Deliverable D5.7

Evaluation report on the second prototypes of the timbral characterisation tools

Grant agreement nr	688382
Project full title	Audio Commons: An Ecosystem for Creative Reuse of Audio Content
Project acronym	AudioCommons
Project duration	36 Months (February 2016 - January 2019)
Work package	WP5
Due date	31 October 2018 (M33)
Submission date	30 November 2018 (M34)
Report availability	Public (X), Confidential ( )
Deliverable type	Report (X), Demonstrator ( ), Other ( )
Task leader	Surrey
Authors	Andy Pearce, Saeid Safavi, Tim Brookes, Russell Mason, Wenwu Wang, and Mark Plumbley.
Document status	Draft (), Final (X)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Table of contents

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 688382

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Executive Summary

This deliverable reports on the evaluation of the timbral models that were developed and documented in Deliverable D5.6 [Pearce et al., 2018].

Deliverable 5.6 presented updated versions of four models: hardness, depth, brightness and roughness. Listener-rated audio datasets were collected for each of these. This report documents the collection of these datasets and the performance of these four models with these datasets. The updated models all performed better on the new datasets than the original models did on the D5.3 data. The performance of the four updated models was then tested against the listener-rating data collected and documented in Deliverable D5.3 [Pearce et al., 2017b]. On this data, the models of hardness, depth, and brightness all performed better than the original models did; however, the updated roughness model performed worse than the original roughness model on this data.

Deliverable 5.6 also presented four new models: warmth, sharpness, boominess and reverb. Listener-rated audio datasets were collected for each of these attributes. The models of warmth and sharpness performed well, with linear correlations of r=0.79 and r=0.78 respectively. The boominess model performed the worst, with a linear correlation of 0.67. The model of reverb was changed to a classification model, able to predict the perceived level of reverberation in two classes, with a prediction accuracy of 75.25 %.

The models discussed in this report relate to version 0.3 of the timbral models on GitHub repository[1].

[1] https://github.com/AudioCommons/timbral_models

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Background

This deliverable is part of the “semantic annotation of non-musical sound properties” work package (WP5). This work package aims to enhance the usefulness of existing content and to facilitate more creative uses by: (i) developing better tools for manually annotating sound effects and soundscapes; and (ii) developing a system to automatically add timbral metadata, such that content can be searched by perceptual sound quality (e.g., piercing, crunchy, rich, etc.).

Towards this aim, Deliverable D5.1 documented a series of experiments that identified the timbral attributes that are used to describe sound effects, and established the frequency of use of each attribute. It was suggested that the most-used attributes would be the most useful to model, as these are likely to add the most value to end-users.

Models of hardness, depth, brightness, reverb, and roughness were then developed, and their implementation was documented in Deliverable D5.2 [Pearce et al., 2017a]. The performance of these models was then assessed in Deliverable D5.3 [Pearce et al., 2017b] and it was shown that improvements could be made to the models’ performances. Updated versions of these models, including a completely revised reverb model, plus three new additional models—of warmth, sharpness, boominess—were documented in Deliverable D5.6 [Pearce et al., 2018].

This deliverable documents the acquisition of listener-rated audio datasets, assesses the performance of the models on this data, and (where applicable) compares the performance of the updated models against that of the originals on both the new datasets and those used in D5.3.

1 Introduction

1.1 Main objectives and goals

Deliverable D5.2 [Pearce et al, 2017a] documented the development and operation of six prototype timbral characterisation models. These were all either implementations from literature, or developed based on a small informal listening test. These models were then evaluated in Deliverable D5.3 using data gathered from formal listening tests. The results of this evaluation indicated that the models of hardness, depth, brightness, roughness, and reverb were all reasonable but improvable, whereas the metallic-nature model showed very poor performance and was deemed unsuitable for continued development.

In the process of improving the models other than metallic-nature, the reverb model was restructured into a classification model and trained differently from the other timbral models. A training dataset of 400 stimuli was generated consisting of subjective classification into two classes of either: (i) contains no or low level perceived reverb; or (ii) contains high level of perceived reverb. This dataset was used for the training and evaluation of the reverb model.

For the models of hardness, depth, brightness, and roughness, larger listener-rated audio datasets were collected and used for retraining. In addition to these four attributes, large datasets were also collected for three new attributes: warmth, sharpness, and boominess. The new datasets were used to develop and evaluate the performance of new timbral models predicting these attributes.

This deliverable documents the collection of these larger datasets, evaluates the performance of the updated timbral models with these datasets, and re-evaluates the performance of the models of hardness, depth, brightness, and roughness with the evaluation data from Deliverable D5.3.

1.2 Methodology

As with Deliverable D5.3, there are no existing databases that give scalar data for each of the timbral attributes under investigation. Therefore, a suitable corpus of audio stimuli was compiled, listening tests were conducted, models were trained and the performance of each model was evaluated.

Sound source types suitable for training and evaluating each timbral model (other than reverb) were identified by analysing one month of the freesound search history (from April 2016). Sound sources were deemed suitable if they had been searched for using the relevant timbral attribute in the search query (Section 2). Sounds of these source types were then collected and listening tests were used to acquire timbral-rating data for these sounds. The test design and implementation are documented in Section 2.5. The training and evaluation of the timbral (non-reverb) models is documented in Sections 3–9.

The collection of the reverb model's classification dataset and the evaluation of the model's performance are documented in Section 10.

1.3 Terminology

AudioCommons: reference to the EC H2020 funded project AudioCommons, with grant agreement number 688382.

Audio Commons Initiative: reference to the AudioCommons project core ideas beyond the lifetime and specific scope of the funded project. The term "Audio Commons Initiative" is used to imply i) our will to continue supporting the Audio Commons Ecosystem and its ideas after the lifetime of the funded project, and ii) our will to engage new stakeholders which are not officially part of the project consortium.

Audio Commons: generic reference to the Audio Commons core ideas, without distinguishing between the concept of the initiative and the actual funded project.

Audio Commons Ecosystem (ACE): set of interconnected tools, technologies, content, users and other actors involved in publishing and consuming Audio Commons content.

Audio Commons content (ACC): audio content released under Creative Commons licenses and enhanced with meaningful contextual information (e.g., annotations, license information) that enables its publication in the ACE.

Content creator: individual users, industries or other actors that create audio content and publish in the ACE through content providers.

Content provider: services that expose content created by content creators to the ACE.

Content user: individual users, industries or other actors that use the content exposed by content providers and created by content creators in their creative workflows.

Tool developer: individual users, industries or other actors that develop tools for consuming (and also potentially publishing) Audio Commons content.

Embeddable tools: tools for consuming Audio Commons content that can be embedded in existing production workflows of creative industries.

2 Subjective data

For each of the timbral models of hardness, depth, brightness, roughness, warmth, sharpness, and depth, a large listener-rated audio dataset was collected. This section describes the methods used for the acquisition of audio files, audio processing, and listening tests.

As with the collection of data documented in D5.3, stimuli were desired that were commonly searched for using each of the timbral attributes in the search query. In this way, the models are trained and evaluated using sounds similar to those that users are likely to search for using these attributes. Datasets were collected in five stages: (1) identifying, from the *freesound.org* search history, source types relevant to each timbral attribute under consideration; (2) calculating the number of audio files of each source type that would constitute a representative (according to search frequency) cross-section; (3) downloading a random selection of audio files for each source type; (4) having an independent expert select the determined number of audio files for each source type for each attribute under consideration; and (5) conducting listening tests to collect timbral ratings of the selected audio files.

The following subsections describe each of these stages in more detail. For the sake of simplicity, the processes are documented for the hardness attribute only. The same procedures were also employed for all other attributes.

2.1 Identifying source types

One month of the *freesound.org* search history (from April 2016) was used to identify the source types that are the most commonly searched for using the hardness attribute in the search query. This was done by identifying the most-searched phrases that contain the timbral terms that constitute hardness. From Deliverable D5.1 [Pearce, Brookes, and Mason; 2016], hardness was found to relate to the terms *hard*, *soft*, and *pillowy*. A list of all unique search phrases containing any of these terms was collated from the search history.

Examination of this list revealed several search queries where the source type was unclear, or the timbral term was not being used in a timbral way (e.g. hardcore). The list of matching searches was given to three independent experts who were asked to mark any search queries where the terms hard, soft, or pillowy were not used in a timbral sense, or where the source type was unclear. Any queries that two or more experts marked were removed.

The remaining search queries were then manually grouped based on their source types (e.g. grouping queries of “kick”, “hard kick”, and “soft kick drum”). The number of uses for each source was then calculated.

2.2 Weighted cross section

A weighted cross-section approach was then taken to identify a suitable number of stimuli to be included in a listening test for each source type. The aim was to limit the maximum number of stimuli of a single source type to ten, to ensure the inclusion of at least two stimuli of each source type, and to gather a total number of stimuli close to 200.

First, any sources searched for fewer than *x* times were removed. The remaining number of searches was then calculated with the equation

$$Search\ number\ =\ min\left(10,\ round\left(\tfrac{2n}{x}\right)\right)$$

where *n* is the number of times each source was searched, and *x* is the 'cutoff' number for searches. The value of *x* was varied to give about 200 stimuli in total across all source types. For the hardness dataset, *x* was set to 31, representing at least one search per month. This resulted in 206 stimuli in total, spread over 37 different source types.

