Deliverable D5.3

Evaluation report on the first 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 July 2017 (M18)
Submission date	30 February 2016 (M18)
Report availability	Public (X), Confidential ( )
Deliverable type	Report (X), Demonstrator ( ), Other ( )
Task leader	Surrey
Authors	Andy Pearce, Tim Brookes, and Russell Mason.
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

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 D5.2 [Pearce et al., 2017].

The timbral models are intended to be used as a method of filtering search results, and as such the assessment method was designed to reflect this. Firstly, five suitable source types were identified for each modelled timbral attribute. These were identified by examining one month's worth the search history of freesound (from April 2016) and finding sources types that are commonly searched for in conjunction with the timbral attribute. For each source type, five stimuli of each attribute/source type combination were identified that differed in the attribute. For example, the attribute hardness is commonly used to search for kick drum samples, and as such five kick drum samples were identified that differ in their perceived levels of hardness.

Subjective data was then obtained by running listening tests where subjects rated each stimulus using each subjective attribute, both within a single source type, and across all sources types. The ratings were then compared to the predicted values from the timbral models to assess the performance of each.

This analysis revealed that the models of metallic-nature and reverb do not predict the subjective data very well, with average Spearman's Rho values (calculated over all source types) of 0.006 and 0.11 respectively. These require further work before implementation. The hardness model performed reasonably well (Rho = 0.54) but the subjective data had substantially more variance than the ratings of other attributes. The models of depth, brightness, and roughness all performed well (Spearman's Rho of 0.78, 0.74, and 0.84 respectively).

The findings presented in this report will guide the future development of timbral models within WP5.

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 described a series of experiments that identified the timbral attributes that are used to describe the differences in sound effect and their relative usage. It was considered that the most used attributes would be the most usefully modelled, as these are likely to add the most value to end-users.

A literature review was then conducted for each attribute in order of their relative use, assessing if there were any existing models of the attribute or at least an indication of the objective parameters that could be extracted from an audio signal to model the attribute. From this, six timbral attributes were identified as suitable and models were coded: hardness, depth, brightness, metallic-nature, reverb, and roughness. The implementation of these models is documented in Deliverable D5.2 [Pearce et al., 2017].

This deliverable documents the acquisition of suitable listening test data, and using this to assess the performance of each of the six timbral models.

1 Introduction

1.1 Main objectives and goals

In Deliverable D5.2, the development and operation of six prototype timbral characterisation models was described. These models predicted the timbral attributes of hardness, depth, brightness, metallic-nature, reverb, and roughness. These were all either implementations from literature, or developed based on a small informal listening test. None of the models have yet been validated.

This deliverable describes the experiments conducted to evaluate the performance of these models, identifying which models perform well, if any, and which models require further work.

1.2 Methodology

There are no existing databases that give scalar data for each of the timbral attributes that have been modelled. Therefore, a suitable corpus of audio stimuli needed to be compiled, listening tests conducted on this set of stimuli, and finally the performance of each timbral model assessed against this subjective data.

The developed timbral models are intended to be used to filter the search results of a specific source type. For example, users of the Audio Commons ecosystem (ACE) may search for samples of a snare drum, and wish to rank order these samples by their relative hardness. In this respect, the model should be able to correctly predict the rank order of the perceived hardness of each source type individually. The models should be evaluated in this manner, akin to the intended use: evaluating the rank order performance of a model with multiple stimuli of a single source type, and the source types should be selected that are representative of each modelled timbral attribute.

The suitable source types for assessing each timbral model were obtained by analysing one month's worth of the freesound search history (from April 2016), to identify sources for which each timbral attribute was commonly searched for, as documented in Section 2.1. Listening tests were then implemented in order to acquire suitable subjective data that could be used to evaluate the performance of the timbral models. The test design and implementation are discussed in Section 2.2. The performance of each timbral model was then evaluated using the relevant subjective data, shown in Section 3 – 8.

1.3 Terminology

AudioCommons: reference to the EC H2020 funded project AudioCommons, with grant agreement nr 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

To assess the performance of the timbral models, a suitable dataset is required: subjective ratings of each of the modelled timbral attributes using stimuli typical to the intended usage of the models. Unfortunately, no existing datasets are suitable for assessing the developed timbral models, and as such, a new dataset needs to be developed. This section of the deliverable describes the acquisition of a suitable dataset to evaluate the performance of the timbral models.

As discussed in Section 1.2, the metadata added by the timbral models is most likely to be used as a filter to sort sound effects of a specific source type. Listening test data was therefore collected for each attribute by directly comparing several stimuli of the same source type.

Additionally, it is beneficial to test the generalisability of the timbral models; if a model performs well both within and across multiple source types, then it is likely that the model will generalise to source types not considered in the evaluation. To test this, subjective data was obtained for each timbral attribute by comparing across multiple different source types.

This section is split into two subsections. Subsection 2.1 describes the acquisition of a suitable corpus of audio stimuli. Subsection 2.2 then documents the listening tests that were conducted to obtain timbral ratings for each of the audio stimuli.

2.1 Suitable stimuli

The developed timbral models are intended to add metadata to Creative Commons licenced audio repositories, such as freesound.org. However, it is not feasible to run controlled listening tests to assess the timbral characteristics of the 349,058 audio files currently hosted on freesound. Therefore, a smaller selection of audio stimuli is required that will be representative of the larger dataset.

This smaller dataset was developed by first identifying source types that that are commonly searched for with each timbral attribute (§2.1.1). A random selection of these source types was then downloaded (§2.1.2), followed by a removal of inappropriate stimuli (§2.1.3) and finally conducting experiments with an independent expert to select a suitable corpus of stimuli that are known to vary in each of the timbral attributes (§2.1.4).

2.1.1 Attribute source types

For each of the six modelled timbral attributes, one month’s worth of the freesound search history (from April 2016) was used to identify the sources that are the most commonly searched for in conjunction with each attribute. This was done by identifying the most-searched phases for the timbral terms that constitute each attribute. For example, the hardness attribute was found in Deliverable D5.1 to be a combination of *hard*, *soft*, and *pillowy*. Although there were no freesound searches which matched the term pilloy, the 10 most-searched phrases containing the terms hard or soft are shown in Table 1.

Table 1: Most-searched terms for the hardness attribute

Search ranking	Hard	Soft
1	hard kick	soft wind
2	hard hit	soft piano
3	hard snare	soft rain
4	hard punch	wind soft
5	hard bass	soft hit
6	hard kicks	soft thud
7	bass hard	soft kick
8	hard drum	soft impact
9	hard piano	soft breeze
10	hard e-guitar	soft whoosh

From Table 1, the sources of kick, piano, and hit are within the top ten most searched terms for both the hard and soft terms. Although the sources of kick and piano are clearly specific types of sources, the source hit does not refer to a specific source. To identify a more useful specific source type, freesound.org was searched for 'hit'. The results when ordered by relevance contained multiple source types; however, the most common was cymbal hits. Therefore, the source type of cymbal was substituted for hit. Two other sources of snare and guitar were also selected to represent the hardness attribute.

For each timbral attribute, five source types were selected in this manner to give a reasonable range of stimuli with different spectral and temporal characteristics: identifying specific source types that represent the most common searches for each of the timbral terms. Each of the selected sources are shown in Table 2, along with a description of the intended source.

