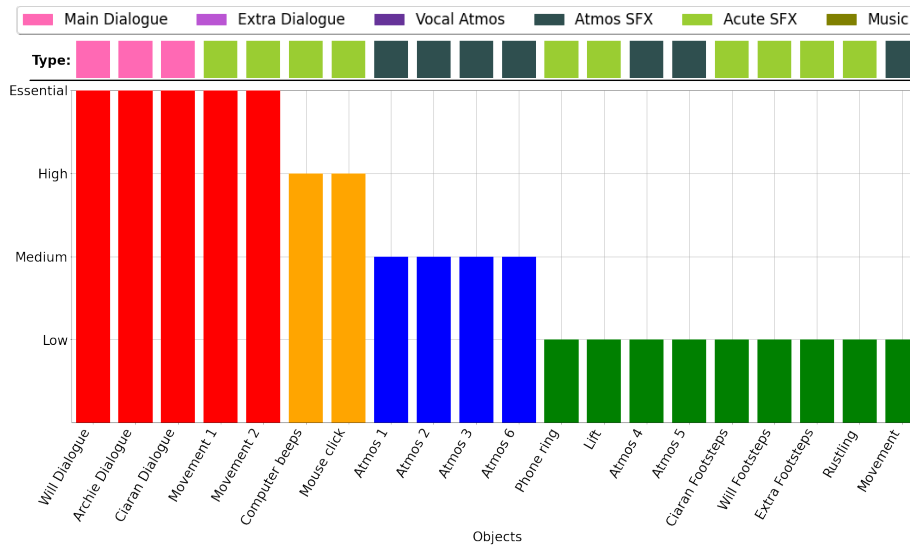
## Scene 46



**Figure 4.9:** A bar chart showing the ground truths of the objects in Scene 46 of Casualty. The Type bar at the top refers to the type of object as defined in Section 4.2.

The object ‘Ruby Walking Towards’ was made up of movement noise picked up in the dialogue track, the ground truth value of essential for this object is subsequently questionable. The audio in this object is incredibly similar to that of ‘Ruby Walking Away’ which was categorised as high importance by the original producer. It is highly probable that both objects would be categorised as high by the producer if they had both been separated from the dialogue originally.

## Scene 49



**Figure 4.10:** A bar chart showing the ground truths of the objects in Scene 49 of Casualty. The Type bar at the top refers to the type of object as defined in Section 4.2.

‘Movement 1’ and ‘Movement 2’ were again movement noise separated out from dialogue tracks. However in this case there were no other objects containing similar audio. The importance of these two objects should be viewed flexibly.

## 4.7 Rejected Content

Several content pieces which purported to be object-based were provided but discarded as they were not suitable for use in this study. This section will talk through the reasons they were deemed unsuitable.

### 4.7.1 European Athletics Championships - Women’s Hurdles

The first piece that was rejected was a clip of the women’s hurdles event from the European Athletics Championships (EAC) 2018. The EAC was run as a trial of several new technologies being implemented live [75]. One of these technologies was Next Generation Audio (NGA) which was used to implement spatial audio, commentaries in different languages and audio description. Microphones were placed around the stadium to capture different areas of the sound field. In theory this could provide some interesting objects for NI implementation. For example microphones were placed around the edge of the track to capture sounds of the athletes. Unfortunately the crowd cheers were so loud that every mic in the stadium picked them up.

This meant that any other sounds were embedded in crowd roar and could not be separated out.

### 4.7.2 Radio Panel Shows

Two radio panel shows were provided, namely a live episode of ‘Just a Minute’ (JAM) and a live episode of ‘Breaking the News’ (BTN). Both of these were discounted for similar reasons related to the nature of the genre. The objects consisted of dialogue for each of the panel, recordings of the crowd, music, and buzzers. The issue with using these pieces for this study arose from the fact that very little overlap occurred between the speech and the other objects. It was decided that it might be an interesting to see how the trained algorithm categorised the objects in one of these pieces given the lack of overlap but that it would not be worth including in the training.

### 4.7.3 The Watches Series

Finally three clips from The Watches series (Springwatch, Summerwatch etc) were provided. Each one contained only three tracks. Two of the three clips contained music, SFX, and narration. The third contained a SFX track and two dialogue tracks, an interviewer and interviewee. The SFX tracks were diegetic sound relating to the video being shown. They were continuous and could not be separated out into separate objects. The music tended to play when the narration was silent. There was no overlap between the two speakers in the third piece.

## 4.8 Content Selection Problems

There were some issues in the selection of the content. Most of these were due to the limited content available to the author but are worth discussing as areas for consideration.

The first oversight is that none of the scenes contained diegetic music. In the content provided there was no diegetic music in any of the scenes. This means that there will be no opportunity for examining how diegetic music is rated compared to non-diegetic music.

The other point for consideration is that of the 9 content extracts, three of the scenes were produced by the same producer as ‘The Turning Forest’ (the piece of content used to originally train the machine learning algorithm in [18]). These are the ‘Protest’ scene and the two ‘The

Vostok-K Incident’ scenes.

Of the remaining six scenes, three of the scenes came from the same producer (‘Casualty’) and two were from another producer (‘Penguins: Spy in the Huddle’). This means that over the 9 scenes, there were only 4 producers. Producers can develop very specific stylistic traits, so it is possible this will affect the generalisation capabilities of the machine learning.

## 4.9 Guidelines for Narrative Importance Content

Ideally all Narrative Importance (NI) content would be made up of clean, separate audio tracks, especially speech tracks. Some genres already have production techniques that are inherently NI friendly. An example of this is animated content, where the audio is recorded separately in studios. Television dramas and sports content are the worst offenders for bleed that have been encountered by the author, though other genres are prone to poor recording conditions.

Traditional radio dramas feature a recording process where the dialogue is recorded at the same time as a large amount of the Foley sounds. A voice actor will often be accompanied by a spotter, who acts out a lot of the physical movements, both are recorded using a single stereo mic. This workflow is necessary to fit within time constraints but leads to dialogue tracks with lots of bleed.

Reduced budgets have resulted in an increased use of lapel mics [1] in order to cut costs of multiple sound recordists and allow for multicam set ups as seen in Casualty. Whilst this improves other areas of the production workflow, there is a resulting negative impact on the quality of the audio. Lapel mics tend to pick up movement of clothing, which for some fabrics can be incredibly loud. This was seen by the author particularly for the paramedics in Casualty as their outfits are waterproof overalls which rustle.

Another issue seen within the recording process is the use of ADR and multiple on set takes to create a dialogue track that is actually in several separate parts. There were instances in the Casualty episode where a single syllable from a sentence was on a separate track to the rest of the audio. When it comes to the application of a machine learning algorithm this could be very problematic depending on how the algorithm is trained.

The next barrier to producing NI content is the post-production process. Effects are often added to stems as a bus, meaning that a collection of stems pass through a single effect channel and the output is added to the mix. This can raise issues within the NI workflow if objects



of different importance are routed through the same bus. This can result in an effects track that is louder than some of the objects being routed to it. In practice this means that, at the extreme of the scale, a low object will be almost inaudible (-48dB) but it could be routed through an effects bus that outputs it at the same level as an essential object (+3dB). The solution here is to have separate effects busses for each of the 4 levels of importance and route objects according to their importance.

When storing object-based content in general, all efforts should be made to retain the granularity of the objects as much as possible. The advent of cloud-based storage makes this far easier, though upload times are still a constraint. Where possible the ‘raw’ objects should be stored, alongside any versions of the content where a selection of objects have been downmixed to create a bed. This is fairly easy for shorter pieces of content but becomes difficult for longer pieces. The Vostok-K files were the simplest to process due to the way they had been stored. The folder contained 90 stems, for a 13 minute piece of content. The reader can imagine that an hour long programme stored in this way (separate stems for separate objects) would become unmanageable very quickly. The audio definition model (ADM) [76] allows for metadata to be stored that describes the start and duration of an object, along with the track ID for objects. This means that non-overlapping objects can be stored in a way that uses far less memory and can share tracks easily.

This boils down to a list of requirements for content that can easily have the NI system applied to it.

1. Record audio in the cleanest way possible
2. Keep different audio events as separate objects
3. Combine the same sounds into one object (e.g. a single character’s dialogue should be a single object)
4. A different effects bus is needed for each importance level to avoid effects being louder than the original sound
5. Do not downmix if storing audio for future use - keep a copy of the ‘raw’ audio objects, or use the ADM system to ensure that objects can be retrieved and changed as needed

## 4.10 Conclusions

This chapter began by defining six categories of audio objects that will be used throughout this thesis to discuss the results of both the survey and the ML algorithm. This was followed by a description of the content used for the survey. Nine excerpts were selected from five different content pieces. This information is summarised in Table 4.2, where the number of objects in each scene is also displayed. The scene names in this table will be used to refer to the scenes henceforth.

Content Name	Scene Name	No. of objects	Duration
Protest	Protest	27	62s
The Vostok-K Incident	Vostok-K Scene 1	27	60s
	Vostok-K Scene 2	21	45s
Casualty Season 33 Episode 38	Casualty Scene 4	27	40s
	Casualty Scene 46	17	87s
	Casualty Scene 49	20	52s
Penguins: Spy in the Huddle	Penguins Opening Credits	23	47s
	Penguins Scene 1	25	55s
Football	Football	11	20s

**Table 4.2:** A table showing the 5 pieces of content and the 9 scenes that were extracted for use with the number of objects contained in each scene

It has been shown that sourcing this content was not a trivial task. The processing required to make the content compatible with the NI systems was extensive. During this work it was found that the term “object-based audio” was not sufficient when describing to colleagues the sort of audio required for the test. This leads directly to the work on OBA definitions in Chapter 7. The ground truths for the scenes from Protest and Casualty were laid out in Section 4.6. Three content pieces that were deemed unsuitable were discussed to demonstrate some of the problems encountered in trying to source appropriate content. Finally, a set of guidelines for NI content were outlined in Section 4.9.

# Chapter 5

## Data Collection Results

### 5.1 Introduction

This chapter will present the results from the data collection task outlined in Chapter 3, using the content described in Chapter 4. In total 50 participants completed the task. The chapter will begin with an overview of the demographics of the participants. Then a brief explanation of some of the data that was removed from the dataset. This will be followed by a review of the agreement between participants. Then the full assignment results will be presented, followed by an exploration of the difference between audio professionals and laypeople.

### 5.2 Survey Demographics

Several demographic questions were asked to participants in the hope that there might be enough who identified as being in each group to do a comparison. Unfortunately there was not enough data for statistical analysis.

Table 5.1 shows the results from the demographic questions asked to all participants. The 21 participants who answered “No” to “Do you work or have you previously worked in audio?” were then asked the questions in Table 5.2. The 29 participants who answered “Yes” to “Do you work or have you previously worked in audio?” were then asked the demographic questions in Table 5.3 and a set of questions on artificial intelligence, OBA, and NI shown in Table 5.4.

Questions	Responses	Counts
Do you have any known hearing loss?	o Yes	4
	o No	42
	o Unsure	4
	o Prefer not to say	0
What is your first language?	o English	35
If you are bilingual list both languages.	o Not English	12
	o Bilingual	3
Do you identify as being neurodivergent (including but not limited to autism, adhd, dyslexia, and dyspraxia)?	o Yes	9
	o No	38
	o Prefer not to say	3
Do you work or have you previously worked in audio?	o Yes	29
	o No	21

**Table 5.1:** A table showing questions asked to all 50 participants with the number who selected each response in the **Counts** column

Table 5.1 shows the demographic questions asked to all 50 participants. It can be seen that the only question where roughly even numbers across possible groupings was observed is “Do you work or have you previously worked in audio?”. For this reason, only these two groups will be compared when discussing the results. The four participants who answered “Yes” to the first question (“Do you have any known hearing loss?”) were asked if they had hearing aids and if they were using them for this task. Of the four, two answered yes to both of these questions.

Questions	Responses	Counts
Have you ever used a digital audio workstation (such as Logic, Protools, Cubase, etc)?	o Yes, I use them regularly	7
	o Yes, but not often	5
	o No	9
Do you have any musical training?	o Yes, as a professional musician	3
	o Yes, as an amateur musician	8
	o I am a self taught musician	3
	o No	7

**Table 5.2:** A table showing questions asked to the 21 laypeople participants to gauge their audio and musical experience with the number of participants who selected each response in the **Counts** column

The 21 participants who had no experience of working in audio were asked the questions in Table 5.2 to ascertain their audio and critical listening experience. In 2014 it was estimated that 74% of adults have played a musical instrument [77]. Fourteen participants (66%) answered “yes” to having music experience, which seems comparable to this figure given the small sample

size. However it is estimated that there are 37,600 working musicians in the UK in 2022 [78], less than 1% of the population. 3 out of 21 participants (14%) responded that they are professional musicians. This indicates that these participants may be skewed towards having greater musical (and therefore critical listening) experience than the general population. This is likely due to the recruitment process and could mean that there is less of a difference between this group and the audio professionals group. Not all people who work in audio are musicians but critical listening skills are developed by both musicians and audio professionals.

The audio professionals were asked a series of questions regarding their professional background. The responses are shown in Table 5.3. Due to recruitment techniques and constraints, the majority of respondents are UK based and there is a bias towards mixing music.

Questions	Responses	Counts
How many years have you worked in audio?	Less than 5	9
	Between 5 and 10	7
	Between 10 and 15	7
	15+	6
In which country do you spend the majority of your time working?	United Kingdom	26
	Ireland	1
	United States of America	1
	India	1
What medium do you most commonly work in?	Television	1
	Radio	6
	Film	1
	Music	11
	Virtual Reality	2
	Other	8 - Research [3], Acoustics [2], Psychoacoustics [1], Spatial Audio [1], Lecturer [1]
What genre/s of content do you most commonly work on? <i>-Select all that apply</i>	Documentary	7
	Drama	5
	Music	20
	News	1
	Other	6 - Vitual Reality [2], Research Stimulii [3], Games [2], Speech [1]
Which of the following best describes the majority of the work you do?	Sound Mixer	4
	Producer	4
	Sound Recordist	2
	Teacher/Lecturer	2
	Editor	1
	Other	16 - Researcher [7], Audio Programmer [2], Acoustician [2], Sound Designer [1], Sound Assistant [1], Mastering Engineer [1], Composer [1], Student [1], Journalist [1]

**Table 5.3:** A table showing the working background of the 29 audio professional participants. The number of participants who selected each response is in the **Counts** column.

Questions	Responses	Counts
Have you ever worked on an object-based audio production before?	○ Yes	11
	○ No, but I am familiar with the concept	10
	○ No and I am unfamiliar with the concept	8
Is the importance of a sound to the narrative something you consider when you mix?	○ Yes	26
	○ No	1
	○ Unsure	2
Rate how you would feel about the audience being able to control the volume balance of objects in the mix?	○ Extremely comfortable	4
	○ Somewhat comfortable	19
	○ Neither comfortable nor uncomfortable	1
	○ Somewhat uncomfortable	5
	○ Extremely uncomfortable	0
To your knowledge, have you ever used any artificial intelligence (AI) in your workflow?	○ Yes, I currently use AI	5
	○ Yes, I have used AI in the past	6
	○ Yes, I have used AI once or twice	4
	○ No	14
Rate how you would feel about using an artificial intelligence based plug-in as part of your production workflow.	○ Extremely comfortable	13
	○ Somewhat comfortable	9
	○ Neither comfortable nor uncomfortable	5
	○ Somewhat uncomfortable	2
	○ Extremely uncomfortable	0

**Table 5.4:** A table showing the questions on OBA, NI, and AI asked to the 29 participants who worked in audio with the number of participants who selected each response in the **Counts** column

Table 5.4 shows the questions the audio professionals were asked regarding OBA, NI and AI. For some of these questions participants were asked to elaborate on their answers using a text box. Most participants had knowledge of OBA. All participants were given a description of OBA at the beginning of the survey. 26 of the 29 said that they consider the importance of a sound when mixing. One participant responded “No” and said that they “don’t mix audio to accompany a narrative, I am either producing audio for teaching/research or creating electronic music”. For the two who responded “unsure”, one said that they were not experienced in mixing, and the other that they mostly mix music, which can be considered to contain narrative elements but not always.

The majority of participants were comfortable with the idea of the audience having some

control over the balance of a mix. Two of the five participants who responded that they were “Somewhat Uncomfortable” with this idea expressed concern about the users’ ability to control a mix at an object level, indicating that the wording of this question was not clear enough. One participant expressed “As an audience member I do not want to have to fuff about with this - give me a good mix in the first place”. The final two commented that the mix is a creative process that should be down to the producers.

Participants had a variety of experience with AI, with roughly half (14/29) of respondents not having used any AI in their workflow. Despite this, the majority of participants were comfortable with the idea of using an AI based plug-in. A description of the plug-in was provided before this question, so that participants with no experience of AI would have an idea of its capabilities. The description read:

You work for a broadcaster and they have implemented a personalisation control allowing the end-user to switch between high, medium and low complexity mixes in their home. To assist in the assignment of the required metadata you have access to a plug-in which uses artificial intelligence to create four Narrative Importance buses and assign each object to one of them. The plug-in allows for fine-tuning by the user.

One of the two participants who responded that they were “Somewhat Uncomfortable” expressed concerns about the capabilities of “AI to handle novel or unusual audio outside of its training dataset”. The other participant who responded this way said that they “remain sceptical of all artificial intelligence-based solutions, given how little we understand them”.

## 5.3 Participant Removal

For the majority of participants, the time they spent on each of the three scenes was tracked. For some of the early participants this wasn’t logged due to an error in the site code. The time taken was examined as a possible exclusion metric. Two ‘relative duration’ metrics were created by dividing the time taken by the length of each content piece and the number of objects in each scene. By examining the time taken as a ratio of the length of the scenes it was possible to check whether any participants couldn’t have listened through to the entire scene with their assignments. It was found that this metric was not very useful for several reasons. The first being that not every participants’ time was tracked, so it was impossible to treat all participants the same. The second was that some participants seemed to have left the page open in the



middle of the task (one participant spent 8 hours on a single scene) indicating that the metric was not reliable even when tracked. The third reason was that some of the audio professionals were incredibly quick to complete that task, presumably because the interface was familiar and they work with sound daily. This meant that these participants could not be treated the same as participants without audio experience. Finally, almost all participants sped up throughout the task. The first scene presented to participants took them longer than the third, especially in cases where they had two scenes from the same content piece. Meaning that if a scene was presented third, it wasn't comparable with that same scene presented first. There was no solid way to make exclusions based on the time stamp.

Participant **4RA435** indicated in their feedback and in an email to the author that they had technical issues in their final content piece. The technical issue meant that they were unable to hear the mix and consequently the effects of their assignments as they worked through them. It was decided to only remove the final asset from this participant's data set, keeping the first two they had completed.

Participant (**8TA556**) expressed in their feedback that they would have liked to redo the first asset they completed as they had misunderstood the task initially. This participant contacted the author about this directly after completing their task. After a discussion they reported that they hadn't used the accessible mix to hear how their assignments were affecting the mix, but that they felt they probably would have similar results with this feature. The author inspected their results and decided that they were logical enough to be included in the dataset.

Some participants' data was more complex to discern. Ultimately it was decided to keep their data, since the point of having a survey is that not everyone will agree. These participants were:

**4JA849** Rated a single dialogue track as "low" in one scene and another as medium in a second scene. Most of the objects in the scenes were rated "low". This was initially thought to be an error due to not scrolling down the page however the participant had rated objects lower down so it was decided this was intentional and the results were kept.

**6EP636** Stated in the feedback they adjusted their volume during their second task, this was decided not to be exclusionary behaviour as they remarked they couldn't hear some of the individual objects.

**2AH357** For Vostok-2 this participant said in their feedback that they had wanted to put 'Tatiana Dialogue' in 'Essential' but in order to get around the rule of one object in each

category they put the object in ‘High’. It was decided to change this assignment in the dataset. This result was the only one that didn’t have an object in all 4 categories.

**2EQ986** Set all Main Dialogue objects in two Casualty scenes as high. This struck the author as odd, but was clearly intentional (since in the third scene they had put Main Dialogue in Essential) and so their results were retained.

## 5.4 Fleiss Kappa

Fleiss’ Kappa [79] was calculated for each content piece. Fleiss’ Kappa is a metric which is used to evaluate the level of agreement amongst raters in categorisation tasks. It ranges from 0-1 with 0 signifying no agreement and 1 signifying complete agreement.

The value  $p_i$  given in Eq. (5.1) is the agreement for each object. The mean of  $p_i$  is  $\bar{p}$ , given in Eq. (5.3). The value  $p_j$  given in Eq. (5.2) is the proportion of agreements in each of the categories. This is then used to calculate the proportion of errors,  $\bar{p}_e$  in Eq. (5.4). Finally  $\kappa$  is given in Eq. (5.5).

$$p_i = \frac{1}{n(n+1)} \left[ \left( \sum_{j=1}^k n_{ij}^2 \right) - n \right], \quad (5.1)$$

$$p_j = \frac{1}{Nn} \sum_{i=1}^N n_{ij}, \quad (5.2)$$

$$\bar{p} = \frac{1}{N} \sum_{i=1}^N p_i, \quad (5.3)$$

$$\bar{p}_e = \sum_{j=1}^k p_j^2, \quad (5.4)$$

$$\kappa = \frac{\bar{p} - \bar{p}_e}{1 - \bar{p}_e}. \quad (5.5)$$

The values in Table 5.5 show the agreement amongst participants for each scene, for the full dataset, the audio professionals, and laypeople. N indicates the number of participants asked to label that scene and  $\kappa$  is the value of agreement. These values are comparable to the value of Fleiss’ Kappa (0.11) found by Ward in [17] for the content piece ‘The Turning Forest’.

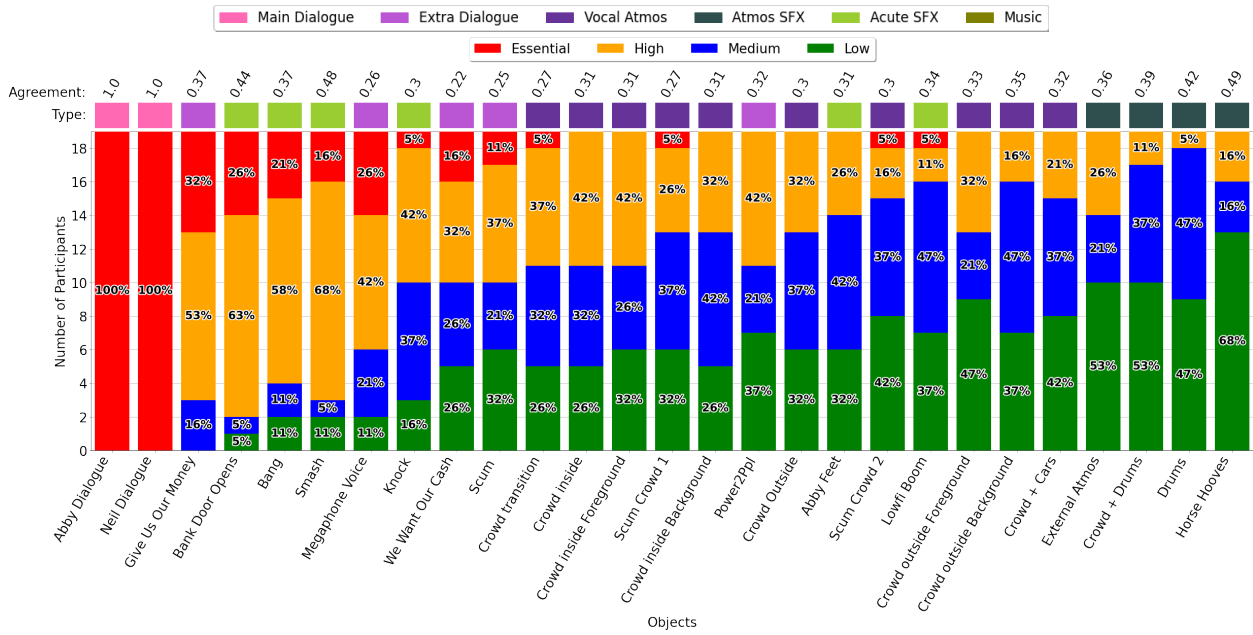
Content	Full Dataset		Audio Pro's		Laypeople	
	N	$\kappa$	N	$\kappa$	N	$\kappa$
Protest	19	0.16	11	0.16	8	0.15
Vostok-K S1	19	0.14	10	0.14	9	0.14
Vostok-K S2	16	0.15	12	0.19	4	0.09
Casualty S4	15	0.18	6	0.23	9	0.14
Casualty S46	15	0.23	6	0.29	9	0.22
Casualty S49	18	0.25	10	0.26	8	0.20
Penguins Credits	16	0.19	11	0.17	5	0.27
Penguins Seal	15	0.13	11	0.11	4	0.22
Football	16	0.14	9	0.16	7	0.08

**Table 5.5:** A table showing N, the number of participants, and  $\kappa$ , the value of Fleiss' Kappa, for each of the content pieces across three groups. The full dataset, the audio professionals, and laypeople.

Some scenes exhibited higher levels of agreement, for example two of the scenes from Casualty achieved agreement of over 0.2. It is not possible to say how significant this difference is due to Fleiss' Kappa being difficult to categorise due to the number of categories and raters influencing the value, for example a larger number of categories, the more likely a low agreement value [80].

## 5.5 The Assignments

In this section the assignments for each of the scenes will be explored. They will be presented in the form of stacked bar charts showing how many participants places the object in each of the four importance categories. Along the top of the bar charts is the **Type** bar which of the seven groups defined in Section 4.2 the object falls under. There is also an agreement value for each object.

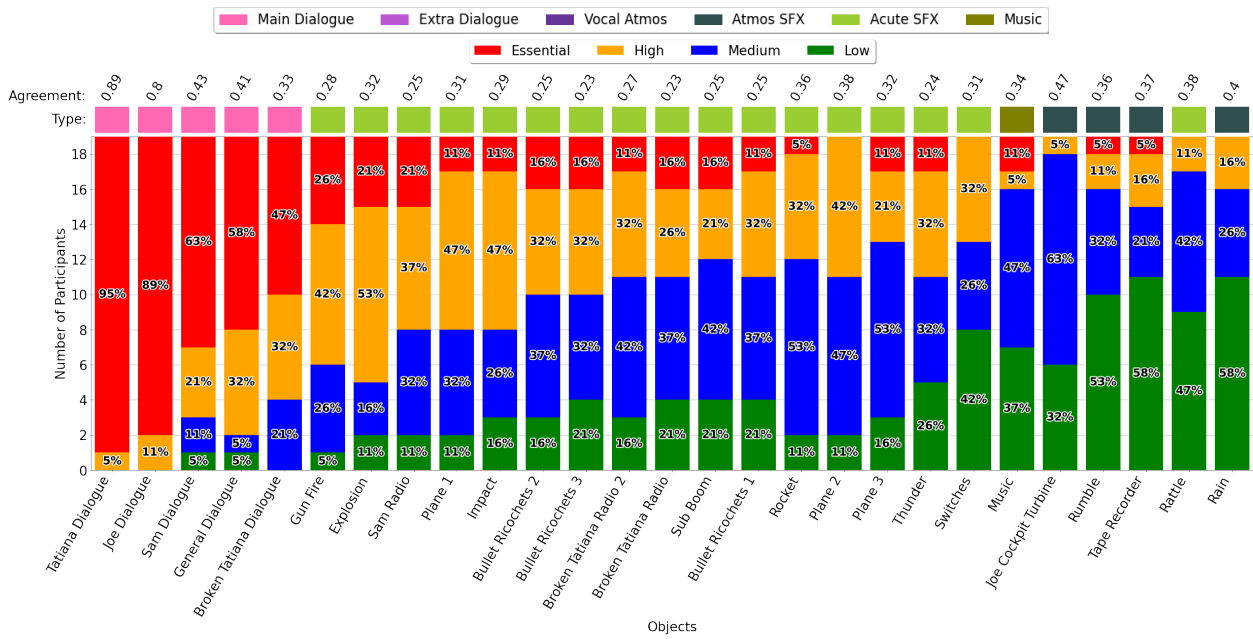


**Figure 5.1:** A bar chart showing the assignments for the Protest scene. Agreement refers to the value of agreement for each object as calculated in Eq. (5.1)

Other than all participants being in agreement that the two main characters’ dialogue is essential the assignments vary greatly. The agreement value shown above the bars is the agreement for each object calculated using Eq. (5.1). Other than the two main dialogue tracks, where the agreement is 1, the range for agreement falls between 0.22 for ‘We Want Our Cash’ and 0.48 for ‘Smash’.

The bars are ordered from left to right by the value of the sum of all assignments where Low was given a value of 0, Medium a value of 1, High a value of 2, and Essential a value of 3. The bars on the left had the highest sum, so were generally considered more important and the bars to the right had the lowest sum so were considered less important. This ordering was chosen for ease of digestion of the data as it shows the patterns of the assignments well.

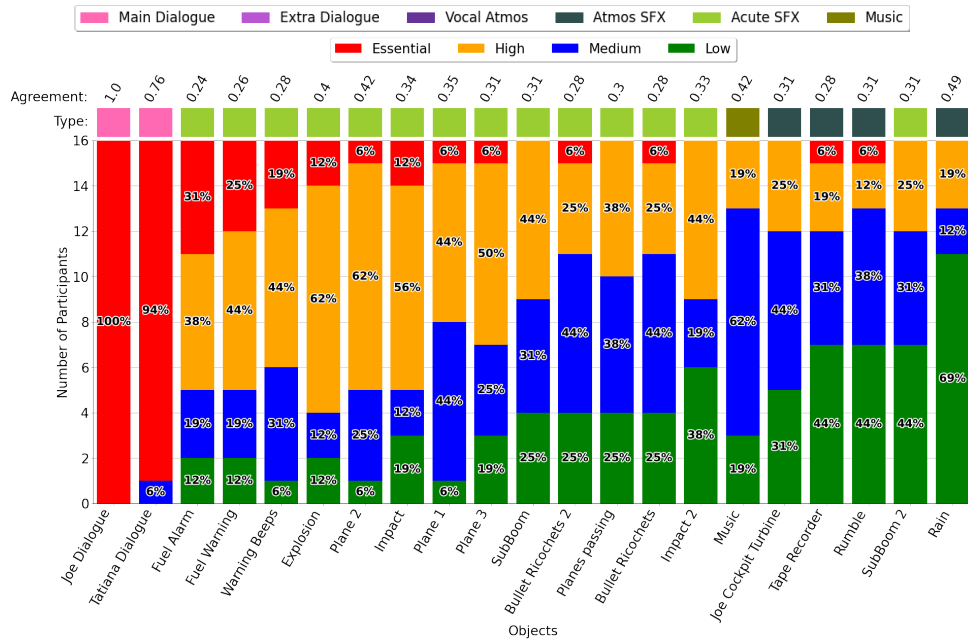
In most cases the Main Dialogue objects, as defined in Section 4.2, are considered to be the most important objects. The objects’ category can be seen in the **Type** bar at the top of the bar charts.



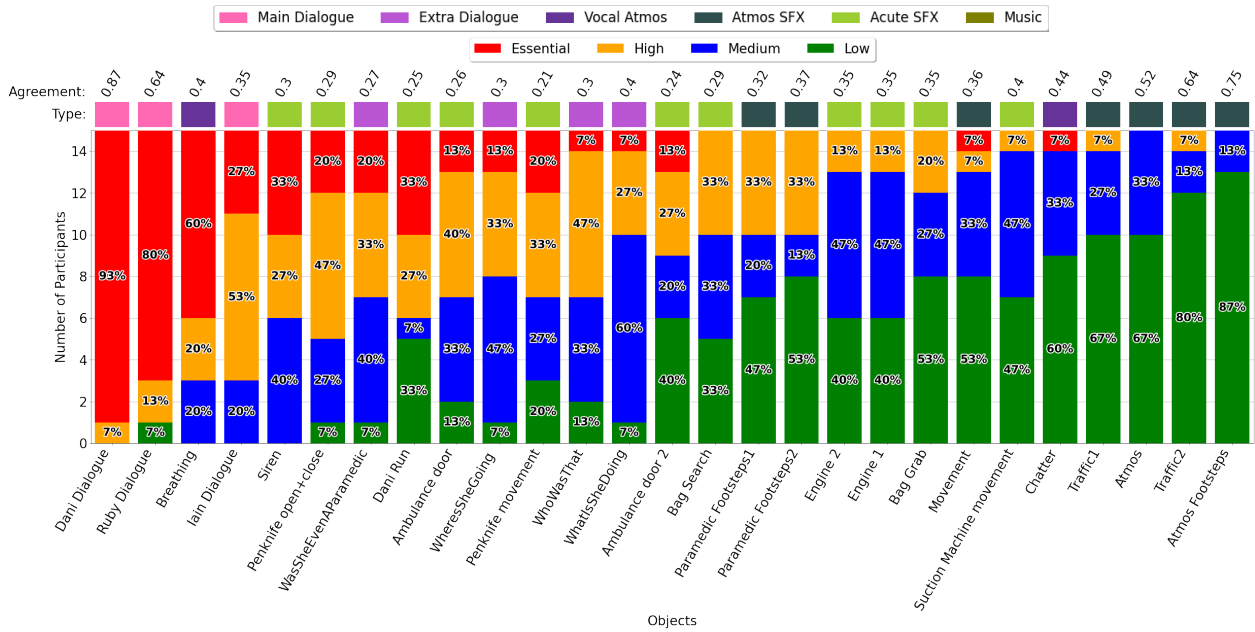
**Figure 5.2:** A bar chart showing the assignments for Vostok-K extract 1. Agreement refers to the value of agreement for each object as calculated in Eq. (5.1)

The bar chart of assignments for Vostok-K extract 1 is shown in Fig. 5.2. The responses for this scene were more varied among the group, particularly the Main Dialogue objects, possibly indicating that the participants struggled more with the assignments. The highest agreement value in this scene was 0.89 for ‘Tatiana Dialogue’, compared with other scenes where the highest level of agreement is 1. A significant number of objects for this scene had assignments in all four of the categories. The object ‘General Dialogue’ (fourth from the left) is interesting since originally it was part of the same object as ‘Tatiana Dialogue’. The two characters are in conversation with each other, although ‘General Dialogue’ only contains a single sentence. These two objects were separated by the author to be congruent with all other scenes, where dialogue for different characters was always separate. It should be noted that “General” is the name of the character speaking, not an indication of the dialogue being generic. It is possible this naming of the object influenced participants and this is why it was only rated Essential by 58% of participants, compared with 95% for ‘Tatiana Dialogue’. It was certainly unexpected that participants would place two parts of a conversation in different levels.