2.3 Downloading and preparing audio files

For each source type, a search was conducted on freesound.org using the freesound API. Each source type was searched individually, and in conjunction with the terms hard and soft. The term pillowy was removed from the hardness evaluation since no queries in the freesound search history contained this term. For example, when searching for kick drum samples to assess the hardness attribute, three searches were made: "kick", "hard kick" and "soft kick". For each search, 50 unique results were randomly selected and downloaded.

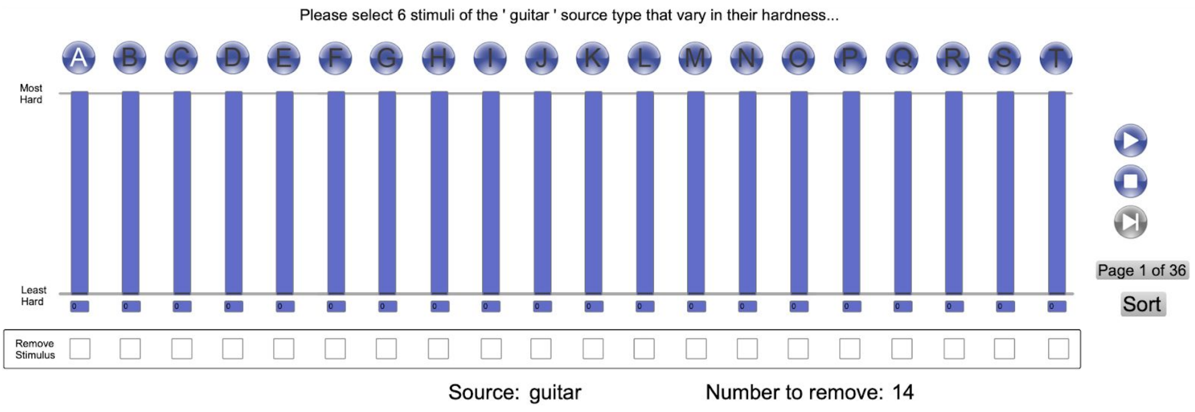
All downloaded stimuli were converted to WAV files, resampled to 44.1 kHz, and loudness normalised using the loudnorm function from ffmpeg to a target loudness of -24 LUFS.

2.4 Expert selection

Using the automated randomised downloading method, there was no guarantee that each of the downloaded audio files were of the intended source type. To overcome this, each of the downloaded files was manually inspected and auditioned by one experimenter. Any audio files that did not match the intended source were removed.

For each source type, 20 stimuli were retained: the first ten audio files downloaded when searching for the source type alone; the first five downloaded when searching for the source type and "hard"; and the first five downloaded when searching for the source type and "soft".

To ensure that the final dataset covered a wide range of hardness, a stimulus selection experiment was conducted. Stimuli (downloaded audio files) were presented on a multi-page test interface as shown in Figure 1. Each page comprised the twenty stimuli of a single source type. An independent expert was asked to: (i) remove any stimuli they considered to not be of the source type specified; (ii) select the most and least hard stimuli; and (iii) select stimuli as required, equally spaced on a hardness scale between the most and least hard, to make the number of each source type as specified in Section 2.2. For the hardness attribute, this method provided the 206 desired stimuli across the 36 source types.



Please select 6 stimuli of the ' guitar ' source type that vary in their hardness...

Source: guitar Number to remove: 14

Figure 1 - Test interface for selection of appropriate stimuli.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

2.5 Listening test design

A multiple stimulus comparison listening test design was used to collect ratings of hardness. Listening tests were conducted in two test sessions. To maximise consistency of ratings across sessions and test pages, an independent expert was asked to select the most and least hard stimuli of all 206. These were used as hidden anchors on each test page. To test subject consistency, the independent expert also selected six stimuli that varied clearly in hardness and source type.

Prior to each test session, subjects were presented with a multi-page familiarisation test interface, as shown in Figure 2. Each page consisted of 20 stimuli. Stimuli were randomly distributed between test pages, but the hidden anchors were always on the first page. Subjects were instructed to listen to all stimuli on the first familiarisation page, then as many other pages as required in order to be familiar with the range of hardness exhibited by the stimuli. Subjects were then presented with the main test interface where they were asked to make ratings relative to the full range of hardness heard during the familiarisation stage.

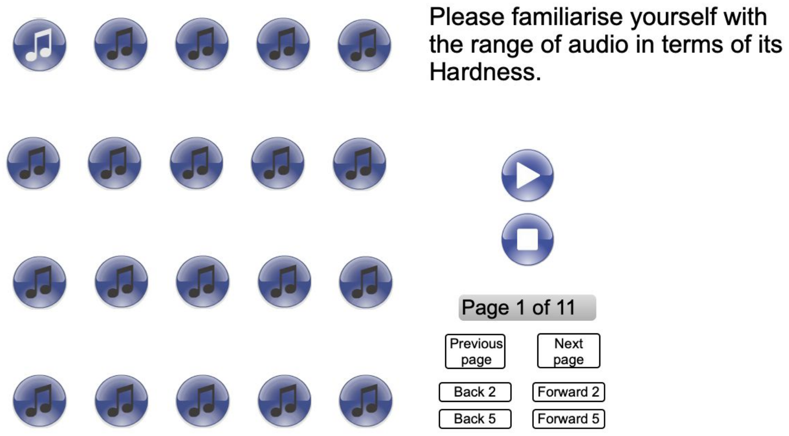
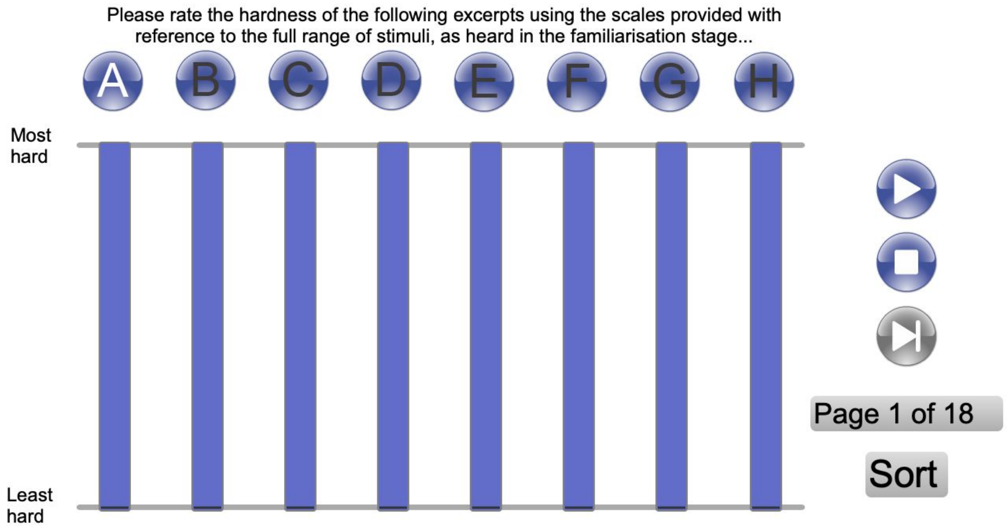

Please familiarise yourself with the range of audio in terms of its Hardness.

Figure 2 - Familiarisation listening test interface.

Each page of the listening test comprised eight stimuli: two hidden anchors, and six stimuli randomly drawn from the remaining 204 (with the exception of a page to test intra-subject consistency, comprised of the two hidden anchors and the six stimuli selected by the independent expert). An example of the test interface is shown in Figure 3. Subjects were asked to rate the perceived hardness of the stimuli, relative to the full range of hardness encountered during the familiarisation stage.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382



Please rate the hardness of the following excerpts using the scales provided with reference to the full range of stimuli, as heard in the familiarisation stage...

Figure 3 - Listening test interface.

It should be noted that for the hardness test, the number of stimuli selected divided evenly into two test sessions of 18 test pages each. For other attributes that did not divide into an integer number of test pages, the first test session had one more test page. For attributes where the total number of stimuli did not divide evenly into six random stimuli per page, the independent expert was asked to select up to five stimuli (depending on the number required) that were repeated on the last test page. This ensured that every test page had eight stimuli.

All listening tests were conducted in acoustically treated edit rooms at the University of Surrey, replaying stimuli with Neumann KH 120A loudspeakers. The gain of the replay system was adjusted so that -14 dBFS pink noise was reproduced at 68 dBASPL. This was found to be a comfortable listening level for participants for test sessions of approximately 25 minutes.

3 Hardness model

3.1 Hardness training dataset

Ratings of hardness were collected as described in Section 2. The hardness dataset comprised 206 sound files over 36 source types. Hardness-rating listening tests were completed by seventeen undergraduate students on the University of Surrey's Music and Sound Recording course, all of whom had completed a module of technical listening and had experience participating in listening tests.

To analyse the performance of the listeners, the intra-subject consistency, inter-subject agreement, and tucker-1 correlation loadings were calculated. From the tucker-1 analysis, it appeared that subjects 4, 7, 9, 14, and 17 had differing decision criteria when rating hardness. This was further evidenced by their lower inter-subject agreement scores. These subjects were removed from subsequent analysis.

3.2 Development of the hardness model

Multiple features were extracted that may relate to the perceived hardness of an audio file. To test which feature metrics would be suitable for modelling hardness, an iterative modelling process was used; creating a linear regression model with the feature which best correlates with the listener ratings of hardness, and iteratively adding the next-best performing feature into a multilinear regression model. The performance of this multilinear regression model at each iteration is shown in Figure 4.

