Proceedings of the 22nd International Conference on Digital Audio Effects (DAFx-19), Birmingham, UK, September 2–6, 2019

MODELLING EXPERTS' DECISIONS ON ASSIGNING NARRATIVE IMPORTANCES OF OBJECTS IN A RADIO DRAMA MIX

Emmanouil Theofanis Chourdakis *
Queen Mary University of London
London, UK
e.t.chourdakis@qmul.ac.uk

Lauren Ward †
University of Salford
Salford, Manchester, UK
l.ward7@edu.salford.ac.uk

Matthew Paradis
BBC Research & Development
London, UK
matthew.paradis@bbc.co.uk

Joshua D. Reiss
Queen Mary University of London
London, UK
joshua.reiss@qmul.ac.uk

ABSTRACT

There is an increasing number of consumers of broadcast audio who suffer from a degree of hearing impairment. One of the methods developed for tackling this issue consists of creating customizable object-based audio mixes where users can attenuate parts of the mix using a simple complexity parameter. The method relies on the mixing engineer classifying audio objects in the mix according to their narrative importance.

This paper focuses on automating this process. Individual tracks are classified based on their music, speech, or sound effect content. Then the decisions for assigning narrative importance to each segment of a radio drama mix are modelled using mixture distributions. Finally, the learned decisions and resultant mixes are evaluated using the Short Term Objective Intelligibility, with reference to the narrative importance selections made by the original producer. This approach has applications for providing customizable mixes for legacy content, or automatically generated media content where the engineer is not able to intervene.

1. INTRODUCTION

Hearing loss is estimated to affect one in six people in the United Kingdom (UK) and North America [1, 2]. This figure is likely to rise given an aging demographic and the prevalence of age-related hearing loss [3]. Further to this, 2017 audience statistics indicate that those over 50 years old in the United States of America and those over 55 in the UK watch more television on average than any other age demographic [4, 5]. Therefore listeners with some degree of hearing loss make up an increasing proportion of television audiences.

Object-based audio offers the potential for personalizable content, and may significantly improve the broadcast experience for this segment of television audiences. Complex auditory scenes, such as those found in television, radio dramas and other genres, contain some objects that are essential to narrative comprehension (e.g., dialogue and certain sound effects), and others that facilitate increased immersion (e.g., background sounds and reverberation). For people with some hearing loss, the latter sounds can impair their comprehension of the narrative [6].

In order to mitigate the issue, an approach to accesible audio has been developed which allows users to control the complexity of those scenes based on their listening needs, using a single dial control [7, 8, 9]. This is achieved by the mix engineer assigning each audio object a rank based on its narrative importance (NI) so that each track can then be amplified or attenuated according to those assignments, as well as the user's desired complexity. This method is currently limited to newly-authored object-based content or old content remixed into an object-based format. For legacy content, this manual process is arduous and prohibitively costly.

This paper investigates an approach to alleviate the issue. It first models the decision processes of mixing engineers when assigning narrative importances to a track in a radio programme mix, then recreates those decisions in unseen tracks and mixes. To do this, features that inform decisions are identified, methods for extracting those features are developed and mixture models for assigning narrative importances are used together with the audio effect developed in [7, 8, 9]. The efficacy of this model is then evaluated using an automatic speech intelligibility metric, the Short Term Objective Intelligibility (STOI) criterion [10]. This provides a first indication of the method's merit. The main contributions of this paper are twofold: we provide a model for discriminating between music/speech and sound effects, and we demonstrate a method for learning to automatically assign narrative importances to older mixes without such metadata. The latter may underpin future tools which enable hard of hearing users greater access to the large amount of legacy content.

2. PREVIOUS WORK

Personalising the balance of audio elements at the user end utilising object-based audio methods has been explored for both normal hearing listeners in noise [11, 12] and hard of hearing listeners [13, 14]. At its most simple, this personalisation provides the end-user with the ability to control the balance between background

* This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and the BBC as part of the BBC Audio Research Partnership.

† Lauren Ward is funded by the General Sir John Monash Foundation.

and foreground elements [11, 12]. A more nuanced approach explored for hard of hearing listeners gives end-users the ability to control four categories of sound: dialogue, foreground sound effects, background sound effects, and music [13]. Whilst feedback for such an approach was overwhelmingly positive, it lacked the ease of use required for large scale adoption.