Fig. 5.3 shows the assignments for Vostok-K extract 2. Again, other than the Main Dialogue objects, participants did not agree on which objects were narratively important. Similarly to extract 1, all objects except for the dialogue were assigned to at least three of the four importance levels by participants.



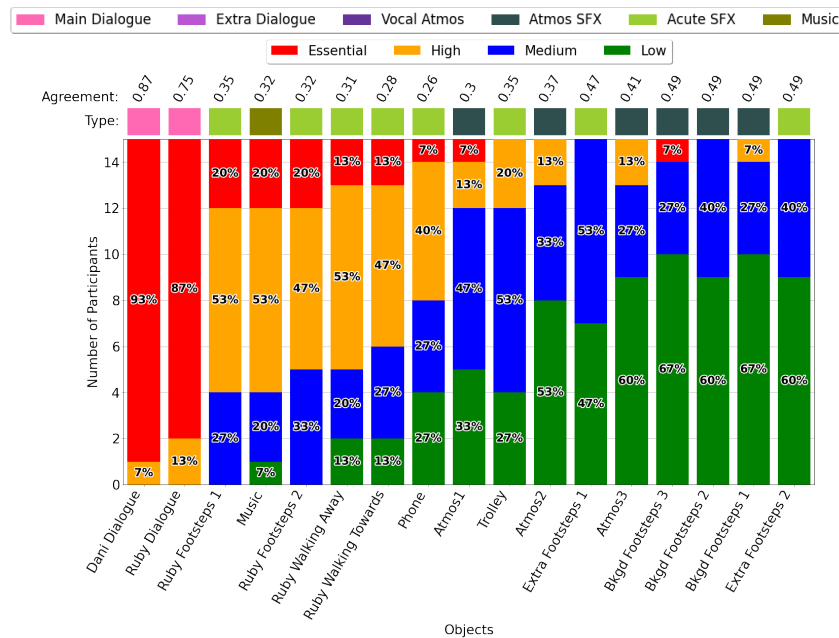
**Figure 5.3:** A bar chart showing the assignments for Vostok-K extract 2. Agreement refers to the value of agreement for each object as calculated in Eq. (5.1)



**Figure 5.4:** A bar chart showing the assignments for Casualty Scene 4. Agreement refers to the value of agreement for each object as calculated in Eq. (5.1)

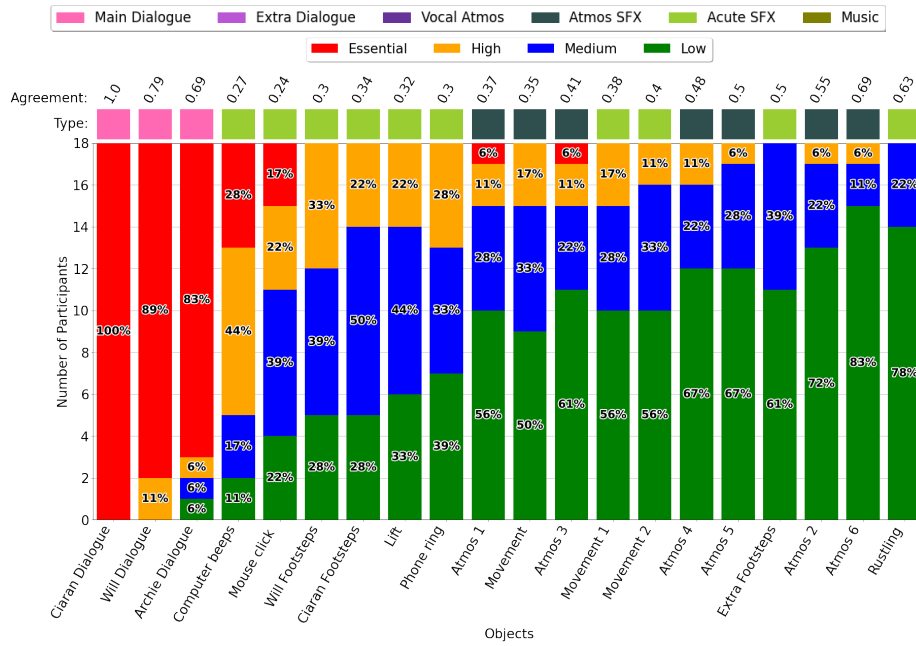
The assignments for Casualty Scene 4 can be seen in Fig. 5.4. The ground truth assignments for this scene placed all Main Dialogue (pale pink) and Extra Dialogue (pale purple) objects in the Essential category. It can be seen that participants' opinions on these objects was significantly varied. This scene is the first of the results that originally accompanied video. It

was anticipated that the lack of video may influence assignments and therefore some variation from the ground truth was expected. However, the Extra Dialogue objects provide context for the character Dani running away when the paramedics arrive, which without the visual would appear to be even more important. It is also interesting that the ‘Breathing’ object was rated as Essential by 60% of participants. In the ground truth this was a Medium object (though originally the first two breaths were placed on a separate track so they could be High). It is suspected that this is a result of the lack of video. With the video viewers can see the character Barbara on the floor having been knocked down by a moped. Without video, the haggard breaths are the only indication that a character is injured and therefore should be considered more important to the narrative.



**Figure 5.5:** A bar chart showing the assignments for Casualty Scene 46. Agreement refers to the value of agreement for each object as calculated in Eq. (5.1)

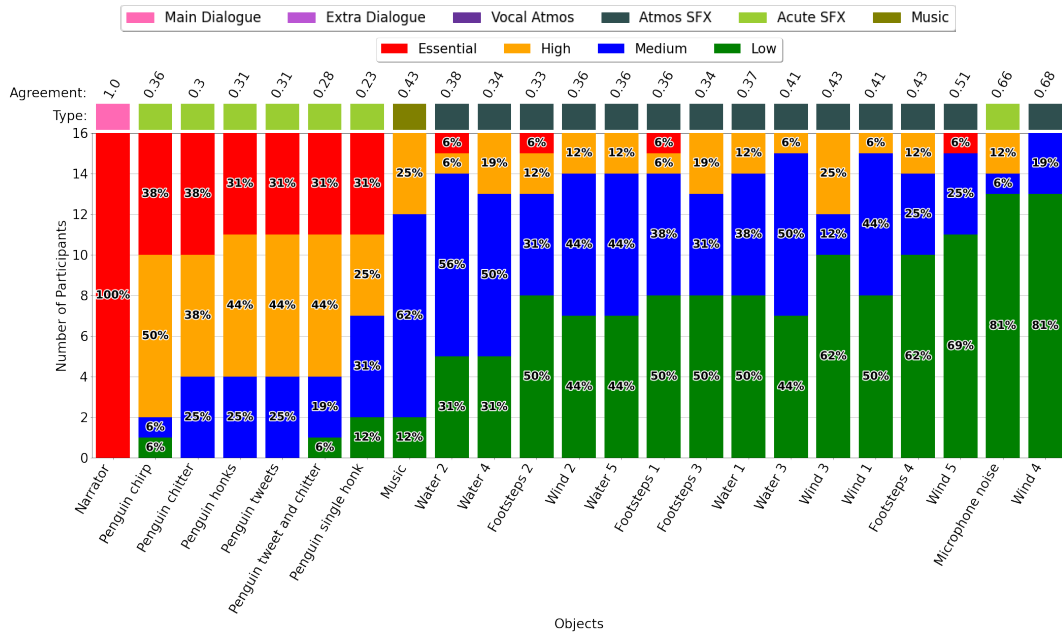
In Fig. 5.5 the results for Casualty Scene 46 are seen. The most notable point of these assignments is the ‘Music’ object (fourth from the left, khaki green in the **Type** bar). This music was labelled as more important than the music in any other scenes, with over 70% of participants rating it as Essential or High importance. This scene was an emotional conversation between the two characters Ruby and Dani, where Ruby is telling Dani they can no longer be friends. The music begins towards the end of the conversation and provides emotive context as Ruby walks away from Dani. The majority of the other objects in the scene build the atmosphere of a busy hospital ward.



**Figure 5.6:** A bar chart showing the assignments for Casualty Scene 49. Agreement refers to the value of agreement for each object as calculated in Eq. (5.1)

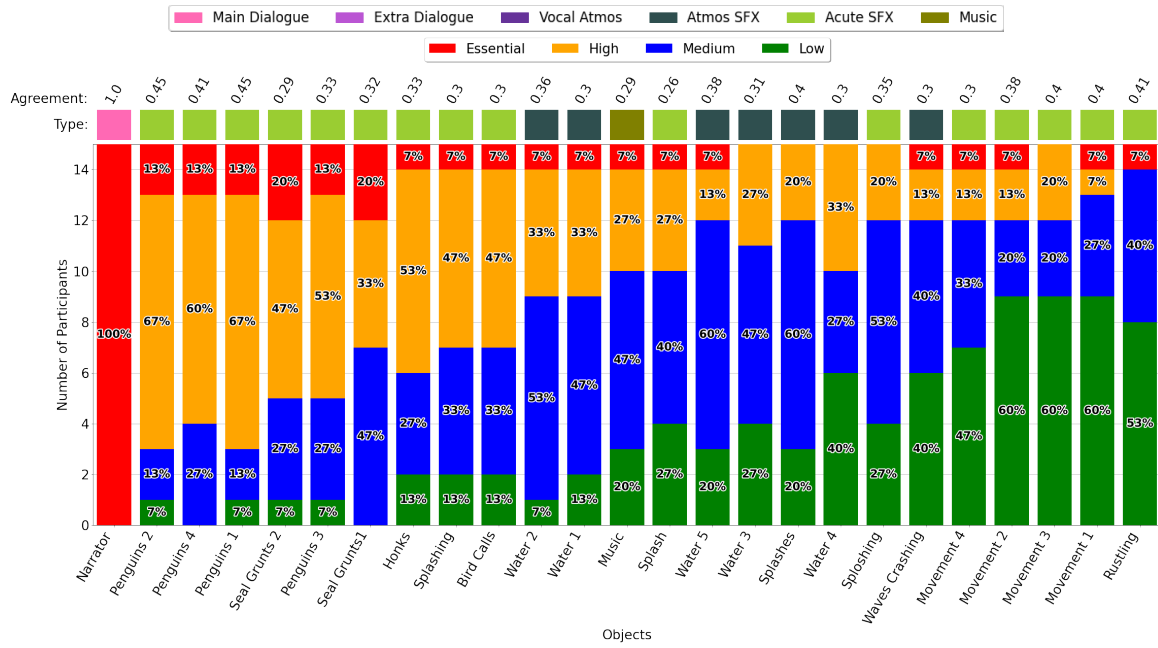
Casualty Scene 49 featured a conversation between three characters after one of the characters (Ciaran), discovers that some of the data on his computer has been deleted. In Fig. 5.6 it can be seen that again some participants did not rate all three characters' dialogue the same, despite them being involved in a conversation with each other. It is also notable that the computer beeps were generally considered more important than the mouse clicks, despite these two sounds contributing to the same plot point. Again the lack of video may have influenced this. In the ground truth both of these objects were originally High. The video showed the mouse being clicked and an error showing on the computer screen so without one or both of these sounds, the visual would have been incongruent with the audio.





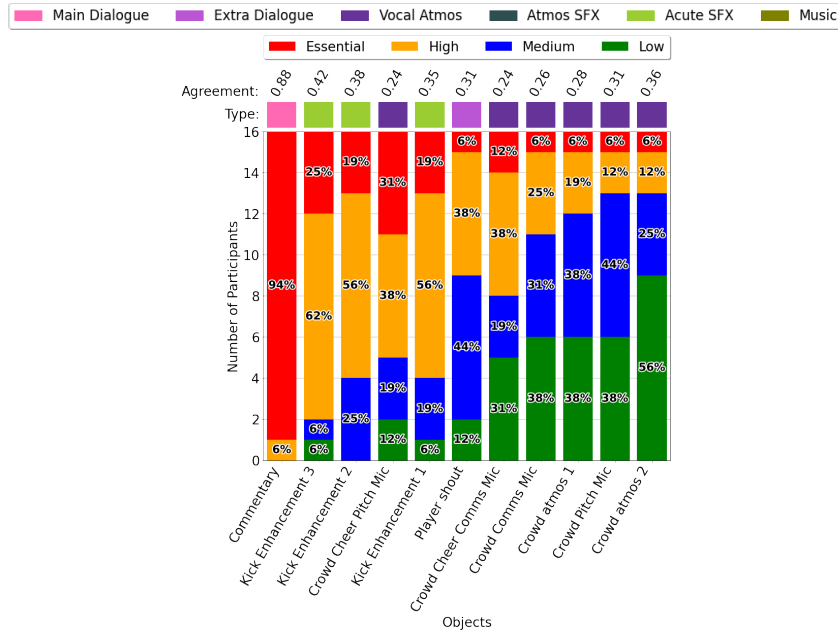
**Figure 5.7:** A bar chart showing the assignments for Penguins Spy In The Huddle: Opening Credits. Agreement refers to the value of agreement for each object as calculated in Eq. (5.1)

Both of the scenes from Penguins Spy In The Huddle exhibited 100% agreement that the narrator should be rated as Essential (Figs. 5.7 to 5.8). Other than the narration objects, participants generally rated the animal vocalisations as the next most important tracks, with the music and atmospheric tracks following. In the Seal and cormorant scene (Fig. 5.8) participants rated sounds relating to the movement of the action on screen lowest (objects like ‘Rustling’, ‘Waves Crashing’, and ‘Movement 1-4’). This again could be attributed to the lack of video. These sounds do not explicitly add to the narrative, but with the inclusion of video some, or all of these objects may be required in order to tally with the on-screen action.



**Figure 5.8:** A bar chart showing the assignments for Penguins Spy In The Huddle: Cormorants and Seals. Agreement refers to the value of agreement for each object as calculated in Eq. (5.1)

Finally the assignments for the Football clip can be seen in Fig. 5.9. This clip was an excerpt from a football match where a player takes a near miss shot on goal. The commentary was considered the most important object by most participants, followed by the kick enhancements. Interestingly the ‘Crowd Cheer Pitch Mic’ object was also rated highly. This object was the crowds reaction to the players shot, as recorded by a microphone on the pitch. The feedback from participants for this scene was enlightening, with some commenting how surprised they were that the crowds reaction told them more than the commentary about what was happening in the game. This feedback will be discussed further in Section 5.7.



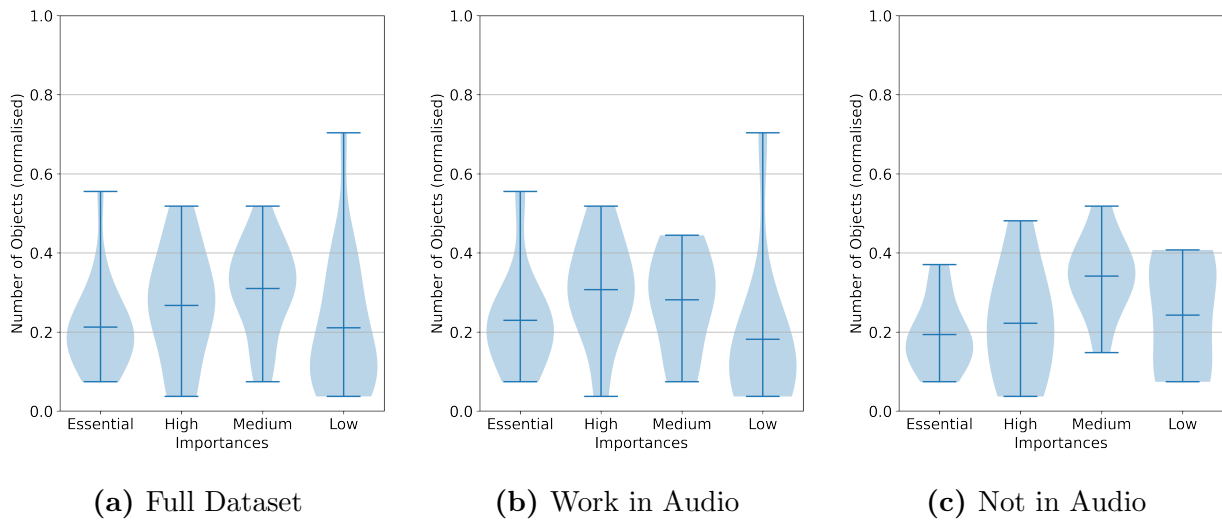
**Figure 5.9:** A bar chart showing the assignments for the Football scene. Agreement refers to the value of agreement for each object as calculated in Eq. (5.1)

## 5.6 Results for Audio Professionals VS. Non-Professionals

The survey run in [17] was only distributed to audio professionals. In this thesis it was decided to open up the participant pool and collect data from people with little or no audio experience. In part, this was done to enable a larger dataset for the ML algorithm to work from. It also provides an opportunity to explore whether there are differences between the two groups. In this section the focus will be on the Vostok-K S1 assignments. Firstly because this content piece had a good split across the two groups, 10 audio professionals (APs) to 9 non audio professionals (NAPs). Secondly, this piece of content had less agreement on dialogue than many of the others. The feedback varied from some saying they found this scene “the easiest to mix”, to others remarking that this scene was “quite difficult”. Presentation order may have influenced some of these comments, since participants generally reported finding the second and third tasks easier than the first, however this does not explain all of the comments as the order of presentation was randomised.

Fig. 5.10 contains three violin plots which show the distributions of the number of objects each participant assigned to each importance level for Vostok-K S1. The plots are normalised between 0 and 1, the total number of objects in this scene was 27. Fig. 5.10a shows the distributions for the full dataset of all participants. Fig. 5.10b and Fig. 5.10c show the distributions for the APs and NAPs respectively. It can be seen from these that for this scene NAPs tended

to rate more objects as Medium or Low than APs. Violin plots showing the distributions for the other scenes can be found in Appendix C.



**Figure 5.10:** Three violin plots showing the distributions of the number of objects (normalised to between 0 and 1) each participant assigned to each importance level for Vostok-K S1.

Moving to a more detailed view of how the assignments varied for specific objects further confirms the difference between the two groups. Figs. 5.11 to 5.12 contain the bar charts showing how each object was assigned. It can be seen that within both groups there were some participants who did not rate all dialogue as Essential. For example the ‘Sam Dialogue’ object was rated High by 40% of the APs with the other 60% rating it Essential. In contrast, 33% of NAPs thought this object should be Low or Medium, but none thought it should be High, and 67% classed it as Essential. Another interesting object is the ‘Impact’ object. For this object 80% of the APs thought it should be in the High category, whereas 66% of the NAPs rated it Medium or Low. This pattern is seen across most of the objects. Where the NAP group tended to rate lower than the AP group as a whole.

This is an interesting phenomenon which suggests that the general public might feel they don’t require as complex a sound scene as many audio professionals do. This would require further investigation to confirm. The pattern was not observed across all scenes, however both of the Vostok-K scenes exhibited this pattern and were arguably the most complex scenes in the set.

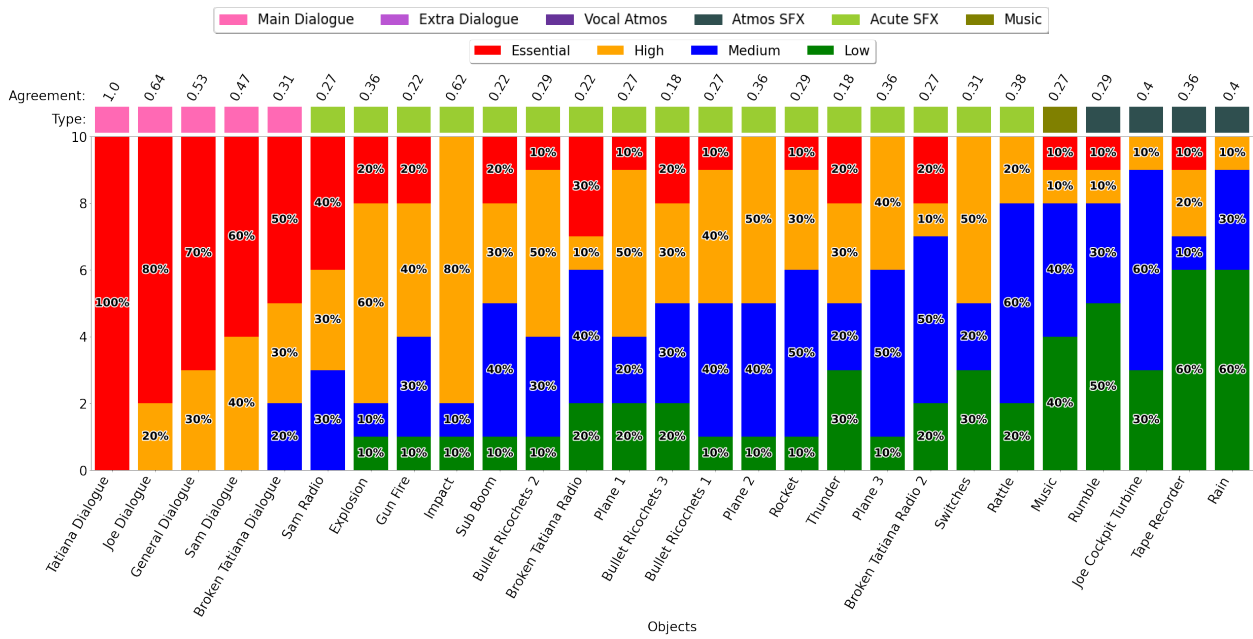


Figure 5.11: A bar chart showing AP assignments for Vostok-K S1

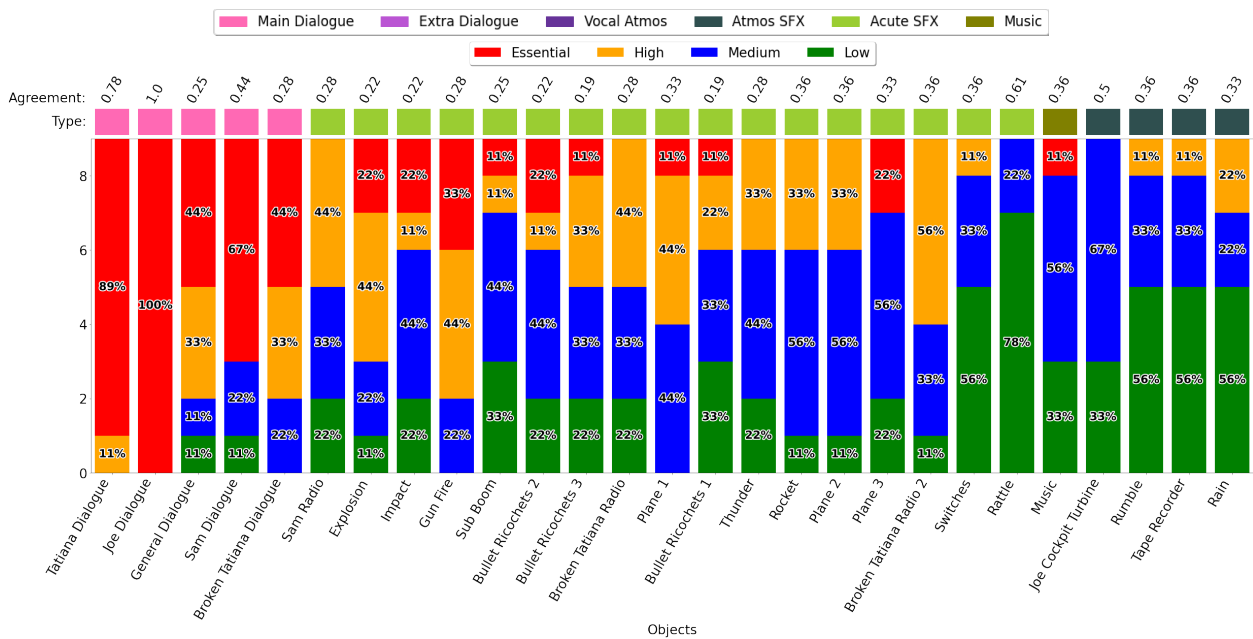


Figure 5.12: A bar chart showing NAP assignments for Vostok-K S1

## 5.7 Feedback Quotes

After each of the three tasks, the participants were presented with a text box and asked to give feedback on how they found the task. Not all participants gave feedback, and many only gave feedback after their first task. This section will look at some of the more noteworthy and detailed feedback from participants.

After completing their assignments for Protest, participant **9TL152** commented:

“Some sound objects - such as crowd sounds - were vital to the narrative - but at a steady volume impaired intelligibility when overlapping with speech.”

They went on to suggest that automation could be used to combat this issue. After completing assignments for Vostok-K extract 1, participant **9NY535** remarked:

“I think some narratively less important things (especially the music) were still important for mood.”

This comment demonstrates one of the anticipated problems with training a ML algorithm to assign NI metadata. That being that the importance to the narrative is not only a decision based on comprehension, but also one of emotional impact. **2JS323**, who was one of two participants that completed the task using hearing aids, said in their feedback for Vostok-K extract 1:

“With my hearing aids working correctly I could not understand anything of the original clip! It was a real revelation to hear the dialogue... I would have been happy to simply have the dialogue alone without any of the background sounds, including the ‘heavy breathing’ in the top track... For me, the visuals of bombs, planes, etc. would have been fine to fill in the detail of what was happening. The audio tracks of these elements completely obscure the voices for me otherwise.”

They also said of Vostok-K extract 2 that

“In the original mix the Fuel Alarm was almost painful to hear... It was blissful to be able to turn it off!”

From these comments it is possible to hypothesise that the NI system may not be suitable for some hearing aid users, since even a reduced complexity mix may still be too much competing audio for them to comprehend speech.

Casualty Scene 4 received a few comments from participants about specific objects that they felt were important (or not). Participant **2AH257** stated that they had prioritised the speech, and then had worked through SFX deciding whether they were important. There were specific objects they found more difficult.

“I had a bit of an issue with the breathing/penknife sounds and found them kind of uncomfortable... so I couldn’t decide if they should be more quiet (because they are

gross) or a bit loader[sic] (because they improve the understanding of the situation)”

The author had considered that some of the objects, when heard in isolation, may have a high annoyance or disgusting factor for some participants, so it is interesting to have feedback stating this. Another participant, **9TL152**, stated:

“For other items, such as suction, the background noise was not as important - probably because it was stated in speech that this was happening. Whether or not something is described in speech may affect the importance of associated foley?”

This is another point where the ML may fall down, because it is not able to discern whether a sound has been referred to by a character in the scene.

For Casualty Scene 46, one participant (**5AQ531**) raised the point of atmosphere. Casualty scenes often have a distinct soundscape, with layers of atmospheric tracks that create the feeling of a busy hospital environment. This scene, which took place at a bedside in a ward, features a lot of objects of this nature.

“I could hear this excerpt in its original format but it did become even clearer when I reduced some sounds. I felt there was a danger of losing all the atmosphere. I raised the voices because I know that when I watch programmes I often miss hearing some of the dialogue due to background sounds being too loud”

The need for these atmospheric cues was referred to again by participants for Casualty Scene 49, again a scene set in the hospital. Participant **9FB134** said:

“...making sure that the contextual sounds specifically the lift and footsteps were included help reduce the awkward silence between dialogue even in the moderate mix... Making sure that it was clear the mix sounds like it was within the hospital with the phone rings and general hubub in the background, and making sure that it was clear they were moving around into difference spaces with the lift and changes in atmospheric sound.”

The point about filling the silences between conversations is a further point where ML may fall down if it is not provided with contextual information about the other sounds in a mix.

After assigning importance for Penguins Credits, participant **0PL822** remarked:

“I had to put the sounds that overlapped with the narrators voice lower, even if they were somewhat important.”

Again, this is something that ML will potentially struggle to do. Participant **5AQ808**, another hearing aid user, made similar comments to the other participant with hearing aids:

“Having to use all the categories was difficult as it was the narrative that I want to hear and all the rest could have been low.”

It would seem that from the feedback here, for hearing aid users, providing a very minimal mix is required.

Similar themes ran across most of the scenes. For Penguins Scene 1, participant **3XQ683** said:

“I initially put music as low, but without something consistent in the background the transition between clips was too harsh, so I put music as med priority.”

Again, demonstrating that the silences between other sounds felt incongruent to many participants. Participant **4RA435** talked about their ambivalence whilst completing the task:

“...it feels uncomfortable to remove so much from the ambience which seems to be central to the experience and narrative, but so threatening to intelligibility for speech impaired listeners.”

For the Football scene, participants responses suggested that they were surprised by how much the crowd sounds informed their understanding of the action. Participant **1DL302** said:

“...even with the commentary it was harder to know what was going on without the [crowd’s] reaction... I was surprised at how much the reactions helped me know what was happening - more than the commentary.”

Participant **3HQ159** similarly observed:

“...the crowd often cheers slightly before the commentator has time to say what has happened, so to not hear it in the accessible version would be to miss out on key information.”

Obviously the inclusion of video may change these views, however football is often accessed via radio and so this has validity in that context. Finally, participant **4RA435** said:

“I felt inclined to remove some audience noise in the accessible mix because this would make the commentary more difficult to understand. But is the energy of the crowd not important to the narrative?... I especially wanted to capture the sound of the ball striking ads behind the net, so the listener could visualize how badly the shot missed and be certain that it did in fact miss. However, this sound



was included on a channel for the pitch crowd mic. I had dubbed this mic low in importance overall... In other words, the bleed between mics and the capturing of different sounds across channels, makes it very difficult to identify which channels can be attenuated as a whole.”

This feedback both remarks on the energy of the crowd, and also the issue that the author had found with sports content. Whereby the bleed between mics means that many of the sounds cannot be increased in level without inadvertently adding crowd noise, thus defeating the aim of the increase in level.

## 5.8 Conclusions

In this chapter the results of the data collection task have been presented. 50 participants took part in the task, 29 of whom had experience of working in audio. The assignments for the 9 scenes have been discussed. It has been demonstrated that participants did not agree on which sounds were important to the narrative, other than the Main Dialogue objects. It is anticipated that the low level of agreement is likely to impact a ML algorithm’s ability to learn robust categorisation of audio objects. It is certainly the case that NI is a subjective concept and therefore very difficult to quantify.

Perhaps the most insightful part of the data collected is the feedback provided by participants. Particularly when respondents discussed the choices they made, and the reasons for those choices. The participants’ reasoning also has provided some insight into the function of some sounds for listeners. The football content for example received several comments that participants needed the crowds reaction to help them follow the action. These comments shed some light on areas where an ML algorithm may struggle to handle the nuance of such decisions.

# Chapter 6

## Machine Learning

### 6.1 Introduction

This chapter will outline the Machine Learning (ML) work undertaken in this thesis. The starting point for the ML algorithm was [18]. In this paper, an ML algorithm was developed to assign NI metadata for a radio drama called ‘The Turning Forest’ which was created as part of the S3A project [67]. This algorithm will be referred to as the ‘legacy algorithm’ for the remainder of this chapter. The work here extends the method to more media content with a diverse range of genres. This chapter will outline the legacy algorithm’s structure and discuss its strengths and weaknesses. The results from retraining the legacy algorithm will then be presented, followed by the results of training a K-Nearest Neighbours (KNN) algorithm trained with the same database.

### 6.2 Method

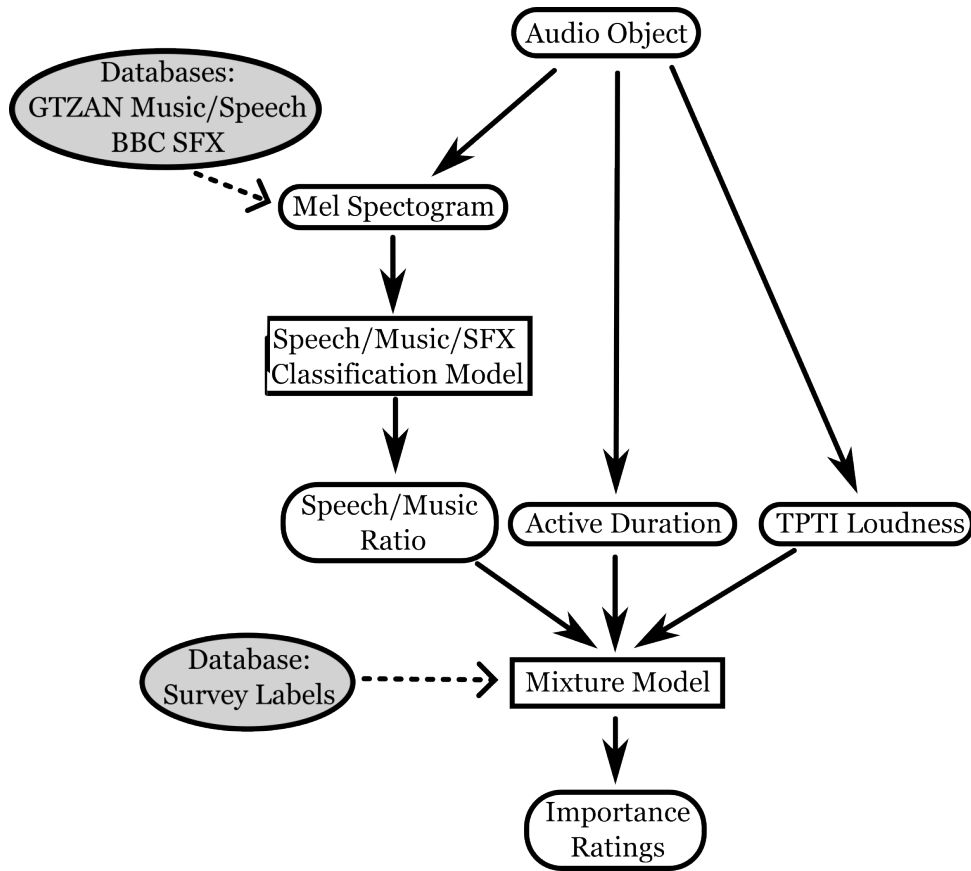
#### 6.2.1 Overall Architecture

The legacy algorithm was written in Python and uses two different ML models. The first model is used to create features which are utilised by the second model. The first model in the algorithm is a classification model which gives music, speech and SFX ratios for an input audio file.

The speech and music ratios are then used as two of four features which are fed into a mixture model, along with the assignment labels from the data presented in Chapter 5. The SFX ratio was not included as a feature, in keeping with the original work done in [18]. The other

two features used were True-Peak-To-Integrated (TPTI) loudness and active duration. The active duration was the duration of all frames with a root mean square (RMS) value greater than 0.001. The term ‘active’ will be used to signify frames with  $\text{RMS} > 0.001$  for the rest of this chapter. The TPTI loudness is the true peak (dBFS) divided by the integrated loudness (LUFS).