Table 2: Selected source types for each timbral attribute

Timbral attribute	Selected source types	Description of intended source
Hardness	Kick Piano Cymbal (Hit) Snare Guitar	A kick drum. A piano. A cymbal being hit. A single snare drum hit, does not include snare rolls. A guitar.
Depth	Bass Kick Drum loop Voice Impact	A synthesiser playing bass sounds. A kick drum. A drum loop. A human voice. The sound of a real-life object being struck, does not include

		synthetic, percussion instruments, or musical sounds.
Brightness	Ambience Piano Bell Voice Swoosh	Synth ambient sounds and keyboard pads. A piano. A bell. A human voice. Sounds that have a 'swoosh'-like characteristic.
Metallic-nature	Clanging Footsteps Creaking Scrape Whoosh	Percussive sounds that have a 'clang'-like characteristic, not including synthetic, percussion instruments, or musical sounds. The sound of footsteps. Sounds that have a 'creaking' characteristic. Sound that sound like they have been generated by scraping an object. Sounds that have a 'whoosh'-like characteristic.
Reverb	Snare Atmosphere Hit Footsteps Vocal	A single snare drum hit, does not include snare rolls. Synth ambient sounds and keyboard pads. The sound of a real-like object being hit, does not include synthetic, percussion instruments, or musical sounds. The sound of footsteps. A human voice.
Roughness	Bell Synth Engine Alarm Guitar	A bell. Synth pads. The sound of a car/bike engine. The sound of an alarm. A guitar.

2.1.2 Acquisition of 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 each of the relevant timbral terms. For example, when searching for kick drum samples to assess the hardness attribute, three searches of 'kick', 'hard kick', and 'soft kick' were made. For each search, 50 of the results were randomly selected and downloaded.

The code was written in a way to ensure that no sounds of the same freesound ID were downloaded. However, this does not necessarily exclude the potential of downloading duplicates, as users could have uploaded the same sample multiple times, each with a unique freesound ID.

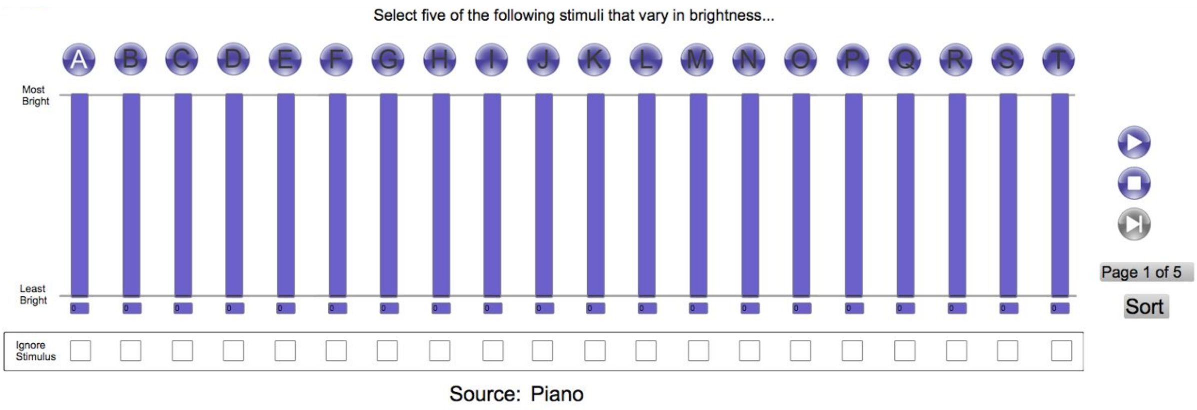
2.1.3 Removal of inappropriate files

Using the automated randomised downloading method, there is no guarantee that each of the downloaded samples will relate to the intended source, described in Table 1. To overcome this, each of the downloaded files was manually inspected and auditioned by one author and any stimuli that did not match the description of the intended source was removed.

2.1.4 Expert selection

Assessing the timbral characteristics of 20 stimuli per attribute/source type is too large of a task for the number of source types used, and as such a more manageable number of stimuli was required. It was deemed that five stimuli per attribute/source type would give a listening test of appropriate length whilst still allowing a reasonable resolution across a wide range of each attribute scale. However, randomly selecting only five of the 20 retained sounds would not guarantee that the stimuli vary on the given timbral attribute. To better assess the models, the stimuli should have a relatively even distribution across the majority of the range of the attribute.

An independent expert, who was experienced in both participating in listening tests and semantic description of audio, was asked to manually assess the sounds for each attribute/source type, and select five that vary in the given attribute. The independent expert was presented with an interface as shown in Figure 1. On each page were the 20 retained stimuli for a given attribute/source type. The expert was asked to select the stimuli that represents the most and least of the attribute, as well as three other stimuli that have a relatively even distribution across the attribute. For example, with the snare source type and hardness attribute, the expert was asked to select the hardest and least hard stimuli, as well as three other stimuli that would be approximately equally spaced on a hardness scale.



Select five of the following stimuli that vary in brightness...

Source: Piano

Figure 1 - Test interface for selection of appropriate stimuli

The results of this experiment were the final corpus of stimuli: for each of the six timbral attributes, five source types were selected; and for each of these source types five stimuli were selected that vary in the given timbral attribute, giving 150 stimuli in total.

2.2 Listening test design

As stated in the introduction to this section, there were two methods of getting listening test data: direct comparisons between stimuli of a single source type (Section 2.2.1); and comparisons across multiple different sources (Section 2.2.2).

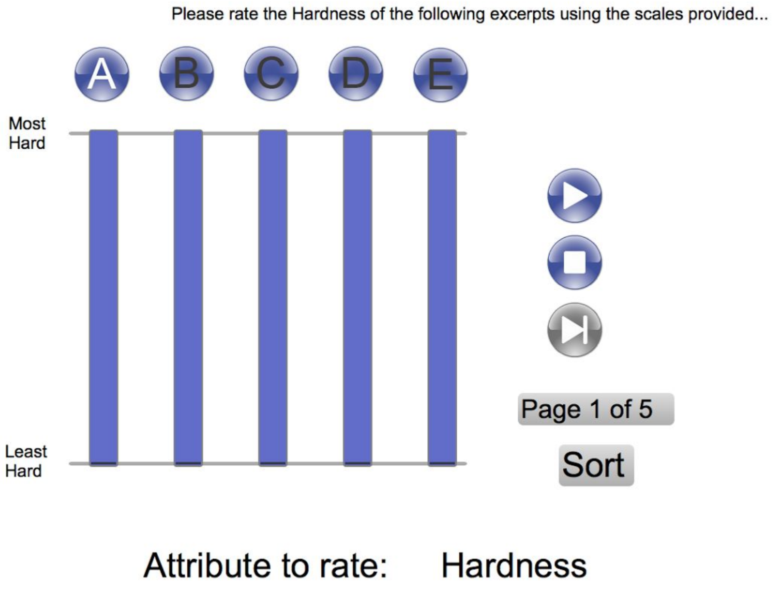
To prepare the stimuli for listening tests, each of the selected stimuli were edited to be of an appropriate length for a listening test (no more than 20 seconds) and converted to 44.1kHz wav files for consistency. Each stimulus was then loudness matched using the Nugen Audio LM Correct to -35.2 LUFS, the lowest loudness level of all normalised stimuli.

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

All listening test were conducted in an acoustically treated editing room, using Neumann KH120A active studio monitors. The playback system was aligned to produce a level of 74 dB_{SPL} at the listening position with -14 dBFS pink noise. This produced a comfortable listening level when the stimuli were reproduced.