A single dial control based on narrative importance metadata has been developed to combine powerful user personalisation with ease of access [7, 8, 9]. NI metadata categorises the audio objects within a soundtrack hierarchially, based on the role each sound plays in conveying the narrative. Each object is assigned an NI value in metadata between 0 (essential) and 3 (least important). Metadata is currently generated and auditioned by the producer in an audio effect plugin [9], in order to ensure that the producer's intent for the content is maintained. Gain adjustments are then applied to each sound category based on the level selected by the user on a single dial control. The control transitions smoothly between a fully immersive mix and a mix containing only the narratively important elements. This effectively allows users to adjust the complexity of the reproduced audio mix based on their needs, whilst ensuring comprehension of the narrative is always maintained. Full details of this implementation can be found in [7]. Early work on this has shown qualitative improvements in intelligibility for hard of hearing listeners whilst maintaining the creative integrity of the producer's work.

Ranking an object according to its NI can be seen as a type of automatic mixing based on gain adjustment, where a gain function of the user's preferences is chosen for each track based on its NI. Automatic gain adjustment works have existed since 1975, initially just for speech [15], and more recently in the generic context of music production [16, 17]. Here the authors optimized gains for ratios of loudness between different tracks in a multitrack live music mix. Our work differs in that we consider that individual track gains have been chosen, e.g. by one of the cited methods or a mixing engineer, and then we apply an additional post-fader gain which is a function of an individual user's preference. A similar approach, which takes individual preference into consideration can be seen in [18] where the proposed method allowed users to adapt the behaviour of a dynamic range compressor to their listening conditions.

Our method learns and models the choices of mixing engineers when ranking an audio object based on importance, as well as important features that can characterize such decisions. The latter is similar to work in [19] where the authors included important musical features as well as domain expert rules for guiding music production decisions using a probabilistic expert system. Finally, we evaluate using the Short Term Objective Intelligibility criterion [10].This metric only indicates objective intelligibility of the resulting mix, rather than the subjective comprehension of the content. However it yields an initial indication of the efficacy of the approach and whether subsequent subjective testing is warranted. A relevant work which used the same criterion to control a dynamic range compressor can be found in [20].

3. METHODOLOGY

Our goal is to assign a narrative importance d to an object based on various features extracted from the object and its role in the mix. We approached the issue as a classification task where the training data comes from a web audio listening experiment where mixing experts assigned NI values to audio objects in a radio drama.

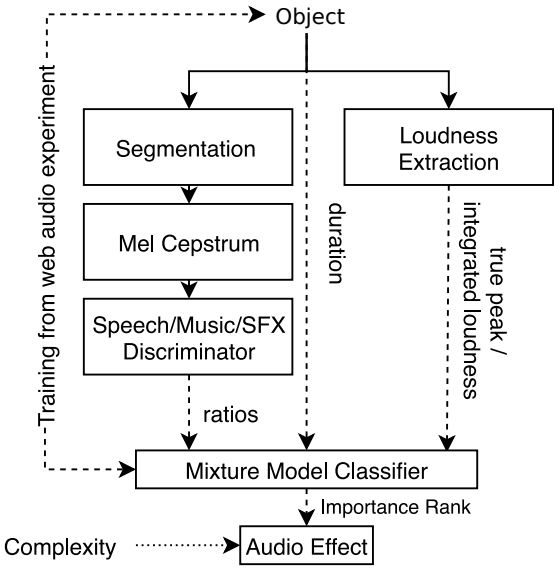


Figure 1: The outline of the process. An object from a radio mix is split into segments and each segment is classified as music/speech or sound effects. The duration of the object and its loudness characteristics are also extracted. Data coming from experiments is used to train a mixture model classifier to assign importances to similar objects. Finally, an audio plugin assigns this importance's gain level based on narrative complexity chosen by the user.

Given similar content, the goal is to make similar decisions. Below we describe each session in developing of the effect an outline of which can be seen in Figure 1.

3.1. Data Acquisition

The data used in this experiment was collected from 34 individuals who identified as audio production or mixing professionals. The majority of participants worked in television production (44%) followed by radio (35%) and film (27%). Most participants work freelance (53%) or for a national broadcaster (35%). Documentary, drama and music were the most common genres they worked in and most identified as a dubbing mixer or sound mixer (58%). 41% of respondents had worked on an object-based production before, and a further 32% were familiar with the concept of object-based audio. On average they had 22.7 years experience (median 21.5 years).

They completed the online task over their own headphones. First they were asked to listen to the radio drama *The Turning Forest* in its entirety [21]. Then they were given a segment of the drama (100s in duration) for which they were asked to assign narrative importance for its constituent 23 audio objects. The web audio interface allowed them to audition their choices at the extremes of the complexity control (fully immersive and narrative only) as well as at the 50% point to ensure their decisions produced mixes they were happy with.