These four features along with the assignment data from the survey were fed into a mixture model based on Stochastic Variational Inference (SVI) [68]. A graphical representation of the model structure can be seen in Fig. 6.1.



**Figure 6.1:** The structure of the legacy algorithm. Rectangular blocks represent the ML models. Shaded ellipses represent the databases used for training.

### 6.2.2 Speech/Music/SFX Classification Model

The Speech/Music/SFX classification model utilises transfer learning to train the VGGish algorithm [69] to classify objects into speech, music, and sound effects (SFX). Transfer learning is a method where an ML model originally trained for a task is adapted with additional training to complete a different task with minimal computational effort [81].

The VGGish network is a convolutional neural network (CNN) trained on a large data set of

manually labelled audio events called *Audio Set* [70]. The dataset contains 632 classes of audio taken from 10 second clips of YouTube videos. The network’s architecture is a variation of VGG [71], an image classification algorithm. The pre-processing for VGGish downsamples the audio to a mono 16kHz file. Then a 64 band (125-7500Hz) log mel spectrogram is computed from a Short-Time Fourier Transform (STFT) (25ms windows with 10 ms overlap, Hann windowing). Non-overlapping frames of 0.96 seconds are created from this log mel spectrogram. Each 0.96s frame contains 64 mel bands and 96 10ms frames.

Transfer learning was applied to the VGGish network to train it to classify frames of audio into speech, music, or SFX. From these frame classifications, the ratio of speech, music, and SFX within an object is calculated by summing each classification and then dividing by the total number of frames in the object. This was done by replacing the final layer of VGGish with a 256-wide fully-connected layer. A final fully connected layer with 3 classes was added with a softmax activation function. A softmax activation function takes an input vector  $z$  of  $K$  real numbers and normalises it to the interval (0,1). Thus, these can be treated as probabilities corresponding to  $K$  categories. The equation for the standard softmax function,  $\sigma$ , can be found in Eq. (6.1). Table 6.1 shows the architecture for the network. Within the table it can be seen that the activation function for other layers is the ReLU (REctified Linear Unit) activation function. This is a function which outputs the input,  $x$ , if and only if  $x$  is positive, otherwise it returns 0. The equation for this is shown in Eq. (6.2).

Layer Type	Layer Shape	Activation Function
Convolutional	64 x 96 x 64	ReLU
Max pooling	64 x 48 x 32	-
Convolutional	128 x 48 x 32	ReLU
Max pooling	128 x 24 x 16	-
Convolutional	256 x 24 x 16	ReLU
Convolutional	256 x 24 x 16	ReLU
Max pooling	256 x 12 x 8	-
Convolutional	512 x 12 x 8	ReLU
Convolutional	512 x 12 x 8	ReLU
Max pooling	512 x 6 x 4	-
<i>Fully-connected</i>	256	ReLU
<i>Fully-connected</i>	3	<i>softmax</i>

**Table 6.1:** Architecture of the VGGish based network. The top layer represents the input with the output at the bottom.

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}, \quad \text{where } i = 1, \dots, K, \text{ and } \mathbf{z} = (z_1, \dots, z_K) \in \mathbb{R}^K \quad (6.1)$$

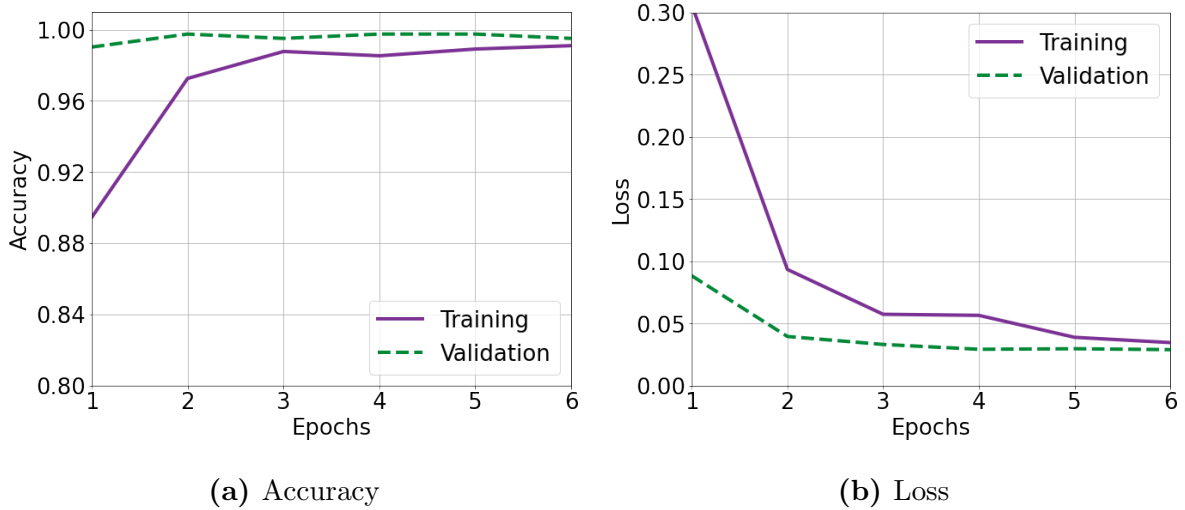
$$f(x) = \begin{cases} x & \text{if } x > 0, \\ 0 & \text{if } x \leq 0 \end{cases} \quad (6.2)$$

The transfer learning was carried out using the GTZAN music/speech discrimination database [82] combined with a selection of tracks from the BBC SFX library [83]. GTZAN is made up of 128 tracks, each 30 seconds in length. There are 64 examples each of speech and music. All of the tracks are 22050 Hz mono 16-bit wav files. The BBC SFX library was processed to match with this format, though some files were shorter than 30 seconds, resulting in 70 total wav files. This meant that the BBC SFX audio was downsampled to 22050Hz to match the GTZAN database, then when it was fed into the VGGish algorithm it was downsampled a second time to 16kHz. This downsampling will be discussed further in Section 6.2.7.

### 6.2.3 Speech/Music/SFX Classification Model Training

The VGGish section of the algorithm was originally trained over 4 epochs with a batch size of 16 and a learning rate (LR) of 0.01. The legacy code was trained on a single CPU, so batch size was kept small to keep computational intensity at a minimum. An investigation of extending the training was undertaken and the results can be seen in Appendix D. Different batch sizes were tested, along with different learning rates, for a greater number of epochs. This was largely done on iOS, which meant that a GPU could not be utilised due to limitations with TensorFlow and iOS compatibility. This limitation was found to not be a problem though as the algorithm appeared to hit optimal accuracy and loss rates at around 6 epochs.

The accuracy and loss plots for the final training are shown in Fig. 6.2. The training accuracy and loss are slightly improved by training for 6 epochs, rather than the original 4. The batch size and LR were kept the same at a batch size of 16 and a LR of 0.01.



**Figure 6.2:** The final training with batch size = 16, LR = 0.01

	Precision	Recall	F1-Score	Support
Music	0.98	0.99	0.98	713
Speech	0.98	0.98	0.98	527
SFX	0.97	0.96	0.96	550
Accuracy			0.98	1790
Macro Avg	0.98	0.98	0.98	1790
Weighted Avg	0.98	0.98	0.98	1790

**Table 6.2:** A table showing the precision, recall, F1 score, and accuracy figures after training the VGGish based model for 6 epochs. The ‘Support’ column refers to the number of samples in each class.

The precision, recall, F1 score, and accuracy figures are shown in Table 6.2. The model achieves very good performance across all three classification categories, with F1-Scores of 96-98%. The equations for calculating the precision, recall, F1 score, and accuracy are shown in Eqs. (6.3)

to (6.7).

$$\mathbf{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (6.3)$$

$$\mathbf{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (6.4)$$

$$\mathbf{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6.5)$$

$$\mathbf{F1-Score} = 2 \left[ \frac{\mathbf{precision} \times \mathbf{recall}}{\mathbf{precision} + \mathbf{recall}} \right] \quad (6.6)$$

$$= \frac{\text{TP}}{\text{TP} + 0.5(\text{FP} + \text{FN})} \quad (6.7)$$

where TP, FP, TN, FN denote the true positive, false positive, true negative, false negative respectively.

#### 6.2.4 Mixture Model

According to [18], a mixture model was chosen due to the low agreement amongst the original dataset of survey assignments from Chapter 9 of [17]. The mixture model allows for the lack of agreement to be treated as stochasticity in the decisions. The four features: speech and music ratios, TPTI loudness and active duration, were chosen to model this stochasticity appropriately and give suitable importance ratings. Mixture models are often chosen to model populations where labels are not available by splitting the dataset into distinctive groups through feature analysis. In this case the objects are labelled from the survey data presented in Chapter 5. However there are multiple labels for each object and the agreement is again low for most objects. These labels are used to train the mixture model along with the four features.

Assuming that each of the four features is sampled from a distribution then the model to make a decision on importance,  $d$ , can be given as in Eq. (6.8).

$$d = \arg \max_i \Pr(I = i) \prod_{\mu \in \{sr, mr, tpti, dur\}} \Pr(x_{\mu, i} | I = i, \theta_{\mu, i}) \cdot \Pr(\theta_{\mu, i} | I = i) \quad (6.8)$$

where  $x_{\mu}$  is each feature when  $\mu \in \{sr, mr, tpti, dur\}$  and  $sr$ ,  $mr$ ,  $tpti$ ,  $dur$  refer to the speech ratio, music ratio, TPTI loudness, and active duration respectively,  $\theta_{\mu}$  are the corresponding distribution parameters, and  $i$  is the level of importance such that  $i \in \{0, 1, 2, 3\}$  where 0 is Low, 1 is Medium, 2 is High and 3 is Essential.

For each of the four features a suitable distribution was chosen to model their probability density functions. For the features that are defined within the interval  $[0, 1]$ , i.e.  $x_{sr}$ ,  $x_{mr}$ ,  $x_{tpti}$ ,

a Beta distribution, Eq. (6.9), was chosen. Since  $x_{dur} \in \mathbb{R}^+$ , a Gamma distribution, Eq. (6.10), was chosen to model this feature. Both distributions are defined by two parameters:  $\alpha, \beta \in \mathbb{R}^+$ , and rely on the Gamma function shown in Eq. (6.11).

$$\text{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)x^{\alpha-1}(1-x)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}, \quad \text{where } 0 \leq x \leq 1 \quad (6.9)$$

$$\text{Gamma}(\alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}, \quad \text{where } x > 0 \quad (6.10)$$

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt, \quad \text{where } \Re(z) > 0 \quad (6.11)$$

To estimate these probability densities the mixture model uses a method called Stochastic Variational Inference (SVI) [68]. SVI is a method for approximating posterior distributions using stochastic optimisation, which allows very large and complex datasets to be modelled better than Variational Inference (VI) is capable of.

### 6.2.5 Mixture Model Training

Originally a 70/30% train/test split was carried out on the dataset of labels from the survey. Despite this being standard practice in many ML training methods, it was not an appropriate method here, as it resulted in there being no unseen objects in the testing set. This will be discussed further in Section 6.2.7.

A new method of splitting the data was required due to there being no unseen objects in the original method. Only 4 of the 9 scenes have a ground truth to refer to, and it is a subjective truth that should be considered flexible, due to the subjective nature of NI. For this reason it was decided that a traditional percentage split would not provide a true picture of how well the algorithm worked. Therefore, Cross-Validation methods were explored.

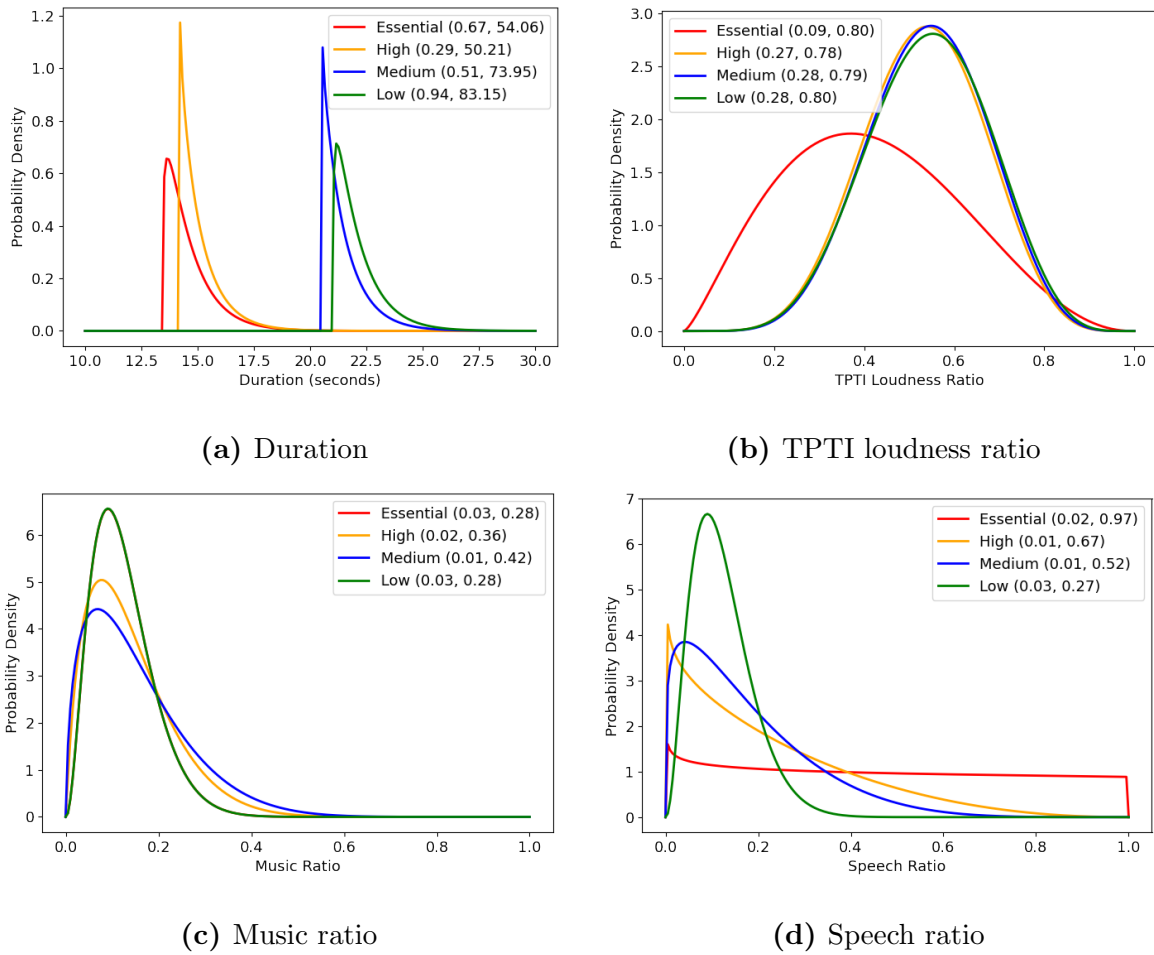
Cross-Validation is a method where the algorithm is trained multiple times. Each time the algorithm is trained a different testing set is extracted from the full dataset. Over the course of training, the algorithm is tested on all of the data. There are several different ways of splitting the data. One method, known as  $k$ -Fold Cross-Validation, involves splitting the dataset into  $k$  non-overlapping partitions, where  $k$  is usually chosen to be 5 or 10 [84]. The algorithm is then trained  $k$  times on  $k-1$  of the partitions, where the final partition (i.e. the  $k$ th partition, not included in the training) is used as a testing set. Another method is Leave-One-Out Cross-Validation (LOOCV). In LOOCV a single data point is left out on each training loop as a test point and the algorithm is trained on the remaining data. This method can take a very long



time to carry out for large datasets, as the algorithm is trained the same number of times as there are points in the dataset. However for training where datasets are small, or computing power is high, it provides a method which maximises the training set, whilst also providing a way of thoroughly testing the performance.

In the specific case of this work, it was decided to do a version of LOOCV. Since the dataset of survey responses can be both considered as 198 distinct audio objects or 9 scenes, it was decided to treat one as a scene. Meaning the algorithm was trained 9 times on 8 of the scenes, with one scene left out each time. The resulting feature distributions for the case where the protest scene was left out of training are shown in Fig. 6.3. The remaining sets of distributions can be seen in Appendix E.

The distributions are not particularly distinct. This could partially arise from the dataset of importance labels. Due to the inherently subjective nature of NI, a lot of the labels are not robust. That is to say that there are many objects with multiple labels attached, meaning that the data for each of the four categories overlaps significantly. In theory an object where half the participants have labelled the object as ‘High’ and half have labelled it as ‘Medium’ will be equally taken into account in both distribution estimations and this will contribute to the similarities between distributions.



**Figure 6.3:** Probability density functions for the four features when trained on all the data except for the Protest content

### 6.2.6 Changes made to the Algorithm

Several changes were made to the algorithm during this project. Some of the changes were made to simplify the code and some were made after discovering errors in the code. One Python library was removed from the code in part because it proved difficult to install on MacOS. The package was Essentia which is a C++ library with Python bindings for audio analysis [85]. It was entirely replaced with the Librosa package, which is another audio analysis tool that was already used in the code. The other reason for the removal of Essentia was to avoid inconsistencies. There were some sections of code where Essentia was used to load in the audio files and calculate features, whereas in other sections Librosa was used. These conflicts could cause variances in the performance of the ML. Due to a typing error the addition of noise was applied twice to the music ratio in training. In the inference code by the original authors, noise had been erroneously added to the TPTI loudness feature. This had not been done in

training. These errors were removed.

### 6.2.7 Critical analysis of Algorithm

There are several issues with the processing and features in the legacy algorithm. The majority of these issues occurred in the mixture model portion of the code.

Firstly the active duration feature was designed with the idea that a shorter active duration might indicate a more important sound. When considering an atmospheric track, such as a track of background birdsong, the active duration is likely to be relatively long as the track will cover an entire scene. Incidental sounds, such as a character falling, may be shorter and so this seems like a logical feature to use. However when considering an impulsive, repetitive sound, such as gunfire, it becomes clear that this feature could cause problems. A piece of content containing gunfire is likely to have more than one shot per audio object. Using active duration for an object such as this would result in the computer seeing it as one long sound, rather than several impulsive sounds. ‘The Turning Forest’, the piece of content originally used for training and testing the legacy algorithm, doesn’t contain any repetitive impulsive sounds, so for this piece, this feature was appropriate. In other content however it is possible that a different feature would give a better result.

The legacy algorithm lacks any interdependencies between objects. This results in the algorithm always assigning the same importance level to objects regardless of how many objects are in the mix. In theory this means that a piece of content with a single object, say a music track, could have that track assigned to low importance despite it being the only sound in the mix. This example is purely hypothetical, as a content piece of this nature would have no need for Narrative Importance, however the reader can extrapolate the implications of this. The same sound in a mix with 5 objects could have a very different meaning within the story than in a mix with 30 objects. This lack of any relational information between objects also means that sounds which don’t overlap dialogue cannot be detected by the legacy algorithm. Sounds like this could potentially be put in higher importance categories since they won’t be masking speech. Adding in similar interdependence has been recently explored in the related field of mixing music [62], [63]. This is a developing field and the relationships between objects in NI are complex and would require different handling than those in music.

The use of two different sample rates was questioned as it adds additional processing. The down-sampling was done by the Librosa [86], and Resampy [87] Python packages. Librosa and Res-

ampy both use bandlimited interpolation to downsample audio [87]. The audio files are loaded in at 22050Hz then downsampled to 16kHz during the processing of the Music/Speech/SFX classifier. The other features are calculated with the 22050Hz sample rate. All of the assets originally had a sample rate of 48kHz. The downsampling to 16kHz from 48kHz would be a straightforward decimation by a factor of 3. However because an additional downsampling to 22050Hz was added in prior to the 16kHz the samples will have been interpolated to 22050kHz and then interpolated again to 16kHz. The GTZAN database used to train the Music/Speech/SFX classifier has a sample rate of 22050Hz.

The train/test split of the dataset of survey assignments was done across the whole dataframe of features. This meant that the repetition of objects due to the labels wasn't accounted for and this split resulted in objects losing several labels, but still being included in training. Rather than 30% of the audio objects being removed as a testing set, 30% of the labels were removed. In the original work with 'The Turning Forest', 21 of the 22 objects were included in the test data and all 22 were in the training data. If the purpose of the train test split is to have an 'unseen' set of data to evaluate the learning then this method of splitting doesn't provide this. If that was not the aim of the split then the removal of this data from the training dataset seems futile, as all it achieved was reducing the data that the model was able to fit distributions to. In order to achieve an unseen dataset the split should have been done over the objects themselves, rather than the dataframe of assignments. This however could have easily led to no dialogue being included in training since 'The Turning Forest' is a narrated story with little other speech sounds.

The random seed had not been correctly set. Normally in ML algorithms best practice is to set the seed to be the same for all libraries, throughout the whole training process. At the start of the mixture model training code the seed for PyTorch [88] was set to 3, and the seed for NumPy [89] was set to 1. Later in the code the seed for Pyro [90], [91] was set to 0. The function that sets the seed in Pyro sets the seed for PyTorch, NumPy, and the Random library, meaning the seed for PyTorch and NumPy was changed mid-way through the code. The setting of the Pyro seed appears to have been tested to find the minimum loss in a range from 0-99. Upon testing this section of code, the results were that, despite changing the seed, the loss always remained the same, which meant that the seed was always set to 0. Upon running the code several times, it was found that the results were different every time, suggesting that there was a seed that wasn't fixed. Fixing the seed for the operating system appeared to resolve this issue, however it meant that the original results from [18] could not be reproduced. Later the original seed

setting was remedied so that the seed was constant throughout the code. The seed was set to 0. Several different seeds were tried but were found to have negligible impact on the performance of the model.

## 6.3 Results

The results from using the test set with each of the trained models was not promising. The model has a bias towards assigning Low importance. Of the 198 objects, 154 were assigned to the Low category, 7 were Medium, 10 High, and 27 Essential. The highest agreement in the survey data was achieved for the ‘Main Dialogue’ objects (as defined in Section 4.2). For these objects the algorithm performed well, assigning 18 of the 20 to the Essential Category.

It was expected that content where multiple scenes were included in the dataset would perform better due to the training method chosen. In other words, Vostok-K, Casualty, and Penguins, may produce better test results due to the inclusion of similar content in the training dataset. This was not the case. Since the Casualty scenes have a ground truth (Section 4.6) to compare with, these scenes have been chosen to discuss in more detail. The results from the other scenes can be found in Appendix F.

The results for the three Casualty scenes are shown in Figs. 6.4 to 6.6. As can be seen two of the three scenes only had objects assigned to Essential and Low. In those two scenes, the resulting mix is effectively a dialogue enhancement, since the objects in the Essential category are all dialogue, except for the ‘Breathing’ object in Scene 4 (Fig. 6.4), which was categorised by the author as Vocal Atmos, but could be considered as a dialogue type object, since it is a sound associated with a specific character. Scene 46 performs best out of the three scenes, and is even comparable with the ground truth. Precision and Recall tables for the three scenes are shown in Tables 6.3 to 6.5.

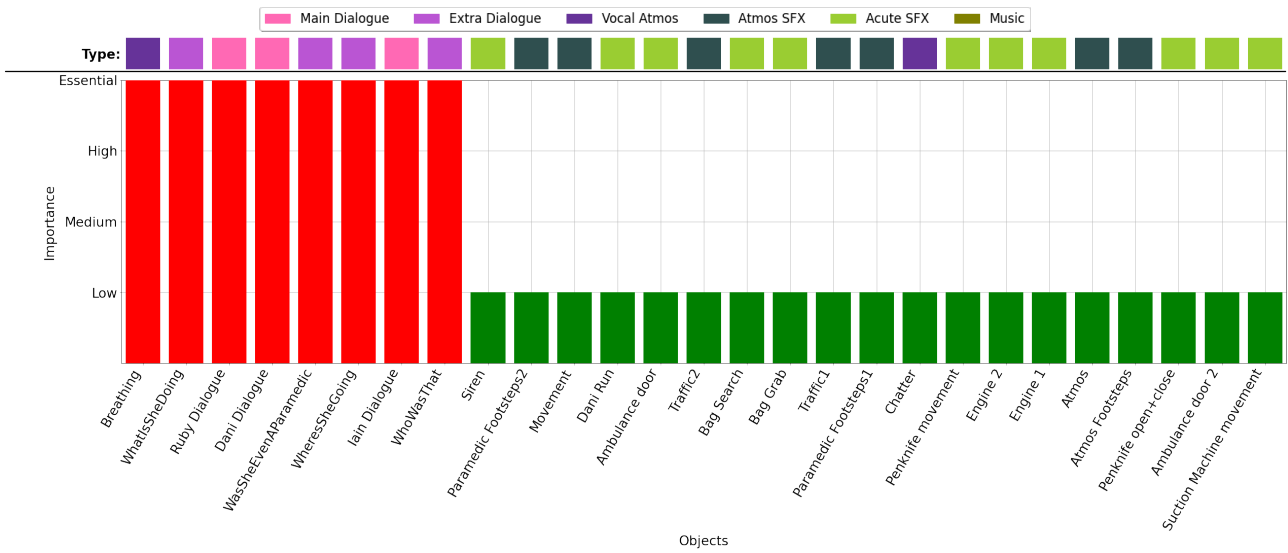


Figure 6.4: Casualty Scene 4 results from training the mixture model with survey data

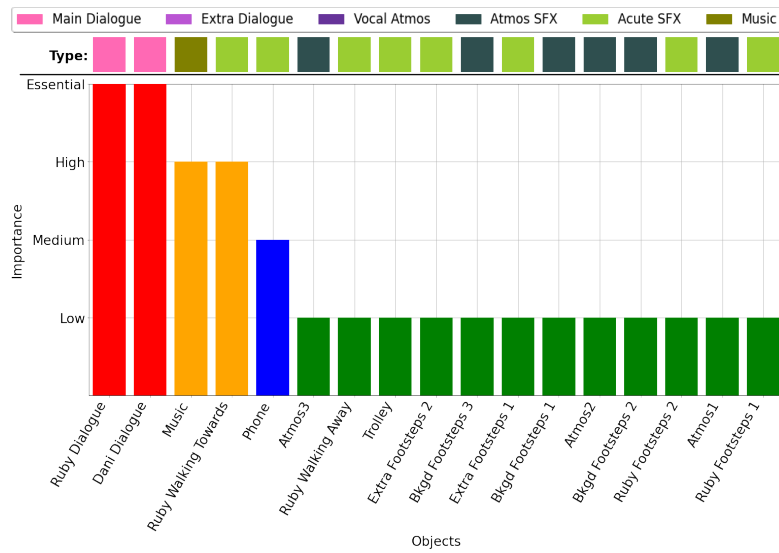


Figure 6.5: Casualty Scene 46 results from training the mixture model with survey data

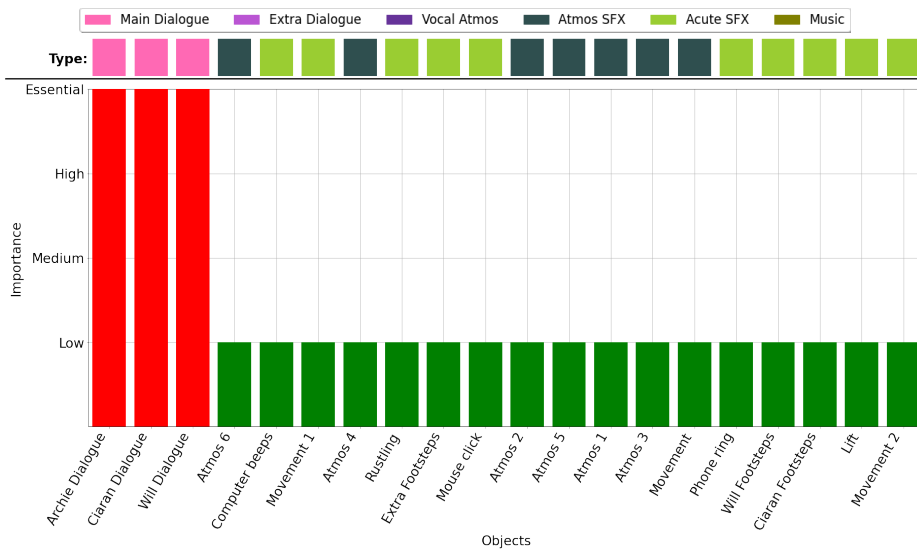


Figure 6.6: Casualty Scene 49 results from training the mixture model with survey data

	Precision	Recall	F1-Score	Support
Low	0.21	1.00	0.35	4
Medium	0.00	0.00	0.00	12
High	0.00	0.00	0.00	4
Essential	0.88	1.00	0.93	7
Accuracy			0.41	27
Macro Avg	0.27	0.50	0.32	27
Weighted Avg	0.26	0.41	0.29	27

**Table 6.3:** A table showing the precision, recall, F1 score, and accuracy for the model when tested with Scene 4 of Casualty. The ‘Support’ column refers to the number of samples in each class.

	Precision	Recall	F1-Score	Support
Low	0.92	1.00	0.96	11
Medium	0.00	0.00	0.00	1
High	0.50	0.50	0.50	2
Essential	1.00	0.67	0.80	3
Accuracy			0.82	17
Macro Avg	0.60	0.54	0.56	17
Weighted Avg	0.83	0.82	0.82	17

**Table 6.4:** A table showing the precision, recall, F1 score, and accuracy for the model when tested with Scene 46 of Casualty. The ‘Support’ column refers to the number of samples in each class.

	Precision	Recall	F1-Score	Support
Low	0.53	1.00	0.69	9
Medium	0.00	0.00	0.00	4
High	0.00	0.00	0.00	2
Essential	1.00	0.60	0.75	5
Accuracy			0.60	20
Macro Avg	0.38	0.40	0.36	20
Weighted Avg	0.49	0.60	0.50	20

**Table 6.5:** A table showing the precision, recall, F1 score, and accuracy for the model when tested with Scene 49 of Casualty. The ‘Support’ column refers to the number of samples in each class.

## 6.4 K-Nearest Neighbours Algorithms

In [18], the legacy algorithm was compared with two  $k$ -Nearest Neighbour ( $k$ NN) algorithms, namely a 3-Nearest Neighbour (3NN) and a 5-Nearest Neighbour (5NN). The findings were that the legacy algorithm outperformed both of these models. Again, these were trained on the same data they were tested on, due to the problems with the train test split discussed in Section 6.2.7. In this thesis it was decided to train 3NN, 5NN, and 7NN models, using the same Cross Validation method and features as in the mixture model.

$k$ NNs are a supervised learning method which can be used for both classification and regression problems. They use the  $k$  nearest training points in order to determine an output for a test point [92].

The results from all three  $k$ NNs were similar. The spread of objects across the four importance categories was more even. This can be seen in Table 6.6. However by looking at the ‘Main Dialogue’ objects it is possible to see that the algorithms are not successfully categorizing these objects as Essential all of the time. Table 6.6 also shows the number of Main Dialogue objects in each category for each  $k$ NN model. It can also be seen that there is little variation in the performance of the  $k$ NNs. This is likely due to the nature of the dataset. Since the features are repeated the number of participants who labelled each object, the ‘neighbours’ for each input are likely the same object with different labels.

	Total objects			Main Dialogue objects		
	3NN	5NN	7NN	3NN	5NN	7NN
Essential	22	19	19	6	6	6
High	31	34	40	4	5	4
Medium	73	74	61	7	7	7
Low	72	71	78	3	2	3

**Table 6.6:** A table showing how many objects (left three columns) and how many Main Dialogue objects (right three columns) were assigned to each importance category by each  $k$ NN

For consistency and ease of comparison the three Casualty scenes will again be looked at in more detail. The bar charts of results can be seen in Figs. 6.7 to 6.9. The remaining 7NN bar charts can be seen in Appendix G.