2.2.1 Within source type ratings

Sixteen subjects participated in the listening tests comparing stimuli of a single source. Participants were undergraduate students on the Tonmeister Sound Recording course at the University of Surrey, all having technical ear training and experience in listening tests. The tests were split into six sessions, with a separate timbral attribute in each session. The order of the sessions for each participant was randomised. An example of the test interface is shown in Figure 2.


Please rate the Hardness of the following excerpts using the scales provided...

Attribute to rate: Hardness

Figure 2 - Within source type listening test interface

In each session, subjects were asked to rate the relative levels of the assessed timbral attribute for the five stimuli of the same source type. On each page, subjects could audition the stimuli as many times as required, rate each stimulus, and rearrange the ordering of the stimuli into ascending order based on their ratings. Subjects were asked to use the full range of the scale (ensuring that one stimulus was rated at the bottom of the scale and another stimulus rated at the top of the scale), and as such, the scale is not consistent across sources.

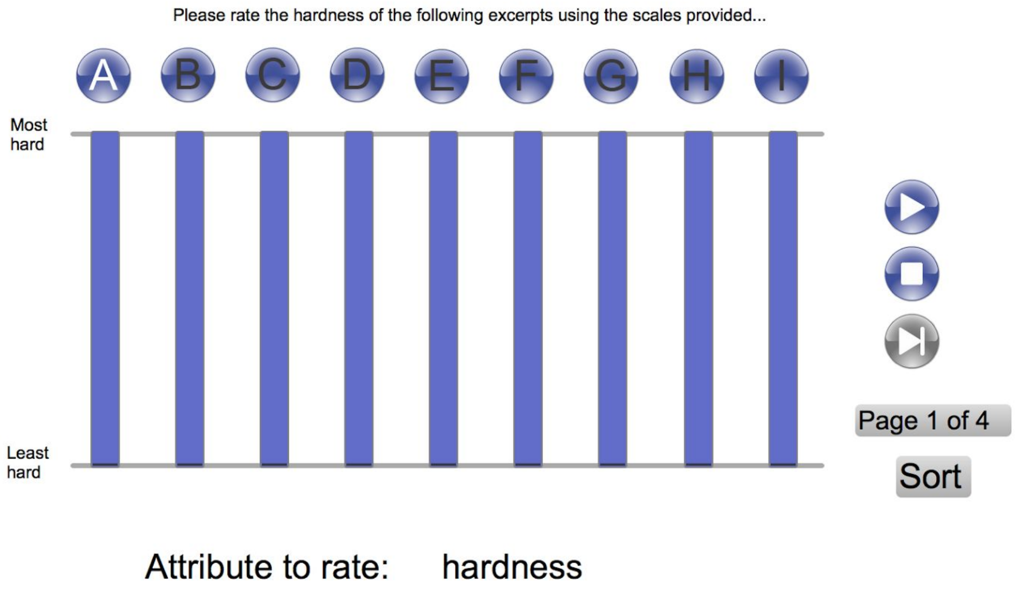
2.2.2 Between source type ratings

After the within source type ratings test had been completed, the results were averaged across all users, and the median stimulus from each attribute/source type was taken. Of these 30 stimuli, an independent expert was asked to select the pair of stimuli representing the extremes for each attribute. These stimulus pairs were used as hidden anchors on each page of the between source type listening tests.

Once again, each test session consisted of a single timbral attribute, the order of which was randomised for each subject. Before each test, subjects were presented with a familiarisation task.

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

The subjects were presented all 30 stimuli and asked to consider the range of the attribute being assessed within the stimuli so that any subsequent ratings of this attribute can be made with respect to this full range. After subjects reported being comfortable with the range of stimuli, they were then presented the test interface as shown in Figure 3. On each page the two hidden anchor stimuli and seven other random stimuli were presented. Subjects were asked to rate the stimuli with respect to the full range of the timbral attribute as heard during the familiarisation stage.



Please rate the hardness of the following excerpts using the scales provided...

Attribute to rate: hardness

Figure 3 - between source type listening test interface

2.3 Summary

For each of the six timbral attributes that have been modelled, five source types were identified as suitable for the assessment of the attribute. For each of these 30 source types, a randomised downloading method followed by manual filtering was used to select five stimuli that varied in the given attribute.

Two types of listening tests were then conducted on these stimuli: within source type; and between source types. The results of each of these experiments can then be used to evaluate the performance of the timbral model in terms of their ability to (i) predict the rank order of a single source type, and (ii) predict the rank order between multiple source types.

3 Hardness model evaluation

In this section, the performance of the hardness model is evaluated on both the within source type data (Section 3.1) and the between source types subjective data (Section 3.2). For each of these assessment methods, first an analysis of the listening test results is performed, identifying any listeners who disagreed with the group. The methods used to perform this analysis, as well as criteria for the removal of outlying subjects are described in this section only, but utilised in the assessment of each timbral attribute.

3.1 Within source type analysis

In this section, the performance of the hardness model is evaluated with the results from the within source type listening test. Prior to this evaluation, the within source type listening test data was assessed to potentially remove any outlier listeners.

3.1.1 Assessment of participants

Since no repeats were included in the design of the listening tests, the performance of the subjects was assessed using the inter-subject agreement. This is a measure of the correlation of each subject's results to the mean of all subjects. Typically, this correlation takes the form of the Pearson's correlation coefficient, *r*.

Figure 4 below shows the correlation of each subject's responses across all five pages to the mean of all subjects. The solid red line on this plot represents the grand mean of all subject's agreement scores, and the dashed red line is one standard deviation below the grand mean. In previous experiments [Pearce, 2017; Francombe, 2014], subjects were considered unreliable if a subject's agreement is less than one standard deviation below the grand mean (dashed red line). However, the overall level of agreement should also be considered. Evans [1996] gives guidelines for verbally describing the strength of correlation as r = 0.8 – 1.0 as *very strong*. Therefore, even if a subject's agreement score is less than one standard deviation below the grand mean, their results will not be excluded unless their agreement with other participants is not *very strong*, i.e. also has an agreement below 0.8.

As shown in Figure 4, subjects 3, 8, and 12 all meet this exclusion criteria and were removed from subsequent analysis. In addition to this, it was noted that the grand mean of the inter-subject agreement was low, shown by the grand mean of 0.576. This low overall agreement may have been caused by subjects disagreeing on what factors relate to hardness, potentially having several different groups of subjects rating hardness differently, or potentially rating differently for each source.

To test the effect of the source on the ratings, the inter-subject agreement was calculated for each source individually. The results showed poor agreement for each source, with no one source in particular having worse overall agreement than others. It is interesting to note that the source of snare and piano had several subjects whose agreement scores were negative. These were subjects 3, 8, 6, and 12, three of whom were planned for exclusions from the listening tests.

To test if there were multiple groups of subjects rating hardness differently, agglomerative hierarchical clustering was performed on the data. This is a form of cluster analysis, where a dendrogram can be created to show the distances between groups of subjects. This dendrogram is shown in Figure 5.