In addition to this, a workshop was undertaken with the original sound designer using an early iteration of the narrative impor-

Layer Type	Layer Shape	Activation function
Convolutional	$64 \times 96 \times 64$	$relu$
Max pooling	$64 \times 48 \times 32$	–
Convolutional	$128 \times 48 \times 32$	$relu$
Max pooling	$128 \times 24 \times 16$	–
Convolutional	$256 \times 24 \times 16$	$relu$
Convolutional	$256 \times 24 \times 16$	$relu$
Max pooling	$256 \times 12 \times 8$	–
Convolutional	$512 \times 12 \times 8$	$relu$
Convolutional	$512 \times 12 \times 8$	$relu$
Max pooling	$512 \times 6 \times 4$	–
Fully-connected	256	$relu$
Fully-connected	3	$softmax$

Table 1: Shapes of the layers of the network. The network is represented with the top layer being the input layers and the bottom giving the output. The first dimension of the convolutional layers refers to the number of extracted features from that layer, and the next two to the shape of those filters. With $softmax$ we denote the softmax function which converts the output of the layer to a discrete probability distribution and with $relu$ the rectified linear unit which allows the network to model non-linearities. Since we do transfer learning, we only train the last two layers.

tance metadata acquisition tool. The sound designer was encouraged to develop her own workflow for authoring the metadata for the entirety of the *The Turning Forest*. Metadata changes could be auditioned by the sound designer in real-time using the full range of the NI control interface. Whilst an objective *ground truth* for the narrative importance assignments is not possible as it is inherently subjective, the assignments by the original producer provide the point of reference for this investigation.

3.2. Features informing decisions

Observing the decisions made by the mixing engineers, an initial hypothesis could be formed based on the type of content of the individual objects. We mainly deal with 3 classes of content; speech, music, and sound effects [22]. We developed a Convolutional Neural Network (CNN) for classification to the above three classes. CNNs have been successfully used for fast classification of images, video, or spectral representations of audio since they have many fewer parameters than fully connected neural networks and can thus be trained much faster [23]. For the task of classification, they are usually constructed using building blocks called "Convolutional Layers" which extract useful features from an image-like input, "pooling" layers which select part of the resulting features, "fully-connected" layers which combine those features, and a final classification layer [23, 24]. Such networks have previously been used to successfully distinguish between speech and music [25]. To develop our network, we used VGGish [26] as a starting point and we applied transfer learning to make it classify between speech, music, and sound effects. Transfer learning is a technique where a network trained for a task can be trained for a different task with minimal computational effort [24]. VGGish is a CNN originally trained to distinguish between 632 classes found in AudioSet [27] which is a dataset consisting of the soundtracks of 8 million Youtube videos. Since speech, music, and sound effects are among those tracks, we can achieve good performance in discriminating between those three "superclasses" by retraining the model to only discriminate between those. We therefore train the model by keeping its convolutional layers with their AudioSet-trained weights intact, since this is the part of the network that does feature extraction, and replacing the fully connected layers with a layer of size 256. Finally, we add a classification fully connected layer of 3 classes with the softmax activation function which converts the output of that layer to probabilities of the input being in one of the three classes. The shapes of the individual layers are listed in Table 1.

Inputs to the CNN are fed into the top convolutional layer in Table 1. Each audio object's track is split into non-overlapping segments of 960ms where each segment consists of the magnitudes of 64 bands of the mel cepstrum computed using a frame size of 25ms and a hop size of 10ms. Training was done by freezing the weights of the convolutional layers and only training the last two feed-forward layers. We used the GTZAN music/speech discrimination dataset[1] which contains 120 tracks with 30 minutes of speech, and 30 minutes of music as 22kHz 16bit audio files to adapt the new model to our task. In addition, we added 30 minutes of randomly sampled sound effects from the recently released online BBC SFx library[2] reformatted to match the examples in the other two classes. To make sure a specific class of sound effects is not over-represented, we used stratified sampling to select the sound effects by sampling first the class of sound effects, and then the sound effect audio file. After training using the augmented dataset we have a model that can classify the content of the object into music, speech, or sound effects and use this classification as a feature which informs the NI assignment.

Track loudness and duration is also measured. We expect that important sound effects which require the attention of the listener to have a high peak-to-integrated loudness ratio [22] as well as short duration. For example the clinking sound of two glasses toasting will signify a more important effect than the sound of frogs croaking repeatedly in the background. For this reason we use both peak-to-integrated-loudness ratio and total duration as features.