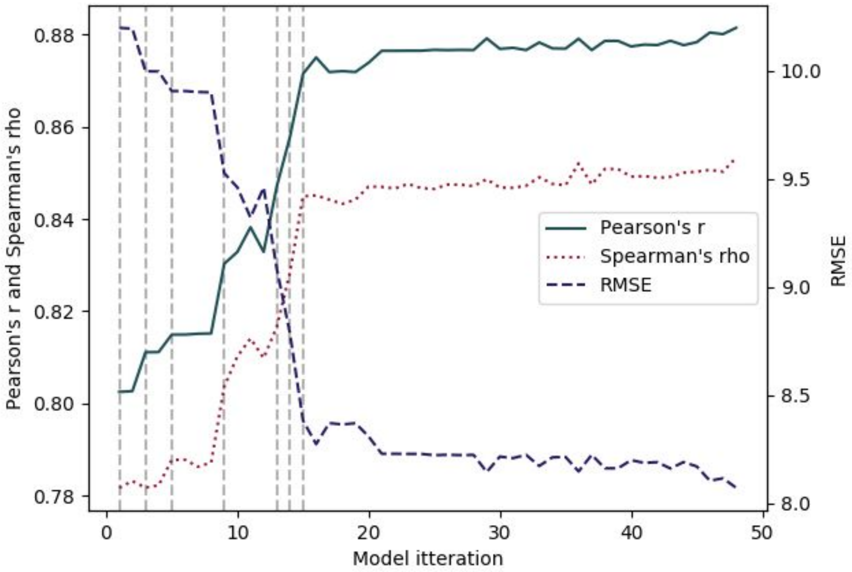


Figure 4 - Iterative modelling process with the hardness model development.

Iterations 1, 3, 5, 9, 13, 14, and 15 (highlighted in Figure 4 with vertical dashed lines) all showed an improvement of the model's performance. These features were retained for the final model.

3.3 Performance of the hardness model

The performance of the updated hardness model with respect to its training dataset is shown in Figure 5. This model gives a linear correlation of r=0.87 and rank order performance of rho=0.84.

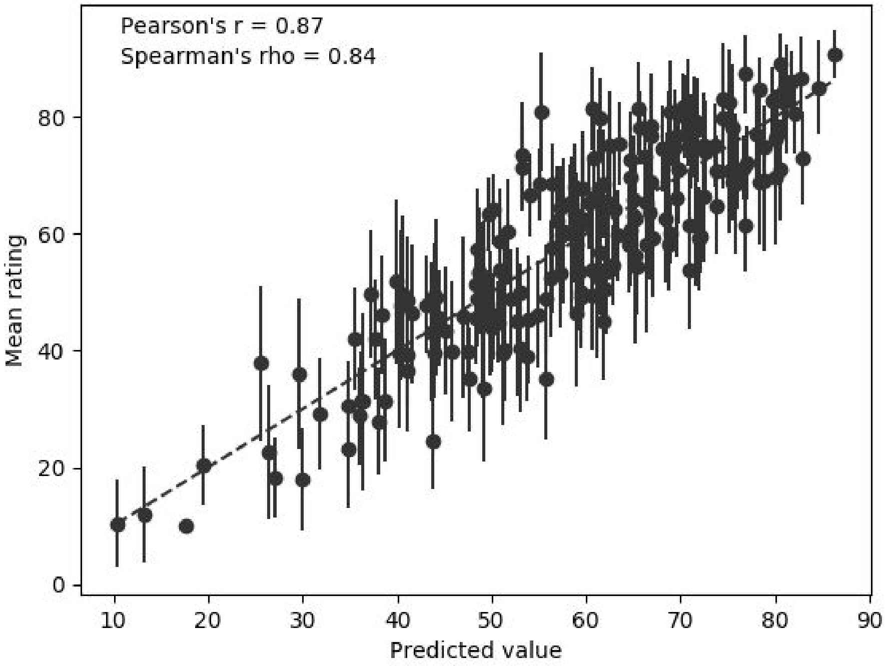


Figure 5 - Hardness model performance with the training dataset.

3.4 Hardness validation

To further validate this model of hardness, its performance was tested against the within-source-type and between-source-type data from Deliverable D5.3 [Pearce, Brookes, and Mason; 2017b]. These datasets did not share any audio files with the training dataset, and as such the end-points of the ratings scale are likely to be different. Therefore, only the correlation and rank order performance are calculated.

3.4.1 Within-source analysis

Figure 6 below shows the performance of the current hardness model with the five source types evaluated in D5.3: cymbal, guitar, kick, piano, and snare. The linear correlation between model-predicted and listener-rated hardness is exact (r=1.0) for the cymbal source type, and is high for the piano, snare, and kick sources (r=0.96, 0.83, and 0.75 respectively). The model performs least well for the guitar source type. However, it should be noted that the confidence intervals for the data are large, and the general trend of the data is still maintained.

To assess the performance of a prediction model considering the variance of the listening test data, the Spearman's Rho* and Kendall's Tau* can be calculated [Pearce et al; publication pending]. According to these metrics, the hardness model can predict the rank order for all source types perfectly.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

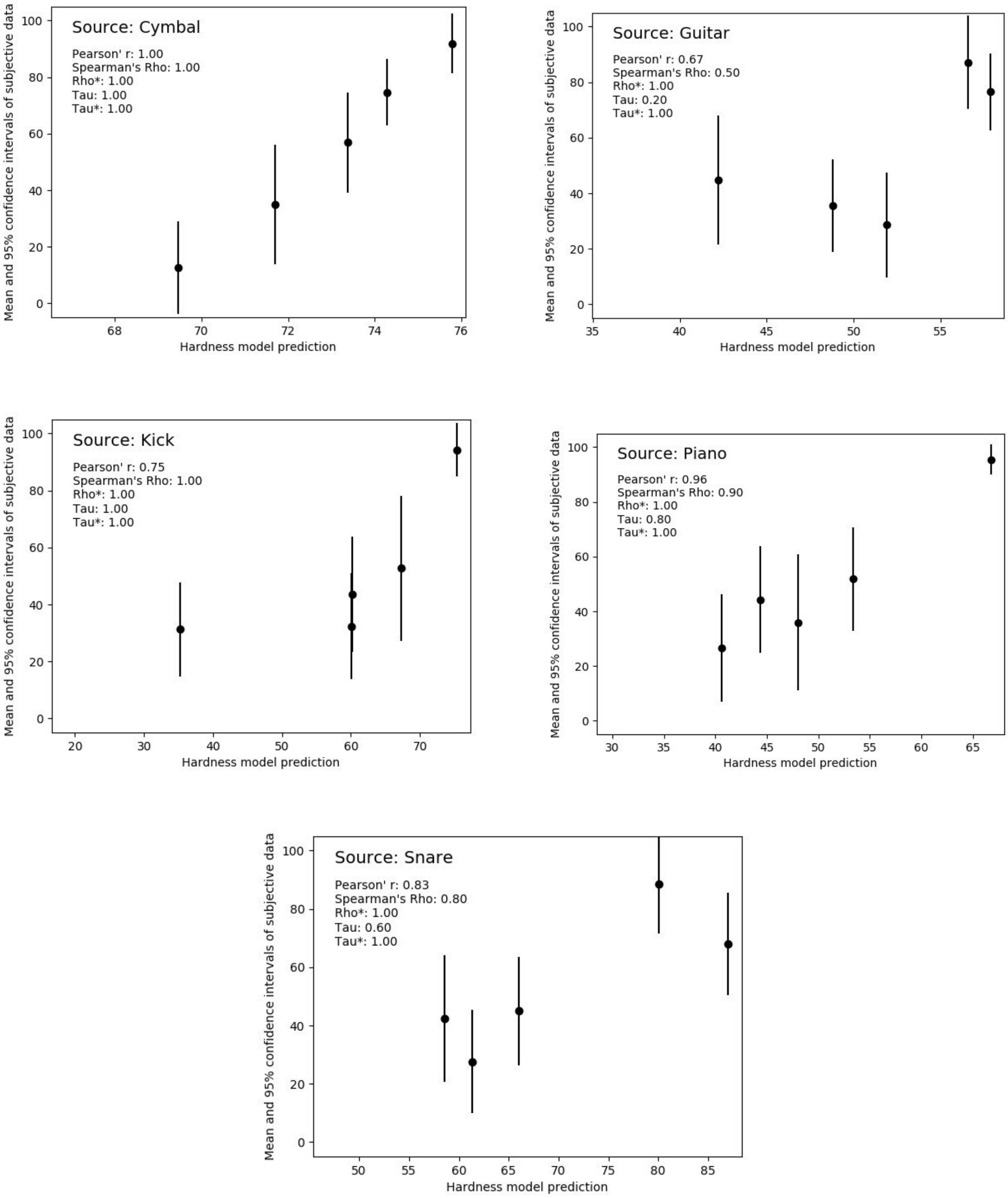


Figure 6 - Within-source-type performance of the hardness model. (a) Cymbal; (b) Guitar; (c) Kick; (d) Piano; (e) Snare.

3.4.2 Between-source analysis

Figure 7 shows the performance of the hardness model with the between-source-type data. The model predictions achieve a linear correlation with listener ratings of r=0.76, less than the training dataset which achieved a correlation of r=0.87. Two stimuli are highlighted in red in Figure 7; both are over-predicted by the hardness model. Both of these stimuli are recordings of voices; one contains excessive sibilance, and the other contains high levels of background noise and clicks. It is likely that subjects rated that hardness of the voice component, and ignored these other components, whereas the feature metrics would measure these components.

When removing these two stimuli, the performance of the model improves to r=0.83, a similar performance to that achieved with the training data.