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

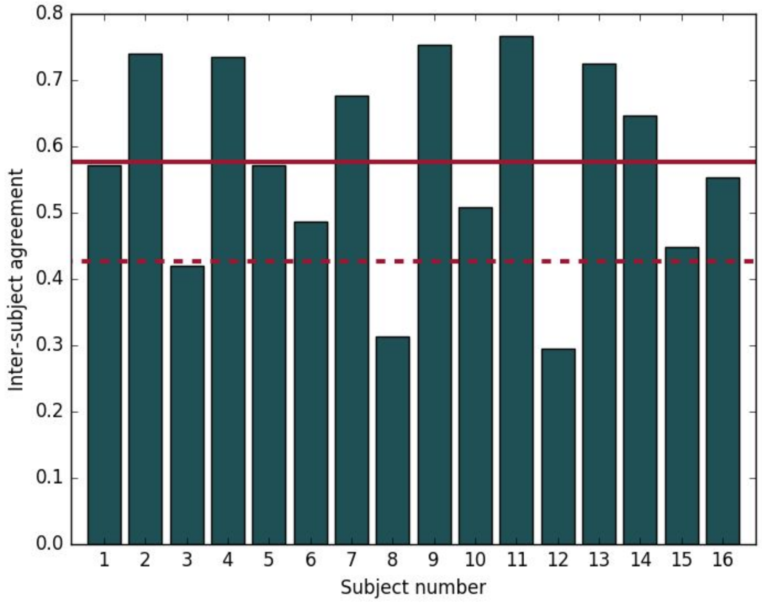


Figure 4 - Inter-subject for the hardness attribute.

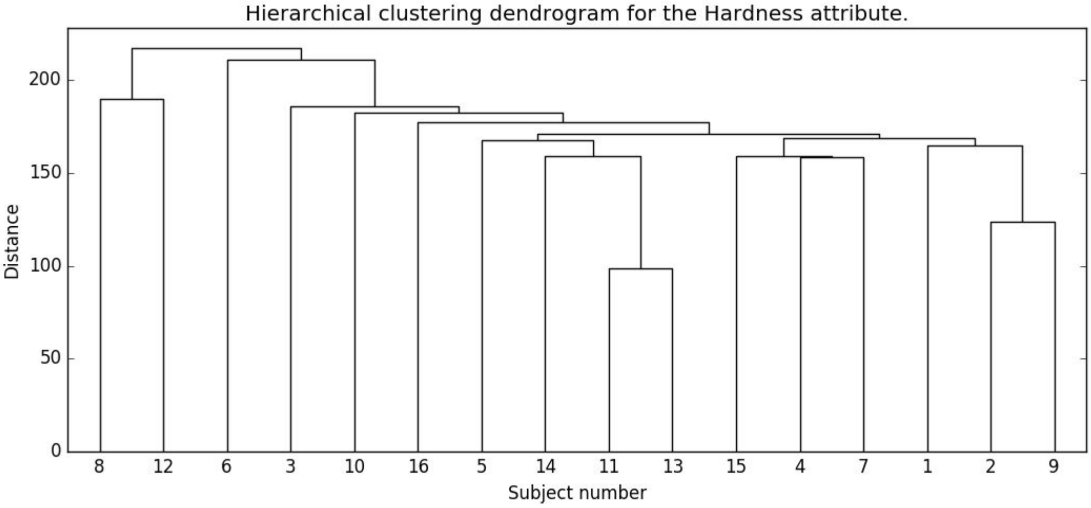


Figure 5 - Dendrogram of subject's differences for the hardness attribute.

From the dendrogram, it can be seen that subjects 8 and 12, identified as having low inter-subject agreement, are in a separate group split at the first clade. However, the distances within the dendrogram are all fairly consistent, with the exception of subjects 11 & 13, and 2 & 9 appearing to be in separate groups. Since this does not coincide with the subjects with low agreement, it was considered that the results could not be put in separate groups.

3.1.2 Model performance - Kick

Figure 6 below shows the scatter plot of the mean and 95% confidence intervals for the kick source type against the hardness model's predicted values.

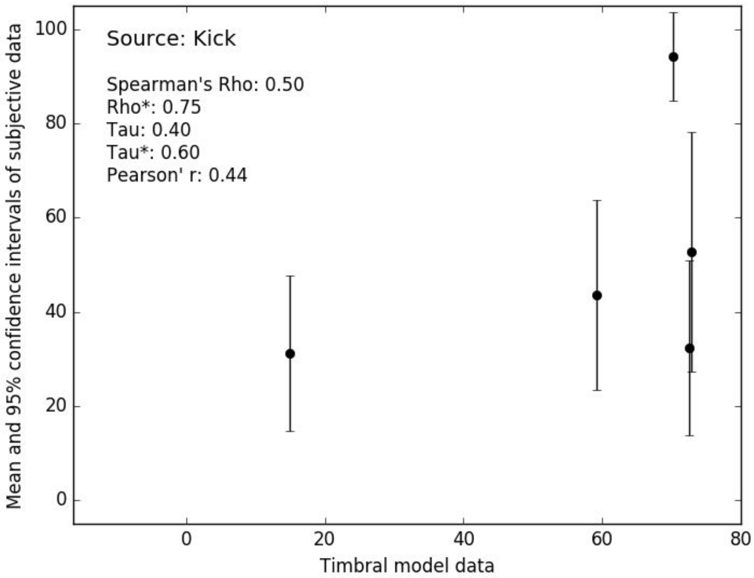


Figure 6 - Evaluation of the hardness model with the kick source type.

With this source type, the model does not fit the subjective data very well. Since the within source type subjective data was collected to span the full scale, the differences in absolute between the subjective data and the model were not of interest, meaning statistical measures of fit such as the Root Mean Square Error (RMSE) were not suitable. For this dataset, it is more important to assess the rank order of the model.

Rank order is commonly assessed with two statistical tests: Spearman's Rho and Kendall's Tau. Both give an indication of the similarity in rank order between two datasets. However, these statistical measures do not account for any variance within either datasets, which is common with subjective data. As such, two new metrics that have been developed, the Epsilon-insensitive Spearman's Rho (Rho*), and the Epsilon-insensitive Kendall's Tau (Tau*), have been used to assess the rank order with respect to the variance in the subjective data.

Looking at these rank order metrics, the model performs fairly poorly, with a Rho of 0.5, only improving to 0.75 when considering the subjective variance (Rho* = 0.75).

3.1.3 Model performance - Piano

Figure 7 shows the performance of the hardness model against the subjective data for the piano source type data.

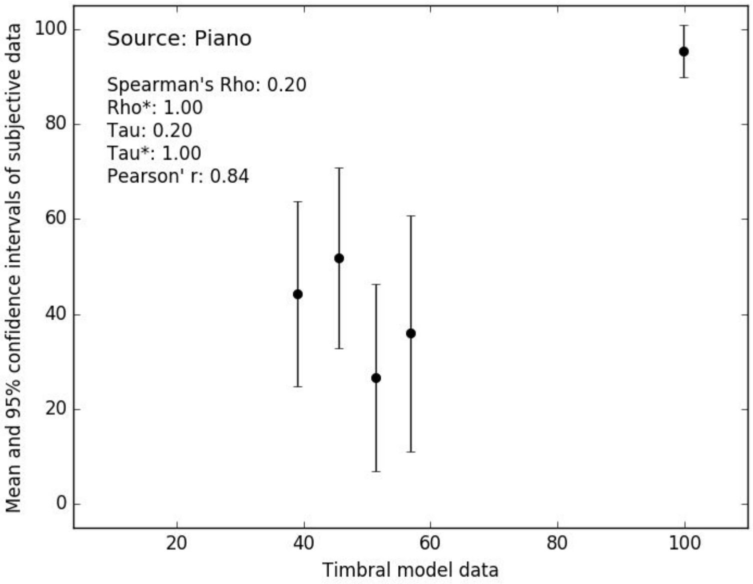


Figure 7 - Evaluation of the hardness model with the piano source type.