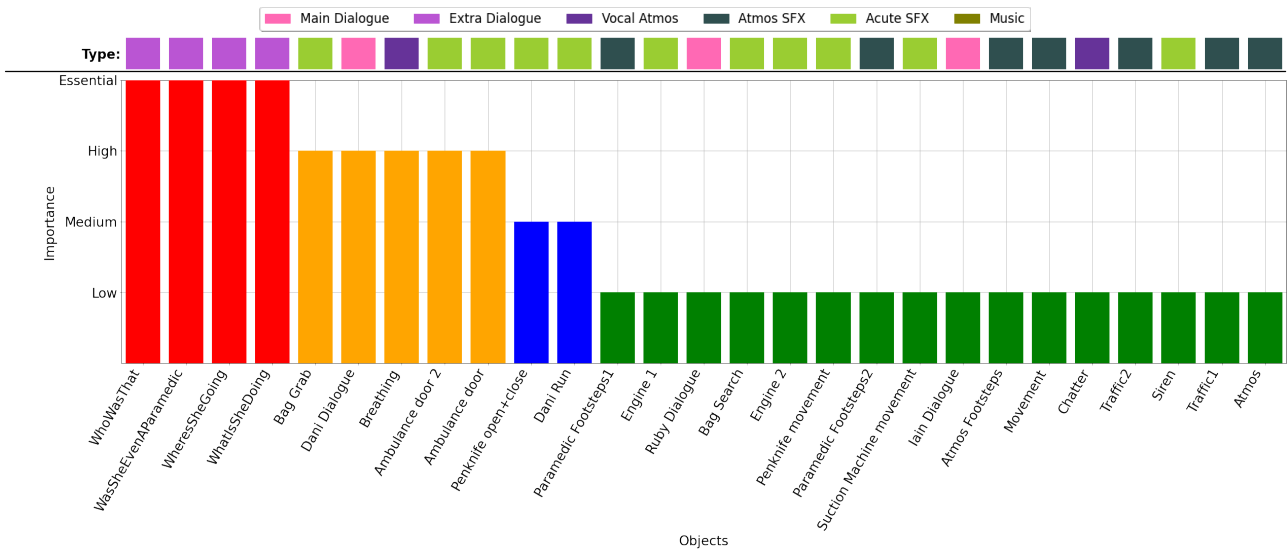


Figure 6.7: Casualty Scene 4 results from training a 7NN model with survey data

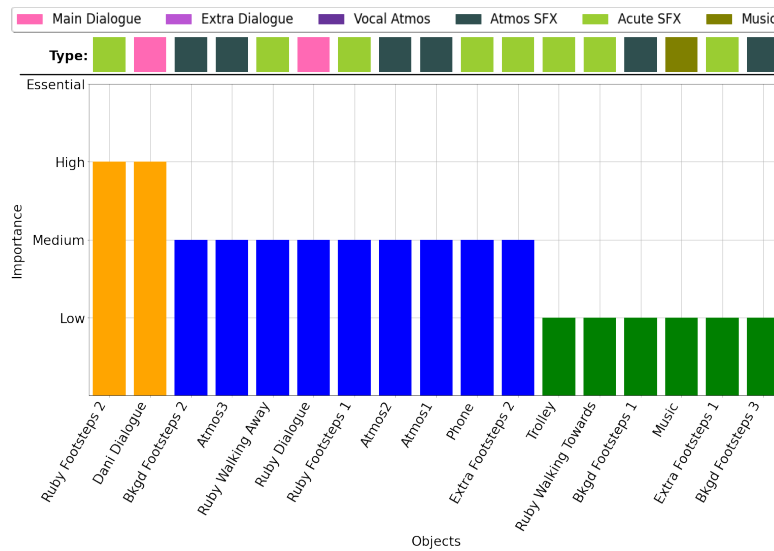


Figure 6.8: Casualty Scene 46 results from training a 7NN model with survey data

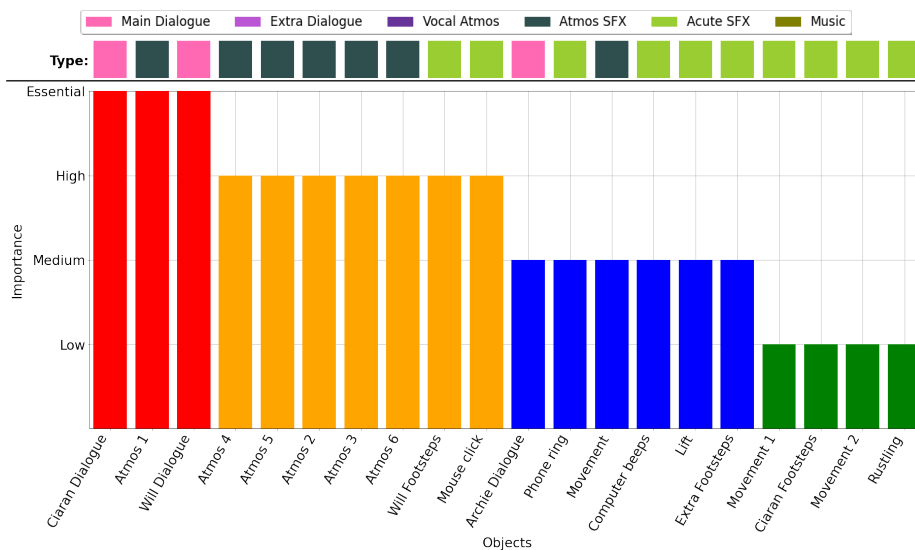
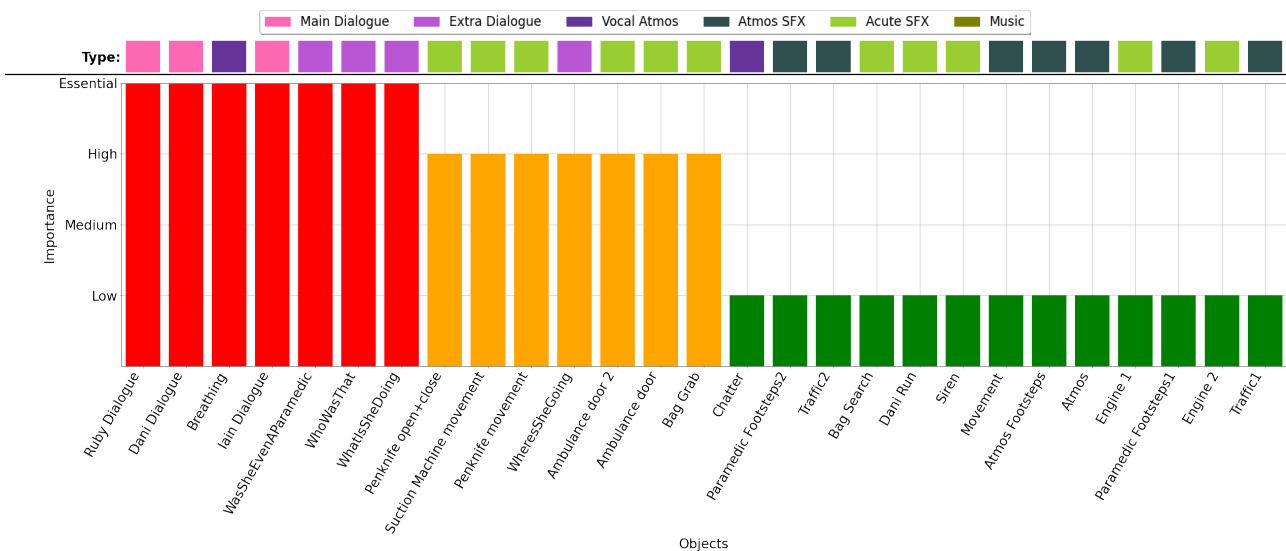


Figure 6.9: Casualty Scene 49 results from training a 7NN model with survey data

## 6.5 Training with Ground Truth Labels

It was suspected that many of the issues within the model are down to the choice to use survey data to train it. In order to investigate this theory, it was decided to carry out the same training but using the ground truth (GT) labels in place of the survey data. This meant that only Protest and the 3 Casualty scenes could be used, as only they had GT data. The cross validation method was retained and all other parameters were kept the same. The testing data resulted in 17 assignments of Essential and High, 11 Medium and 46 Low. All of the 10 Main Dialogue objects were assigned Essential.

Again, the three Casualty scenes will be looked at in depth to allow comparison with the other models. The results for Protest can be found in Appendix H. The bar charts showing the assignments for the three Casualty scenes are seen in Figs. 6.10 to 6.12 and the accuracy tables are in Tables 6.7 to 6.9. The training was more successful using single labels for each of the objects. The algorithm exhibited some bias against the Medium category, which appeared to occur when either Protest or Casualty Scene 4 were removed from training. These two scenes contained the majority of the Medium objects, with only 5 in total across the other two scenes. With a larger dataset this issue is likely to be remedied.



**Figure 6.10:** Casualty Scene 4 results from training the mixture model with GT data

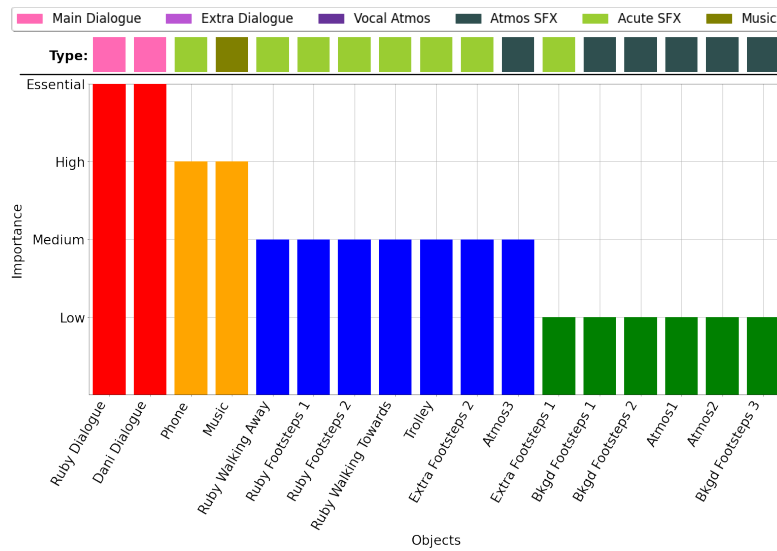


Figure 6.11: Casualty Scene 46 results from training the mixture model with GT data

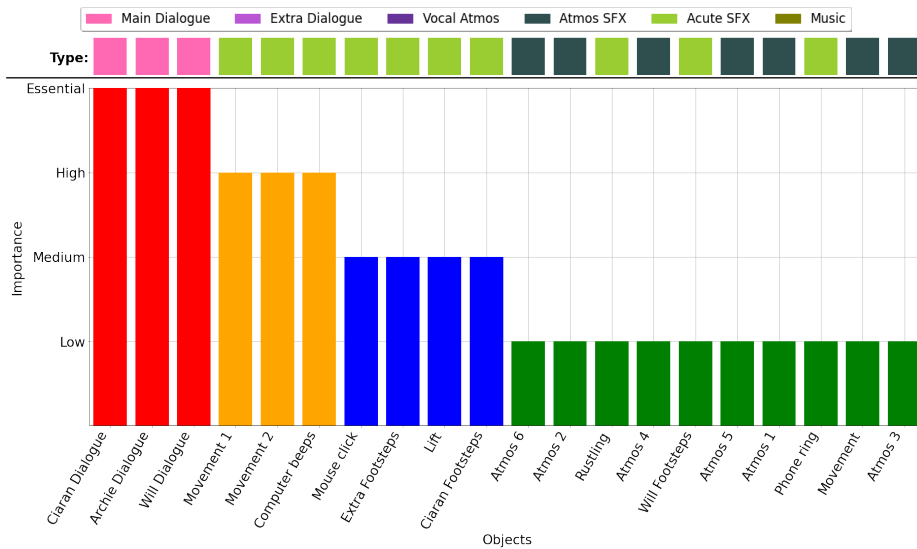


Figure 6.12: Casualty Scene 49 results from training the mixture model with GT data

	Precision	Recall	F1-Score	Support
Low	0.31	1.00	0.47	4
Medium	0.00	0.00	0.00	12
High	0.57	1.00	0.73	4
Essential	1.00	1.00	1.00	7
Accuracy			0.56	27
Macro Avg	0.47	0.75	0.55	27
Weighted Avg	0.39	0.56	0.44	27

Table 6.7: A table showing the precision, recall, F1 score, and accuracy for the model, trained on the GT data, where Scene 4 of Casualty was the test data. The ‘Support’ column refers to the number of samples in each class.

	Precision	Recall	F1-Score	Support
Low	1.00	0.55	0.71	11
Medium	0.14	1.00	0.25	1
High	0.50	0.50	0.50	2
Essential	1.00	0.67	0.80	3
Accuracy			0.59	17
Macro Avg	0.66	0.68	0.56	17
Weighted Avg	0.89	0.59	0.67	17

**Table 6.8:** A table showing the precision, recall, F1 score, and accuracy for the model, trained on the GT data, where Scene 46 of Casualty was the test data. The ‘Support’ column refers to the number of samples in each class.

	Precision	Recall	F1-Score	Support
Low	0.90	1.00	0.95	9
Medium	0.75	0.75	0.75	4
High	0.33	0.50	0.40	2
Essential	1.00	0.60	0.75	5
Accuracy			0.80	20
Macro Avg	0.75	0.71	0.71	20
Weighted Avg	0.84	0.80	0.80	20

**Table 6.9:** A table showing the precision, recall, F1 score, and accuracy for the model when tested with Scene 49 of Casualty. The ‘Support’ column refers to the number of samples in each class.

## 6.6 Conclusions

In this chapter the results of the ML work have been reported. The transfer learning from VGGish to the Music/Speech/SFX classifier was successful and this model categorises audio well. That being said, very few ML systems achieve 100% reliability and this is no exception. There were cases where some of the objects in this dataset were misclassified. Notable ones include the ‘Joe Dialogue’ object in Vostok-K scene 1, which contained a lot of heavy breathing and was classed as SFX. Some of the SFX in the Vostok-K scenes were classed as music, and the ‘Scum’ object in Protest was classified as SFX. At the point of recording, all of these sound sources were known, and could have been tagged with metadata. This raises the question, why use ML to backwork information that was once known? Work focused on metadata creation and retention would likely provide a more robust tagging system.

The generation of importance metadata for NI has been less successful. It has been shown that the use of survey data gave unreliable results when the algorithm was tested with unseen data. The most promising results came from training the mixture model with the singular set of GT labels. This removed a lot of the uncertainty and overlap of distributions for the mixture model that was created by the use of survey data. Unfortunately, due to the limited availability of ground truth data, it was not possible to explore whether some genres lend themselves better to the use of ML. Further work in this area should be considered and some of the ways this could be carried out will be discussed in Chapter 8.

Whilst the GT training objectively performed better than most of the other models, assessing its performance is complicated by the subjective nature of NI. A listening test to assess the performance of the algorithm's output in a subjective manner could prove more informative. Due to time constraints this wasn't possible in this thesis.

It is worth noting that, despite there being some show of potential for this algorithm, the amount of work required to get the audio into objects that could be handled by the algorithm, and the time needed to calculate all of the features, outweighs any benefit. Assigning NI metadata manually would be quicker than using the algorithm in its current form, even if it produced the perfect metadata tags every time. Given how much variation there is between audio recording practices, this problem is not likely to be resolved soon. Any algorithm which is developed in the future either needs to be far more flexible (i.e. trained with a wider variation of audio) or it must be less dependent on the audio itself. One idea could be to use track names as a feature for training, as often there is an indication of the content of the track in the track name (dial, sync, ADR, music, etc.).

# Chapter 7

## Further Understanding of Object-Based Audio

### 7.1 Introduction

The work in this thesis so far has focused on the development of a machine learning algorithm to generate Narrative Importance metadata. During the preparation of content several barriers were encountered, one of which was the provision of notionally object-based audio that couldn't be used. Simply requesting 'object-based audio' from project partners and other researchers resulted in a wide variety of content. The author found that a much more detailed description of the content format was needed to streamline requests. This raised the question that this chapter will aim to explore through a literature review and a survey of audio professionals. That question being 'what is object-based audio?'

During the course of this work, it became apparent that the question "what is object-based audio?" is one that has multiple answers depending on who is being asked and what the end-use of the audio is. In its most basic definition an audio object is simply audio with metadata attached to it. The author's understanding at the beginning of this work was that audio objects should be isolated recordings of specific sounds with no bleed or sharing of tracks.

### 7.2 The Problem/Motivation

The starting point for the work in this thesis was sourcing object-based audio (OBA) content. Initially, it was thought that this would be a simple task. As the search for appropriate content

unfolded it became apparent that not all OBA is created equally. The problems encountered have been discussed in detail in Chapter 4, to avoid repetition this section will briefly recap one example.

‘The Vostok-K Incident’ was the best example of an OBA production provided for this thesis. The piece was originally written for a device orchestration experiment. Two sets of stems had been saved, the delivered stems and the raw stems. Both sets of stems had a metadata file containing panning information and information of mapping from the raw to the delivered stems. Had only the delivered stems been provided then the content would not have been usable for this work, as a large number of the sound sources had been downmixed to a single stereo track. This track was the “stationary” track in the original production; that is, it was the track that always stayed on the main device being used by the listener. Some work was done by the author to adapt the raw stems for use in the survey, though most of this was necessary because of the limitations of the online task, and to make it accessible to people with little or no audio experience. The audio had all been recorded in well isolated conditions with no overlap or bleed. Both the recording and the storage of the assets serve as an example of good practice. No other piece of content provided met the expectations of the author due to the reasons outlined Chapter 4. Briefly these included bleed between objects, segmentation of sound sources, and sound sources sharing objects. The confusion of the variety of OBA is what motivated the work in this chapter.

### 7.3 OBA in the Literature

The literature on the matter is as varied as individuals’ opinions. The recently developed European Broadcasting Union (EBU) Audio Definition Model (ADM) appeared a logical location for a firm definition of OBA. However, even in this standard the EBU give two definitions for OBA in their ADM guidelines [93], demonstrating their awareness of the differing definitions used within the field. They are:

‘This leads us on to defining the term ‘object-based audio’ ... when we attach that metadata to some audio, it becomes object-based audio. So as long as we keep this metadata tied to the audio it is describing, we should be able to handle that audio correctly. However, it does mean we need to carry the metadata with the audio. So, this becomes our first definition of what object-based audio is (yes, I did say first, there’s another definition coming along later...)

and

‘Another approach came along called object-based audio (yes, this is the second definition of this; another name should have been used!), where each audio channel has some positional properties attached to it. These positional properties can then be interpreted by a renderer which attempts to position the sounds in space within the limitations of the location of the speakers.’

Both of these definitions focus on the attachment of metadata to audio. Nothing is said in either one about the content of the audio. The ADM also contains two definitions for an audio object:

- (1) ‘A set of tracks of a finite duration with a particular pack and channel configuration. The metadata includes start time and duration.’
- (2) ‘A sound located in a particular location in 3D space.’

Dolby have produced their own version of ADM for their Atmos system. This ADM describes an object as

‘an audio signal plus its associated object audio metadata that contains individually assigned object properties, content description, or interactivity limitations for personalization.’

It goes on to say that

‘the object properties more explicitly specify how the content creator intends the audio content to be rendered to loudspeakers.’

With the knowledge that Dolby Atmos is a 3D audio technology this proviso of the metadata becomes clear, and goes a long way to explaining the double definition of OBA given by EBU.

Fraunhofer [94] have an OBA system known as MPEG-H which says

‘Object-Based Audio (OBA) is a broad term that refers to the production and delivery of sound based on audio objects. In this context, an audio object represents a component of the audio mix delivered separately to the receiver and to which metadata have been added.’

They go on to say that OBA supports three main innovations; immersion, universal delivery, and advanced user-interactivity. The definitions of these boil down to spatial audio, playback device compatibility, and accessibility/personalisation features respectively.



The EBU, Fraunhofer, and Dolby ADMs reference the ITU standards on ADM. In Recommendation ITU-R BS.2051-3 [76] OBA is described as:

‘Object-based audio is an audio representation in which elements of the content are separate and accompanied by metadata which describe their relationships and allow a renderer to generate signals most appropriate to the playback system. The metadata may vary over time, for example to change the spatial position of an element of the content. An object-based approach also may allow users to interact with the audio content.’

Turning to other areas of the literature it is possible to see that the the dual ADM definitions are a result of long standing ambiguity around OBA which has existed for far longer than the ADM. Looking through the literature the impression is that OBA became assumed knowledge relatively quickly, with many papers on OBA neglecting to outline a clear definition. Often the definitions that are stated rely heavily on the use of the word ‘object’ to describe OBA, without giving an explicit definition of an ‘object’.

One of the earliest papers found (1996) referring to OBA is on object-based media (OBM) with a focus on video, published by the Massachusetts Institute of Technology (MIT) Media Lab [95]. It discusses how the audio was handled in the work.

‘We have also handled audio in an object-based fashion: rather than channels corresponding to speakers, sound was represented as a set of localized sources and an acoustical environment in which they are placed. These sound sources were then linked to the visual objects ... the audio is “rendered” for the speakers associated with the video display.’

A masters thesis from the MIT Media Lab in 1999 [96] focused on the capture of OBA and the use of blind source separation and deconvolution (BSSD) algorithms to aid in the clean capture. The thesis doesn’t explicitly define OBA, but the implication of the work is that, at the time, OBA focused on clean, separate audio capture, and a large focus of the thesis is the removal of room effects from the sources.

Despite a proliferation of work in the area of OBM in the 2000s, often described as object-based audio-visual, the author struggled to find research where OBA is explicitly defined. Frequently authors define the concept of OBM for video and talk about an “equivalent” for audio. In 2005 [97], the use of source separation techniques continues, again implying that OBA is dependent on clean, separate sources.

‘In this paper, we present an object-based 3D audio broadcasting system. This system consists of an authoring tool, a streaming server, and a client. The authoring tool generates an MPEG-4 file, made of multiple audio objects, after adapting several kinds of acoustical effects to the audio objects. Each audio object is recorded by multiple directional microphones and a multichannel 3D microphone. To increase the degree of source separation, source separation technology is applied to each object-based audio signal.’

By 2014 the definitions mention metadata, as seen in the quote from [98] below. In this definition, the objects are described as being ‘separate’ which could be interpreted as simply meaning ‘not downmixed’ or it could be that the intention is to specify that there is no bleed/overlap between objects.

‘In object-based audio the sound is represented by a number of separate audio objects that consist of audio tracks or sound events (e.g. a talker, an airplane, a guitar) and associated side information as metadata.’

In [50], the authors specify that the sound sources should be recorded separately. The metadata is explicitly discussed as having spatial information, along with ‘other’ metadata.

‘A more transparent method of representing a sound scene is to utilise an object-based approach where each sound source in the scene is recorded separately along with its position in space and some associated metadata describing other source characteristics.’

In 2015 a paper presenting a source separation method for the creation of OBA was published [99].

‘An emerging alternative... is object-based spatial audio, in which the auditory scene is represented by audio objects, with each audio object containing an audio stream as well as associated metadata. A typical audio stream is a sound source, and the metadata describes properties of the sound source and the acoustic ambience, e.g. the 3D position of the sound source and the reverberation level of the environment. At the rendering (reproduction) stage, to reconstruct a sound scene, these audio objects are mixed down based on the reproduction system setup as well as the metadata. A listener may interact with the listening environment by manipulating the metadata.’

A common theme seen in the literature is that the explanation of OBA focuses more on its benefits than outlining the technical details. In the ORPHEUS audio project [100], a project on OBA, the closest to a definition is:

‘Object-based audio is a revolutionary approach for the creation and deployment of interactive, scalable, immersive and cross media applications for any type of media content. It enables:

- (a) Multi-dimensional and multi-lingual features;
- (b) Novel interactive user experiences and personalized audio content;
- (c) The delivery of audio content in a format-agnostic manner.’

In Chapter 8 of ‘Immersive sound: The art and science of binaural and multi-channel audio’ (2017) [101] audio objects are described as:

‘We define sound objects as audio waveforms (audio elements) and associated parameters (metadata) that embody the artistic intent by specifying the translation from the audio elements to loudspeaker signals.’

In [102] from 2018, OBA is outlined as:

‘In the object-based audio paradigm, the content is represented as a virtual sound scene, which is a collection of sound-emitting objects. The audio for each individual object is transmitted, together with metadata describing how it should be rendered. The renderer, part of the end user’s sound reproduction equipment, interprets the object-based scene and derives the audio to be played out of each loudspeaker or headphone channel’

In 2019 the Journal of the Audio Engineering Society published a special issue on OBA. The guest editors’ note [103] contains an in-depth explanation of OBA, including its benefits and applications. This definition is by far the most detailed found in the literature, however it seems likely to be overlooked due to it being published as a guest editors’ note, rather than in a standalone paper.

‘Object-based audio is a concept that changes how audio is captured (recorded), how audio is stored, mixed, and produced (workflow), how audio is transmitted (broadcast), and how audio is rendered (reproduced) to the listener. Object-based audio differs significantly from scene-based (e.g., Ambisonic) or channel-based (e.g., ste-

reo) paradigms. In its purest form, object-based audio handles each separate sound or source as a separate object. A scene might be made up of dozens of objects. Each audio object is stored separately, along with its own metadata that describes features of the object such as its type, intended position in space, and so on. This might be accompanied with metadata concerning the environment. Production decisions made about how the objects are to be combined (mixed) are typically encoded into the metadata. This allows the production to be stored and transmitted as separate objects, so that the audio scene is only put together (rendered) at the point of listening. Rendering the scene at the listener gives advantages such as one production being rendered optimally on many different systems (stereo, Ambisonic, binaural, etc.), sometimes called format-agnostic reproduction, as well as the optimization to the listening environment (e.g., non-ideal stereophonic setups) and interactive scene manipulation.’

This text was included in the call for papers, which may explain why only 2 of the 10 published papers included a definition of OBA. Those two definitions were [104]:

‘We define sound objects as audio waveforms (PCM audio elements) and associated parameters (metadata) that embody the artistic intent by specifying the translation from the audio elements to loudspeaker signals.’

and [105]:

‘Object-based audio is becoming an increasingly important paradigm for producing, delivering, and reproducing (spatial) audio. It represents audio scenes as collections of objects — that is, audio signals and corresponding metadata that describe how the object is to be reproduced.’

Of the remaining 8 papers in the special issue, one quotes the ORPHEUS project and list the applications of OBA [47]. The remaining 7 papers [16], [42], [106]–[110] neglect to stipulate what they consider OBA to be.

One commonly used analogy when discussing object-based media(OBM) is that of baking a cake. A recent EBU magazine article [111] used this analogy to define OBM to the reader.

‘In tech terms the ingredients are objects: audio, video or graphics assets. These might be partially combined already, a cake that just needs the icing on top, or can be raw ingredients. The recipe is the metadata: data about the objects, which is

required to assemble them correctly. Finally, the chef is the renderer, which uses the metadata to assemble and play back the assets.’

### 7.3.1 Discussion

As seen in the literature, the two EBU ADM definitions are scattered throughout research. It can be argued that the spatially focused definitions are a subset of the more generic definition since the positional information is merely a type of metadata. However, this reinforces that the belief that multiple definitions for OBA coexist. The key themes in the literature reviewed here are ‘metadata’, ‘positional information’, ‘separation’, ‘playback systems’, and ‘user interaction’.

One consideration not yet touched upon is that OBA falls within the broader family of OBM which is defined across all forms of media as ‘a set of individual assets together with metadata describing their relationships and associations’ [112]. Within other forms of media there is no spatial application. What would a spatial application of a 2D video consist of? With this in mind, the use of OBA to refer to a specific spatial application is inconsistent with the wider industry of media.

Much of the work to date has focused on the formatting of the associated metadata. The ADM is the outcome of this, with the main aim being to standardise the metadata formatting and storage to allow transfer of object-based audio between establishments. There seems to be an assumption that the recording and storing of the audio itself is not a concern, though the authors’ experience has shown that the variability of these methods can render some object-based audio unsuitable for specific use-cases. This re-purposing of audio for different cases is mostly seen in research work, and so may not be a large barrier to the roll-out of OBA. The true barrier arising from the conflicting definitions comes when designing a training scheme for OBA. It is a reasonable assumption that these disparities will be perpetuated throughout the roll-out of OBA, leading to an industry that uses a single term to refer to refer to two distinct things.

## 7.4 Survey

A survey targeted towards people who work with object-based audio was undertaken. This was motivated by the ambiguous and inconsistent definitions seen in the literature. The aim was to establish whether professionals working with OBA shared these discrepancies or if there is an agreed working definition that has been overlooked by the literature. The objective is that

if there isn't a consistent definition, that this data will help inform clearer practical definitions in the future.

### **7.4.1 Method**

The survey was hosted on qualtrics.com. It began with a brief introduction and consent form. Participants who answered "no" to giving consent were directed to the end of the survey and asked no further questions.

The introduction explained that the survey's intention was to explore how people who work with audio define object-based audio. The introduction was kept very brief to avoid biasing the participants' views. As participants were recruited through organisational mailing lists for audio staff, participants were asked not to discuss their responses with colleagues to encourage individual responses.

The consent form explained how participants' data would be stored. Since no personal data was being collected participants were not able to withdraw their data. The participants were informed that their data would not be collected unless they clicked finish on the final page of the survey.

#### **Survey Questions**

Participants were then directed to answer 10 questions, most of which were open text responses. The first three were designed to establish their experience in audio. The fourth question asked if they were familiar/worked with OBA. Participants who answered "no" to this were directed to the end of the survey and thanked for their time. Participants were then asked how long they had been familiar/worked with OBA. The next 5 questions aimed to get a view on each participants individual understanding of OBA and how they work with it. The exact questions can be seen in Table 7.1.

	Question	Response
Q1	Which of the following best describes your work in audio?	<input type="radio"/> Technical <input type="radio"/> Research <input type="radio"/> Content Creator <input type="radio"/> Student <input type="radio"/> Other
Q2	Please give more detail (e.g. ‘I am a sound recordist’ or ‘I am a PhD student’). <i>Please do not give any identifying details such as your employer’s name.</i>	Text Response
Q3	How many years have you worked in audio?	Text Response
Q4	Are you familiar with object-based audio?	<input type="radio"/> Yes, I have worked with it <input type="radio"/> Yes, I know of it <input type="radio"/> No
Q5	How long have you been aware of/worked with object-based audio?	Text Response
Q6	Please give your definition of what ‘object-based audio’ is (give as much detail as you wish).	Text Response
Q7	How would you define an ‘audio object’?	Text Response
Q8	What differences are there in the recording process for object-based audio compared to traditional audio?	Text Response
Q9	If you have worked on object-based audio productions, what implementation of object-based audio have you worked with most commonly (e.g. for spatial audio or for accessible content)?	Text Response
Q10	When working on OBA productions, what tools do you most commonly use?	Text Response

**Table 7.1:** A table of the 10 questions in the OBA survey

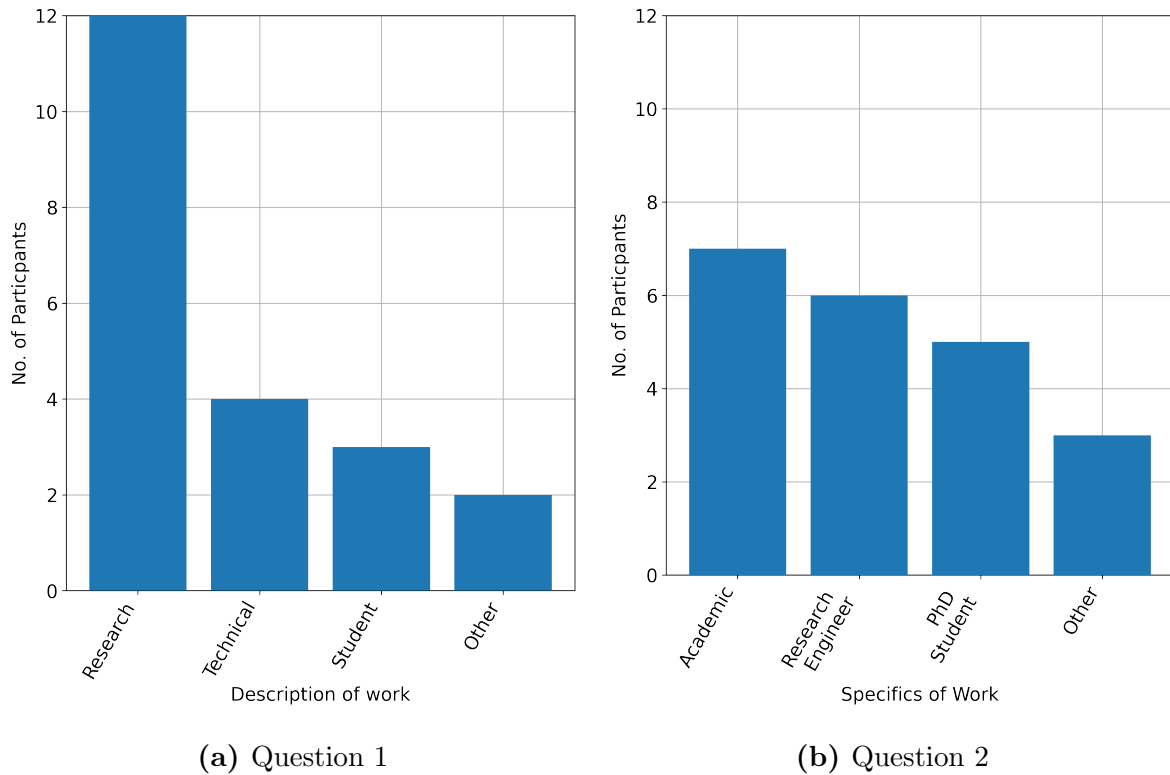
## 7.4.2 Results

### Demographics

The survey was distributed to the BBC R&D Audio Team and through multiple university Audio Research departments, including University of York and University of Salford. 21 participants completed the survey; 13 of those answered that they have worked with OBA and 8 said they were familiar with it. The participants were given integer IDs from 0 to 20 which will be used to discuss specific participant responses in the following sections.

The majority of participants felt that their work was best described as research as can be seen in Fig. 7.1a, which shows the responses to Question 1. No participants identified with the title ‘Content Creator’, which isn’t surprising given the recruitment pool. Participants were asked to give more detail on their work in their next answer. Of the two respondents who ticked ‘Other’, one said they were a lecturer and the other an educator. All three participants who answered ‘Student’ said they were PhD students. Of the four ‘Technical’ responses, 1 declined to give more information but later said that their work involved training and advocacy, 1 said they were a software developer, 1 a research engineer, and 1 a mastering engineer. The 12 ‘Research’ responses were grouped into 2 PhD students, 5 research engineers and 5 academics. Overall this sums to 5 PhD students, 6 research engineers, 7 academics, 1 software developer, 1 mastering engineer, and 1 unspecified technical job as shown in Fig. 7.1b.

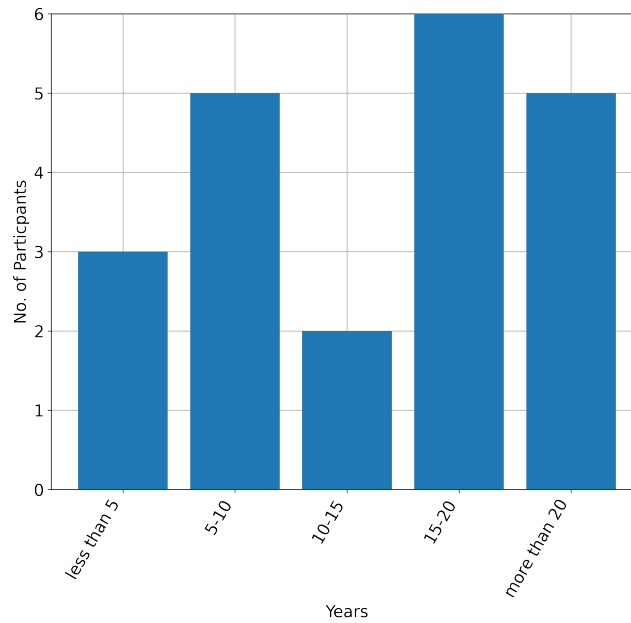




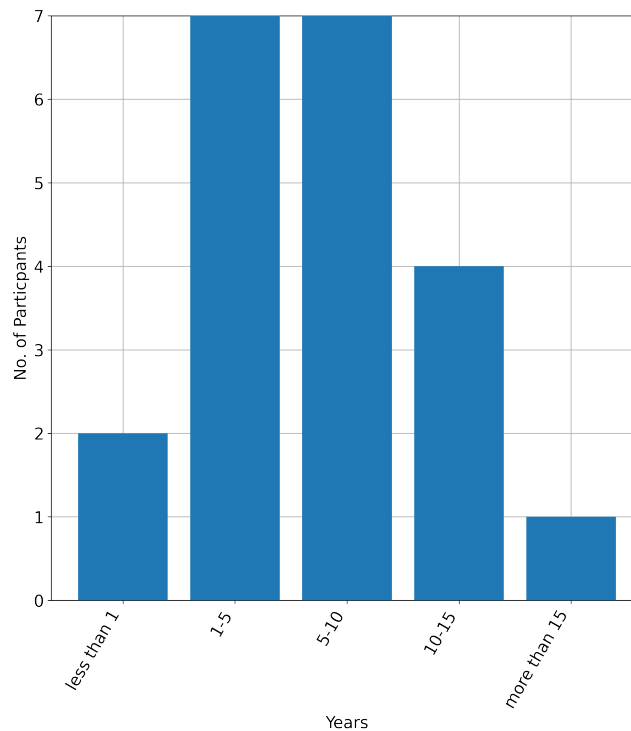
**Figure 7.1:** Bar charts showing the responses to question 1 (which of the following best describes your work in audio?) and 2 (please give more detail) from the survey. No participants answered that they were a content creator for question 1. For question 2 ‘Other’ includes the software developer, mastering engineer, and the unspecified technical job

The number of years participants have worked in audio is shown in Fig. 7.2. The range of experience spanned 38 years, with the least experienced participant saying they had worked in audio for 2 years and the most experienced saying they had worked for 40 years in the field. The mean fell between 15 and 16 years and the median was 16 years.