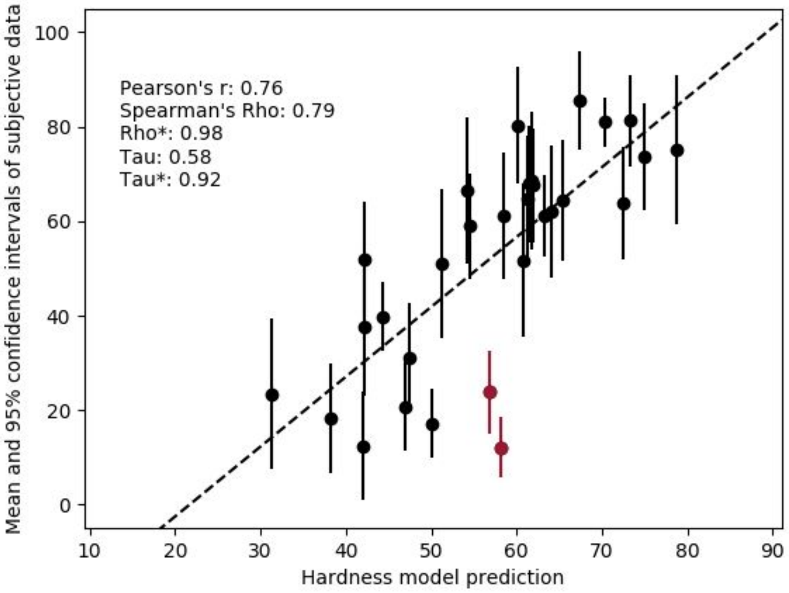


Figure 7 - Between-source-type performance of the hardness model.

3.5 Performance summary

Comparing the performance of the current hardness model to the performance of the model documented in D5.3, the current model tends to perform better. The performance of the model for the within-source-type data is summarised in Table 1.

Table 1: Summary of the hardness model's performance with the within-source-type data.

Source type	Pearson's r	Spearman's Rho	Rho*	Kendall's Tau	Tau*
Cymbal	1.00	1.00	1.00	1.00	1.00
Guitar	0.67	0.50	1.00	0.20	1.00
Kick	0.75	1.00	1.00	1.00	1.00
Piano	0.96	0.90	1.00	0.80	1.00
Snare	0.83	0.80	1.00	0.60	1.00

These results show that the hardness model performs well (and better than the original version) with all source types, with the exception of the guitar source type.

Table 2 compares the performance of the current hardness model against the performance of the initial implementation with the between-source-type data. In D5.3, the initial hardness model had an outlying data-point that significantly skewed the scale. The current implementation doesn't suffer from this outlying data-point and outperforms the initial model. When removing this data-point, the current model still outperforms the initial implementation, although the performance is similar.

Table 2: Summary of the hardness model's performance with the between-source-type data.

Source type	Pearson's r	Spearman's Rho	Rho*	Kendall's Tau	Tau*
Current model	0.76	0.79	0.98	0.58	0.92
Current model (highlighted stimuli removed)	0.83	0.83	1.00	0.62	0.97
D5.3 implementation	0.22	0.76	0.98	0.52	0.92
D5.3 implementation (outlier removed)	0.75	0.75	0.98	0.51	0.92

4 Depth model

4.1 Depth training dataset

Ratings of depth were collected in the manner described in Section 2. The depth dataset comprised 218 stimuli over 34 source types. The listening tests were completed by fifteen undergraduate students on the University of Surrey's Music and Sound Recording course, all of whom had completed a module of technical listening and had experience participating in listening tests.

To analyse the performance of the listeners, the intra-subject consistency, inter-subject agreement, and tucker-1 correlation loadings were calculated. From the tucker-1 analysis, it appeared that subjects 4 and 9 had differing decision criteria when rating depth, further evidenced by lower inter-subject agreement scores. These subjects were removed from subsequent analysis.

4.2 Development of the depth model

The original depth model was tested against the new dataset. From examining stimuli that were not predicted well, several additional features were identified as not accounted for by the original model. New extraction techniques were developed to estimate the note duration and pitch. These were incorporated into an updated model.

4.3 Performance of the depth model

Figure 8 shows the performance of the updated depth model with the training dataset. This model has a linear correlation of r=0.89 with the listener ratings and rank order performance of rho=0.86.

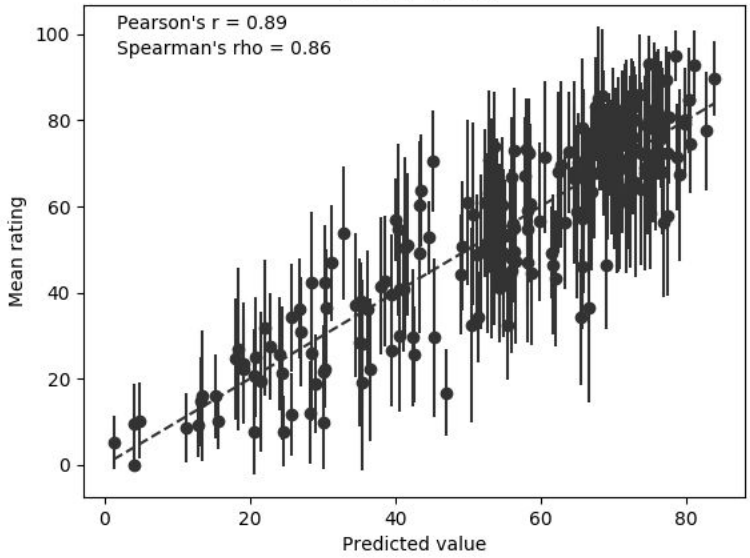


Figure 8 - Depth model performance with the training dataset.

4.4 Depth validation

To further validate this model of depth, its performance was tested against the within-source-type and between-source-type data from Deliverable D5.3.

4.4.1 Within-source analysis

Figure 9 below shows the performance of the updated depth model with the five source types evaluated in D5.3: bass, drums, impact, kick, voice.

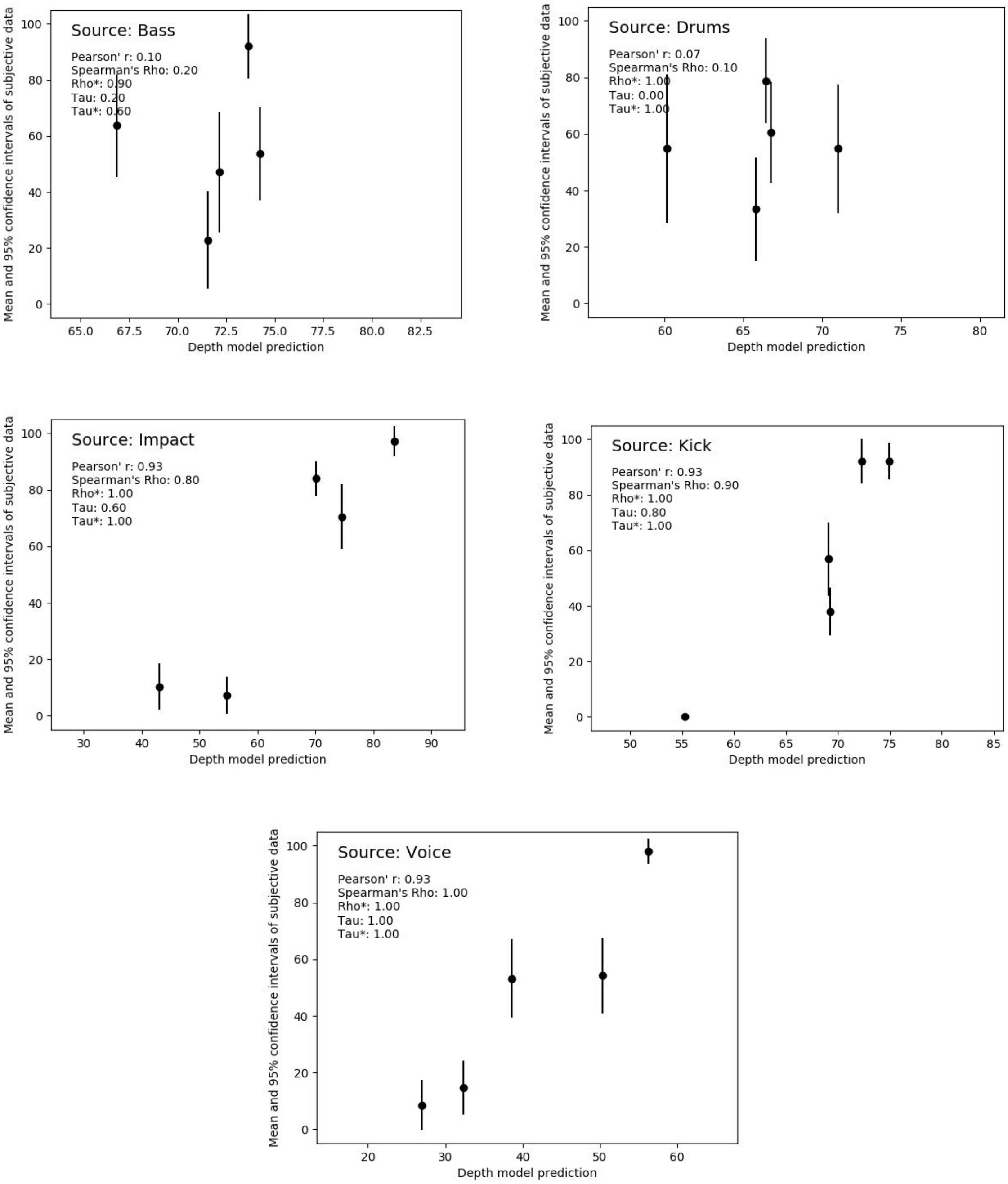


Figure 9 - Within-source-type performance of the depth model. (a) Bass; (b) Drums; (c) Impact; (d) Kick; (e) Voice.