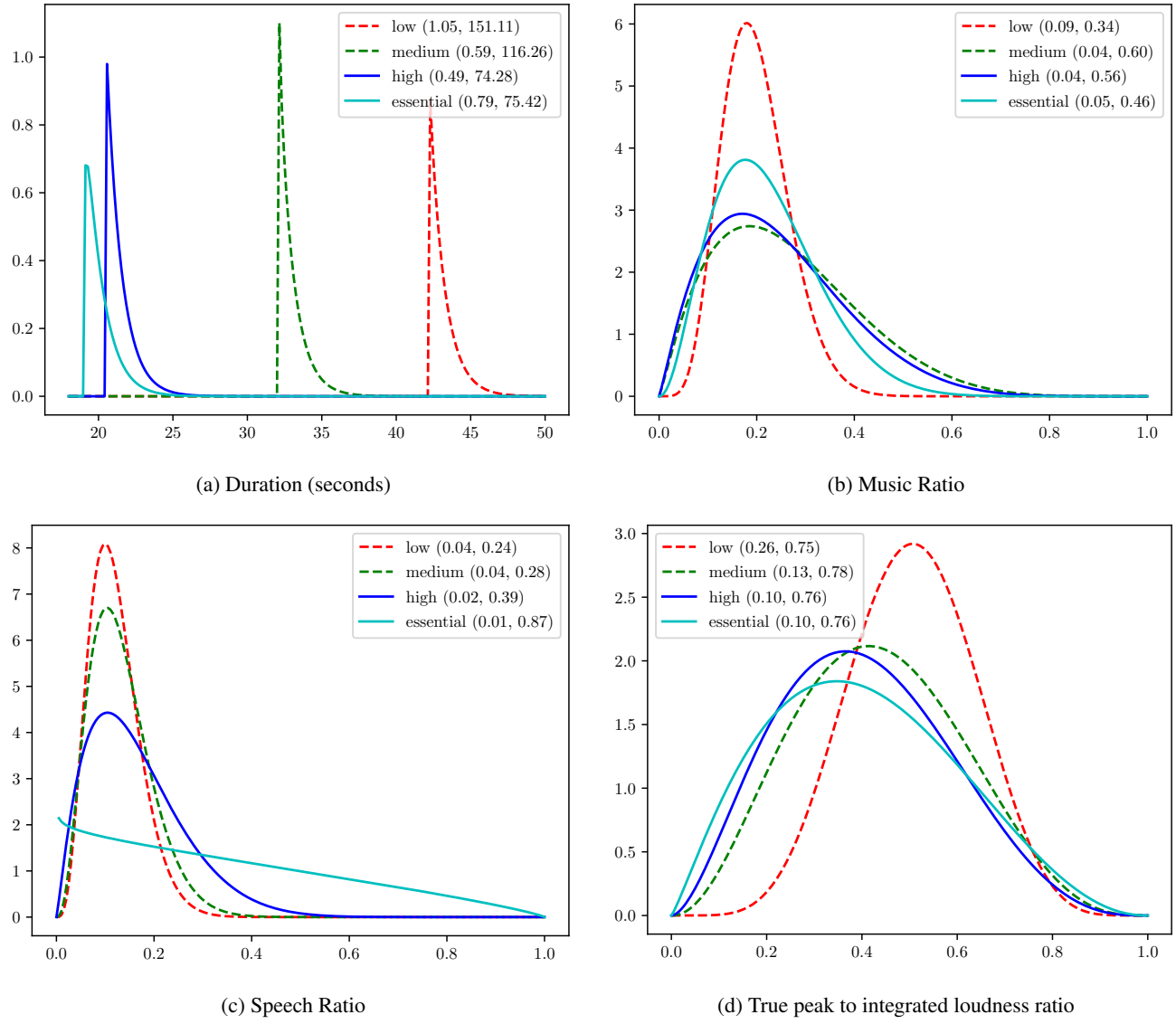
3.3. Decision Modeling

Our goal was to create a model based on the decisions from Section 3.1. An inherent challenge in the data described in Section 3.1 is that the practitioners would disagree quite a lot when ranking objects according to their importances. To quantify this disagreement, we can use Fleiss' kappa [28]:

$$\kappa = \frac{\hat{P} - \hat{P}_c}{1 - \hat{P}_c} \tag{1}$$

$\hat{P}$ is the degree of agreement between raters and $\hat{P}_c$ the degree of agreement attributed to chance. It is defined in the interval $[0, 1]$ where $\kappa = 1$ signifies total agreement. We found $\kappa = 0.008$ which denotes a low degree of agreement. Despite the low level of agreement, from Figure 3 we observe that for most objects, importance assignments are concentrated around 2 neighbouring values. This can be particularly observed for non-narration objects containing speech (*Girl Voice*, *Boy Voice*)). We decided to use a mixture model to treat this uncertainty as stochasticity in the model's decisions. In order to use such a model, we need to determine appropriate features that can give correct decisions and define their

[1] `http://marsyas.info/downloads/datasets.html`
[2] `http://bbcsfx.acropolis.org.uk/`



(a) Duration (seconds)

(b) Music Ratio

(c) Speech Ratio

(d) True peak to integrated loudness ratio

Figure 2: Estimated densities for four different narrative importance levels. In parentheses are the 95% lowest and highest respectively confidence intervals.