Participants’ experience with OBA is shown in Fig. 7.3. The range spanned 21 years, from 8 months to 22 years. The mean fell between 7 and 8 years and the median was 6 years. Two participants had worked with OBA for their entire careers, one who had been working for 4 years and one for 22 years. One participant said they had been familiar with OBA for 6 years but they had only worked in audio for 4 years.



**Figure 7.2:** A bar chart showing the responses to question 3 (how many years have you worked in audio?) of the survey. Participants' answers were grouped into categories during analysis. The range was from 2 to 40 years.



**Figure 7.3:** A bar chart showing the responses to question 5 (how long have you been aware of/worked with object-based audio?) of the survey. Participants' answers were grouped into categories during analysis. The range was from 8 months to 22 years.

## Definitions of Object-Based Audio

Throughout the analysis of the qualitative responses the chosen technique is deductive analysis with an unconstrained matrix [113], [114]. This method uses established theory or concepts to build a framework for analysis, in this case the review of the literature in Section 7.3 was used to inform the choice of key words to search for in the definitions. An unconstrained matrix approach allows for additional categories to be added to the framework if the responses contain any recurring themes not seen in the literature.

The first question simply asked participants for their definition of OBA. Firstly, mentions of metadata were counted. Given that the broadest OBA definition in literature is often “audio with attached metadata” this seemed like an obvious place to begin. 13 out of 21 participants were counted as referring to metadata. 12 of the 13 explicitly used the word metadata and 1 participant referred to ‘parameters that allow the file to be decoded and delivered to a wide range of platforms.’ This was deemed a reference to metadata by the author.

The next theme that was analysed were references to separate sounds. This included participants who talked about individual sound sources, storing sounds as ‘their own objects’, discrete signals, and distinct objects. One participant talked about ‘breaking audio into different parts’ and another who said that OBA ‘compartmentalizes sounds’. These were counted as a reference to separation. There were 15 responses that mentioned distinct sounds in some way.

Themes for applications mentioned in the responses were analysed. 9 participants talked about OBA enabling formatting for different playback systems. 9 participants made references to spatial audio applications. 9 participants made reference to user personalisation, including accessibility, user preferences, variable content length, and interactive narratives. 3 participants didn’t mention any specific use-cases in their definitions. 7 participants mentioned two or more of the above use-cases in their answers. 11 participants only referenced one use case, though some of these listed personalisable and accessible audio as separate uses. The author has grouped these because usually when talking about OBA accessibility implementations the focus is on personalisation.

Answers ranged greatly in length amongst participants. No minimum or maximum length was set by the author, though the question was presented with an answer box that was several lines long, to encourage participants to be as verbose as they wished, and participants were encouraged to ‘give as much detail as you wish’. Participant 7 gave the most succinct answer,

and the least strict definition which was

“Any form of audio which can be varied in presentation or sequence by data and/or interaction.”

The longest definition was given by participant 2 as

“The practice of producing media in such a way that foregoes traditional channel based formats and allows control over individual audio ‘objects’, usually rendered on the listener’s device at the point of consumption. Use cases for object-based audio media tend to fall into two main categories; firstly control over the mix of individual audio objects (which could potentially be used to provide the listener with control over elements of the mix for accessibility or perhaps access to completely different stream of audio e.g. access to different commentary or crowd mix on a sports event. Then secondly, explicit or implicit personalisation of the order or inclusion/exclusion of individual audio ‘objects’ which would allow interactive/personalised narratives.”

### **Definitions of an ‘Audio Object’**

The next question asked participants to explain what they understand an audio object to be. One participant wrote ‘see above’, referring to their answer of the previous question. Their answer to the previous question was considered as an answer to this question.

Metadata was referred to by 15 participants in these answers, 5 of these hadn’t mentioned metadata in the definition of OBA they gave. Discerning further themes to analyse from these answers proved difficult. Some answers were vague and brief such as ‘Anything it needs to be’ and ‘A bit of a programme that has information attached to it, but is not a whole show’. There were some themes within the more detailed responses, for example a few participants mentioned channels or beds, however their answers were contradictory. Participant 3 said:

“An audio object can be anything from a single speech ‘track’ to a high-order ambisonic bed made up of multiple sounds. It is the smallest unit from which an object-based mix can be created. It is very flexible.”

Whereas Participant 13 answered:

“A single audio event tagged with metadata. It should be differentiated from channels or beds, which carry multiple audio objects already pre-mixed into a defined format.”

This was a theme throughout the answers, participants generally agreed that an audio object was a ‘small’ unit of sound but there was little agreement on how small that unit should be. Some participants explained that they felt this was context dependant. Participant 14 said:

“An audio object is a an individual element that is recorded, mixed and transmitted separately. What an audio object includes is dependent upon the context and could include a mixture of different sounds. E.g. a string section could be considered an object in some situations, whereas in others, each string instrument within the section would be considered an object.”

This logic does follow and is reflected in the work in this paper, where all of the music objects in scenes were presented as a single downmixed object, rather than individual instruments or sections. If those same pieces of music were being mixed for use in a spatial application independent of the narrative content then having them downmixed to a stereo track would render them unusable.

### **Recording differences**

Participants were asked what differences there were between OBA and traditional audio. 11 participants mentioned metadata in their responses, 2 of whom had not talked about it in the previous 2 questions, meaning only 1 participant didn’t talk about metadata in any of their answers.

8 participants mentioned ambisonics or 3D sound. Many of these participants were referencing the use of ambisonics microphones in the recording process. Some said that this was the only difference in recording OBA.

8 participants said there was no difference between traditional audio and OBA. 3 of these also talked about ensuring isolation of sounds (which is not always done in traditional recording). 3 of these participants also mentioned ambisonics some said that ambisonics recording was different from traditional recording practices but not all OBA was. Some participants said that ambisonics recording wouldn’t require a different type of recording process.

### **OBA Implementations and Tools**

Participants were asked to list the kind of OBA productions they had worked on and the tools they generally use in their OBA work. Of the 21 participants, 6 said they had not worked on OBA productions. This is interesting when considering that the first of the OBA

questions allowed participants to respond “Yes, I have worked with it” or “Yes, I know of it” and 8 participants opted for the second answer. It would be expected, based on this, that 8 participants would then say they hadn’t worked on OBA productions.

Of the 15 participants who did give answers to these two questions, 9 of them said their work involved spatial audio applications and 8 mentioned personalisable or accessible applications. There were a couple of other applications mentioned, such as device orchestration, ADM, and interactive OBA for XR applications. One participant said their main area of work was training and advocacy. Several participants discussed that their work centres on research of OBA applications.

The tools used by participants were varied. In terms of DAWs 7 participants said they used Reaper, 2 Protools, 2 Audacity, and 1 Nuendo. 5 participants talked about the BBC EAR production suite [115], 2 of these participants mentioned the ADM in their answer, no other participants explicitly said ADM. Three said they used MPEG-H authoring tools. Some participants talked about using self developed plug-ins for their work, and others mentioned specific plug-ins such as SPAT [116], VISR [117], and IEM Plug-in Suite [118]. Two participants included Dolby Atmos in their answer, one of these said this was what they used in Protools and the other was a person who said they hadn’t worked on any OBA productions but that they have assessed other people’s Atmos content.

### 7.4.3 Discussion

An exploration of any patterns between participants’ demographic information and their responses was carried out but no correlation was found. This could be in part due to the small number of participants and the limited recruitment of participants. Since the participant pool was largely made up of people working in research areas and didn’t include many Dolby users it is possible that there is a bias in these answers. It will be informative to continue this work with a wider reach on participant recruitment. Due to time constraints a conscious choice was made to keep the recruitment small for this thesis as a starting point for the investigation.

The survey does show that there is no agreed upon definition amongst professionals. This was the expected outcome after the exploration of the literature. The main finding is that the large majority of participants made reference to metadata in one of their answers (20 out of 21). Only 3 of the 21 didn’t discuss the isolation or separation of sounds, so it would seem that this is another key theme.

## 7.5 Conclusions

This chapter set out to explore the discrepancies within the definitions of OBA. The chapter began by looking at technical documentation around the ADM and found that within a single document two different OBA definitions are given. Other areas of the literature were explored and found to contain similar conflicts. This led to the design of a survey on OBA, sent out to various research departments. The survey results support the findings in the literature, despite a limited participant pool. The main themes agreed upon across both the literature and the survey participants are metadata, and isolation and applications of spatial audio, playback systems, and user interaction/personalisation.

This chapter highlights an oversight in recent work done to standardise OBA metadata formatting - namely the lack of distinction between different types of OBA. If OBA is the term to describe all audio with attached metadata, then it would be wise to develop subsets or levels of OBA that describe the objects' granularity.

The author proposes that OBA should become an umbrella term to encompass all applications and should be defined as:

OBA is an audio production technique where ascribed metadata allows aspects of the mix to be rendered at the user end rather than at the production stage. This requires sounds to be left as separate sound events rather than mixed into beds. Each sound event, along with its respective metadata, is referred to as an audio object.

From this definition, different applications and their requirements can be defined. Guidance on how the audio is stored for ease of transfer to other applications should be outlined. One example of this could be ambisonics OBA, where the audio should be recorded with specialist ambisonics microphones, and the metadata should contain the positional information. In this application, isolation of sounds is less important since the purpose of the multiple objects is to build up a sense of space, not to give control of those objects to the user.

In cases where control is given to the user, i.e. personalisation or interactive use cases, the isolation and quality of recordings are likely to be more important. Taking NI as the obvious example, poor isolation of sounds can lead to the system working inefficiently if low importance sounds are audible on high importance tracks.

# Chapter 8

## Summary and Key Contributions

### 8.1 Introduction

This chapter will provide a summary of the work carried out in this thesis. The novel contributions to knowledge will be outlined. From these key contributions, suggestions for areas of further work will be proposed. Finally the chapter will finish with concluding remarks on the work.

### 8.2 Summary

At the beginning of this thesis, the aim was to determine whether ML could be used to generate NI metadata for audio objects within different broadcast genres. The work predominantly consists of three parts. The first part takes the form of a survey that was run to provide a database for training the ML, the ML algorithm itself constitutes the second part, and the third part is an investigation into the definition of OBA.

#### 8.2.1 The Data Collection Survey

In Chapters 3 to 5, a survey, which comprised of a questionnaire and an assignment task was designed and deployed. The main aim of this survey was to provide a database of labels for training the ML algorithm.

In Chapter 3 the methodology for the survey was laid out. The design was based on a similar survey run in Chapter 10 of [17]. After the questionnaire, which collected demographic information, participants were presented with a DAW-like interface and asked to assign each object



in a mix an importance value based on the NI system. There were four possible importance levels - Essential, High, Medium, and Low.

Chapter 4 gave an extensive overview of the content which was used for the survey. This chapter discussed the reasons for content selection, and gave some examples of content which was considered but ultimately deemed unsuitable. The result was 9 scenes sourced from 5 pieces of broadcast content. Chapter 4 ends with a set of guidelines for sourcing or creating and storing NI-friendly content.

In Chapter 5 the results from the survey were reported. The main findings were that importance is largely subjective. Most participants agreed that main dialogue tracks (main character dialogue, narration, commentary, etc) are essential to the narrative. However all other objects were varied in their assignments, with most objects being assigned to at least three of the four possible categories.

### 8.2.2 Machine Learning

In Chapter 6, the ML approach was described and the results from the algorithm were presented. The work done in [18], was the basis of the ML portion of this thesis. The algorithm developed in that paper was extended here to other genres of broadcast audio. Several issues with the training of that algorithm were uncovered. The main one being that the algorithm had unintentionally been trained on the same data it was tested with. This algorithm was trained on both the survey data and ground truth labels using a cross validation method. The best results came from training the algorithm on a single set of labels (the ground truth data) rather than the survey data, where each object had a minimum of 15 labels. The algorithm was compared with a KNN algorithm trained on the survey labels, which also performed poorly.

### 8.2.3 Object-Based Audio

Finally Chapter 7, raised the question “What is Object-Based Audio?”. A literature review showed that the definition of OBA is dependent on the context of the paper, and what application of OBA is being presented. Similarly, the survey showed that individuals’ ideas of OBA are diverse. There was a general agreement that OBA involves metadata, with 20 out of 21 participants mentioning this at some point in their responses. The key finding from this work is that OBA is poorly defined, both in the literature and amongst professionals who work with it.

## 8.3 Contributions to Knowledge

This section will outline the main contributions to knowledge that have come from this thesis.

### 8.3.1 Narrative Importance Content Guidelines

The first contribution to knowledge in this thesis is the set of NI content guidelines outlined at the end of Chapter 4. These guidelines lay out the requirements of content which is to be used with the NI system. The guidelines aim to streamline the creation and/or sourcing of appropriate content for NI.

### 8.3.2 Narrative Importance - What's Important?

The second key contribution of the thesis is the survey undertaken to collect NI metadata. Whilst the main aim of undertaking the survey was to build a database for the training of a ML algorithm, it was also an interesting experiment in how people value broadcast audio objects. This serves as an extension to the similar work done in Chapter 10 of [17], where audio production staff were surveyed. The main findings are that opinions on what is considered important are a very subjective matter. It is possible to conclude that the majority of people valued Main Dialogue objects - that is dialogue spoken by named characters, narration and commentary. Aside from this, categorisation was varied, and classification was dependent on subjects and genre of content. It was seen in the feedback text from the survey that participants found the task difficult. Some of the feedback provided valuable insights into how participants made their decisions, and shed light on areas where a machine would struggle to identify the nuance of what makes a sound relevant within a storyline.

### 8.3.3 Machine Learning of Importance

This thesis has contributed to the investigation of whether ML can be used for NI metadata generation. The main methods investigated were the use of a mixture model with survey data, a mixture model with ground truth labels, and a KNN with survey data. The main finding is that using survey data to train these algorithms is not an effective method. The use of multiple labels for this kind of data muddies the dataset and leads to generalisation of the distributions for a mixture model. In the case of a KNN, the algorithm is only able to pull from neighbours which are effectively the same object but with different labels from the survey. The use of the ground truth labels gave better results, confirming that the survey data was the main issue.

Any future work in this area should be carefully designed to avoid repeating these errors. Ideas for how this could be continued will be discussed in Section 8.4.

### **8.3.4 What is Object-Based Audio?**

The final contribution of this thesis is the raising of the question “What is OBA?”. In Chapter 7, a literature review and survey revealed that the definition of OBA is unclear. A suggestion of how OBA should be defined was made by the author. This definition should be used as an umbrella term for all types of OBA and further specifications of subsections of OBA based on use-cases (spatial, personalisation, accessibility etc.) should be made.

## **8.4 Areas for Further work**

In this section, areas for further work will be proposed. The suggestions here are not exhaustive, rather are indicative of the key issues the author has encountered throughout this thesis.

### **8.4.1 Narrative Importance - What’s Important?**

As was shown by the work done in Chapter 5, the importance of a sound is largely a subjective decision. Further work could be done to develop a set of guidelines and examples for how importance should be decided by producers. This work could take the form of in depth interviews with audio production staff. A series of listening tests could accompany these interviews, where participants are members of the general public. The test could present different NI mixes to participants and ask them to rate various metrics like enjoyment, understanding, and immersion.

### **8.4.2 Alternatives to Machine Learning**

Work should be carried out to investigate alternative ways to streamline the creation of NI metadata. One possibility is that tagging audio at the creation point with some basic labels (such as ‘dialogue’ or ‘Atmos SFX’) would prevent the need for categorisation after the fact. This would require robust tagging systems, that ideally would be standardised across the industry. This information could help producers to streamline their choices when working through the NI metadata creation, or it could be used in conjunction with a different ML algorithm.

### 8.4.3 Machine Learning Development

Another option could be to develop a plug-in that learns the assignments for a particular content piece or producer. This would be worthwhile for long-running shows such as *Casualty* and other soap operas. These types of programmes tend to have the quickest turnaround, as some air multiple times a week. Dramas are consistently reported as being the most troublesome genre for audio, and for this reason it would be valuable to provide solutions for this genre. Focusing on a single production may also allow for a more specific algorithm, with non-audio based features to be developed. Features such as track names or script parsing may prove to be very informative, and would take less time to process than audio.

Some genres lend themselves better to ML as they are formulaic. Sports is a good example of this, often having crowd sounds and commentary, with a few other discrete sounds. Currently, most sports audio contains few tracks, however with the development of additions such as the kick enhancement seen in this work, the sound scenes are becoming more complex and NI will complement this. Another area where ML could be implemented is in detected points where the crowd is reacting to action on the pitch, as was seen in the football scene in this work. The crowds reaction was commented on by participants as being important to help them follow the action of the game. An algorithm which detected this points and separated them from the general hubbub of crowd could be useful in assisting with NI, since this points may need to be in a higher category than the other crowd sounds.

### 8.4.4 Object-Based Audio Definitions

Further work should be done to better define the term object-based audio. Currently the term object-based is being used to describe different, specific types of object-based audio, which lack distinct names. The author has proposed that the term object-based audio should continue to be used as an umbrella term for all audio with attached metadata and that further names be developed for different branches within that. Without specific titles for different types of OBA, ascertaining whether a piece of content will easily map to another OBA implementation requires examination of the stems. Different divisions of OBA would also allow for guidelines for each type to be drawn up, which could lead to easier creation of flexible content which maps to different use cases.

The work done in this thesis has provided evidence that this issue does exist. In order to fully understand its prevalence, an extension of the survey should be undertaken. The participant

recruitment for this thesis was intentionally limited due to time constraints. Expanding the survey to reach a more varied participant pool is the most logical next step.

### 8.4.5 The Recording Process

Finally for some genres, the way the tracks are recorded is the biggest barrier to accessibility and implementing accessibility tools like NI. Most sports commentary, for example, is saturated with crowd sound, due to the location of the commentators in the stadium. Often, dialogue for dramas contains movement noise or bleed from the background noise of the set. Issues like these are the reason why providing clean audio tracks is so costly and can cause speech intelligibility issues that are difficult to rectify in post-processing.

## 8.5 Conclusions

The advent of OBM is opening up a wealth of opportunities for better media experiences for everyone. NI is one area of many potential accessible formats that are possible with OBM. The question of how these new technologies can be best implemented at scale is still in the process of being answered. At the start of this Master's thesis, using ML to streamline production workflows for NI seemed possible. Throughout the course of this work the method that was previously believed to give promising results, has been shown to be less effective than first thought. It is certainly still possible that ML could be used to assist with workflow time constraints. However it has been shown that using survey data is not an effective method of training, and that future work should be more focused on working with producers to create robust label sets.

The survey that collected NI labels showed that participants struggled to agree on the importance of most objects. Other than Main Dialogue tracks, most objects were assigned to at least three of the four importance levels. There was some indication that audio staff may generally rate a larger number of objects higher, indicating that the general public may be more content with simpler mixes than audio professionals. This would require further investigation to confirm.

In the course of this investigation, it became apparent that another barrier to the roll-out of NI (and other object-based technologies) is how the industry defines OBA. This issue extends beyond the NI system and into many other areas of audio research, which are currently being deployed into the mainstream. The issue of ambiguity is one that has the potential to cause

great confusion within the audio industry.

This work concludes with three main points. The first being that this method of ML is not as effective as first thought and other methods should be the focus of any future ML work. The second being that it is worth continuing with work to understand how decisions on the importance of a sound are made, and whether this importance is different for production staff and the general public. Finally, this work has shown that OBA needs a robust definition. Such a definition has been proposed in this thesis.

# Bibliography

- [1] H. Baumgartner, R. van Everdingen, B. Schreiner, M. Kahsnitz and U. Krämer, *Speech Intelligibility in TV*, EBU Technical Review, 2022.
- [2] H. Ellis-Petersen, ‘BBC’s Jamaica Inn drama loses quarter of audience after sound quality issues,’ *The Guardian*, Apr. 2014. [Online]. Available: <https://www.theguardian.com/media/2014/apr/23/bbc-jamaica-inn-audience-sound-transmission-du-maurier> (visited on 24/04/2023).
- [3] J. Plunkett, ‘BBC TV chief: Source of mumbling problem is ‘incredibly hard’ to isolate,’ *The Guardian*, Apr. 2016. [Online]. Available: <https://www.theguardian.com/media/2016/apr/19/bbc-tv-mumbling-problem-happy-valley> (visited on 24/04/2023).
- [4] O. Strelcyk and G. Singh, ‘TV listening and hearing aids,’ *PLoS ONE*, vol. 13, no. 6, pp. 1–21, 2018.
- [5] H. Behrends, W. Bradinal and C. Heinsberger, ‘Loudspeaker Systems for Flat Television Sets,’ in *123rd Audio Engineering Society International Convention*, New York, NY, USA, Oct. 2007, Paper Number: 7302.
- [6] L. Kelly, H. Wey, S. Cho and J. Kim, ‘In room acoustic simulation of television type reproduction environments,’ in *153rd Audio Engineering Society International Convention*, Oct. 2022.
- [7] J.-H. Jung, J.-W. Choi, S. Park and Y.-H. Kim, ‘Equalization filter design for downfiring flat television speakers,’ *Applied Acoustics*, vol. 76, pp. 66–81, Feb. 2014.
- [8] *RNID - National hearing loss charity*. [Online]. Available: <https://rnid.org.uk/> (visited on 19/04/2023).

- [9] *Subtitle it! - RNID*. [Online]. Available: <https://rnid.org.uk/get-involved/campaign-with-us/subtitle-it/> (visited on 19/12/2022).
- [10] ‘Television channels required to provide access services in 2021,’ Ofcom, Tech. Rep., 2021.
- [11] J. M. Schneider, B. Gopinath, C. M. McMahon, S. R. Leeder, P. Mitchell and J. J. Wang, ‘Dual Sensory Impairment in Older Age,’ *Journal of Aging and Health*, vol. 23, no. 8, pp. 1309–1324, Dec. 2011. [Online]. Available: <https://doi.org/10.1177/0898264311408418> (visited on 05/04/2023).
- [12] F. R. Lin, K. Yaffe, J. Xia *et al.*, ‘Hearing loss and cognitive decline in older adults,’ *JAMA Internal Medicine*, vol. 173, no. 4, pp. 293–299, Feb. 2013.
- [13] B. G. Shirley, ‘Improving Television Sound for People with Hearing Impairments,’ Ph.D. dissertation, Acoustics Research Centre, University of Salford, 2013.
- [14] M. Armstrong, ‘From Clean Audio to Object Based Broadcasting,’ *BBC Research & Development White Paper*, 2016.
- [15] *Prime Video launches a new accessibility feature that makes it easier to hear dialogue in your favorite movies and series*, Apr. 2023. [Online]. Available: <https://www.aboutamazon.com/news/entertainment/prime-video-dialogue-boost> (visited on 19/04/2023).
- [16] L. A. Ward and B. G. Shirley, ‘Personalization in object-based audio for accessibility: A review of advancements for hearing impaired listeners,’ *Journal of the Audio Engineering Society*, vol. 67, no. 7-8, pp. 584–597, Aug. 2019.
- [17] L. A. Ward, ‘Improving Broadcast Accessibility for Hard of Hearing Individuals : Using object-based audio personalisation and narrative importance.,’ Ph.D. dissertation, Acoustics Research Centre, University of Salford, 2020.
- [18] E. T. Chourdakis, L. Ward, M. Paradis and J. D. Reiss, ‘Modelling experts’ decisions on assigning narrative importances of objects in a radio drama mix,’ in *Proceedings of the 22nd International Conference on Digital Audio Effects (DAFx-19)*, Birmingham, UK, Sep. 2019.



- [19] Action on Hearing Loss, ‘Hearing Matters,’ Tech. Rep., 2020.
- [20] L. M. Haile, K. Kamenov, P. S. Briant *et al.*, ‘Hearing loss prevalence and years lived with disability, 1990–2019: Findings from the Global Burden of Disease Study 2019,’ *The Lancet*, vol. 397, no. 10278, pp. 996–1009, Mar. 2021.
- [21] A. M. Goman and F. R. Lin, ‘Prevalence of hearing loss by severity in the United States,’ *American Journal of Public Health*, vol. 106, no. 10, pp. 1820–1822, Oct. 2016. [Online]. Available: <https://ajph.aphapublications.org/doi/10.2105/AJPH.2016.303299> (visited on 22/12/2022).
- [22] ‘Addressing the rising prevalence of hearing loss,’ World Health Organization, Tech. Rep., 2018.
- [23] W. J. Strawbridge, M. I. Wallhagen, S. J. Shema and G. A. Kaplan, ‘Negative Consequences of Hearing Impairment in Old Age: A Longitudinal Analysis,’ *The Gerontologist*, vol. 40, no. 3, pp. 320–326, Jun. 2000. [Online]. Available: <https://academic.oup.com/gerontologist/article/40/3/320/605349> (visited on 21/12/2022).
- [24] A. Shukla, M. Harper, E. Pedersen *et al.*, ‘Hearing Loss, Loneliness, and Social Isolation: A Systematic Review,’ <https://doi.org/10.1177/0194599820910377>, vol. 162, no. 5, pp. 622–633, Mar. 2020. [Online]. Available: <https://journals.sagepub.com/doi/10.1177/0194599820910377> (visited on 22/12/2022).
- [25] C. J. Plack and H. H. Guest, ‘The Psychology of Hearing Loss,’ *Oxford Research Encyclopedia of Psychology*, Nov. 2022. [Online]. Available: <https://doi.org/10.1093/acrefore/9780190236557.013.894> (visited on 22/12/2022).
- [26] D. G. Loughrey, M. E. Kelly, G. A. Kelley, S. Brennan and B. A. Lawlor, ‘Association of Age-Related Hearing Loss With Cognitive Function, Cognitive Impairment, and Dementia: A Systematic Review and Meta-analysis,’ *JAMA Otolaryngology–Head & Neck Surgery*, vol. 144, no. 2, pp. 115–126, Feb. 2018. [Online]. Available: <https://jamanetwork.com/journals/jamaotolaryngology/fullarticle/2665726> (visited on 21/12/2022).

- [27] R. K. Sharma, A. Chern and J. S. Golub, ‘Age-Related Hearing Loss and the Development of Cognitive Impairment and Late-Life Depression: A Scoping Overview,’ *Seminars in Hearing*, vol. 42, no. 1, pp. 10–25, Feb. 2021.
- [28] C. Füllgrabe, B. C. J. Moore, M. A. Stone *et al.*, ‘Age-group differences in speech identification despite matched audiometrically normal hearing: Contributions from auditory temporal processing and cognition,’ *Frontiers in Aging Neuroscience*, vol. 6, 2015.
- [29] S. Schelinski and K. Von Kriegstein, ‘Brief Report: Speech-in-Noise Recognition and the Relation to Vocal Pitch Perception in Adults with Autism Spectrum Disorder and Typical Development,’ *Journal of Autism and Developmental Disorders*, vol. 50, pp. 356–363, Oct. 2019.
- [30] A. Sturrock, H. Guest, G. Hanks, G. Bendo, C. J. Plack and E. Gowen, ‘Chasing the conversation: Autistic experiences of speech perception,’ *Autism and Developmental Language Impairments*, vol. 7, pp. 1–12, Feb. 2022.
- [31] ‘The Total Audience Report,’ The Nielsen Company (US), LLC, Tech. Rep., 2021.
- [32] Ofcom, *Television and on-demand programme services: Access services report*, 2021. [Online]. Available: <https://www.ofcom.org.uk/research-and-data/multi-sector-research/accessibility-research/television-and-odps-access-services-jan-dec-2021> (visited on 19/12/2022).
- [33] ‘Broadcast Centre Incident,’ Ofcom, Tech. Rep., 2022. [Online]. Available: [https://www.ofcom.org.uk/\\_data/assets/pdf\\_file/0032/238964/incident-review-red-bee-media.pdf](https://www.ofcom.org.uk/_data/assets/pdf_file/0032/238964/incident-review-red-bee-media.pdf) (visited on 14/04/2023).
- [34] *Ofcom finds Channel 4 breached licence conditions over subtitle problems - Ofcom*. [Online]. Available: <https://www.ofcom.org.uk/news-centre/2022/channel-4-breached-licence-conditions-over-subtitle-problems> (visited on 19/12/2022).
- [35] E. Zwicker and H. Fastl, *Psychoacoustics, Facts and models*. Berlin: Springer, 2006.
- [36] J. Archer, *Soundbars vs surround sound speakers: Which is best to boost your home theater?* Apr. 2022. [Online]. Available: <https://www.techradar.com/news/soundbars->

- vs-surround-sound-speakers-which-is-best-to-boost-your-home-theater (visited on 24/04/2023).
- [37] *Can a Good Soundbar Rival a True Surround-Sound System in a Blind Listening Test?* Apr. 2020. [Online]. Available: <https://www.nytimes.com/wirecutter/blog/soundbar-vs-surround-sound-system/> (visited on 24/04/2023).
- [38] D. Geary, M. Torcoli, J. Paulus *et al.*, ‘Loudness Differences for Voice-Over-Voice Audio in TV and Streaming,’ *Journal of the Audio Engineering Society*, vol. 68, no. 11, pp. 810–818, Nov. 2020.
- [39] M. Torcoli, A. Freke-Morin, J. Paulus, C. Simon and B. Shirley, ‘Background Ducking to Produce Esthetically Pleasing Audio for TV with Clear Speech,’ in *146th Convention of the Audio Engineering Society*, Dublin, Ireland, Mar. 2019, Paper Number: 10175.
- [40] M. Armstrong, ‘Audio Processing and Speech Intelligibility: A literature review,’ *BBC Research & Development White Paper*, 2011.
- [41] L. Ward, B. Shirley, Y. Tang and W. J. Davies, ‘The effect of situation-specific non-speech acoustic cues on the intelligibility of speech in noise,’ in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017-August, Stockholm, Sweden, Aug. 2017, pp. 2958–2962.
- [42] J. Paulus, M. Torcoli, C. Uhle, J. Herre, S. Disch and H. Fuchs, ‘Source separation for enabling dialogue enhancement in object-based broadcast with MPEG-H,’ *Journal of the Audio Engineering Society*, vol. 67, no. 7-8, pp. 510–521, Aug. 2019.
- [43] N. L. Westhausen, R. Huber, H. Baumgartner, R. Sinha, J. Rennie and B. T. Meyer, ‘Reduction of Subjective Listening Effort for TV Broadcast Signals With Recurrent Neural Networks,’ *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3541–3550, Nov. 2021.
- [44] A. Bidner, J. Lindberg, O. Lindman and K. Skorupska, ‘Deploying Enhanced Speech Feature Decreased Audio Complaints at SVT Play VOD Service,’ in *Lecture Notes in Networks and Systems*, vol. 440, Springer Science and Business Media Deutschland GmbH, 2022, pp. 277–285.

- [45] T. Walton, M. Evans, D. Kirk and F. Melchior, ‘Does environmental noise influence preference of background-foreground audio balance?’ In *141st Audio Engineering Society International Convention*, Los Angeles, CA, USA, Oct. 2016, Paper Number: 9637.
- [46] A. Master and H. Müsch, ‘A Model to Predict the Impact of Dialog Enhancement or Mix Ratio on a Large Audience,’ in *149th Audio Engineering Society International Convention*, Online, Oct. 2020, eBrief: 637.
- [47] A. Silzle, R. Schmidt, W. Bleisteiner, N. Epain and M. Ragot, ‘Quality of Experience Tests of an Object-based Radio Reproduction App on a Mobile Device,’ *Journal of the Audio Engineering Society*, vol. 67, no. 7/8, pp. 568–583, Aug. 2019.
- [48] J. Francombe, J. Woodcock, R. J. Hughes, K. Hentschel, E. Whitmore and A. Churnside, ‘Producing audio drama content for an array of orchestrated personal devices,’ in *145th Audio Engineering Society International Convention*, New York, NY, USA, Oct. 2018, eBrief: 461.
- [49] R. Bleidt, A. Borsum, H. Fuchs and S. M. Weiss, ‘Object-Based Audio: Opportunities for Improved Listening Experience and Increased Listener Involvement,’ *SMPTE Motion Imaging Journal*, vol. 124, no. 5, pp. 1–13, Jul. 2015.
- [50] R. Oldfield, B. Shirley and J. Spille, ‘An object-based audio system for interactive broadcasting,’ in *137th Audio Engineering Society International Convention*, Los Angeles, CA, USA, Oct. 2014, Paper Number: 9148.
- [51] L. Ward, M. Paradis, B. Shirley, L. Russon, R. Moore and R. Davies, ‘Casualty Accessible and Enhanced (A&E) Audio: Trialling object-based accessible TV audio,’ in *147th Audio Engineering Society International Convention*, New York, NY, USA, Oct. 2019, eBrief: 563.
- [52] I. Mcclenaghan, L. Pardoe and L. Ward, ‘The next generation of audio accessibility,’ in *152nd Convention of the Audio Engineering Society*, The Hague, Netherlands, May 2022, Paper Number: 10598.