The depth model performs well on the voice, kick, and impact sources. With the bass source, there appears to be one stimulus that is under-predicted, resulting in poor correlation with the listener

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

ratings. With the drums source, the performance is very poor. However, the confidence intervals of the ratings are very large, suggesting a degree of listener inconsistency/uncertainty.

4.4.2 Between-source analysis

Figure 10 shows the performance of the depth model with the between-source-type data from D5.3. This shows the model performing very well, with a linear correlation of r=0.93 with listener ratings and a rank order performance of rho=0.91.

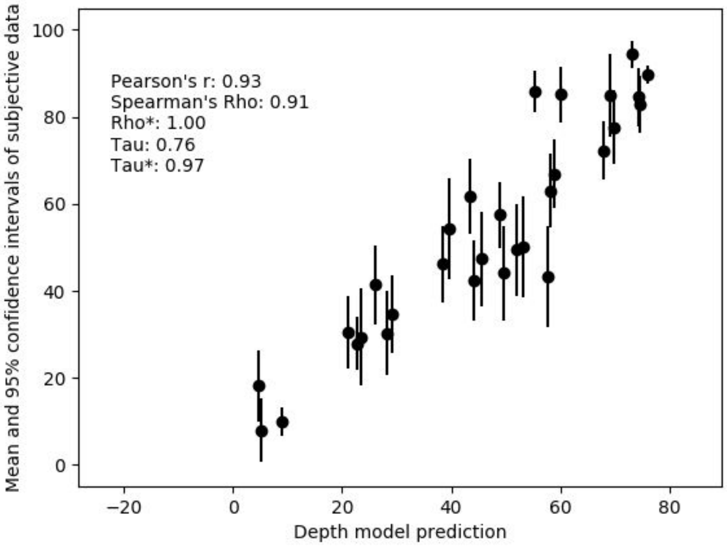


Figure 10 - Between-source-type performance of the depth model.

4.5 Performance summary

The performance of the updated depth model with the within-source-type data is summarised in Table 3.

Table 3: Summary of the depth model's performance with the within-source-type data.

Source type	Pearson's r	Spearman's Rho	Rho*	Kendall's Tau	Tau*
Bass	0.10	0.20	0.90	0.20	0.60
Drums	0.07	0.10	1.0	0.00	1.00
Impact	0.93	0.80	1.0	0.60	1.0
Kick	0.93	0.90	1.0	0.80	1.0
Voice	0.93	1.0	1.0	1.0	1.0

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

The depth model performs well with the impact, kick and voice source types. However, the model shows poor performance with the bass and drums source types. Inspection of these plots reveals that the model performs poorly with the bass source type due to a single stimulus being under-predicted. For the drums source type, the confidence intervals are very large, and when considering these the model can still predict the rank order.

Table 4 compares the performance of the updated depth model against that of the original model with the between-source-type data. From this table it can be seen that the updated model outperforms the original model in all metrics. Interestingly, the updated model performs better with the D5.3 data than with the training data, achieving a linear correlation of r=0.93 compared to 0.89.

Table 4: Summary of the depth model's performance with the between-source-type data.

Source type	Pearson's r	Spearman's Rho	Rho*	Kendall's Tau	Tau*
Current model	0.93	0.91	1.00	0.76	0.97
D5.3 implementation	0.59	0.72	0.94	0.53	0.81

5 Brightness model

5.1 Brightness training dataset

Ratings of brightness were collected in the manner described in Section 2. The brightness dataset comprised 210 stimuli over 33 source types. Listening tests were completed by thirteen undergraduate students on the University of Surrey's Music and Sound Recording course, all of whom had completed a module of technical listening and had experience participating in listening tests.

To analyse the performance of the listeners, the intra-subject consistency and inter-subject agreement were calculated. Subject 11 obtained a relatively poor agreement score and low consistency. From this it was decided to remove this subject's results from subsequent analysis.

5.2 Development of the brightness model

The original brightness model consisted of two metrics: the ratio of high-frequency energy to the total energy; and the spectral centroid above a cut-off frequency. This model performed reasonably well with the D5.3 evaluation data, so no additional features were sought.

The initial feature extraction implementation used a hard cut-off. This may produce strange results when analysing signals with significant energy close to this cut-off frequency. To improve this model, feature extraction was re-implemented with 6th order filters. Crossover frequencies for the metrics were recalculated to maximise correlation.

5.3 Performance of the brightness model

The performance of the updated brightness model with the training data is shown in Figure 11. This model achieves a linear correlation of r=0.86 with listener ratings, and rank order performance of rho=0.83.

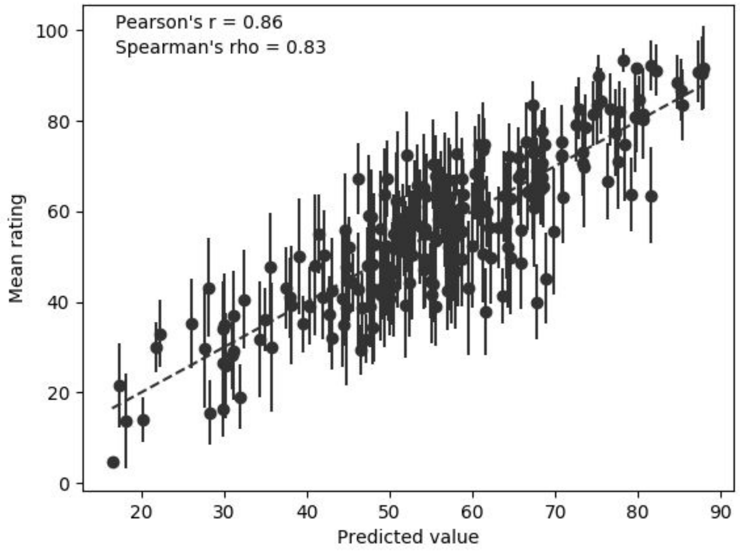


Figure 11 - Brightness model performance with the training dataset.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

5.4 brightness validation

The performance of the updated brightness model was also assessed with the two datasets from Deliverable D5.3: the within-source-type and between-source-type data.

5.4.1 Within-source analysis

Figure 12 below shows the performance of the updated brightness model with the five source types evaluated in D5.3: ambience, bell, piano, swoosh, and voice.

The model predicts listener brightness ratings for all sound source types very well, with the exception of one outlier in the voice source type. Auditioning of this file reveals some high-frequency noise/distortion that would most likely be ignored by listeners, but measured by the models.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

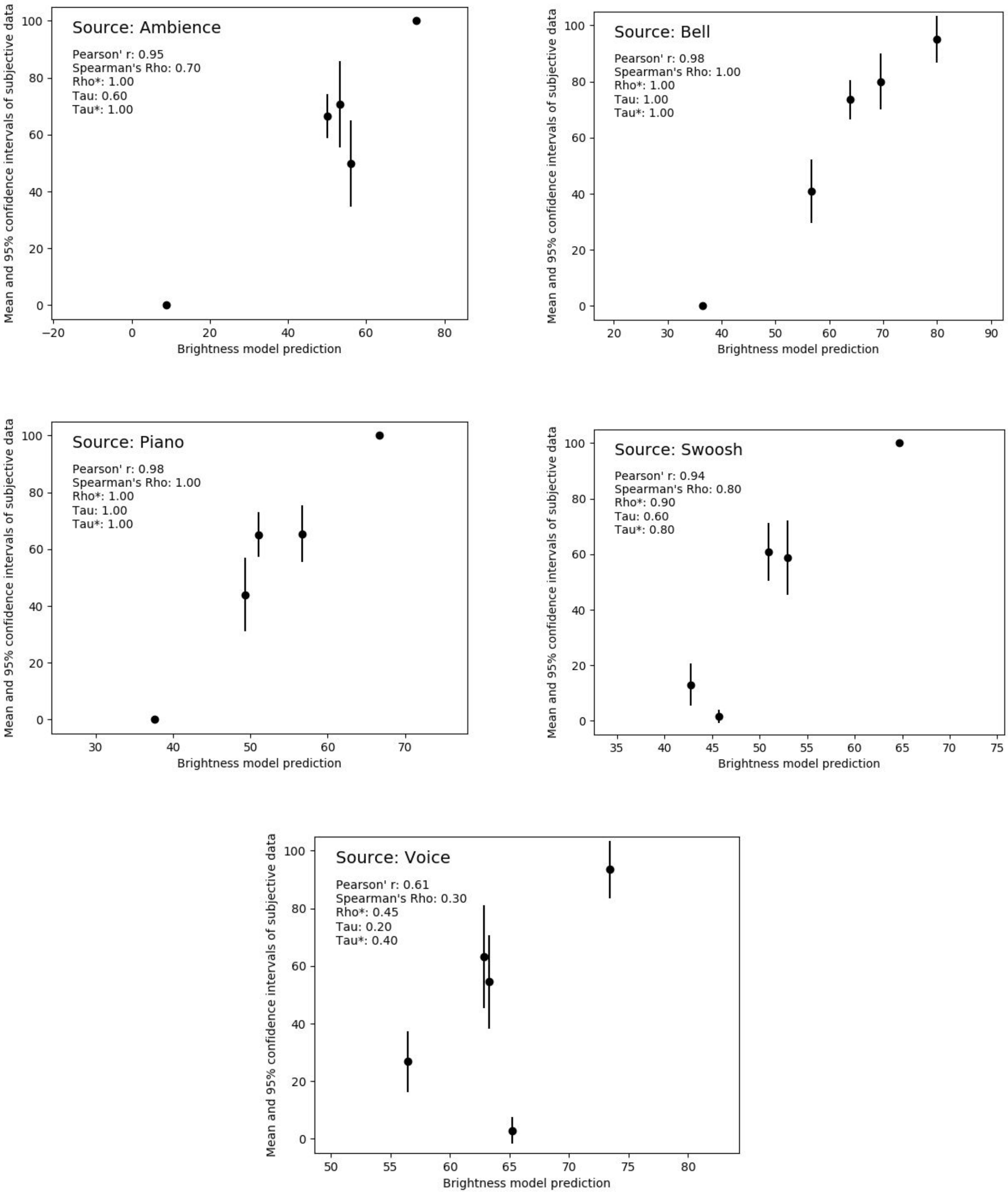


Figure 12 - Within-source-type performance of the brightness model. (a) Ambience; (b) Bell; (c) Piano; (d) Swoosh; (e) Voice.