Looking at the metrics, the model performs well when considering the subjective variance (Rho* = 1.0, Tau* = 1.0). However, without considering this, the performance if poor, with a Rho of 0.2. This is due to the subjective data having large confidence intervals, caused by disagreement between subjects. However, the model can predict the hardest stimulus, which was fairly consistently rated as the hardest, as indicated by the smaller confidence intervals.

3.1.4 Model performance - Cymbal (hit)

Figure 8 shows the performance of the hardness model against the subjective data for the cymbal (hit) source type. The hardness model seems to perform very well on this dataset, not only achieving perfect rank ordering accounting for the subjective variance (Rho* - 1.0, Tau* = 1.0), but also achieving good rank order and linearity without considering this variance (Rho = 0.9, Tau = 0.8). The model is also fairly linearly related with the subjective data (r = 0.9).

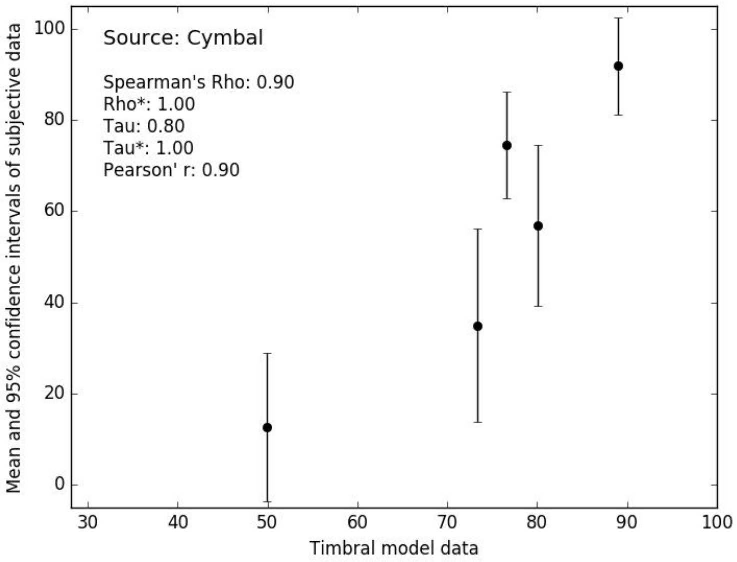


Figure 8 - Evaluation of the hardness model with the cymbal source type.

3.1.5 Model performance - Snare

Figure 9 shows the performance of the hardness model against the subjective data for the snare source type.

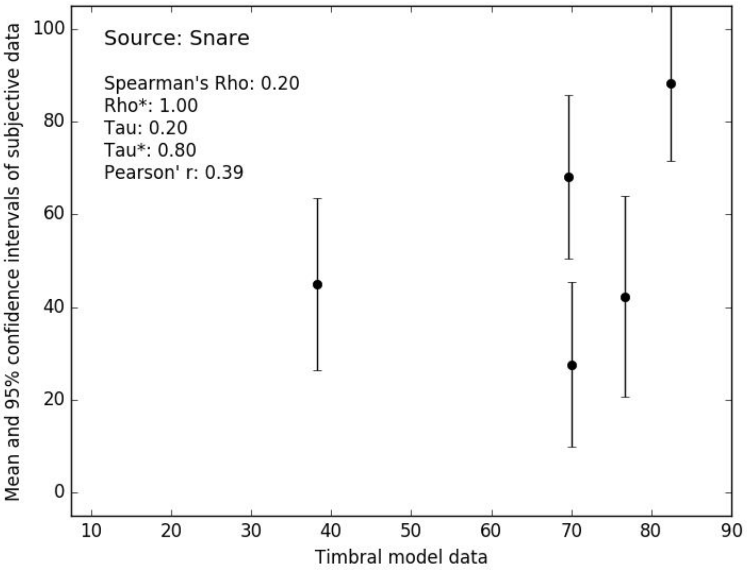


Figure 9 - Evaluation of the hardness model with the snare source type.

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

The hardness model performs poorly when considering only the mean datapoints (Rho = 0.2). This improves when considering the variance (Rho* = 1.0), but this is due to the large subjective 95% confidence intervals on the subjective data.

3.1.6 Model performance - Guitar

Figure 10 shows the performance of the hardness model against the subjective data for the guitar source type.

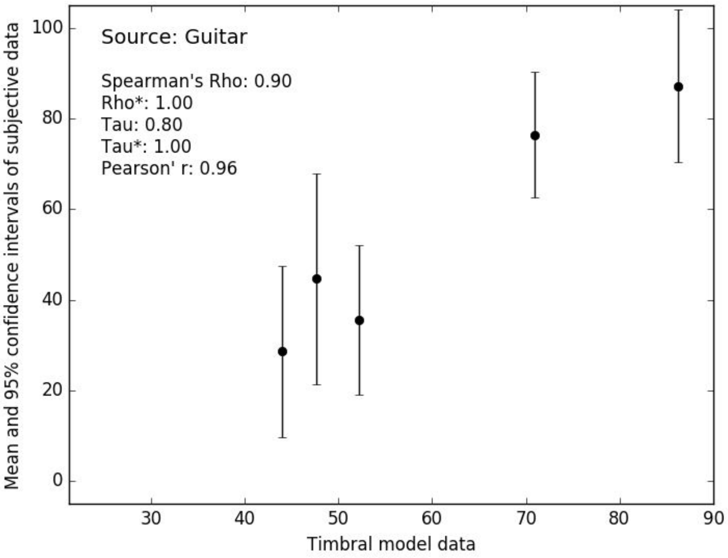


Figure 10 - Evaluation of the hardness model with the snare source type.

From figure 10, it can be seen that the hardness model performs very well, achieving good rank order and linearity without even considering the subjective variance (Rho = 0.9, r = 0.96). Additionally, the model can perfectly predict the rank order when considering the subjective variance of the data.

3.1.7 Summary

Table 3 below shows the Spearman's Rho, Rho* and Pearson's r for each source type.

Table 3: Summary of the hardness model's performance for each source type

Source type	Spearman's Rho	Rho*	Kendall's Tau	Tau*
Kick	0.5	0.75	0.4	0.6
Piano	0.2	1.0	0.2	1.0
Cymbal (hit)	0.9	1.0	0.8	1.0
Snare	0.2	1.0	0.2	0.8
Guitar	0.9	1.0	0.8	1.0

These results show that the hardness model performed well with the guitar and cymbal source types. However, the model only performed reasonably with the snare and piano source type data, and poorly with the kick source type data. This poorer performance may have been due to the large variance within the data, especially for the piano and snare stimuli whose performance increased to perfect rank ordering when considering this variance.

3.2 Between source types analysis

As with the within source type analysis, firstly the subjective data is analysed to remove any listeners whose ratings deviated from the means before assessing the performance of the model.

3.2.1 Assessment of participants

Figure 11 below shows the inter-subject agreement for the between source type data.

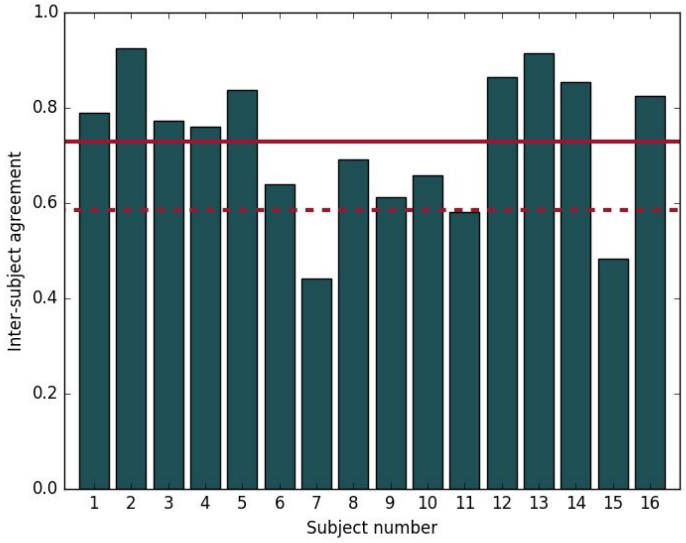


Figure 11 - Inter-subject agreement for the between source type analysis of the hardness attribute

As with the within source type analysis, the criteria for removal is having an agreement score less than 0.8, and less than one standard deviation below the grand mean (red dashed line). From figure 11, subjects 7, 11, and 15 all met this rejection criteria and were removed from subsequent analysis.