- [53] S. Pauletto, R. Selfridge, A. Holzapfel and H. Frisk, ‘From Foley professional practice to Sonic Interaction Design: Initial research conducted within the Radio Sound Studio Project.,’ in *Nordic Sound and Music Computing Conference*, Online, Nov. 2021.
- [54] L. Ward and B. G. Shirley, ‘Television Dialogue; Balancing Audibility, Attention and Accessibility,’ in *Accessibility in Film, Television and Interactive Media*, York, UK, Oct. 2017.
- [55] J. Gillick and D. Bamman, ‘Telling Stories with Soundtracks: An Empirical Analysis of Music in Film,’ in *Proceedings of the First Workshop on Storytelling*, New Orleans, LA, USA, Jun. 2018, pp. 33–42.
- [56] B. Hoeckner, E. W. Wyatt, J. Decety and H. Nusbaum, ‘Film Music Influences How Viewers Relate to Movie Characters,’ *Psychology of Aesthetics, Creativity, and the Arts*, vol. 5, no. 2, pp. 146–153, May 2011.
- [57] M. Kock and C. Louven, ‘The Power of Sound Design in a Moving Picture: An Empirical Study with emoTouch for iPad,’ *Empirical Musicology Review*, vol. 13, no. 3-4, pp. 132–148, Apr. 2019.
- [58] James Woodcock, William J. Davies, Trevor J. Cox and Frank Melchior, ‘Categorization of broadcast audio objects in complex auditory scenes,’ *Journal of the Audio Engineering Society*, vol. 64, no. 6, Jun. 2016.
- [59] Google Developers, *Machine Learning Glossary*. [Online]. Available: <https://developers.google.com/machine-learning/glossary> (visited on 25/04/2023).
- [60] G. Sharma, K. Umapathy and S. Krishnan, ‘Trends in audio signal feature extraction methods,’ *Applied Acoustics*, vol. 158, Jan. 2020.
- [61] S. Venkatesh, D. Moffat and E. R. Miranda, ‘Word Embeddings for Automatic Equalization in Audio Mixing,’ *Journal of the Audio Engineering Society*, vol. 70, no. 9, pp. 753–763, Sep. 2022.
- [62] M. A. Martínez-Ramírez, W.-H. Liao, G. Fabbro, S. Uhlich, C. Nagashima and Y. Mitsufoji, ‘Automatic music mixing with deep learning and out-of-domain data,’ in *23rd*

- International Society for Music Information Retrieval Conference (ISMIR)*, Bengaluru, India, Aug. 2022.
- [63] M. A. Martínez Ramírez, D. Stoller and D. Moffat, ‘A Deep Learning Approach to Intelligent Drum Mixing With the Wave-U-Net,’ *Journal of the Audio Engineering Society*, vol. 69, no. 3, pp. 142–151, Mar. 2021.
- [64] S. Venkatesh, D. Moffat and E. R. Miranda, ‘You Only Hear Once: A YOLO-like Algorithm for Audio Segmentation and Sound Event Detection,’ *Applied Sciences*, vol. 12, no. 7, Apr. 2022.
- [65] D. Griffiths, S. Cunningham, J. Weinel and R. Picking, ‘A multi-genre model for music emotion recognition using linear regressors,’ *Journal of New Music Research*, vol. 50, no. 4, pp. 355–372, 2021.
- [66] T. Ciborowski, S. Reginis, D. Weber, A. Kurowski and B. Kostek, ‘Classifying emotions in film music—a deep learning approach,’ *Electronics (Switzerland)*, vol. 10, no. 23, Dec. 2021.
- [67] J. Woodcock, C. Pike, F. Melchior, P. Coleman, A. Franck and A. Hilton, ‘Presenting the S3A Object-Based Audio Drama Dataset,’ in *140th Convention of the Audio Engineering Society*, Paris, France, Jun. 2016, eBrief: 255.
- [68] M. D. Hoffman, D. M. Blei, C. Wang, J. Paisley, J. Edu and T. Jaakkola, ‘Stochastic Variational Inference,’ *Journal of Machine Learning Research*, vol. 14, pp. 1303–1347, May 2013.
- [69] S. Hershey, S. Chaudhuri, D. P. Ellis *et al.*, ‘CNN architectures for large-scale audio classification,’ in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 131–135.
- [70] J. F. Gemmeke, D. P. W. Ellis, D. Freedman *et al.*, ‘Audio Set: An ontology and human-labeled dataset for audio events,’ in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 776–780.

- [71] K. Simonyan and A. Zisserman, ‘Very Deep Convolutional Networks for Large-Scale Image Recognition,’ in *ICLR 2015*, San Diego, CA, USA, May 2015.
- [72] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang and M. D. Plumbley, ‘PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition,’ *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 28, pp. 2880–2894, 2020.
- [73] G. Costantini, A. Franck, C. Pike *et al.*, ‘A dataset of high-quality object-based productions,’ in *147th Audio Engineering Society International Convention*, New York, NY, USA, Oct. 2019, eBrief: 542.
- [74] *ATK for Reaper*. [Online]. Available: <https://www.ambisonictoolkit.net/documentation/reaper/> (visited on 21/03/2023).
- [75] F. De Jong, D. Driesnack, A. Mason, M. Parmentier, P. Sunna and S. Thompson, ‘European Athletics Championships: Lessons from a Live, HDR, HFR, UHD and Next-Generation Audio sports event,’ *BBC Research & Development White Paper*, 2019.
- [76] *Advanced sound system for programme production*, Recommendation ITU-R BS.2051-3, 2022.
- [77] ABRSM, *Making Music*. [Online]. Available: <https://gb.abrsm.org/en/making-music/4-the-statistics/> (visited on 12/04/2023).
- [78] Statista, *Musicians in the UK 2022*. [Online]. Available: <https://www.statista.com/statistics/319278/number-of-musicians-in-the-uk/> (visited on 12/04/2023).
- [79] J. L. Fleiss, ‘Measuring nominal scale agreement among many raters,’ *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.
- [80] J. Sim and C. C. Wright, ‘The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements,’ *Physical Therapy*, vol. 85, no. 3, pp. 257–268, Mar. 2005.
- [81] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*. MIT Press, 2016.
- [82] Kaggle, *GTZAN music/speech collection*. [Online]. Available: <https://www.kaggle.com/datasets/lncalo/gtzan-musicspeech-collection> (visited on 07/03/2023).

- [83] BBC, *BBC Sound Effects*. [Online]. Available: <https://sound-effects.bbcrewind.co.uk/> (visited on 07/03/2023).
- [84] K. J. Grimm, G. L. Mazza and P. Davoudzadeh, 'Model Selection in Finite Mixture Models: A k-Fold Cross-Validation Approach,' *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 24, no. 2, pp. 246–256, Mar. 2017.
- [85] D. Bogdanov, N. Wack, E. Gómez Gutiérrez *et al.*, 'Essentia: An audio analysis library for music information retrieval,' in *International Society for Music Information Retrieval Conference (ISMIR 13)*, 2013.
- [86] B. McFee, C. Raffel, D. Liang *et al.*, 'Librosa: Audio and Music Signal Analysis in Python,' Austin, Texas, 2015, pp. 18–24.
- [87] Julius O. Smith, *Digital Audio Resampling Home Page*. [Online]. Available: <https://ccrma.stanford.edu/~jos/resample/> (visited on 27/03/2023).
- [88] A. Paszke, S. Gross, F. Massa *et al.*, 'PyTorch: An Imperative Style, High-Performance Deep Learning Library,' in *33rd Conference on Neural Information Processing Systems*, Vancouver, Canada, Dec. 2019.
- [89] C. R. Harris, K. J. Millman, S. J. van der Walt *et al.*, 'Array programming with NumPy,' *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020.
- [90] D. Phan, N. Pradhan and M. Jankowiak, *Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro*, arXiv:1912.11554, Dec. 2019. [Online]. Available: <http://arxiv.org/abs/1912.11554> (visited on 27/03/2023).
- [91] E. Bingham, J. P. Chen, M. Jankowiak *et al.*, 'Pyro: Deep Universal Probabilistic Programming,' *Journal of Machine Learning Research*, vol. 20, no. 28, pp. 1–6, 2019.
- [92] Z.-H. Zhou, 'Dimensionality Reduction and Metric Learning,' in *Machine Learning*, Z.-H. Zhou, Ed., Singapore: Springer, 2021, pp. 241–264. DOI: 10.1007/978-981-15-1967-3\_10.
- [93] European Broadcasting Union, 'EBU ADM Guidelines,' [Online]. Available: <https://adm.ebu.io/> (visited on 15/12/2022).



- [94] C. Simon, M. Torcoli and J. Paulus, ‘MPEG-H Audio for Improving Accessibility in Broadcasting and Streaming,’ *Fraunhofer White Paper*, 2019.
- [95] V. M. Bove, ‘Multimedia based on object models: Some whys and hows,’ *IBM Systems Journal*, vol. 35, no. 3-4, pp. 337–348, 1996, Publisher: IBM Corporation.
- [96] A. G. Westner, ‘Object-based audio capture : Separating acoustically-mixed sounds,’ Thesis, Massachusetts Institute of Technology, 1999.
- [97] I. Jang, K. Kang, T. Lee and G. Y. Park, ‘An Object-based 3D Audio Broadcasting System for Interactive Services,’ in *118th Audio Engineering Society International Convention*, Barcelona, Spain, May 2005, Paper Number: 6384.
- [98] S. Füg, A. Hölzer, C. Borß, C. Ertel, M. Kratschmer and J. Plogsties, ‘Design, Coding and Processing of Metadata for Object-Based Interactive Audio,’ in *137th Audio Engineering Society International Convention*, Los Angeles, CA, USA, Oct. 2014, Paper Number: 9097.
- [99] Q. Liu, W. Wang, P. J. Jackson and T. J. Cox, ‘A source separation evaluation method in object-based spatial audio,’ in *23rd European Signal Processing Conference, EUSIPCO*, Publisher: Institute of Electrical and Electronics Engineers Inc. ISBN: 9780992862633, Nice, France, Dec. 2015, pp. 1088–1092.
- [100] A. Silzle, M. Weitnauer, O. Warusfel *et al.*, ‘ORPHEUS Audio Project: Piloting an End-to-end Object Based Audio Broadcasting Chain,’ Amsterdam, Netherlands, Sep. 2017.
- [101] A. Roginska and P. Geluso, Eds., *Immersive sound: The art and science of binaural and multi-channel audio*. Routledge, 2017.
- [102] P. Coleman, A. Franck, J. Francombe *et al.*, ‘An Audio-Visual System for Object-Based Audio: From Recording to Listening,’ *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 1919–1931, Aug. 2018, ISSN: 15209210.
- [103] W. Davies and S. Spors, ‘Guest editors’ note : Special issue on object-based audio,’ *Journal of the Audio Engineering Society*, vol. 67, no. 7/8, pp. 484–485, Aug. 2019.

- [104] J. Breebaart, G. Cengarle, L. Lu, T. Mateos, H. Purnhagen and N. Tsingos, ‘Spatial Coding of Complex Object-Based Program Material,’ *Journal of the Audio Engineering Society*, vol. 67, no. 7/8, pp. 486–497, Aug. 2019.
- [105] J. Francombe, J. Woodcock, R. Hughes *et al.*, ‘A System Architecture for Semantically Informed Rendering of Object-Based Audio,’ *Journal of the Audio Engineering Society*, vol. 67, no. 7/8, pp. 498–509, Aug. 2019.
- [106] A. Wilson and B. M. Fazenda, ‘User-guided Rendering of Audio Objects Using an Interactive Genetic Algorithm,’ *Journal of the Audio Engineering Society*, vol. 67, no. 7/8, pp. 522–530, Aug. 2019.
- [107] M. C. Heilemann, D. A. Anderson and M. F. Bocko, ‘Near-Field Object-Based Audio Rendering on Flat-Panel Displays,’ *Journal of the Audio Engineering Society*, vol. 67, no. 7/8, pp. 531–539, Aug. 2019.
- [108] H. Kon and H. Koike, ‘Estimation of Late Reverberation Characteristics from a Single Two-Dimensional Environmental Image Using Convolutional Neural Networks,’ *Journal of the Audio Engineering Society*, vol. 67, no. 7/8, pp. 540–548, Aug. 2019.
- [109] D. Menzies and F. M. Fazi, ‘Multichannel Compensated Amplitude Panning, An Adaptive Object-Based Reproduction Method,’ *Journal of the Audio Engineering Society*, vol. 67, no. 7/8, pp. 549–556, Aug. 2019.
- [110] J. Woodcock, W. J. Davies and T. J. Cox, ‘Influence of Visual Stimuli on Perceptual Attributes of Spatial Audio,’ *Journal of the Audio Engineering Society*, vol. 67, no. 7/8, pp. 557–567, Aug. 2019.
- [111] L. Ward, ‘Satisfy the appetite for personalized media experiences with object-based production,’ *EBU tech-i magazine*, vol. 51, pp. 6–6, Mar. 2022.
- [112] A. Churnside, M. Melchior, M. Armstrong, M. Shotton and M. Brooks, ‘Object-based broadcasting - curation, responsiveness and user experience,’ in *International Broadcasting Convention (IBC) 2014 Conference*, Amsterdam, Netherlands, Sep. 2014.
- [113] S. Elo and H. Kyngäs, ‘The qualitative content analysis process,’ *Journal of Advanced Nursing*, vol. 62, no. 1, pp. 107–115, Apr. 2008.

- [114] A. Assarroudi, F. Heshmati Nabavi, M. R. Armat, A. Ebadi and M. Vaismoradi, ‘Directed qualitative content analysis: The description and elaboration of its underpinning methods and data analysis process,’ *Journal of Research in Nursing: JRN*, vol. 23, no. 1, pp. 42–55, Feb. 2018.
- [115] BBC R&D and IRT, *EAR Production Suite*. [Online]. Available: <https://ear-production-suite.ebu.io/> (visited on 23/03/2023).
- [116] T. Carpentier, ‘A new implementation of Spat in Max,’ in *15th Sound & Music Computing Conference (SMC)*, Limassol, Cyprus, 2018, pp. 184–191.
- [117] A. Franck and F. M. Fazi, ‘VISR-A versatile open software framework for audio signal processing,’ in *AES International Conference on Spatial Reproduction*, Tokyo, Japan, Aug. 2018, pp. 278–288.
- [118] IEM, *IEM Plug-in Suite*. [Online]. Available: <https://plugins.iem.at/> (visited on 23/03/2023).

# Appendix A

## Object Tables

The objects in each of the nine scenes are listed in the tables below. The objects from Protest and the two Vostok-K scenes can be seen in Table A.1, those from the three Casualty scenes are shown in Table A.2, and the objects for the two Penguins extracts and the Football scene are in Table A.3.

<b>Protest</b>	<b>Vostok-K Extract 1</b>	<b>Vostok-K Extract 2</b>
Neil Dialogue	Joe Dialogue	Joe Dialogue
Abby Dialogue	Tatiana Dialogue	Tatiana Dialogue
Abby Feet	General Dialogue	Tape Recorder
Give Us Our Money	Tape Recorder	Fuel Alarm
Scum	Sam Dialogue	Fuel Warning
Power2Ppl	Sam Radio	Warning Beeps
Megaphone Voice	Broken Tatiana Dialogue	Joe Cockpit Turbine
We Want Our Cash	Broken Tatiana Radio	Rumble
Bang	Broken Tatiana Radio 2	Rain
Smash	Switches	Plane 1
Knock	Joe Cockpit Turbine	Plane 2
Bank Door Opens	Rumble	Plane 3
Horse Hooves	Thunder	Planes passing
Crowd inside	Rain	Bullet Ricochets
Crowd inside Background	Plane 1	Bullet Ricochets 2
Crowd inside Foreground	Plane 2	Explosion
Crowd transition	Plane 3	SubBoom
Crowd + Cars	Bullet Ricochets 1	SubBoom 2
Crowd Outside	Bullet Ricochets 2	Impact
Crowd + Drums	Bullet Ricochets 3	Impact 2
Drums	Explosion	Music
Crowd outside Background	Gun Fire	
Crowd outside Foreground	Sub Boom	
Scum Crowd 1	Impact	
Scum Crowd 2	Rocket	
Lowfi Boom	Rattle	
External Atmos	Music	

**Table A.1:** A table showing all of the object names for Protest and the two Vostok-K scenes.

Casualty Scene 4	Casualty Scene 46	Casualty Scene 49
Ruby Dialogue	Ruby Dialogue	Will Dialogue
Dani Dialogue	Dani Dialogue	Archie Dialogue
Iain Dialogue	Ruby Footsteps 1	Ciaran Dialogue
WasSheEvenAParamedic	Ruby Footsteps 2	Phone ring
WhoWasThat	Ruby Walking Towards	Lift
WhatIsSheDoing	Ruby Walking Away	Computer beeps
WheresSheGoing	Music	Mouse click
Paramedic Footsteps1	Trolley	Movement 1
Paramedic Footsteps2	Phone	Movement 2
Dani Run	Extra Footsteps 1	Atmos 1
Bag Search	Extra Footsteps 2	Atmos 2
Bag Grab	Bkgd Footsteps 1	Atmos 3
Suction Machine movement	Bkgd Footsteps 2	Atmos 4
Movement	Bkgd Footsteps 3	Atmos 5
Penknife open+close	Atmos1	Atmos 6
Penknife movement	Atmos2	Ciaran Footsteps
Siren	Atmos3	Will Footsteps
Engine 1		Extra Footsteps
Engine 2		Rustling
Ambulance door		Movement
Ambulance door 2		
Breathing		
Chatter		
Traffic1		
Traffic2		
Atmos Footsteps		
Atmos		

**Table A.2:** A table showing all of the object names in each of the Casualty scenes.

Penguins Extract 1	Penguins Extract 2	Football
Music	Music	Commentary
Narrator	Narrator	Crowd Comms Mic
Wind 1	Water 1	Crowd atmos 1
Wind 2	Water 2	Crowd atmos 2
Wind 3	Water 3	Crowd Pitch Mic
Wind 4	Water 4	Crowd Cheer Pitch Mic
Wind 5	Water 5	Crowd Cheer Comms Mic
Water 1	Waves Crashing	Player shout
Water 2	Splashing	Kick Enhancement 1
Water 3	Splash	Kick Enhancement 2
Water 4	Splashes	Kick Enhancement 3
Water 5	Rustling	
Footsteps 1	Movement 1	
Footsteps 2	Movement 2	
Footsteps 3	Movement 3	
Footsteps 4	Movement 4	
Microphone noise	Sploshing	
Penguin chitter	Bird Calls	
Penguin honks	Penguins 1	
Penguin tweets	Penguins 2	
Penguin tweet and chitter	Penguins 3	
Penguin single honk	Penguins 4	
Penguin chirp	Honks	
	Seal Grunts1	
	Seal Grunts 2	

**Table A.3:** A table showing all of the object names in the two Penguins scenes and Football.

# Appendix B

## Object Creation Tables

The tables included in this appendix show the work done to adapt the two publicly available content pieces, ‘Protest’ and ‘The Vostok-K Incident’. These tables are designed to give the reader more idea of the level of work done. The open source assets have been chosen since these are the assets the reader will be able to access the original stems for.

Both of the tables show the original track names in the leftmost column. The next columns show the objects that came from the respective tracks. Many tracks were split into multiple objects - mainly because there were different sounds at different times along the track, though there was one case where the left and right channels of a stereo track contained different sounds at the same time. There were also cases where several tracks were combined into a single object, this can be seen particularly in the ‘Protest’ example since the spatial nature of the content meant that sounds happened on several tracks at the same time.

Table B.1 also shows the ‘ground truth’ for the objects taken from the work done in [17]. A reminder to the reader that 0 signifies Low Importance, 1 is Medium, 2 is High, and 3 is Essential Importance. Where two options are given this is because the objects in this work differed slightly from the objects used in [17]. As a result the objects in this work were an amalgamation of two differently assigned ground truth objects.



Original	Object	Ground Truth
Neil	‘Neil Dialogue’	3
Abby	‘Abby Dialogue’	3
Dialogue 1	‘Give Us Our Money’	2
	‘Power2Ppl’	2
Dialogue 2	‘Megaphone Voice’	1
	‘Power2Ppl’	2
Dialogue 3 Abby Feet	‘Abby Feet’	1
Dialogue 4	‘Scum’	2
	‘Power2Ppl’	2
	‘We Want Our Cash’	2
FX 1	‘Bank Door Opens’	2
	‘Horse Hooves’	1
	‘Crowd + Cars’	1
FX 2	‘Bank Door Opens’	2
	‘Horse Hooves’	1
	‘Crowd Outside’	1
FX 4	‘Bank Door Opens’	2
	‘Smash’	2
	‘Knock’	2
	‘Drums’	2
	‘Crowd transition’	2
FX 5	‘Bang’	2
	‘Smash’	2
	‘Crowd + Cars’	1
Music	‘Lowfi Boom’	1
Crowd 1	‘Crowd inside’	1 or 0
	‘Crowd + Drums’	0
Crowd 3	‘Scum Crowd 1’	1
	‘Crowd inside’	1 or 0
Crowd 4	‘Scum Crowd 2’	1
	‘Crowd inside’	1 or 0
Crowd BG 1-9	‘Crowd inside Background’	0
	‘Crowd outside Background’	0
Crowd FG 1-9	‘Crowd inside Foreground’	1 or 0
	‘Crowd outside Foreground’	1 or 0
Atmos 1-8	‘External Atmos’	0

**Table B.1:** A table showing the creation of objects for the Protest scene. The first column shows the original track names, the second column shows the objects that came from the original tracks. Some objects were a down mix of several tracks, similarly some tracks were separated into multiple objects. The final column shows the ‘ground truth’.

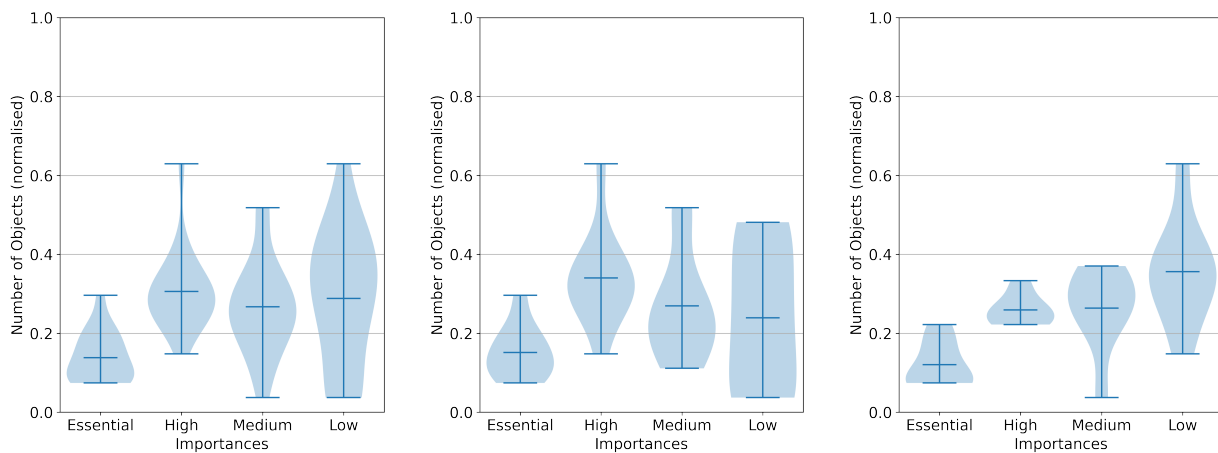
<b>Original</b>	<b>Extract 1</b>	<b>Extract 2</b>
Joe	‘Joe Dialogue’	‘Joe Dialogue’
TatianaGeneralFX	‘Tatiana Dialogue’ ‘General Dialogue’	‘Tatiana Dialogue’
NagraFX	‘Tape Recorder’	‘Tape Recorder’
Sam	‘Sam Dialogue’	
SamRadioFX	‘Sam Radio’ ‘Broken Tatiana Radio 2’	
BrokenTatiana	‘Broken Tatiana Dialogue’	
BrokenTatianaRadioFX	‘Broken Tatiana Radio’	
Switches	‘Switches’	‘Fuel Alarm’ (left channel) ‘Fuel Warning’ (right channel)
SwitchesMiddle		‘Warning Beeps’
JoeCockpitTurbine	‘Joe Cockpit Turbine’	‘Joe Cockpit Turbine’
RumbleFX	‘Rumble’	‘Rumble’
Thunder	‘Thunder’ ‘Bullet Ricochets 1’	‘Thunder’
Rain	‘Rain’	‘Rain’
PlaneExtFX1	‘Plane 1’	‘Plane 1’ ‘Planes passing’
PlaneExtFX2	‘Plane 2’	‘Plane 2’
PlaneExtFX3	‘Plane 3’	‘Plane 3’
HitFX1	‘Bullet Ricochets 1’ ‘Bullet Ricochets 2’ ‘Bullet Ricochets 3’	‘Bullet Ricochets’ ‘Bullet Ricochets 2’ ‘Planes passing’
HitFX2	‘Bullet Ricochets 1’ ‘Bullet Ricochets 2’ ‘Bullet Ricochets 3’	‘Bullet Ricochets’ ‘Bullet Ricochets 2’
HitFX3	‘Bullet Ricochets 1’ ‘Bullet Ricochets 2’ ‘Bullet Ricochets 3’	‘Bullet Ricochets’ ‘Bullet Ricochets 2’ ‘Impact 2’
HitFXMiddle	‘Bullet Ricochets 1’ ‘Bullet Ricochets 2’ ‘Bullet Ricochets 3’	‘Bullet Ricochets’
FXExplosion	‘Explosion’ ‘Bullet Ricochets 1’ ‘Gun Fire’	‘Explosion’ ‘Bullet Ricochets 2’
FXSubBoom	‘SubBoom’ ‘Bullet Ricochets 1’	‘SubBoom’ ‘Bullet Ricochets 2’
FXImpact	‘Impact’ ‘Rocket’ ‘Rattle’	‘Impact’ ‘Bullet Ricochets’ ‘Bullet Ricochets 2’
Music	‘Music’	‘Music’

**Table B.2:** A table showing the creation of objects for the two Vostok-K scenes. A blank cell indicates that the track wasn’t included in that scene. Tracks that didn’t feature in either scene have been omitted from the table.

# Appendix C

## Violin Plots

In Section 5.6, three violin plots were presented for Vostok-K Scene 1. This appendix contains the violin plots for the 8 remaining content pieces. The plots are normalised between 0 and 1 for ease of comparison. The number of objects and participants in each group is given in the figure captions.

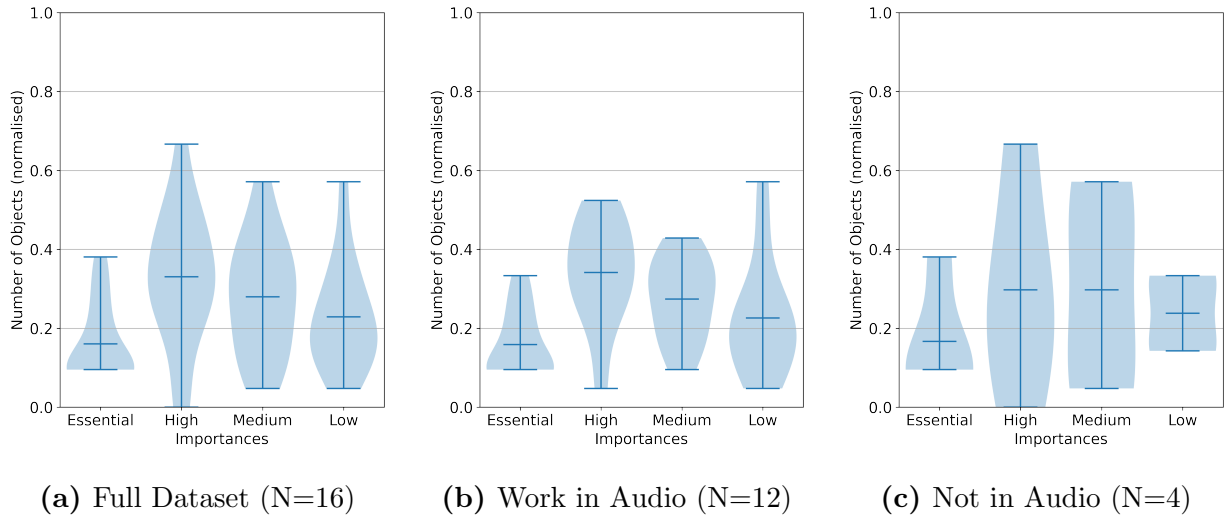


(a) Full Dataset (N=19)

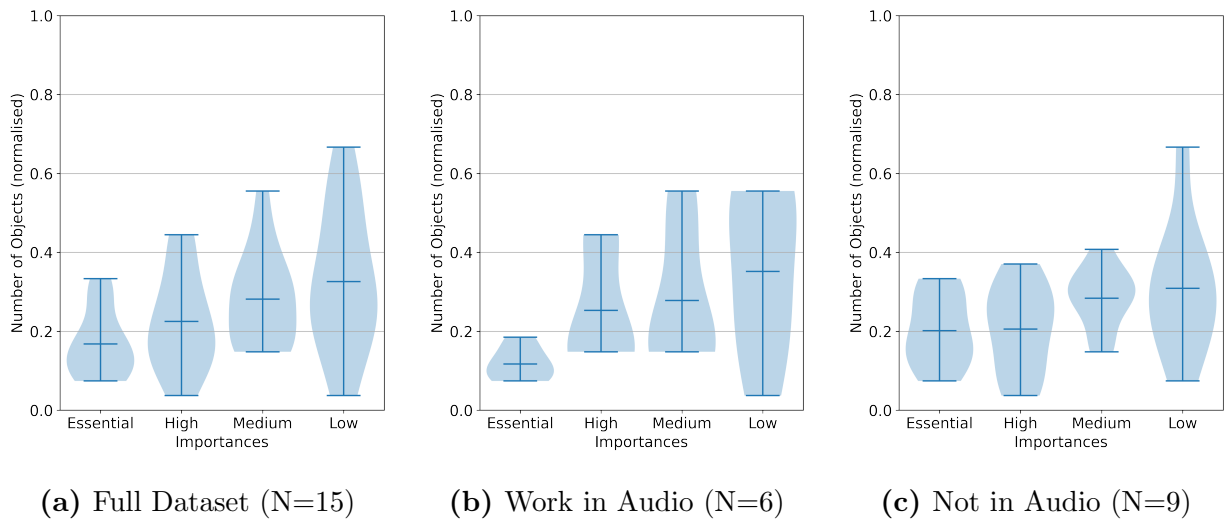
(b) Work in Audio (N=11)

(c) Not in Audio (N=8)

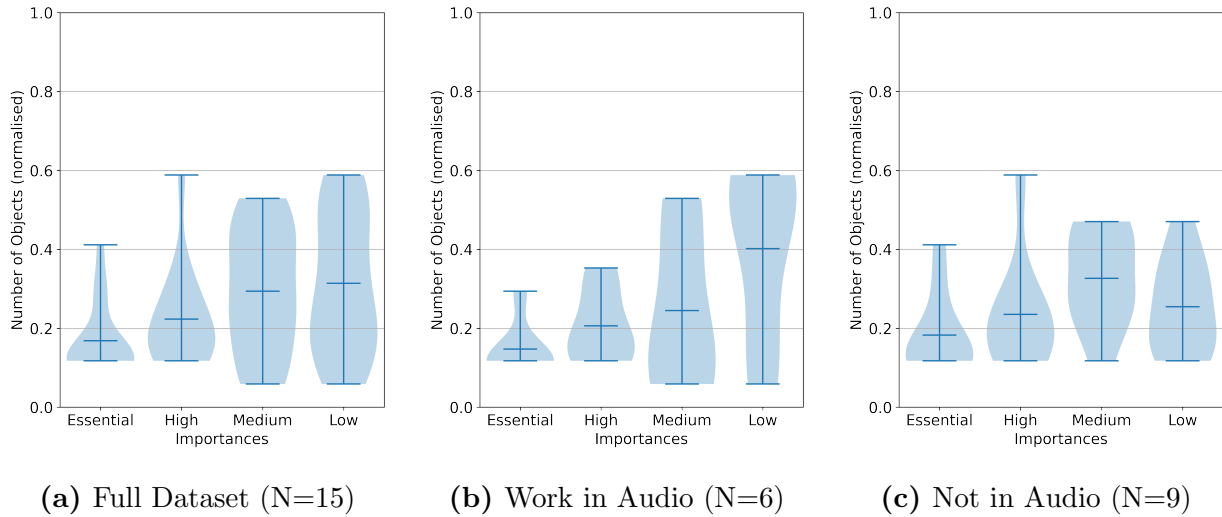
**Figure C.1:** Three violin plots showing the distributions of the number of objects (normalised to between 0 and 1) each participant assigned to each importance level for Protest. Protest contained 27 objects.



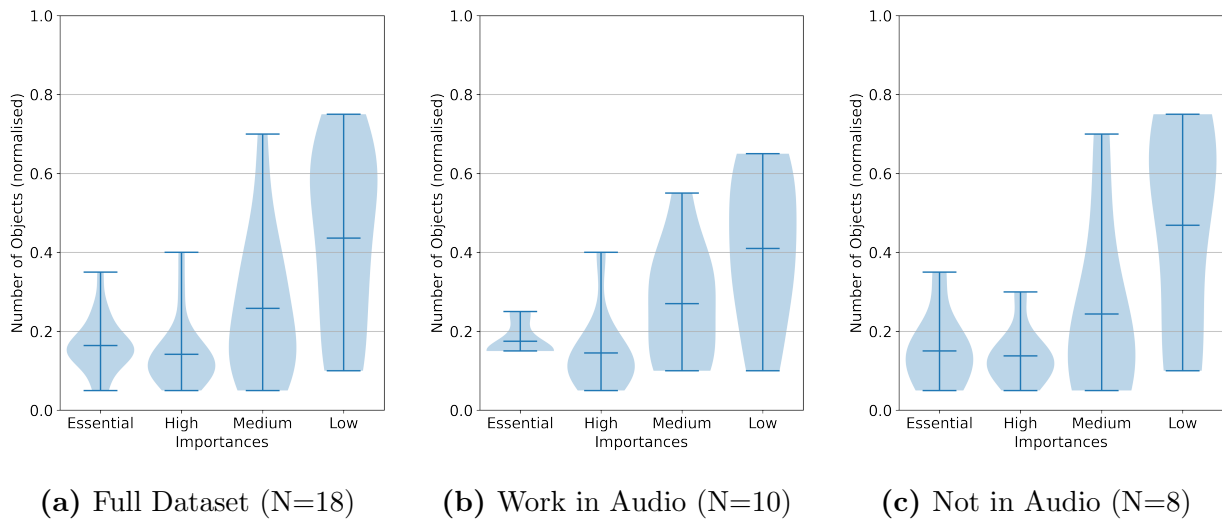
**Figure C.2:** Three violin plots showing the distributions of the number of objects (normalised to between 0 and 1) each participant assigned to each importance level for Vostok-K Scene 2. Vostok-K Scene 2 contained 21 objects.