probability densities. From the features selected in Section 3.2, we found that ratio of speech ($p \ll 0.05$) and music ($p < 0.05$) in an object, true-peak-to-integrated-loudness ratio ($p \ll 0.05$) and total duration ($p \ll 0.05$) are good features. If we represent the values of the features above as x_{sr}, x_{mr}, x_{tpti}, and x_{dur}, the goal of decision modelling is to decide an importance d given those values. If we furthermore assume that each feature is a sample from a respective independent feature distribution, this decision can be given as:

$$\begin{aligned} d &= \arg\max_i \Pr\left(I = i | x_{sr,i}, ..., x_{dur,i}, \boldsymbol{\theta}_{sr,i}, ..., \boldsymbol{\theta}_{dur,i}\right) \\ &= \arg\max_i \Pr\left(I = i\right) \prod_{\mu \in \{sr,mr,tpti,dur\}} \Pr\left(x_{\mu,i}, \boldsymbol{\theta}_{\mu,i} | I = i\right) \\ &= \arg\max_i \Pr\left(I = i\right) \prod_{\mu \in \{sr,mr,tpti,dur\}} \Pr\left(x_{\mu,i} | I = i, \boldsymbol{\theta}_{\mu,i}\right) \\ &\quad \cdot \Pr\left(\boldsymbol{\theta}_{\mu,i} | I = i\right) \end{aligned} \tag{2}$$

Where $\boldsymbol{\theta}_\mu$ is the parameter vector for the distribution that corresponds to $x_{\mu,i}$ and i is a level of narrative importance ($essential$, $high$, $medium$, low). Music x_{mr}, speech x_{sr}, and true-peak-to-loudness x_{tpti} ratios are defined in the interval $[0, 1]$ and thus

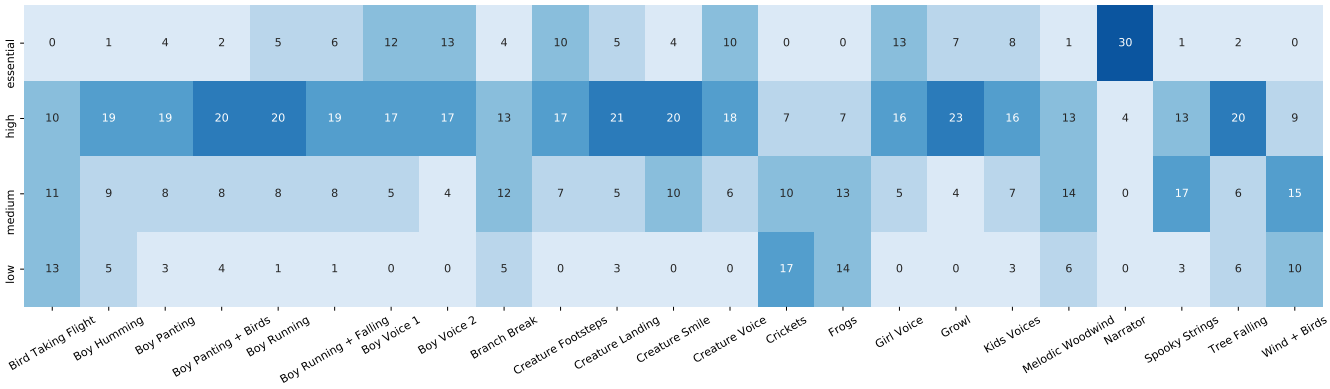


Figure 3: Heatmap of importance assignment decisions made by the experts for *The Turning Forest* radio drama. Horizontal axis shows the object names and vertical axis the assigned importances. The numbers inside individual cells correspond to the frequency of each importance value assigned to each object where darker cells correspond to higher frequencies. We observe that for most objects, there are 1 or 2 "most preferred" importance assignments.

make Beta distributions suitable for modelling their values. On the other hand the Gamma distribution is suitable for modelling total duration x_{dur} since it is defined in positive numbers. Both Beta and Gamma distributions are defined by two parameters α and β. Finally, the prior distribution of importances $\Pr(I = i)$ can be modelled as a categorical distribution, since it can take one of 4 distinct values:

$$x_{\nu,i} \sim \text{Beta}(\alpha_{\nu,i}, \beta_{\nu,i}), \quad \nu \in \{sr, mr, tpti\} \tag{3}$$
$$x_{dur,i} \sim \text{Gamma}(\alpha_{dur,i}, \beta_{dur,i}) \tag{4}$$
$$i \sim \text{Categorical}(4) \tag{5}$$

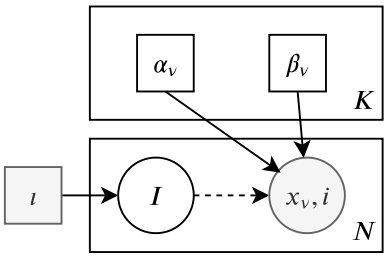


Figure 4: The mixture model used. In the diagram above, circles represent random variables, squares parameters of those variables. Shaded shapes represent observed variables. $K = 4$ represents the number of mixtures which is the same as the assigned importances, and N is the number of samples in the training data. The goal of the estimation process is to estimate the parameters represented in the non-shaded boxes (parameters of feature densities) given the observations represented in the shaded boxes and given from a training dataset (importance assignment and feature values).