3.2.1 Model assessment

Figure 12 shows the performance of the hardness model against the between source types data. As can be seen in this plot, one stimulus is predicted as having a hardness greater than 1000, possibly due to shortcomings of the model, predicting the attack time being much shorter than possible. This will be examined in future work.

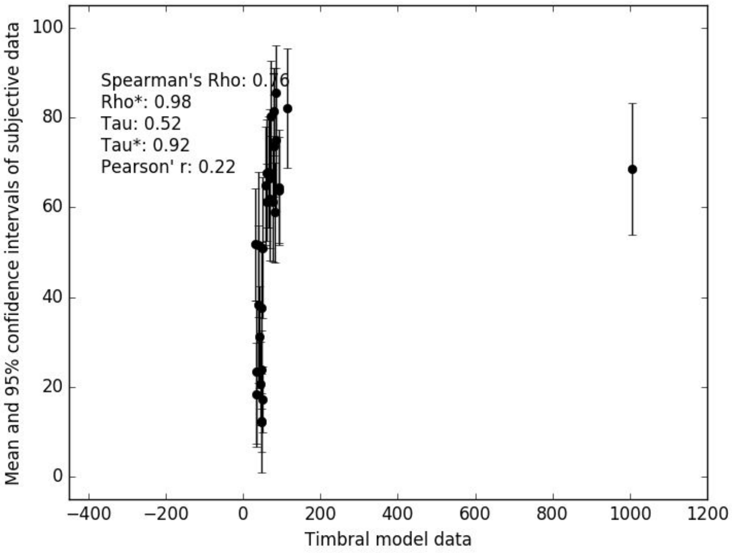


Figure 12 - Evaluation of the hardness model against the full between source type dataset

Figure 13 shows the same plot but this this point removed. This data point was still included within the measures of statistical fit.

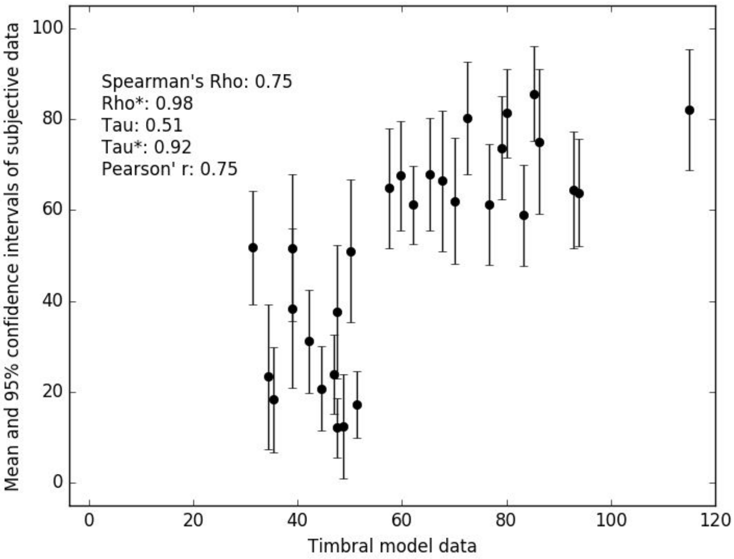


Figure 13 - Evaluation of the hardness model against the between source type dataset, with outlying data point removed.

Without this outlying data point, the model shows a moderate linear trend (r = 0.75) and reasonable rank order performance (Rho = 0.75, Tau = 0.52). The model appears to perform even better when considering the variance of the subjective data (Rho* = 0.98, Tau* = 0.91); however, as with the within source type analysis, this was considered likely to be due to the large confidence intervals.

Additionally, it should be noted that the data point that was removed did not lie at the top of the scale of the subjective results.

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

4 Depth model evaluation

The evaluation of the depth model, and all subsequent timbral models, was completed in the same manner as described in Section 3: first evaluating the performance for each source type on the within source type evaluation data; then evaluating the performance on the between source type dataset.

4.1 Within source type analysis

As with the hardness model, the within source type listening test data was assessed to potentially remove any outlier listeners before evaluating the model's performance.

4.1.1 Assessment of participants

The performance of the listening test participants was assessed by measuring the inter-subject agreement, shown in Figure 14.

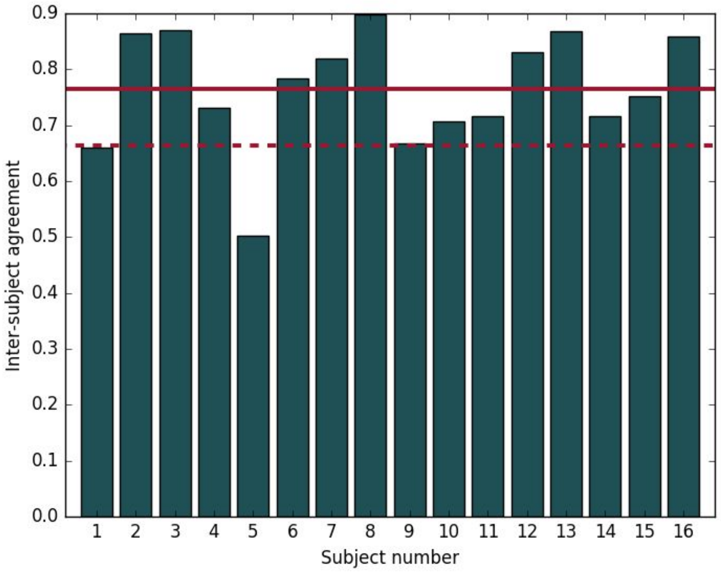


Figure 14 - Inter-subject agreement for the depth attribute.

From this figure, it can be seen that the grand mean of the inter-subject agreement is reasonably high. However, subject 5 had a much lower agreement than all other subjects, well below one standard deviation below the grand mean. Subject 1 also had an agreement less than one standard deviation below the grand mean. Since subjects 1 and 5 met the exclusion criteria (agreement less than 0.8, and less than one standard deviation below the grand mean) these were removed from subsequent analysis.

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

4.1.2 Model performance – Bass

Figure 15 shows a scatter plot of the subjective data from the bass source type against the depth model. From this plot, the model does not perform exceptionally well, with a poor rank order (Rho = 0.5). Although this improves when considering the variance of the subjective data (Rho* = 0.95), this is likely due to the large confidence intervals. Visually, it appears that the model performs reasonably well with the exception of the stimulus rated as the most deep. Although the rank order is not predicted for other stimuli, the wide overlapping confidence intervals indicate that listeners were not sure as to which sounded the most deep.