5.4.2 Between-source analysis

Figure 13 shows the performance of the current brightness model with the between-source-type data. The model performs well, giving linear correlation of r=0.81 with listener ratings. This performance is similar to that achieved with the training dataset (r=0.86).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

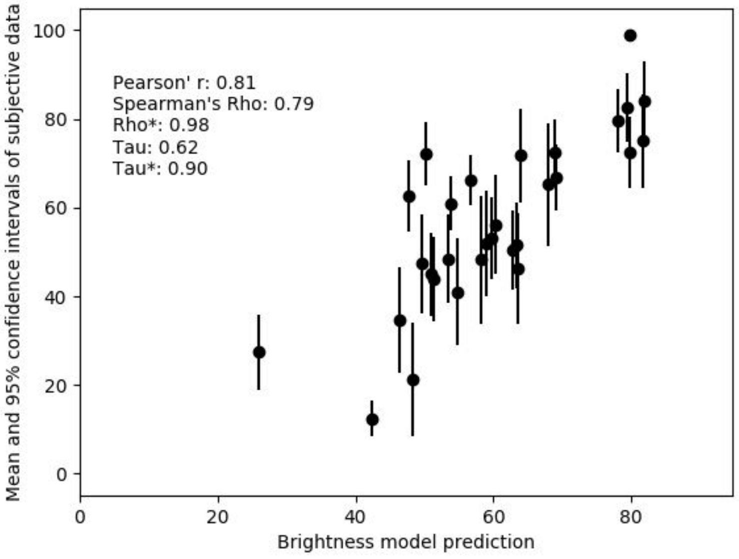


Figure 13 - Between-source-type performance of the brightness model.

5.5 Performance summary

The performance of the brightness model with the within-source-type data is summarised in Table 5.

Table 5: Summary of the brightness model's performance with the within-source-type data.

Source type	Pearson's r	Spearman's Rho	Rho*	Kendall's Tau	Tau*
Ambience	0.96	0.70	1.0	0.60	1.0
Bell	0.99	1.0	1.0	1.0	1.0
Piano	0.98	1.0	1.0	1.0	1.0
Swoosh	0.98	1.0	0.90	1.0	0.80
Voice	0.57	0.30	0.45	0.20	0.40

These results show that the brightness model performs well with all source types with the exception voice. Inspection of this source type reveals one sound that is overpredicted in its brightness. Auditioning this sound reveals some high frequency distortion components that were most likely ignored by listeners but captured by the feature extraction algorithms.

Table 6 compares the performance of the current brightness model against that of the original model with the between-source-type data. From this table it can be seen that the updated model out-performs the original model according to all metrics.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Table 6: Summary of the brightness model's performance with the between-source-type data.

Source type	Pearson's r	Spearman's Rho	Rho*	Kendall's Tau	Tau*
Current model	0.81	0.79	0.98	0.62	0.90
D5.3 implementation	0.69	0.65	0.96	0.49	0.84

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

6 Roughness model

6.1 Roughness training dataset

Listener ratings of roughness were collected in the manner described in Section 2. The roughness dataset comprised 212 audio files over 40 source types. Listening tests were completed by fifteen undergraduate students on the University of Surrey's Music and Sound Recording, all of whom had completed a module of technical listening and had experience participating in listening tests.

Listener performance was analysed in terms of their inter-subject agreement and intra-subject consistency, and no listener performed significantly worse than any other, so all data was retained for analysis.

6.2 Development of the roughness model

The original model of roughness was a direct implementation of the Vassilakis model [Vassilakis; 2007]. This model performed reasonably well on the D5.3 training data. To improve the model, several other implementations of roughness models were also tested, all of which performed worse than the Vassilakis model for the sounds tested [Daniel and Weber, 1997; Cabrera et al., 2008; Vecchi et al, 2016]. Additional metrics based on octave-band decomposition and envelope strength in the 15-300 Hz region were also tested. All of these were found to produce a worse correlation with the data than the standard metric.

Some small feature optimisations were implemented to maximise correlation with the training data. These included changes to the spectral peak-picking algorithm and a logarithmic transformation.

6.3 Performance of the roughness model

The performance of the roughness model with the training dataset is shown in Figure 14. This model gives a linear correlation of r=0.70 with listener ratings and a rank order performance of rho=0.71.

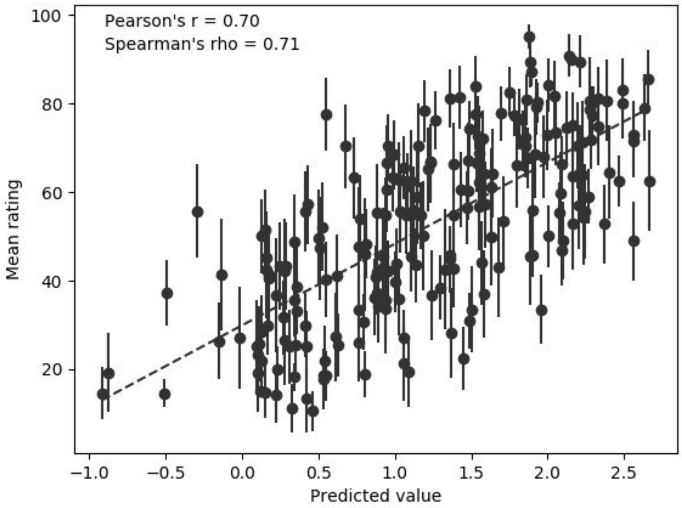


Figure 14 - Roughness model performance with the training dataset.

6.4 Roughness validation

The performance of the roughness model was also assessed with the two datasets from Deliverable D5.3: the within-source-type and between-source-type data.

6.4.1 Within-source analysis

Figure 15 shows the performance of the current model with the five source types evaluated in D5.3: alarm, bell, engine, guitar, and synth. The model produces a good linear correlation with listener ratings for all source types, with r ranging from 0.76 to 0.87. However, it should be noted that this performance is worse than that of the original model for all source types with the exception of the engine, which performed better with the updated model.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