The model for each feature can be seen as a diagram in Figure 4 and its distributions in Figure 2. Finally, the Beta and Gamma distributions in Eq. 3 are defined according to:

$$\text{Gamma}(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, x > 0 \tag{6}$$
$$\text{Beta}(\alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)}, 0 \leq x \leq 1 \tag{7}$$
$$\Gamma(t) = \int_0^{+\infty} dx \cdot x^{t-1} e^{-x} \tag{8}$$

What the above essentially mean is that when we know the importance of an object (differently stylized lines in Figure 2) we expect the values of the features given in Section3.2 to be samples from the distributions given in Eqs. 6, and 7. What is left is to decide on the exact shapes of those distributions, which are defined by parameters α and β above. We estimate those by using a training set of observations of importances and corresponding features and using Stochastic Variational Inference [29]. After having defined the model that model the values of the object features, the decision can be taken as in Eq. 2 where in this case:

$$\boldsymbol{\theta}_{\mu,i} = \begin{bmatrix} \alpha_{\mu,i} \\ \beta_{\mu,i} \end{bmatrix} \tag{9}$$

In this case, Eq. 2 gives a decision on an importance level d that maximizes the probability that an object belongs to that importance level given the values of its features.

3.4. Applying gains

The importance assignment process in the previous sections controls the audio effect described in [7]. This is a mixing effect with 4 stereo inputs and 2 stereo outputs:

$$\boldsymbol{y}_n = \underbrace{\begin{bmatrix} 1 & \dots & 0 \\ 0 & \dots & 1 \end{bmatrix}}_{\text{Downmixes to 2 channels}} \begin{bmatrix} \boldsymbol{I}_3(c) & \dots & & 0 \\ \vdots & 1 & & \vdots \\ & & \boldsymbol{I}_{-12}(c) & \\ 0 & & \dots & \boldsymbol{I}_{-48}(c) \end{bmatrix} \boldsymbol{x}_n \tag{10}$$

where $\boldsymbol{x}_n$ is the 8 channel input at time n corresponding to the inputs at the 4 importance levels, $\boldsymbol{y}_n$ the corresponding output, $\boldsymbol{I}_3$, 1, $\boldsymbol{I}_{-12}$, and $\boldsymbol{I}_{-48}$ are the mixing coefficients corresponding to

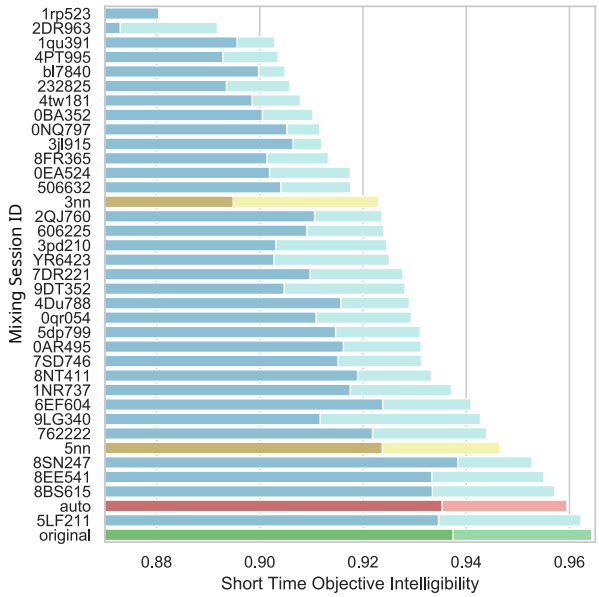


Figure 5: Results using the Short Term Objective Intelligibility measure. With lighter colors a complexity of 0 has been selected and with darker colors a complexity of 50. STOI values of the mixing sessions of the practitioners (Section 3.1) are shown as cyan colored bars. STOI of automatic assignment of narrative importances is labeled as 'auto', and STOI of the mix done by the author of the radio drama as 'original'. STOI values for decisions made using K-Nearest Neighbour classifiers for $K = 3$ and 5 are also included as '3nn' and '5nn'.

essential, *high*, *medium*, and *low* respectively (the subscripts are the gain values for each importance level in dB):

$$\boldsymbol{I}_M = \begin{bmatrix} 10^{M\left(\frac{-c}{2000}+20\right)} \\ 10^{M\left(\frac{-c}{2000}+20\right)} \end{bmatrix} \tag{11}$$

where c is a complexity number between 0 and 100 chosen by the user during playback.