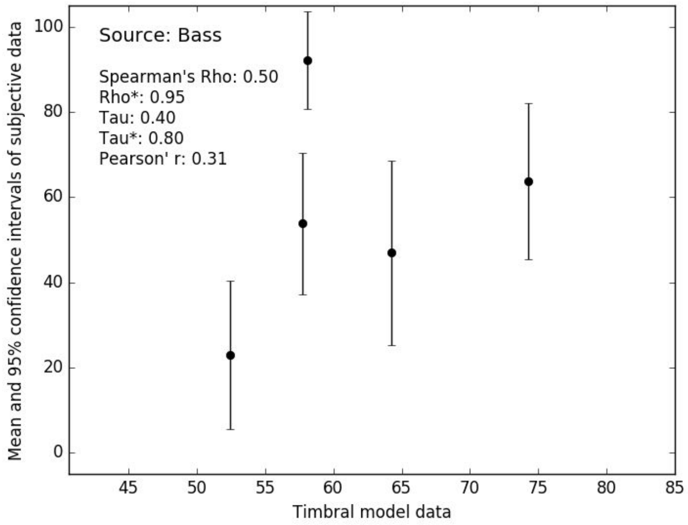


Figure 15 - Evaluation of the depth model with the bass source type.

4.1.3 Model performance – Kick

A scatter plot of the depth model against the kick programme item data is shown in Figure 16. From this figure, it can be seen that the model performs very well, with perfect rank ordering without even considering the subjective variance (Rho = 1.0, Tau = 1.0).

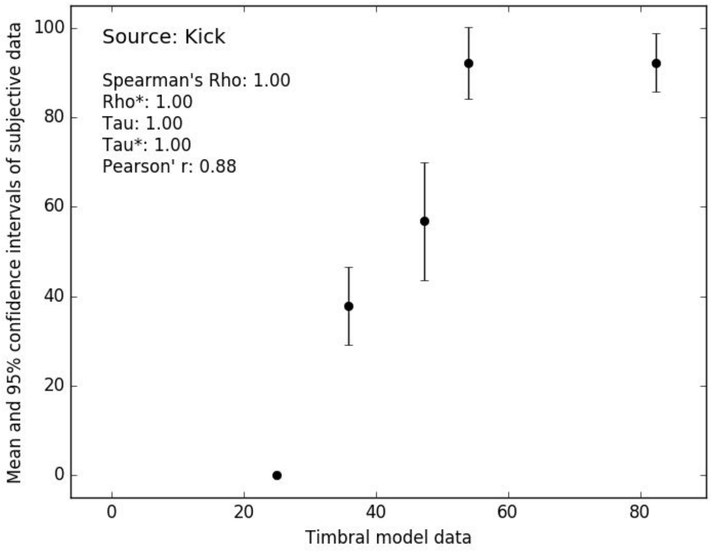


Figure 16 - Evaluation of the depth model with the kick source type.

4.1.4 Model performance – Drum loop

Figure 17 shows a scatter plot of the subjective data from the drum loop source type against the depth model.

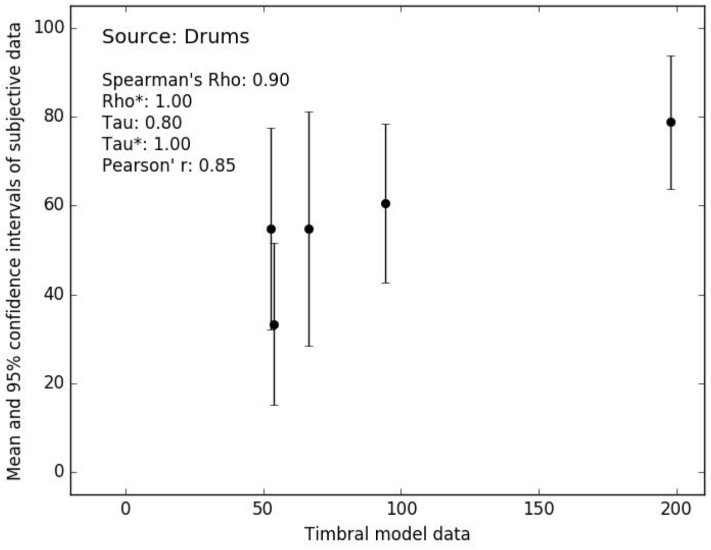


Figure 17 - Evaluation of the depth model with the drum loop source type.

The statistical analysis shows that the depth model performs well, with good rank ordering (Rho = 0.9, Tau = 0.8). However, there are large confidence intervals on the data indicating subjects were not consistent in their ratings.

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

4.1.5 Model performance – Voice

Figure 18 shows a scatter plot of the subjective data from the voice source type against the depth model.

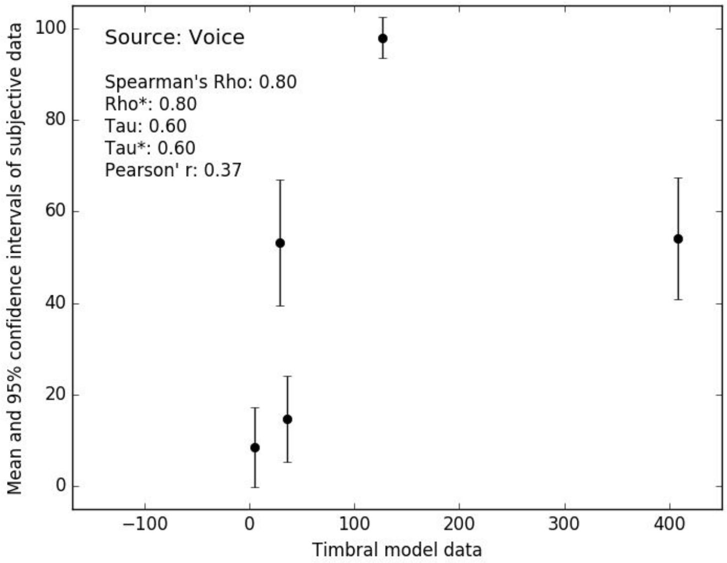


Figure 18 - Evaluation of the depth model with the voice source type.

The model has reasonable rank ordering (Rho = 0.8), but a poor linear relationship (r = 0.37). This is likely due to the stimulus with a depth value over 400.

4.1.6 Model performance – Impact

Figure 19 shows a scatter plot of the subjective data from the impact source type against the depth model. Visually, the model seems to perform well except for one stimulus not being predicted in the appropriate rank order. Additionally, the confidence intervals for this dataset are relatively small, indicating good agreement between subjects.

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

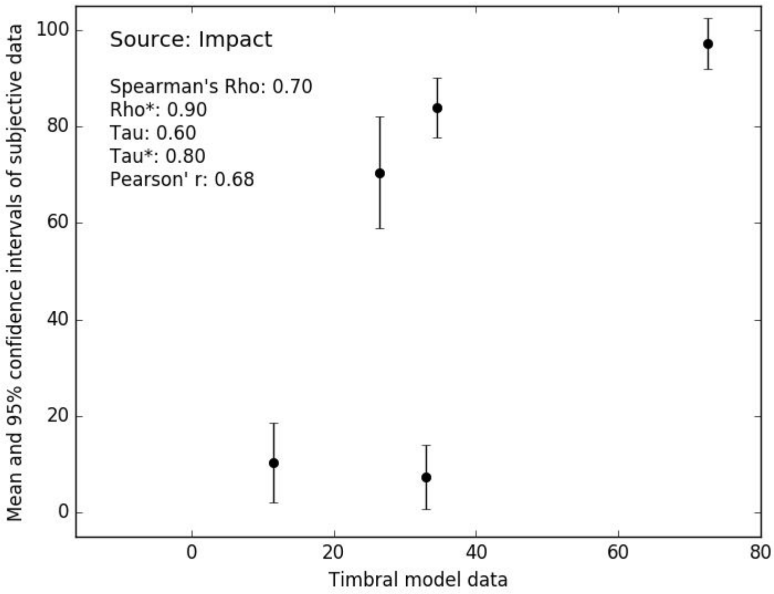


Figure 19 - Evaluation of the depth model with the impact source type.