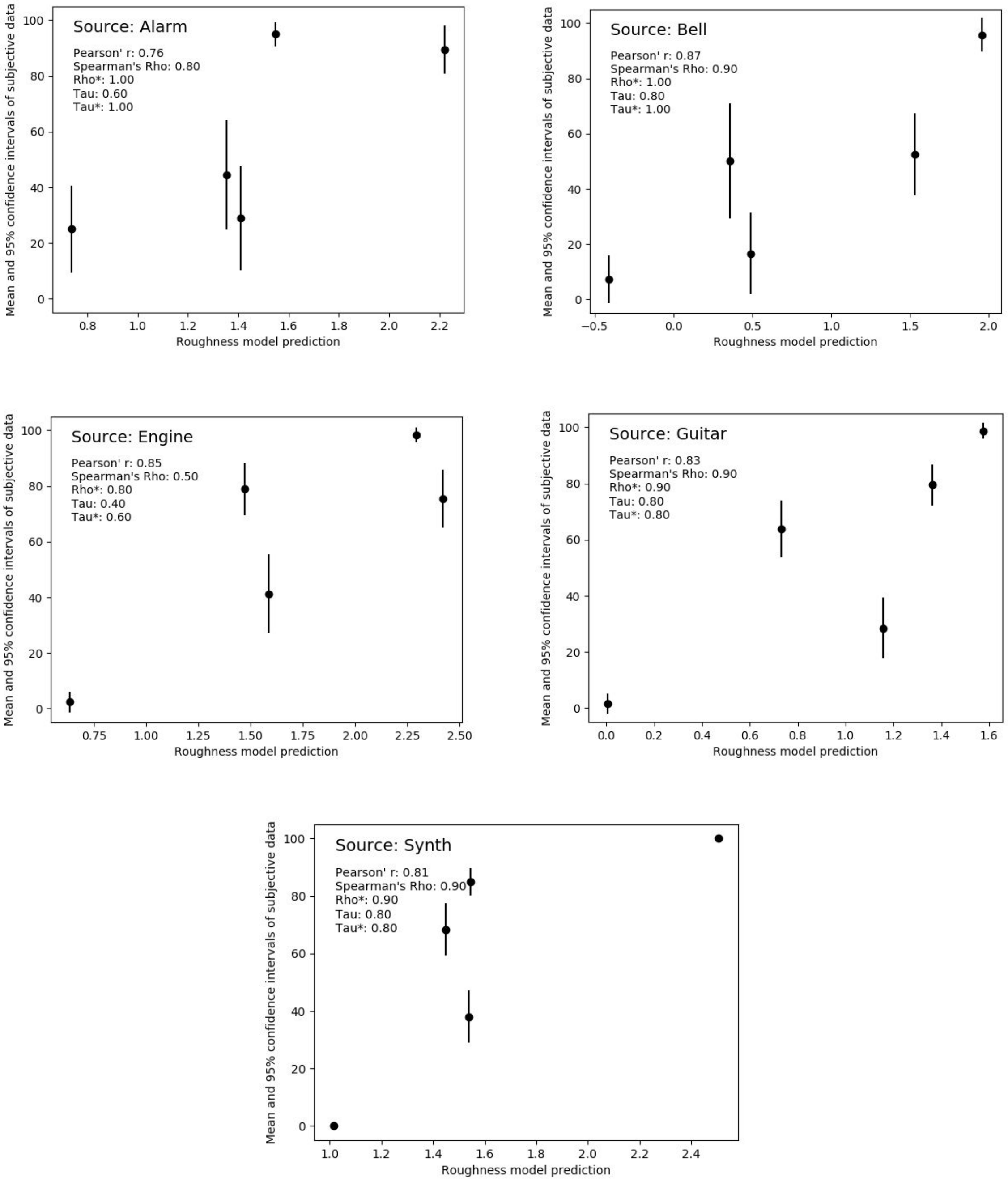


Figure 15 - Within-source-type performance of the roughness model. (a) Alarm; (b) Bell; (c) Engine; (d) Guitar; (e) Synth.

6.4.2 Between-source analysis

The performance of the updated roughness model with the D5.3 data is shown in Figure 16. The updated model achieves a linear correlation of r=0.56 with listener ratings and a rank order performance of rho=0.53.

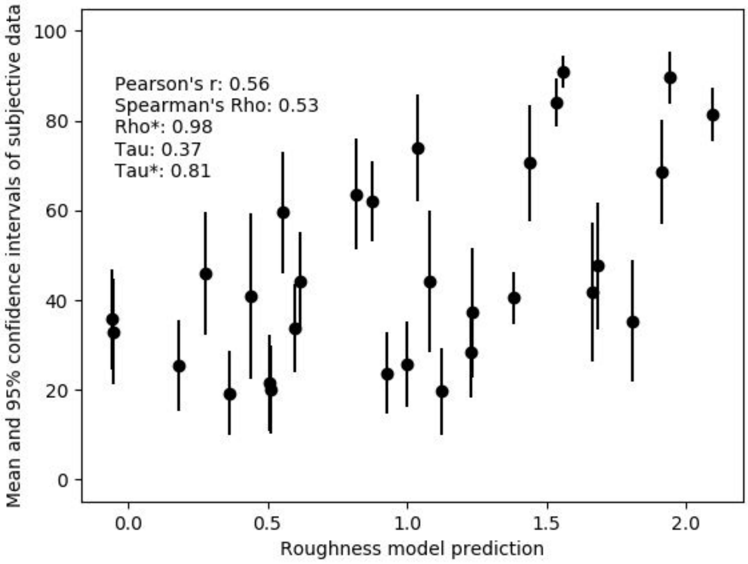


Figure 16 - Between-source-type performance of the roughness model.

6.5 Performance summary

The performance of the roughness model with the within-source-type data is summarised in Table 7.

Table 7: Summary of the roughness model's performance with the within-source-type data.

Source type	Pearson's r	Spearman's Rho	Rho*	Kendall's Tau	Tau*
Alarm	0.76	0.80	1.00	0.6	1.00
Bell	0.87	0.90	1.00	0.8	1.00
Engine	0.85	0.50	0.80	0.40	0.60
Guitar	0.83	0.90	0.90	0.80	0.80
Synth	0.81	0.90	0.90	0.80	0.80

Although the model achieves good linear correlation and good rank order performance for many of the sound source types, this is worse than the original model's performance from D5.3 for all source types with the exception of the engine. This is interesting as the optimisations included to better predict the new training dataset have caused a reduction in performance.

Table 8 compares the performance of the updated roughness model against that of the original implementation with the between-source-type data. From this table it can be seen that the current model does not perform as well. This suggests that the roughness model would benefit from additional development.

Table 8: Summary of the roughness model's performance with the between-source-type data.

Source type	Pearson's r	Spearman's Rho	Rho*	Kendall's Tau	Tau*
Current model	0.56	0.53	0.98	0.37	0.81
D5.3 implementation	0.63	0.60	0.95	0.41	0.85

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

7 Warmth model

The warmth, sharpness, and boominess models were newly added for Deliverable D5.6, and as such, were not evaluated in Deliverable D5.3. Therefore, the performance of these models was evaluated with the newly collected data only.

7.1 Warmth training dataset

Ratings of warmth were collected in the manner described in Section 2. The warmth dataset comprised 185 audio files over 32 source types. Listening tests were completed by thirteen undergraduate students on the University of Surrey's Music and Sound Recording course, all of whom had completed a module of technical listening and had experience participating in listening tests.

7.2 Development of the warmth model

Several features likely to relate to warmth perception were extracted. A subset of features was selected using the iterative modelling method described in Section 3.2, and a multilinear regression modelling method was employed.

7.3 Performance of the warmth model

The developed linear regression model is shown in Figure 17. The model achieves a linear correlation of r=0.79 with the listener ratings and rank order correlation of rho=0.79. This is a reasonable performance, but not as good as that achieved by the hardness, depth, or brightness models.

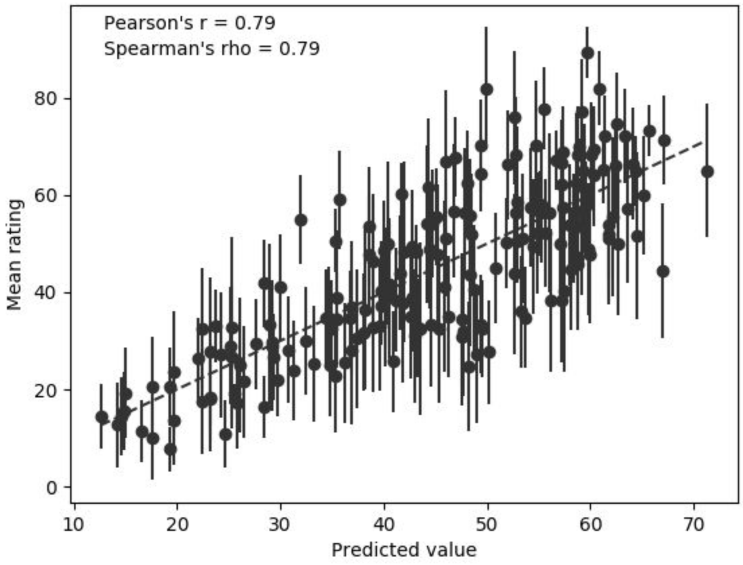


Figure 17 - Warmth model performance with the training dataset.

The warmth model achieves a reasonable performance of r=0.79. This is not as good as that of the models of hardness, depth, or brightness, but still a reasonable performance. There are no specific stimuli that are over- or under-predicted by this model.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

8 Sharpness model

8.1 Sharpness training dataset

Ratings of sharpness were collected in the manner described in Section 2. The sharpness dataset comprised 165 audio files over 29 source types. Listening tests were completed by six undergraduate students on the University of Surrey's Music and Sound Recording course, all of whom had completed a module of technical listening and had experience participating in listening tests.

8.2 Implementation of the sharpness model

The sharpness model is a direct implementation of the Klippel sharpness model [Fastl and Zwicker, 1991; Churchill, 2004]. The new dataset was used for validation only, not training.

8.3 Performance of the sharpness model

The performance of the sharpness model is shown in Figure 18. The model achieves a linear correlation of r=0.78 with listener ratings, and rank order performance of rho=0.83. From visual inspection of this plot, it seems there may be a nonlinear relationship between the sharpness ratings and model performance. Additionally, since the model does not contain a linear regression component, the scale bounds are unknown.

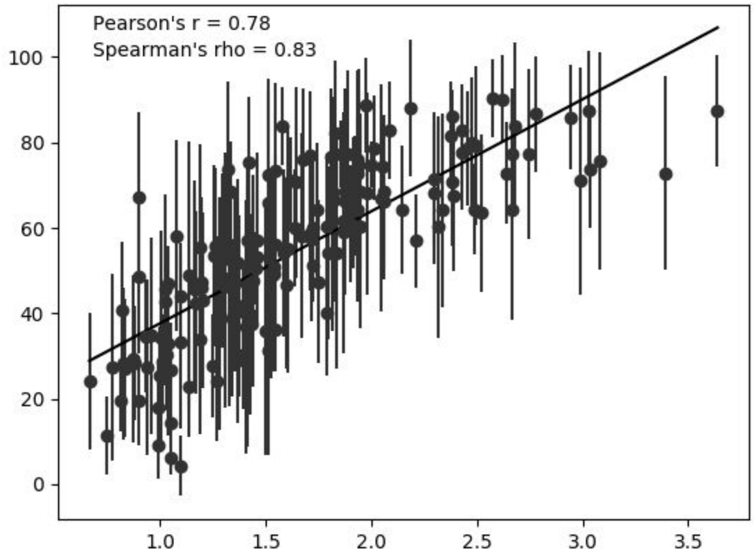


Figure 18 - Sharpness model performance with the training dataset.