4. RESULTS

For the discriminator model, 70% of the data was used as training/validation and 30% was kept aside for testing. The split was performed with a predefined random seed to guarantee reproducibility. For evaluating the model, we calculated precision p, Recall r, and f_1 on the test set:

$$p = \frac{tp}{tp+fp} \tag{12}$$

$$r = \frac{tp}{tp+fn} \tag{13}$$

$$f_1 = 2\frac{p \cdot r}{p+r} \tag{14}$$

where tp is the number of true positives, fp the number of false positives and fn the number of false negatives. The calculated metrics can be seen in Table 2.

We tested how the importance assignment model works compared to the data acquired from practitioners (Section 3.1) as well as the original radio drama author. We also included a K-Nearest neighbour classifier with $K \in \{3, 5\}$. We evaluated according to the Short Term Intelligibility criterion [10] and more specifically the PYSTOI implementation[3]. In Figure 5 we can see that when fully attenuating non-essential narrative elements our model outperformed all but one of the mixings done using the online platform (0.959 vs 0.962) and scored close to the mix by the original author (0.964). A more interesting result is when choosing a complexity value of 50, which keeps some less-important narrative elements as well, the STOI is equal to the original author's mix (0.937), even if the latter was not included in the training set. This suggests that our classifier managed to model "good" decisions from the practitioners despite the high level of disagreement. In comparison, the K-NN classifiers which do not account for uncertainty performed worse, although the classifier using the 5 nearest neighbours was still ranked above the third quartile regarding STOI.

class	precision	recall	f_1
music	1.00	0.97	0.99
speech	0.98	0.98	0.98
sfx	0.96	0.99	0.98

Table 2: Results on the test set for the music/speech/sfx discriminator model

5. DISCUSSION

This paper presented a method for modelling decisions made by mixing engineers with the goal of allowing the listener to alter the complexity of a radio drama while retaining speech intelligibility. The method relies on modelling the mixing engineers' behaviour using mixture models even when those have a large degree of disagreement. The model was tested against decisions made by the original mixing engineer of a radio drama mix and it was found that it could perform comparatively well when evaluated with an objective intelligibility metric. In the process we developed a simple music/speech/sound effects discriminator that works well for this application and is provided freely to those interested [4] and a plugin based on the VISR [30] environment is planned in order to automate the process. The current work is limited however to a single radio drama and also to a single intelligibility metric. More metrics should be considered that also measure quality and immersion. We also assume that each object is assigned a single importance value that does not change for the duration of the drama. This is an assumption that does not necessarily hold. For example we expect the footsteps of a monster approaching the main character to have higher narrative importance than the footsteps of a monster when it is further away doing something irrelevant to the story. Using our method however those two different scenarios would be ranked the same. A simple solution to this issue employed in the current work is to manually assigns the footsteps in the two senarios in distinct objects. Further work could also consider characteristics of an object that change throughout the duration of the drama when ranking them based on importance. In this paper we also consider gains after the fader stage in the mix. Gain effects

[3] https://github.com/mpariente/pystoi.git

[4] The models and other supplementary material can be found at: https://github.com/bbc/audio-dafx2019-automatic/

that synergize with our current work in applying appropriate gains pre-fader can also be examined as well as other automated mixing techniques such as in EQ [16, 31], Compression [18], or Reverberation [32]. Finally, subjective listening tests should be undertaken such that the overall quality and comprehension of the automatically assigned mixes can be evaluated by human subjects.

6. ACKNOWLEDGMENTS

We would like to thank the reviewers for their suggestions in clarifying parts of this paper. We would also like to thank Andrew Mason for his advice on measuring loudness and also his help in proofreading the final draft of the paper. This work was supported by BBC R&D as part of BBC Audio Research Partnership. Lauren Ward is supported by the General Sir John Monash Foundation.

7. REFERENCES

[1] Action on Hearing Loss, "Hearing Matters Report," November 2015.

[2] Y. Agrawal, E. A. Platz, and J. K. Niparko, "Prevalence of hearing loss and differences by demographic characteristics among US adults: data from the National Health and Nutrition Examination Survey, 1999-2004," *Archives of internal medicine*, vol. 168, no. 14, pp. 1522–1530, July 2008.

[3] Office for National Statistics, "National population projections: 2014-based statistical bulletin," October, 2015.

[4] The Nielsen Company (US), "The total audience report q1, 2017.," November 2017.