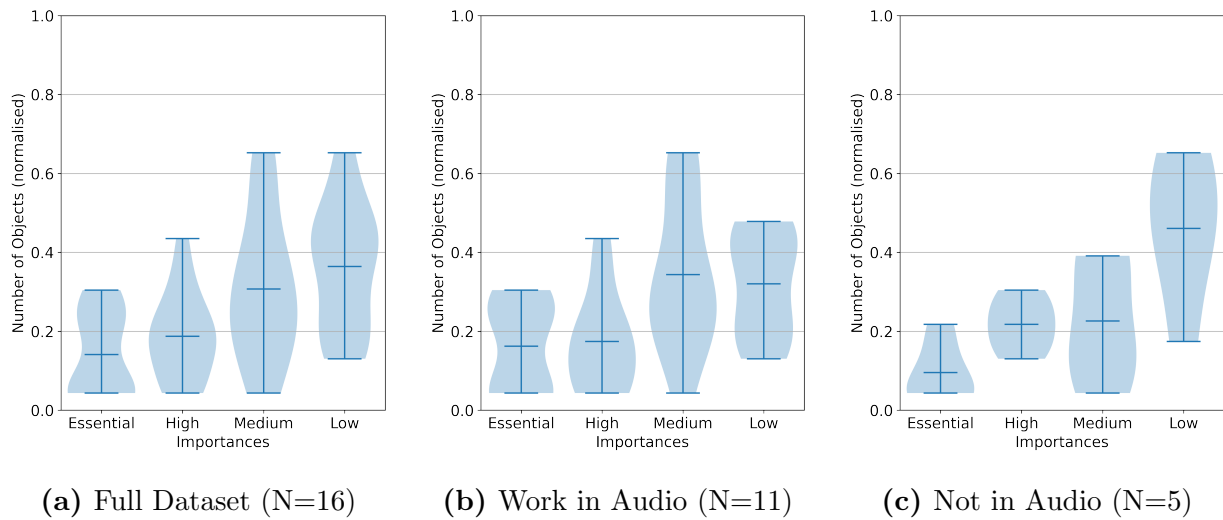
**Figure C.3:** Three violin plots showing the distributions of the number of objects (normalised to between 0 and 1) each participant assigned to each importance level for Casualty Scene 4. Casualty Scene 4 contained 27 objects.



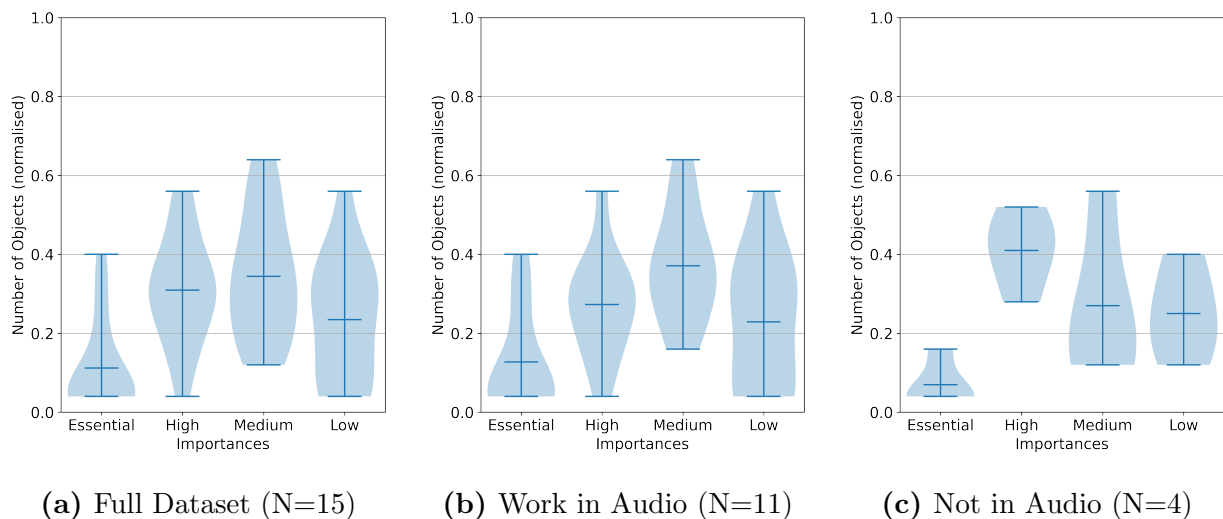
**Figure C.4:** Three violin plots showing the distributions of the number of objects (normalised to between 0 and 1) each participant assigned to each importance level for Casualty Scene 46. Casualty Scene 46 contained 17 objects.



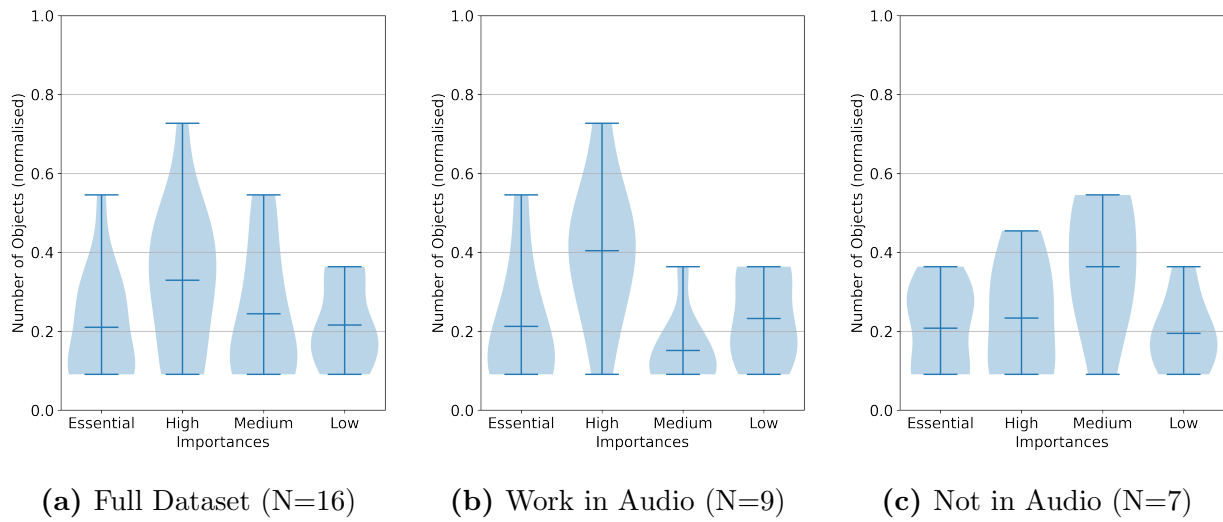
**Figure C.5:** Three violin plots showing the distributions of the number of objects (normalised to between 0 and 1) each participant assigned to each importance level for Casualty Scene 49. Casualty Scene 49 contained 20 objects.



**Figure C.6:** Three violin plots showing the distributions of the number of objects (normalised to between 0 and 1) each participant assigned to each importance level for Penguins Opening Credits. Penguins Opening Credits contained 23 objects.



**Figure C.7:** Three violin plots showing the distributions of the number of objects (normalised to between 0 and 1) each participant assigned to each importance level for Penguins Scene 1. Penguins Scene 1 contained 25 objects.



**Figure C.8:** Three violin plots showing the distributions of the number of objects (normalised to between 0 and 1) each participant assigned to each importance level for Football. Football contained 11 objects.

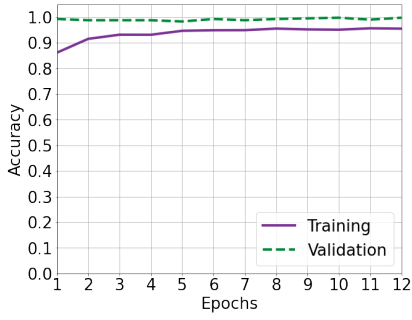
# Appendix D

## Speech/Music/SFX Classification Model Parameter Sweep

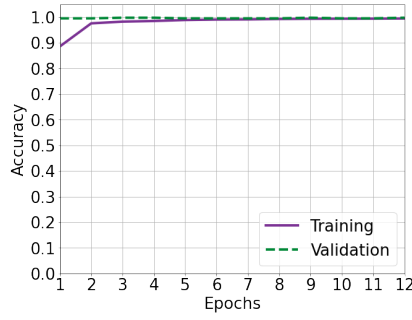
A small parameter sweep was run over the batch size and learning rate to find optimal values. This was kept small since the original algorithm gave good results with batch size = 16 and learning rate = 0.01 for 4 epochs. The algorithm was trained for 12 epochs with batch sizes of 16, 32, and 64 and learning rates of 0.05, 0.01, and 0.005. The resulting accuracy and loss plots are shown in Figs. D.1 and D.2. As can be seen in the graphs most of the parameter changes didn't affect the training significantly. In most cases the optimal training was achieved within 6 epochs. The original parameters of batch size = 16 and LR = 0.01 (Figs. D.1b and D.2b) gave the best results, marginally, and that increasing the number of epochs to 6 would improve the training.

Training using a GPU was investigated on another machine but appeared to give worse results and couldn't be run on the compatible keras libraries. Due to time constraints, and the initial poor results, this investigation was abandoned.

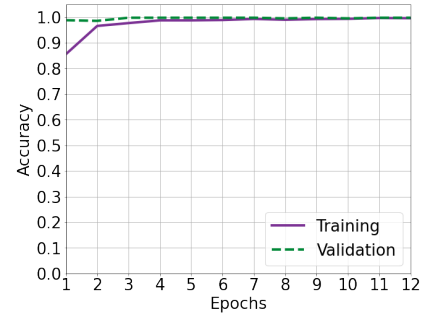




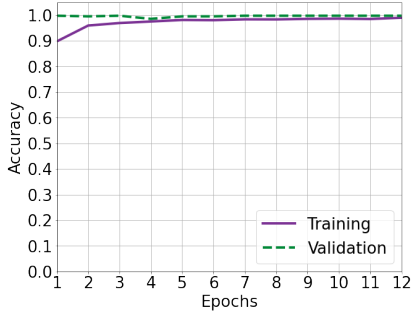
(a) Batch size = 16, LR = 0.05



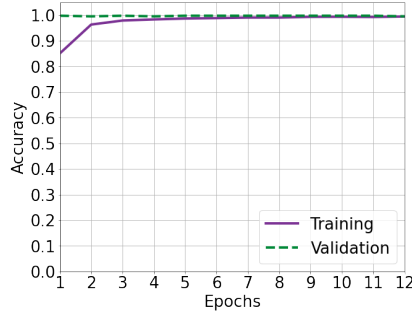
(b) Batch size = 16, LR = 0.01



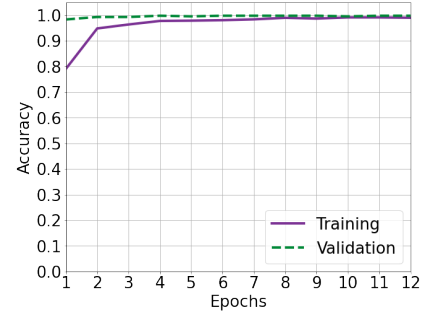
(c) Batch size = 16, LR = 0.005



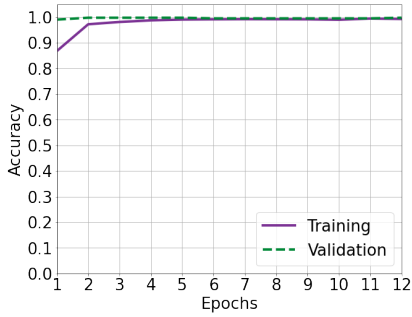
(d) Batch size = 32, LR = 0.05



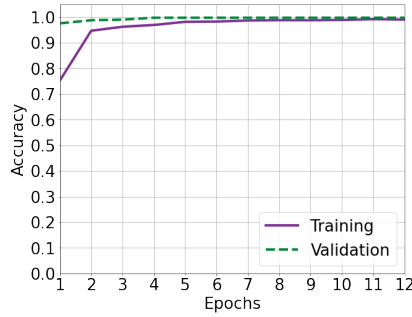
(e) Batch size = 32, LR = 0.01



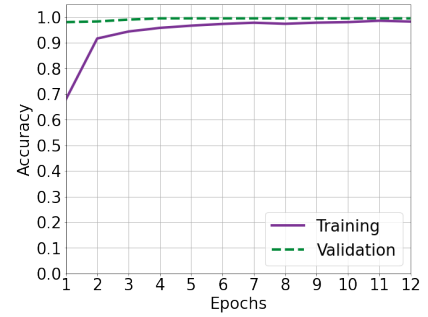
(f) Batch size = 32, LR = 0.005



(g) Batch size = 64, LR = 0.05

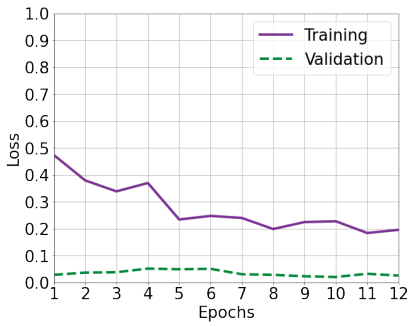


(h) Batch size = 64, LR = 0.01

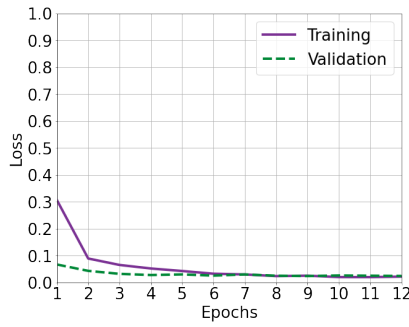


(i) Batch size = 64, LR = 0.005

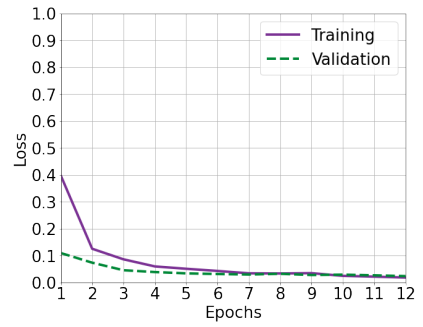
**Figure D.1:** Accuracy plots for the parameter sweep



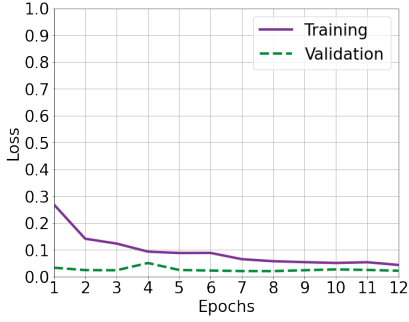
(a) Batch size = 16, LR = 0.05



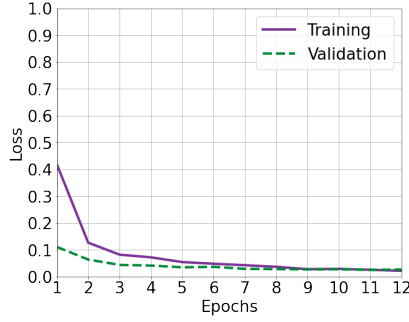
(b) Batch size = 16, LR = 0.01



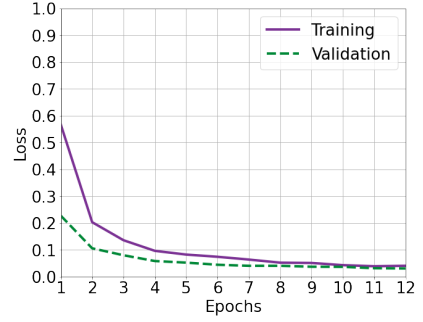
(c) Batch size = 16, LR = 0.005



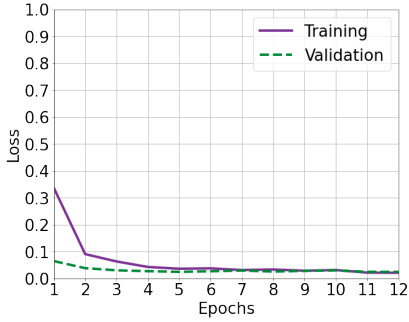
(d) Batch size = 32, LR = 0.05



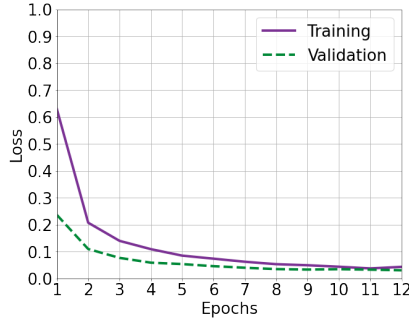
(e) Batch size = 32, LR = 0.01



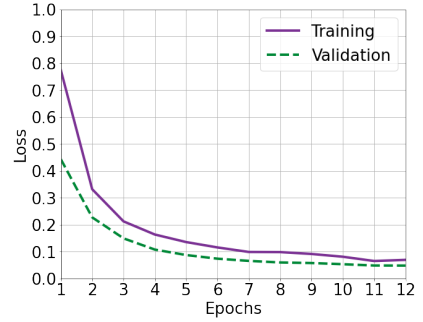
(f) Batch size = 32, LR = 0.005



(g) Batch size = 64, LR = 0.05



(h) Batch size = 64, LR = 0.01

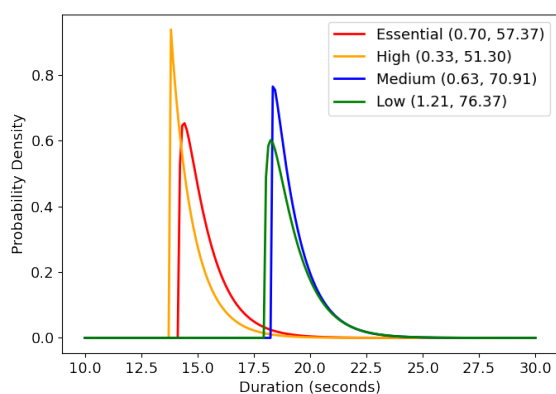


(i) Batch size = 64, LR = 0.005

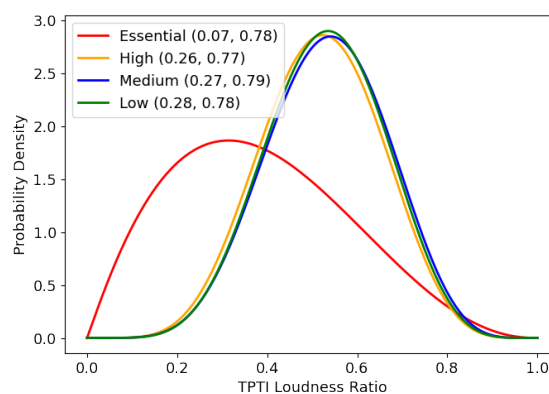
**Figure D.2:** Loss plots for the parameter sweep

# Appendix E

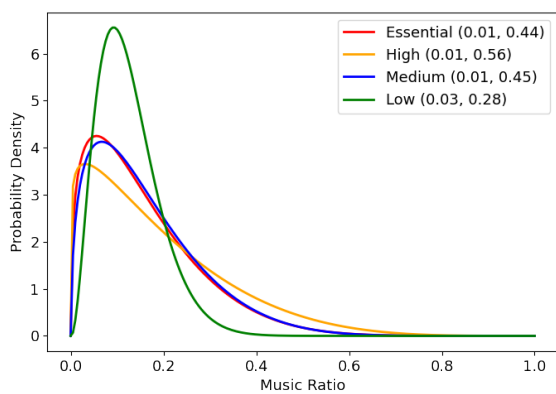
## Mixture Model Distribution Plots



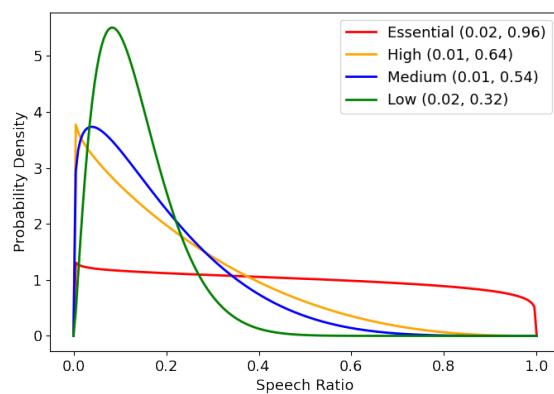
(a) Duration



(b) TPTI loudness ratio

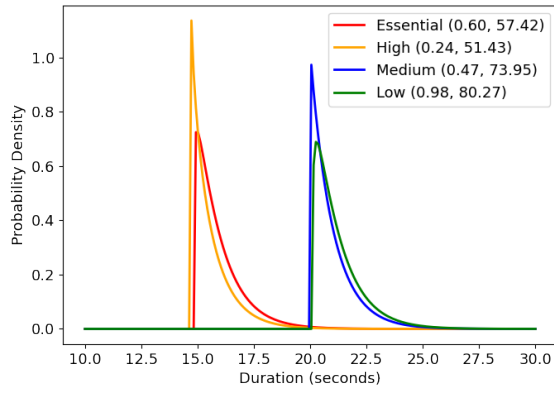


(c) Music ratio

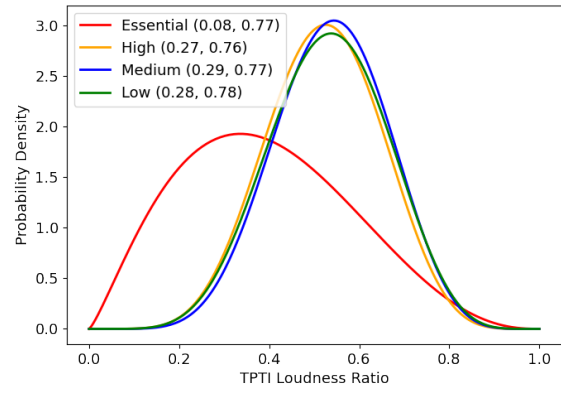


(d) Speech ratio

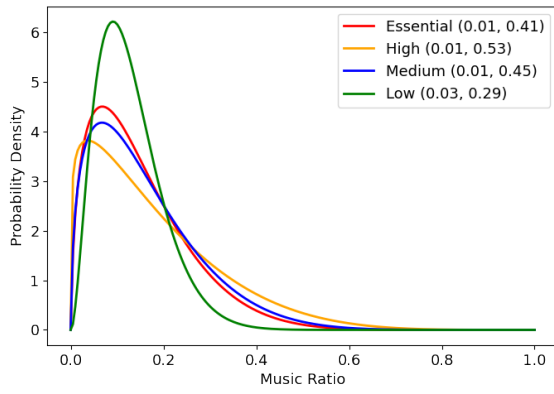
**Figure E.1:** Probability density functions when the mixture model is trained on all the data except for Vostok-K Scene 1



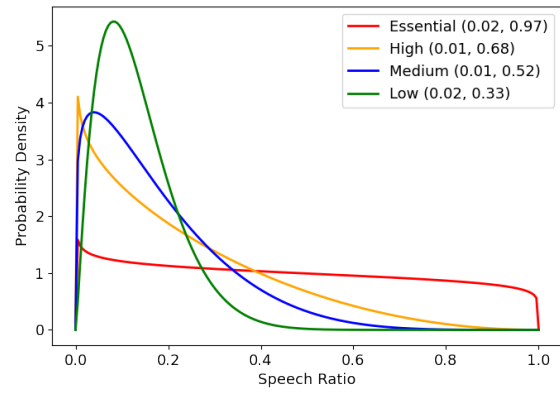
(a) Duration



(b) TPTI loudness ratio

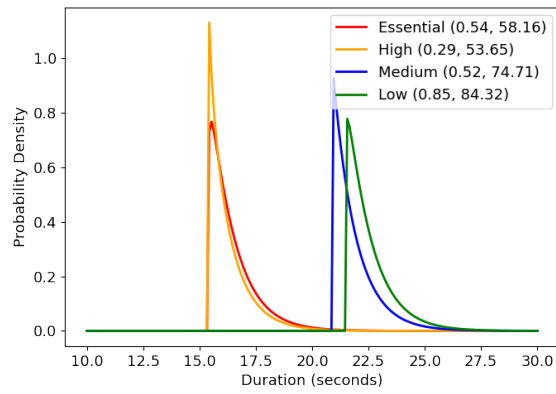


(c) Music ratio

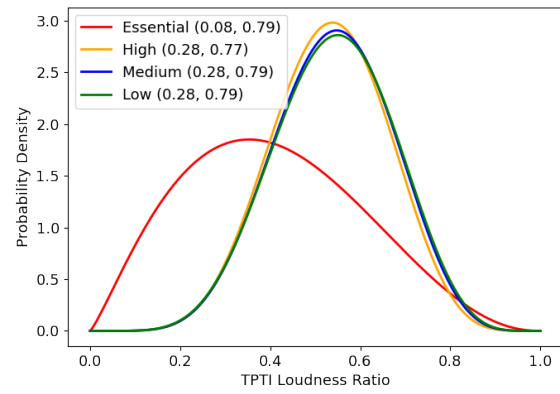


(d) Speech ratio

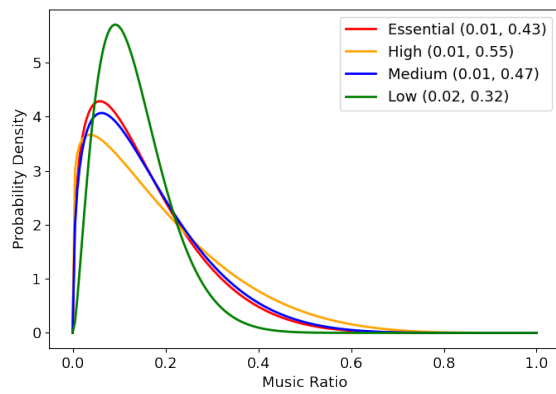
**Figure E.2:** Probability density functions when the mixture model is trained on all the data except for Vostok-K Scene 2



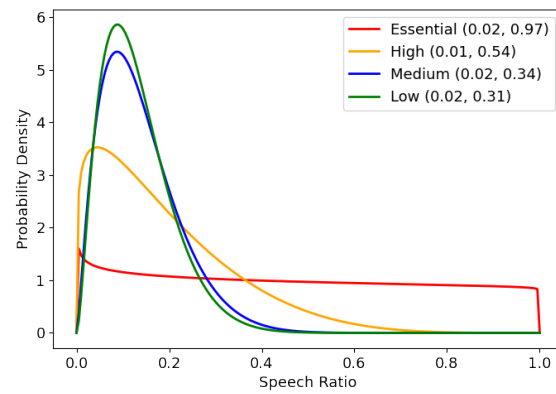
(a) Duration



(b) TPTI loudness ratio

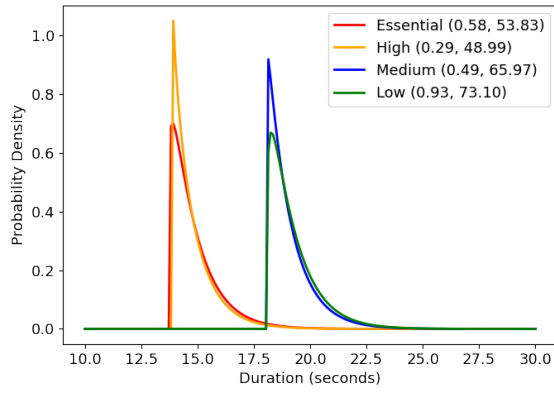


(c) Music ratio

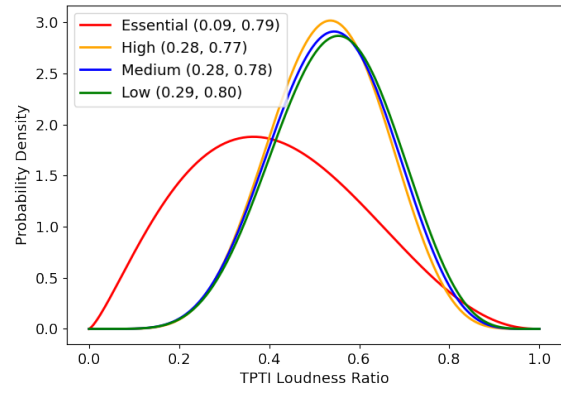


(d) Speech ratio

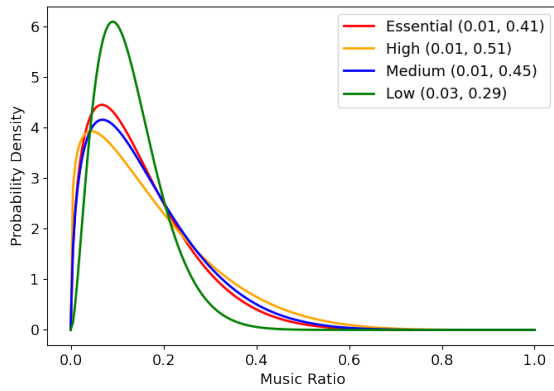
**Figure E.3:** Probability density functions when the mixture model is trained on all the data except for Casualty Scene 4



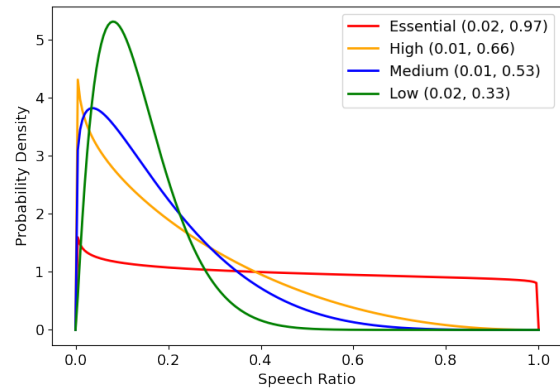
(a) Duration



(b) TPTI loudness ratio

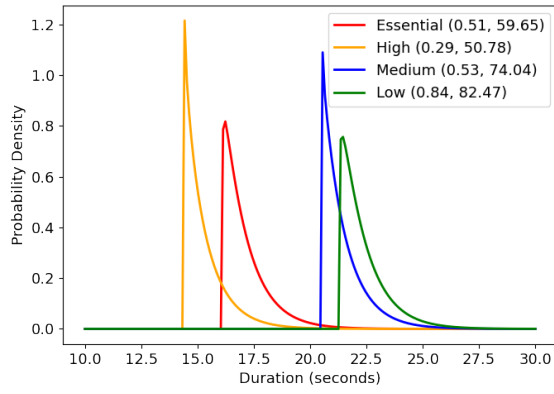


(c) Music ratio

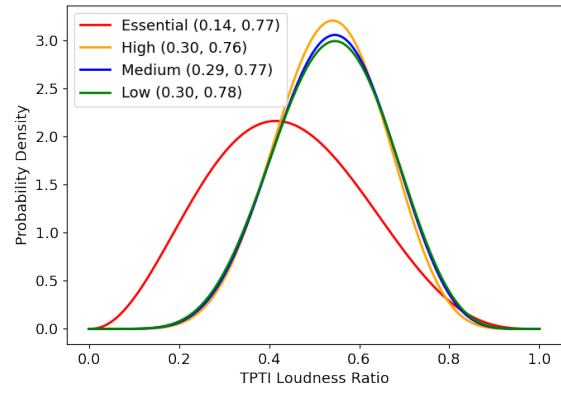


(d) Speech ratio

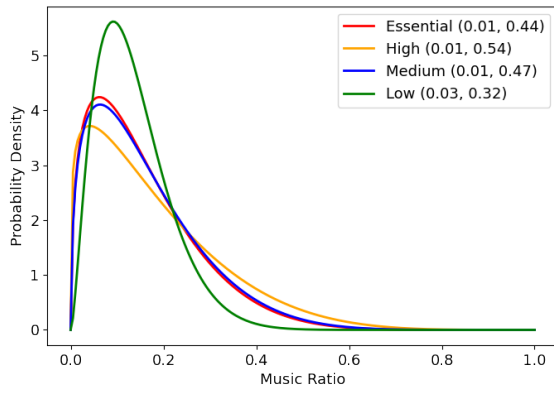
**Figure E.4:** Probability density functions when the mixture model is trained on all the data except for Casualty Scene 46



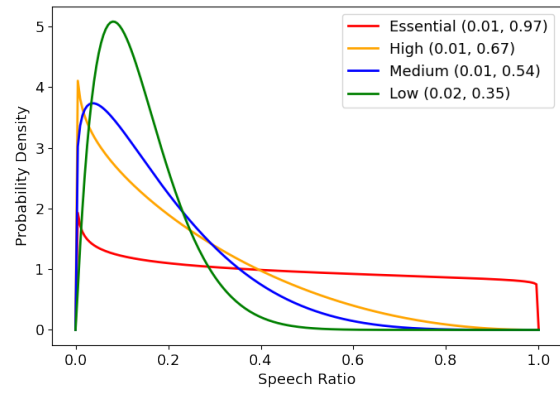
(a) Duration



(b) TPTI loudness ratio

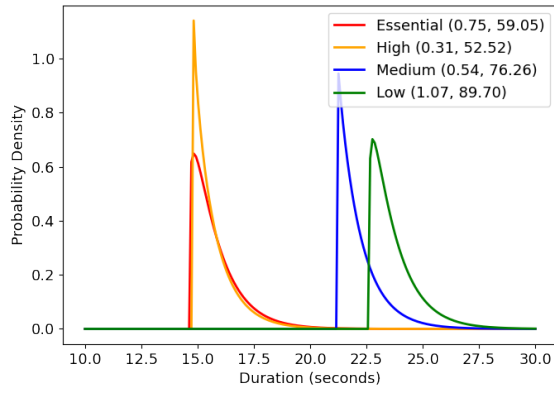


(c) Music ratio

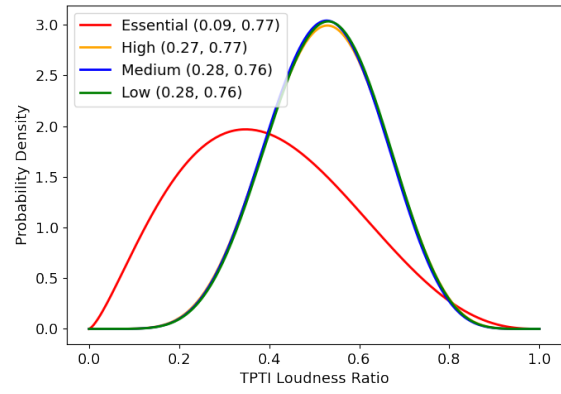


(d) Speech ratio

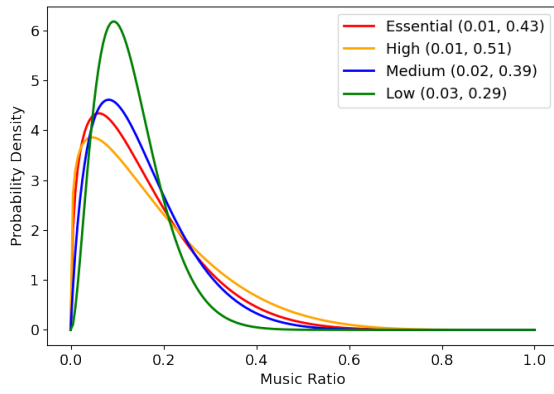
**Figure E.5:** Probability density functions when the mixture model is trained on all the data except for Casualty Scene 49