9 Boominess model

9.1 Boominess training dataset

Ratings of boominess were collected in the manner described in Section 2. The boominess dataset comprised 102 audio files over 20 source types. Due to the smaller number of stimuli, tests were completed in a single test session.

Listening tests were completed by six undergraduate students on the University of Surrey's Music and Sound Recording course, all of whom had completed a module of technical listening and had experience participating in listening tests.

9.2 Implementation of the boominess model

The boominess model was a direct implementation of the Hatano and Hashimoto's boominess index [Hatano and Hashimoto, 2000]. As with sharpness, ratings were collected from six listeners only, to validate (but not train) this model.

9.3 Performance of the boominess model

The performance of the boominess model is shown in Figure 19. This model achieves a linear correlation of r=0.67 with listener ratings and rank order performance of rho=0.67. This is the worst performing of the timbral models. However, it should be noted that the confidence intervals on the data are very large, suggesting that subjects may have been unsure of their ratings of boominess.

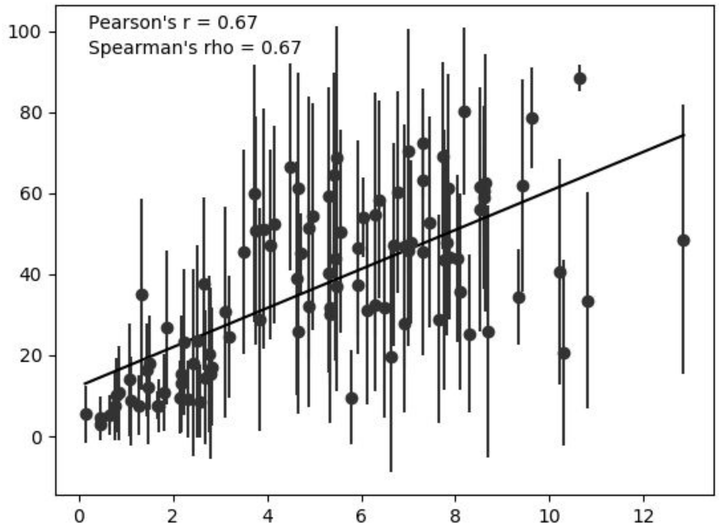


Figure 19 - Boominess model performance with the training dataset.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

10 Reverb model

Most existing models of perceived level of reverberation require the impulse response for prediction. The designed model, however, predicts the perceived levels of reverberation from audio recordings as input alone. Unlike the timbral models described in previous sections, the presented model of reverb is a two-class classification model, predicting stimuli as having no perceived reverberation, or some perceived reverberation.

10.1 Reverb training dataset

The training dataset for the reverb comprised 400 stimuli taken from the hardness and depth datasets, described in Sections 3.1 and 4.1. Each of these 400 stimuli was auditioned and categorised into one of two classes: none or low level of perceived reverberance; or high level of perceived reverberance.

Due to the size of the dataset, a three-fold cross validation approach was taken. In this, the dataset is randomly split into three folds of approximately equal size. Two of these fold are grouped together and used for training of the model, and the third fold used for evaluating the performance. This process is repeated three times so that each fold is utilized for evaluating the performance.

10.2 Implementation of the reverb model

Seven feature extraction algorithms were found suitable for the prediction of perceived level of reverberation: RT60; level of foreground stream (LFS); level of background stream; interaural time differences fluctuation in the background (ITDb); interaural time differences fluctuation in the foreground (ITDf); level of low frequency stream (LLS); and reverberance (REV). RT60 is blindly calculated by the method proposed by Vandrop et al. [2013], and the other six features extracted using the nonlinear auditory model introduced by Jan et al. [2012]. Further detail regarding the implementation of these features are documented by Savavi et al. [2018].

10.3 Performance of the reverb model

As discussed in Section 10.1, a three-fold cross validation approach was taken to model training and validation. Three distinct approaches to binary classification were used for modelling of the perceived reverberance: multinomial logistic regression, decision tree, and multilayer perceptron (MLP). Table 9 shows the summarised results with each of these modelling methods. The performance is the mean performance from each fold. These results indicate that the best performing method is the multilayer perceptron classifier, achieving a mean accuracy of 75.25%.

Table 9: Prediction accuracy of the reverb model.

Logistic regression	Decision tree	MLP
67.75%	72.75%	75.25%

To test the relative importance of each feature in the MLP model, an MLP model was created using each feature. The prediction accuracy for each feature is shown in Table 10. The feature with the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

highest prediction accuracy is the RT60, with a prediction accuracy of 63.75%, followed by REV with an accuracy of 58.75%.

Table 10: Prediction accuracy for MLP models of each reverb-related feature.

Feature	Prediction accuracy
RT60	63.75 %
LFS	54.00 %
LBS	49.50 %
ITDf	50.25 %
ITDb	51.25 %
LLS	57.00 %
REV	58.75 %

11 Conclusion

In this deliverable, large training datasets comprised of listener ratings of audio stimuli of multiple sound source types were collected for the timbral attributes of hardness, depth, brightness, roughness, warmth, sharpness, and boominess. These datasets were used to retrain the models of hardness, depth, brightness, and roughness, as well as to assess the performance of the newly added models of warmth, sharpness, and boominess.

The updated hardness, depth and brightness models all showed improvement over the original models, and all performed better than the originals when re-evaluated on the data from Deliverable D5.3.

The updated roughness model performed less well with the new dataset than the original model did with the D5.3 data. However, the original model performed worse than the updated model with the new dataset. Further work on the roughness model might improve its performance.

The newly added model of warmth performed reasonably well, with performances nearing that of the hardness, depth, and brightness models. The sharpness model also showed good performance, although it appeared there was a nonlinear relationship between the listener ratings and the model's predictions. This can be examined further in the future.

The boominess model performed the worst of the three new models. However, it should be noted that this model was a direct implementation from literature that was designed to identify the booming sensation of road noise when in a car. The current use case is different and so the model may require some modifications to work effectively. Additionally, the ratings of boominess had the largest confidence intervals of all datasets, implying that there was a significant degree of disagreement between listeners when rating boominess.

The reverb model has been redesigned as a classification model. Therefore, it is not possible to compare the performance of the initial linear regression model to this classification model. This updated model achieves a prediction accuracy of 75.25 %.

All datasets are publically available at DOI: 10.5281/zenodo.1697212.

12 References

Cabrera, D., Ferguson, S., and Schubert, E., 2008: 'PsySound3: an integrated environment for the analysis of sound recordings', Acoustics 2008, Proceedings of the Australian Acoustic Society Conference, Geelong, Australia.

Churchill, C, 2004: 'MATLAB Codes, Calculating the Metrics', Salford Innovation Research Centre, https://www.salford.ac.uk/research/sirc/research-groups/acoustics/psychoacoustics/sound-quality-making-products-sound-better/accordion/sound-quality-testing/matlab-codes.

Daniel, P., and Weber, R., 1997: 'Psychoacoustic roughness: Implementation of an optimized model' Acustica, No. 83, pp.113-123.

Fastl, E., and Zwicker, H., 1991: 'Psychoacoustics, Facts and Models, Springer.

Hatano, S., and Hashimoto, T., 2000: 'Booming index as a measure for evaluating booming sensation', 29th International congress and Exhibition on Noise Control Engineering.

Jan, T., and Wang, W., 2012: 'Blind reverberation time estimation based on Laplace distribution', EUSIPCO, pp: 2050-2054, Romania.

Pearce, A., Brookes, T., and Mason, R., 2016: 'Deliverable D5.1 – Hierarchical ontology of timbral semantic descriptors', available: http://www.audiocommons.org/materials/.

Pearce, A., Brookes, T., and Mason, R., 2017a: 'Deliverable D5.2 – First prototype of timbral characterisation tools for semantically annotating non-musical content', available: http://www.audiocommons.org/materials/.

Pearce, A., Brookes, T., and Mason, R., 2017b: 'Deliverable D5.3 – Evaluation report on the first prototypes of the timbral characterisation tools', available: http://www.audiocommons.org/materials/.

Pearce, A., Safavi, S., Brookes, T., Mason, R., Wang, W., and Pumbley, M., 2018: 'Deliverable D5.6 – Second prototype of timbral characterisation tools for semantically annotating non-musical content', available: http://www.audiocommons.org/materials/.

Pearce, A., Isabelle, S., Francois, H., and Oh, E., publication pending: 'Methods of Assessing the Rank Order of Prediction Models with Respect to Variance of Listening Test Ratings', Journal of the Audio Engineering Society, Publication accepting, awaiting print.

Safavi, S., Choobbasti, A., Wang, W., Plumbley, and M., Fazekas, G., 2018: 'Predicting the perceived level of reverberation using features from nonlinear auditory model', ISAI (FRUCT), Italy.

Van Drop, J., Vries, D., and Lindau, A., 2013: 'Deriving content-specific measures of room acoustic perception using a binaural nonlinear auditory model', The Journal of the Acoustical Society of America, Vol. 133, pp. 1572–1585.

Vassilakis, P., 2007: 'Sra" A web-based research tool for spectral and roughness analysis of sound signals', in Proceedings of 4th Sound Music Computing (SMC), pp. 319–325.

Vecchi, A., Leon, R., and Kohlrausch, A., 2016: 'Modelling the sensation of fluctuation strength', Proceedings of the 22nd international congress on acoustics, Buenos Aires.