[5] Broadcasters Audience Research Board, "Trends in television viewing 2017," February 2018.

[6] O. Strelcyk and G. Singh, "Tv listening and hearing aids," *PloS one*, vol. 13, no. 6, June 2018.

[7] L. Ward, B. Shirley, and J. Francombe, "Accessible object-based audio using hierarchical narrative importance metadata," in *Audio Engineering Society Convention 145*, New York, USA, October 2018.

[8] L. A. Ward, "Accessible broadcast audio personalisation for hard of hearing listeners," in *Adjunct Publications of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video*, Hilversum, Netherlands, June 2017, pp. 105–108.

[9] B. Shirley, L. A. Ward, and E. T. Chourdakis, "Personalization of object-based audio for accessibility usingnarrative importance.," in *ACM International Conference on Interactive Experiences for Television and Online Video, Workshop on In-Programme Personalisation*, Manchester, UK, June 2019.

[10] C. H. Taal et al., "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, USA, March 2010, pp. 4214–4217.

[11] T. Walton, M. Evans, D. Kirk, and F. Melchior, "Exploring object-based content adaptation for mobile audio," *Personal Ubiquitous Comput.*, vol. 22, pp. 707–720, August 2018.

[12] W. Bleisteiner et al., "D5.6: Report on audio subjective tests and user tests," July 2018.

[13] B. Shirley, M. Meadows, F. Malak, J.S. Woodcock, and A. Tidball, "Personalized object-based audio for hearing impaired tv viewers," *J. Audio Eng Soc.*, vol. 65, no. 4, pp. 293–303, April 2017.

[14] H. Fuchs and D. Oetting, "Advanced clean audio solution: Dialogue enhancement," *SMPTE Motion Imaging J.*, vol. 123, no. 5, pp. 23–27, July 2014.

[15] D. Dugan, "Automatic microphone mixing," *J. Audio Eng Soc.*, vol. 23, no. 6, pp. 442–449, August 1975.

[16] E. Perez-Gonzalez and J.D. Reiss, "Automatic gain and fader control for live mixing," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, USA, October 2009, pp. 1–4.

[17] D. Ward, J. D. Reiss, and C. Athwal, "Multitrack mixing using a model of loudness and partial loudness," in *Audio Engineering Society Convention 133*, October 2012.

[18] A. Mason, N. Jillings, Z. Ma, J. D. Reiss, and F. Melchior, "Adaptive audio reproduction using personalized compression," in *AES 57th Conf. on The Future of Audio Entertainment Technology*, Hollywood, California, USA, March 2015.

[19] G. Bocko, M. F. Bocko, D. Headlam, J. Lundberg, and G. Ren, "Automatic music production system employing probabilistic expert systems," in *Audio Engineering Society Convention 129*, San Fransisco, USA, November 2010.

[20] H. Schepker, J Rennies, and S. Doclo, "Speech-in-noise enhancement using amplification and dynamic range compression controlled by the speech intelligibility index," *J. Acoustical Soc. of America*, vol. 138, no. 5, pp. 2692–2706, November 2015.

[21] J. Woodcock et al., "Presenting the s3a object-based audio drama dataset," in *Audio Engineering Society Convention 140*, Paris, France, June 2016.

[22] "Guidelines for production of programmes in accordance with EBU R 128," Tech. Rep., European Broadcasting Union, January 2016.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, December 2012, pp. 1097–1105.

[24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*, MIT press, 2016.

[25] M. Papakostas and T. Giannakopoulos, "Speech-music discrimination using deep visual feature extractors," *Expert Systems with Applications*, vol. 111, pp. 334–344, December 2018.

[26] S. Hershey et al., "CNN architectures for large-scale audio classification," in *Int. Conf. on acoustics, speech and signal processing*, New Orleans, USA, March 2017, pp. 131–135.

[27] Jort F. G. et al., "Audio set: An ontology and human-labeled dataset for audio events," in *Int. Conf. on acoustics, speech and signal processing*, New Orleans, USA, March 2017.

[28] J. L. Fleiss, "Measuring nominal scale agreement among many raters.," *Psychological bulletin*, vol. 76, no. 5, pp. 378, November 1971.

[29] M. D. Hoffman et al., "Stochastic variational inference," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1303–1347, January 2013.

[30] 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, July 2018.

[31] Y. Tang and M. Cooke, "Optimised spectral weightings for noise-dependent speech intelligibility enhancement," in *13th Annual Conference of the International Speech Communication Association*, September 2012.

[32] E. T. Chourdakis and J. D. Reiss, "A machine-learning approach to application of intelligent artificial reverberation," *Journal of the Audio Engineering Society*, vol. 65, no. 1/2, pp. 56–65, February 2017.