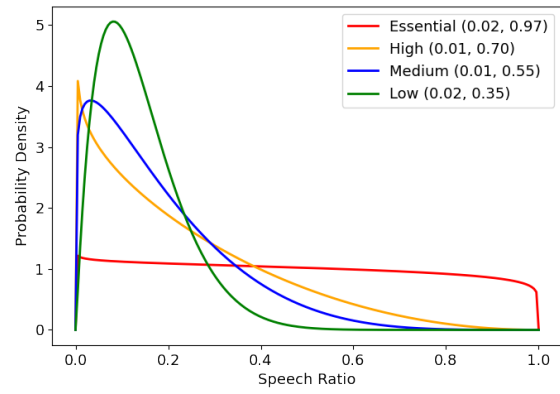
(a) Duration



(b) TPTI loudness ratio



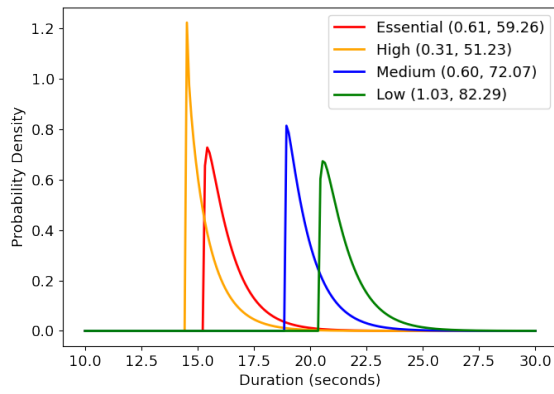
(c) Music ratio



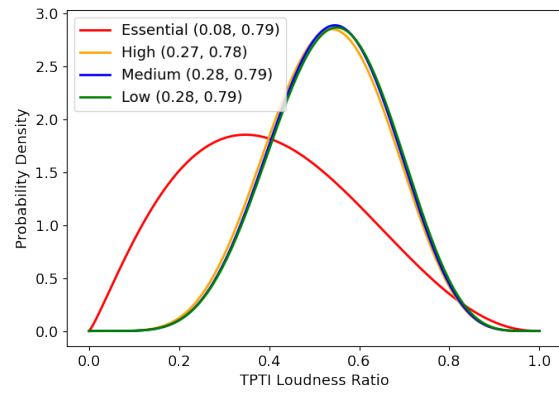
(d) Speech ratio

**Figure E.6:** Probability density functions when the mixture model is trained on all the data except for Penguins Opening Credits

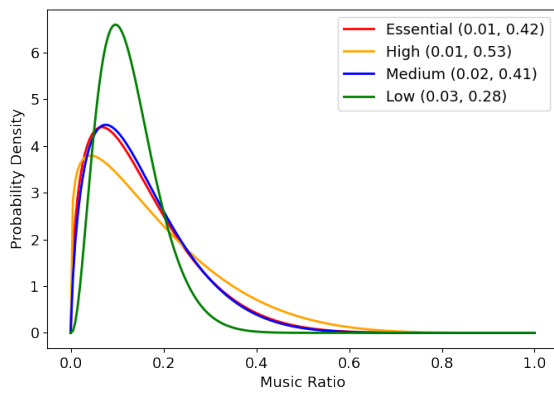




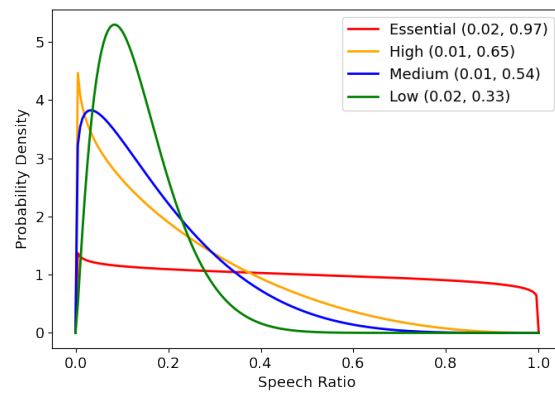
(a) Duration



(b) TPTI loudness ratio

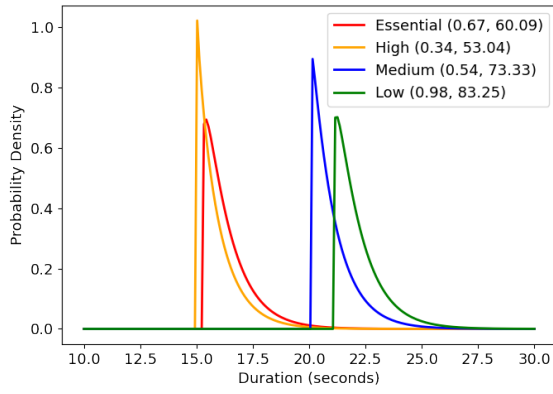


(c) Music ratio

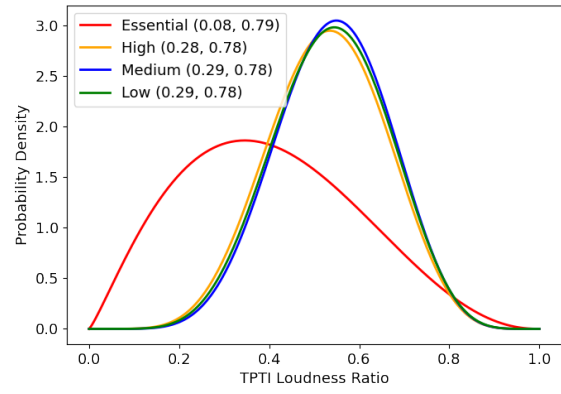


(d) Speech ratio

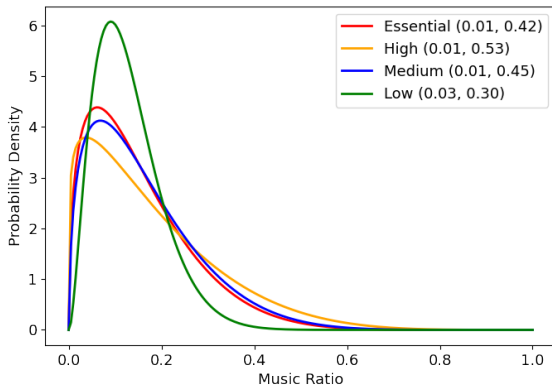
**Figure E.7:** Probability density functions when the mixture model is trained on all the data except for Penguins Scene 1



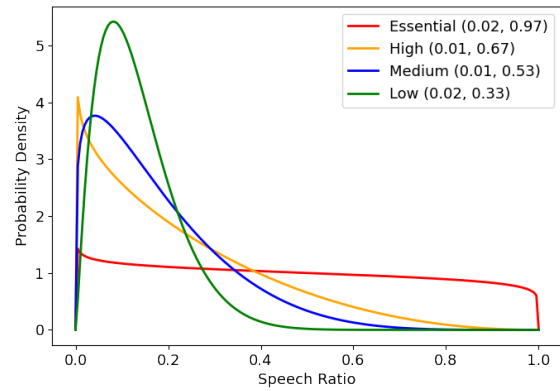
(a) Duration



(b) TPTI loudness ratio



(c) Music ratio

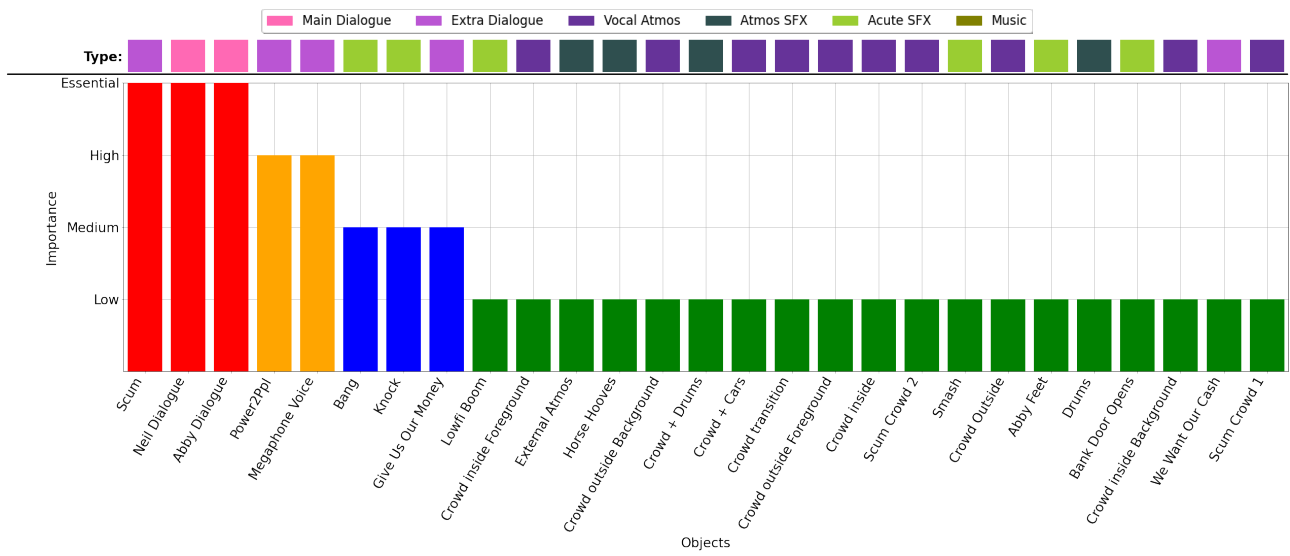


(d) Speech ratio

**Figure E.8:** Probability density functions when the mixture model is trained on all the data except for Football

# Appendix F

## Mixture Model Results



**Figure F.1:** A bar chart showing the mixture model's assignments with Protest as the test data

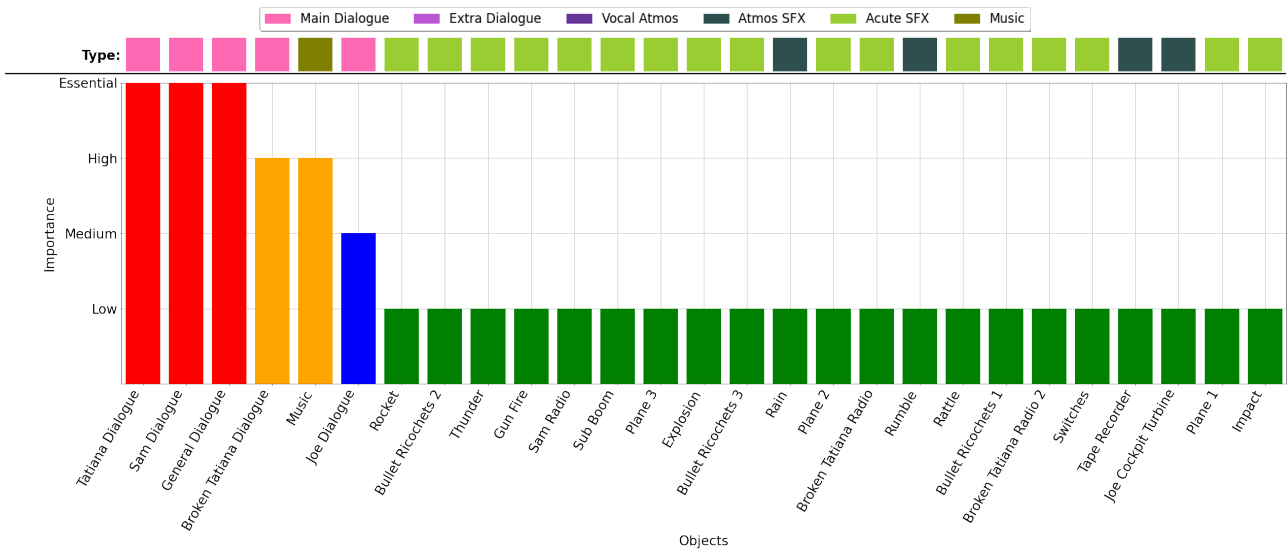


Figure F.2: A bar chart showing the mixture model’s assignments with Vostok-K Scene 1 as the test data

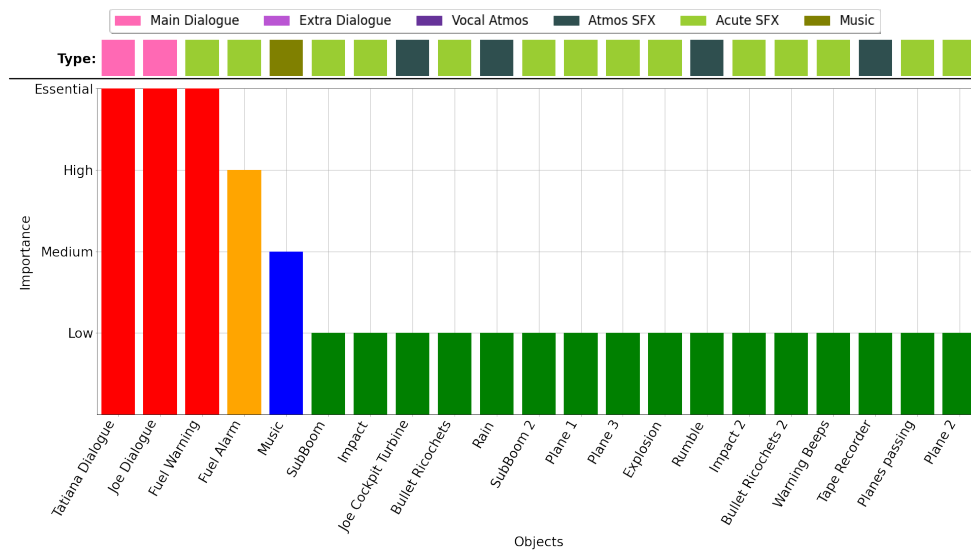
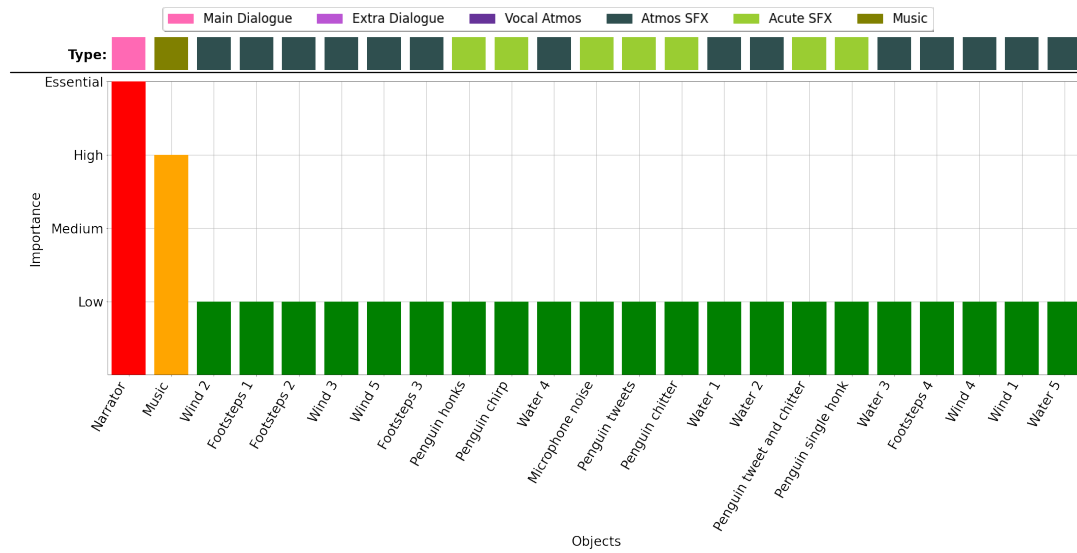
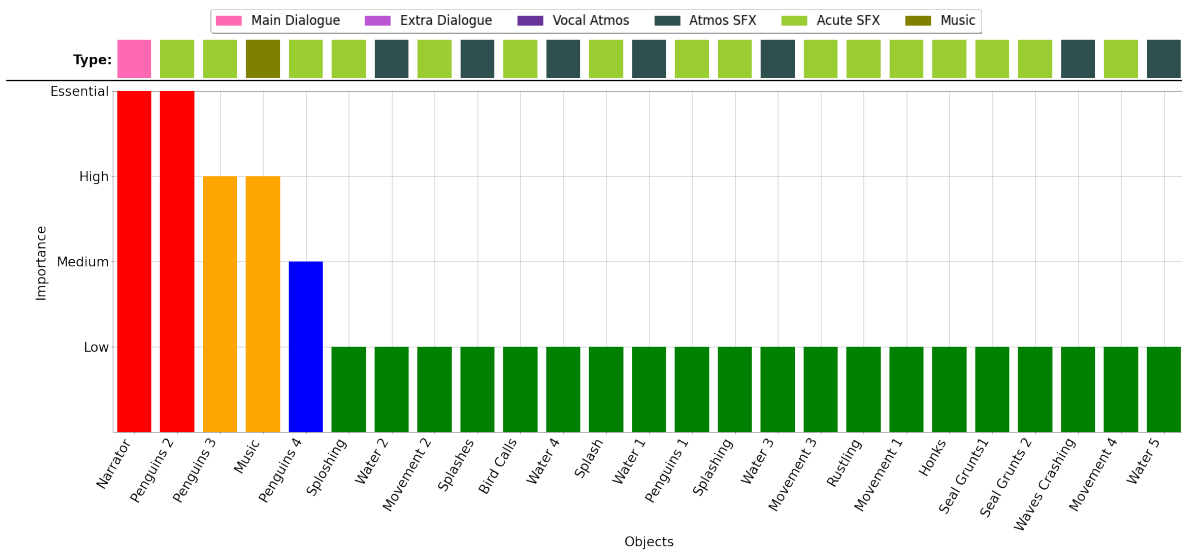


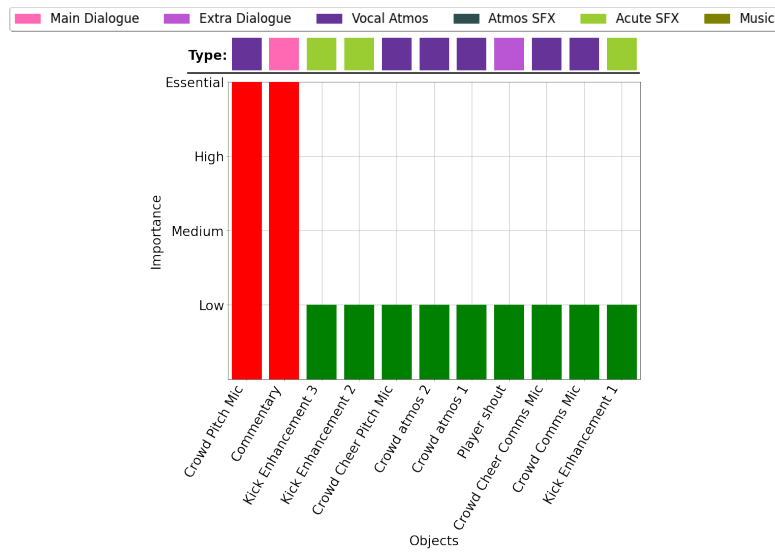
Figure F.3: A bar chart showing the mixture model’s assignments with Vostok-K Scene 2 as the test data



**Figure F.4:** A bar chart showing the mixture model’s assignments with Penguins Opening Credits as the test data



**Figure F.5:** A bar chart showing the mixture model’s assignments with Penguins Scene 1 as the test data



**Figure F.6:** A bar chart showing the mixture model's assignments with Football as the test data

# Appendix G

## 7NN Results

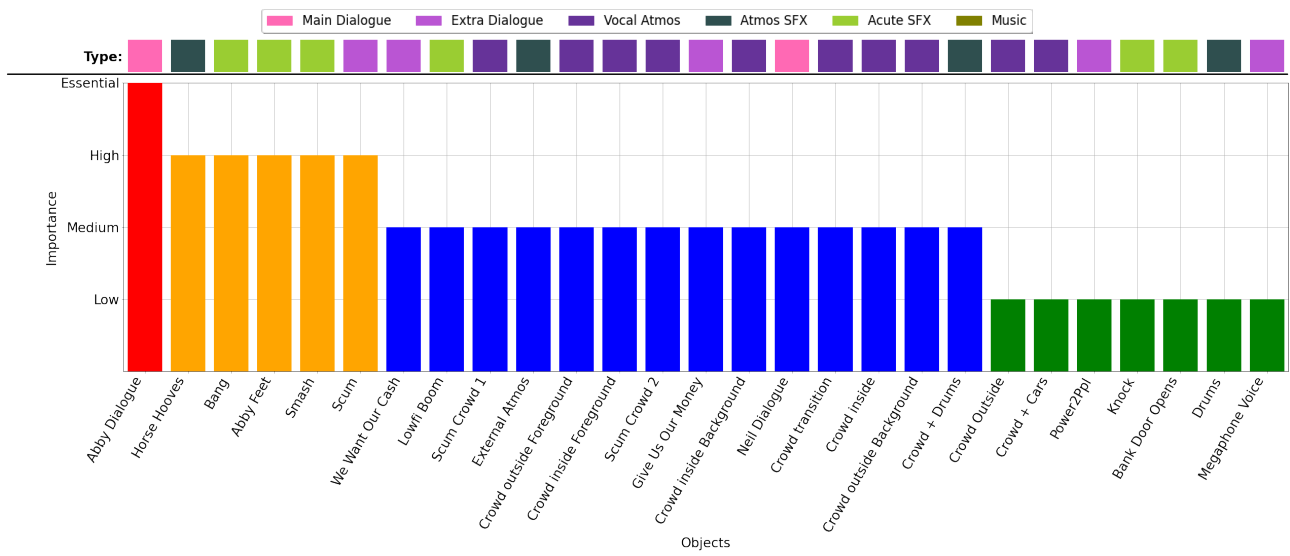
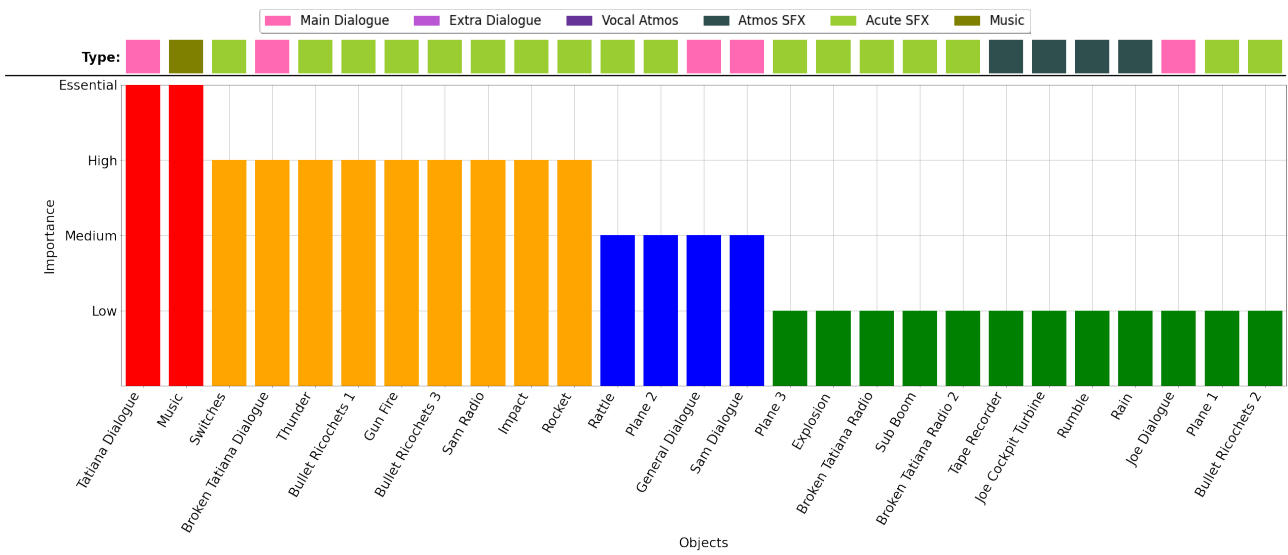
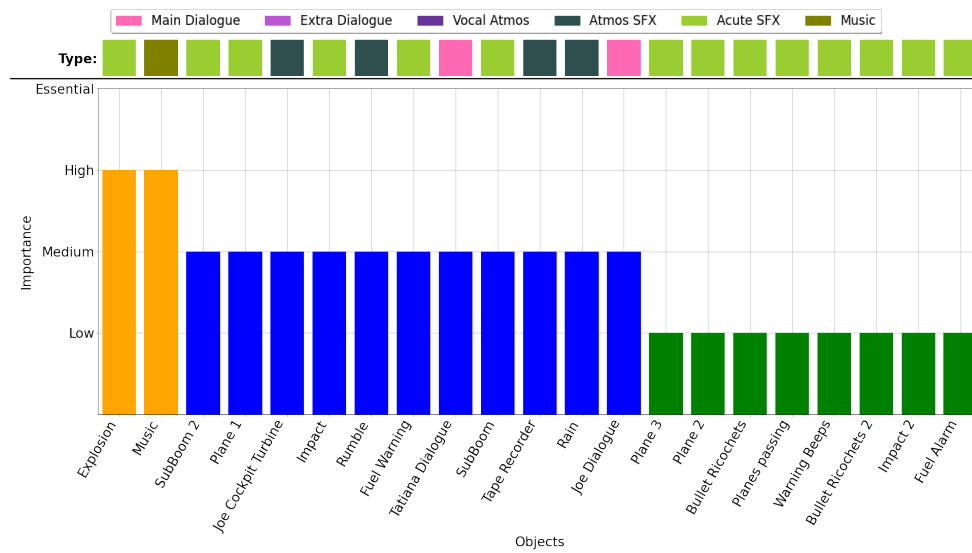


Figure G.1: A bar chart showing the 7NN model's assignments with Protest as the test data

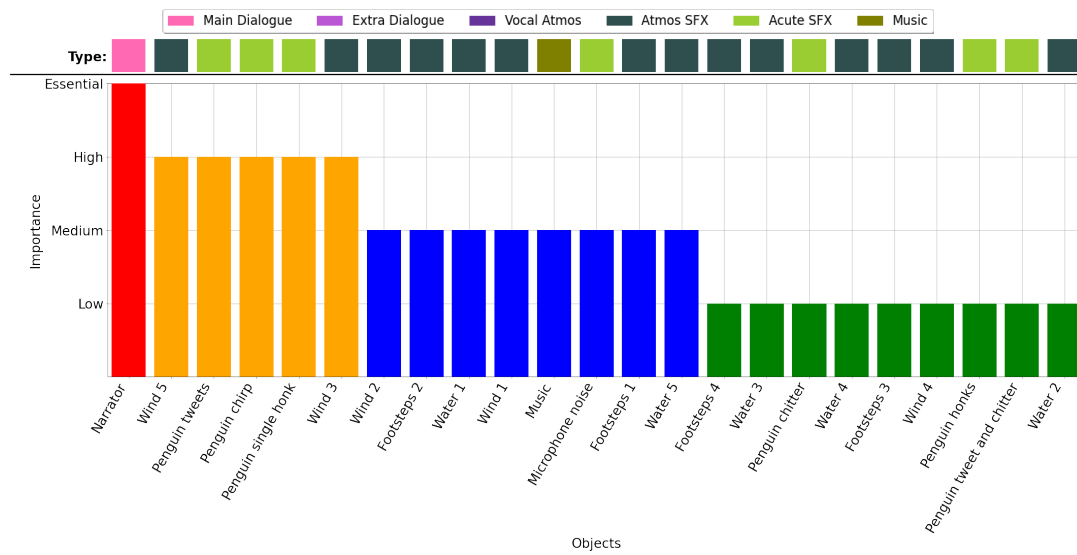


**Figure G.2:** A bar chart showing the 7NN model’s assignments with Vostok-K Scene 1 as the test data

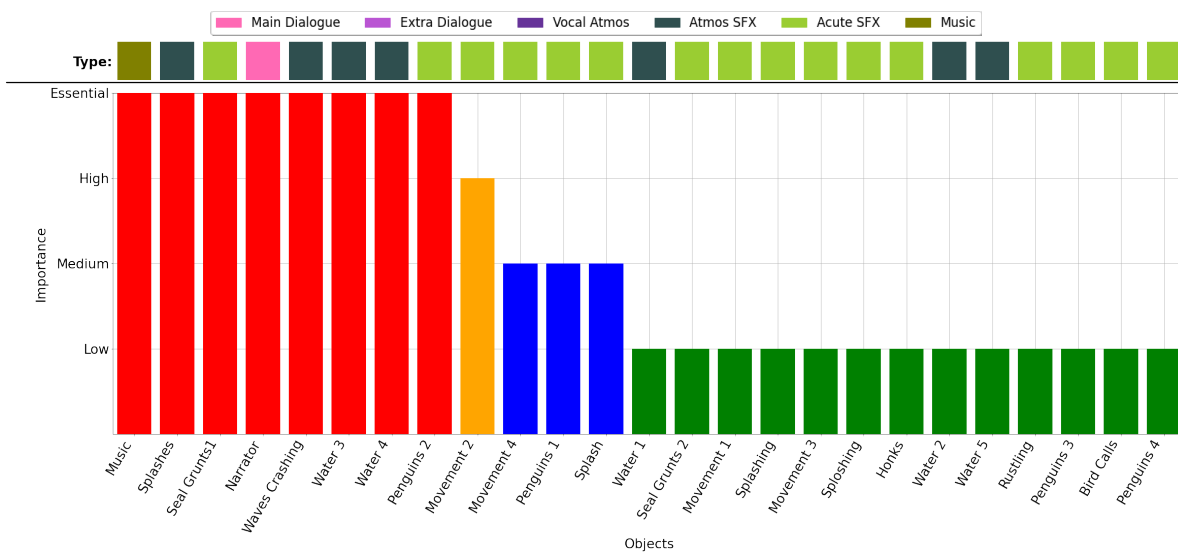


**Figure G.3:** A bar chart showing the 7NN model’s assignments with Vostok-K Scene 2 as the test data





**Figure G.4:** A bar chart showing the 7NN model’s assignments with Penguins Opening Credits as the test data



**Figure G.5:** A bar chart showing the 7NN model’s assignments with Penguins Scene 1 as the test data

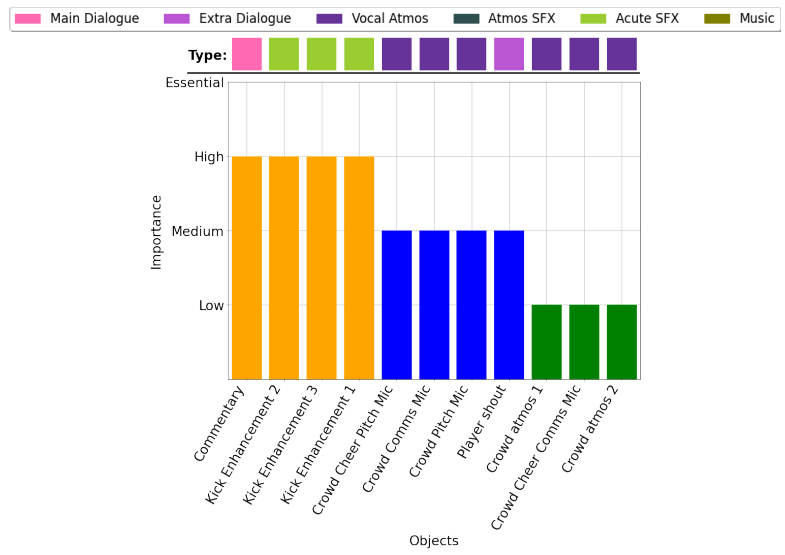
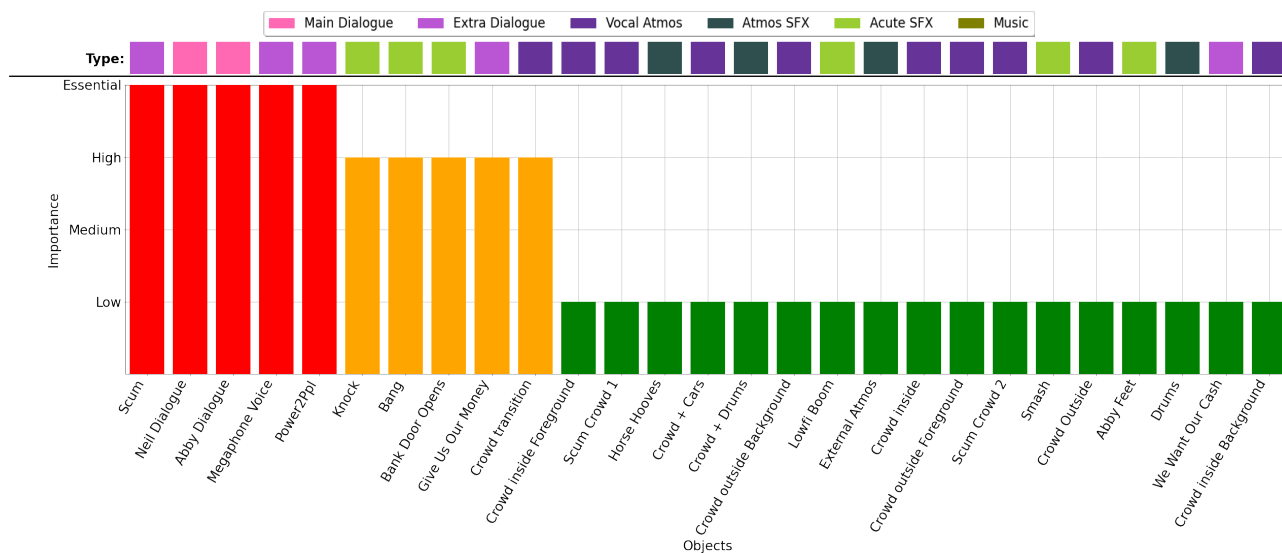


Figure G.6: A bar chart showing the 7NN model's assignments with Football as the test data

# Appendix H

## Mixture Model Training with Ground Truth Data



**Figure H.1:** A bar chart showing the mixture model’s assignments when it was trained with the ground truth data, with Protest as the test data

---

	Precision	Recall	F1-Score	Support
Low	0.35	1.00	0.52	6
Medium	0.00	0.00	0.00	9
High	1.00	0.50	0.67	10
Essential	0.40	1.00	0.57	2
Accuracy			0.48	27
Macro Avg	0.44	0.62	0.44	27
Weighted Avg	0.48	0.48	0.41	27

**Table H.1:** A table showing the precision, recall, F1 score, and accuracy when the model is trained with only ground truth data, where Protest was the test data. The ‘Support’ column refers to the number of samples in each class.