4.1.7 Summary

Table 4 summarises the performance of the depth model for each of the five sources types.

Table 4: Summary of the depth model's performance for each source type

Source type	Spearman's Rho	Rho*	Kendall's Tau	Tau*
Bass	0.5	0.95	0.4	0.8
Kick	1.0	1.0	1.0	1.0
Drum loop	0.9	1.0	0.8	1.0
Voice	0.8	0.8	0.6	0.6
Impact	0.7	0.9	0.6	0.8

The bass source type shows the poorest performance of the depth model, with a Spearman's Rho of 0.5. however, other source types, such as the kick and impact, show very good performance of the model. In many cases, the model appears to poorly predict one or more stimuli, and resolving this would significantly improve the performance of the model. With the Voice source type, there is one stimulus which is over predicted by the model and should be investigated further.

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

4.2 Between source types analysis

As with the within source type analysis, the subjective data was first analysed to identify any subjects who made ratings drastically differently from the rest of the group.

4.2.1 Assessment of participants

The inter-subject agreement is shown in Figure 20.

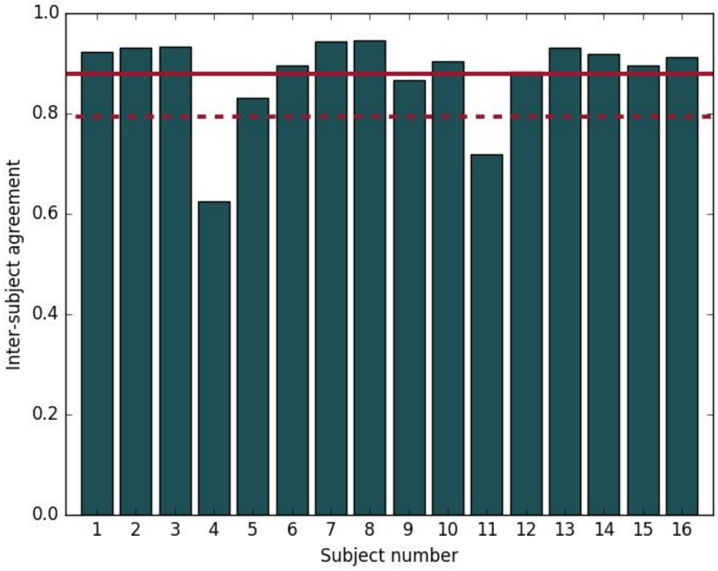


Figure 20 - Inter-subject agreement for the between source type analysis of the depth attribute.

This shows good agreement between listeners except for subjects 4 and 11. These both met the exclusion criteria and were removed from subsequent analysis.

4.2.2 Model assessment

Figure 21 below shows a scatter plot of the mean between source types results against the depth model's predicted values. As can be seen from this plot, the model appears to work fairly well, with the exception of a few stimuli. The trend is not particularly linear, r = 0.59, but has reasonable rank order performance, rho = 0.72. This increases further when consider the subjective variance Rho* = 0.94.

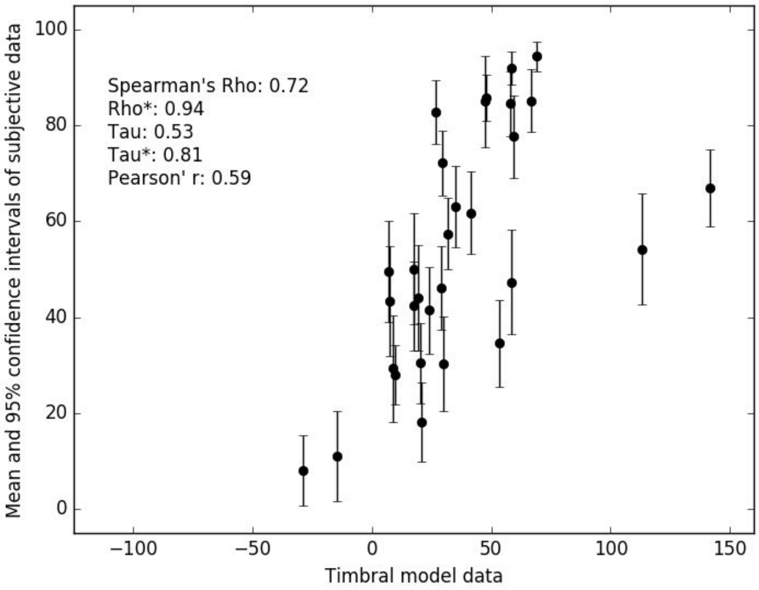


Figure 21 - Evaluation of the depth model against the between source type dataset.

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

5 Brightness model evaluation

5.1 Within source type analysis

Prior to assessing the performance of the brightness model, the within source type listening test data was assessed to potentially remove any outlier listeners.

5.1.1 Assessment of participants

The inter-subject agreement is shown in Figure 22. This shows very good agreement between subjects, with a grand mean of 0.88. However, subjects 4, 10, and 11 all met the exclusion criteria and were removed from subsequent analysis.

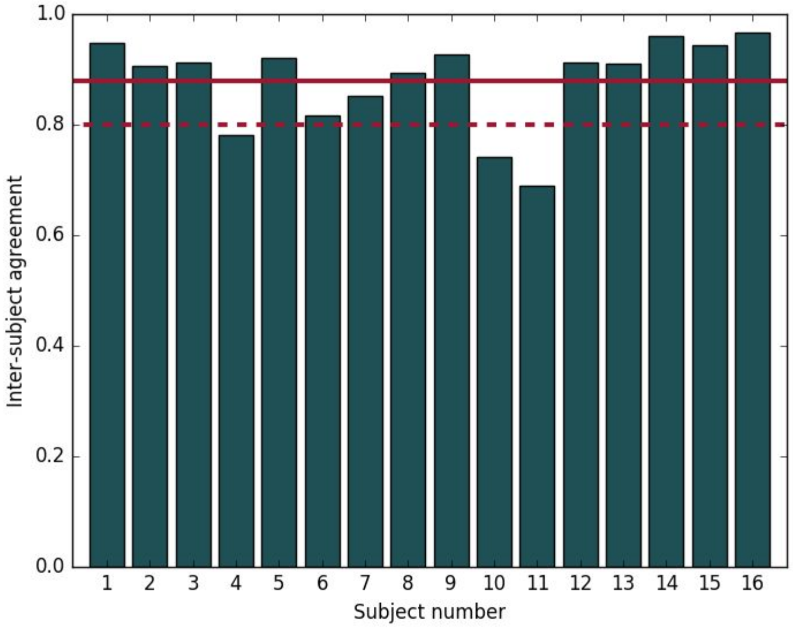


Figure 22 - Inter-subject agreement for the brightness attribute.

5.1.2 Model performance – Ambience

A scatter plot of the brightness model's predictions against the subjective data for the ambience source type is shown in Figure 23. This plot shows that the model performs fairly well, the most and least bright stimuli being predicted correctly. However, the middle three stimuli, whose confidence intervals all overlap, were not predicted in the correct rank order. Considering the subjective variance, however, results in a perfect rank ordering.

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

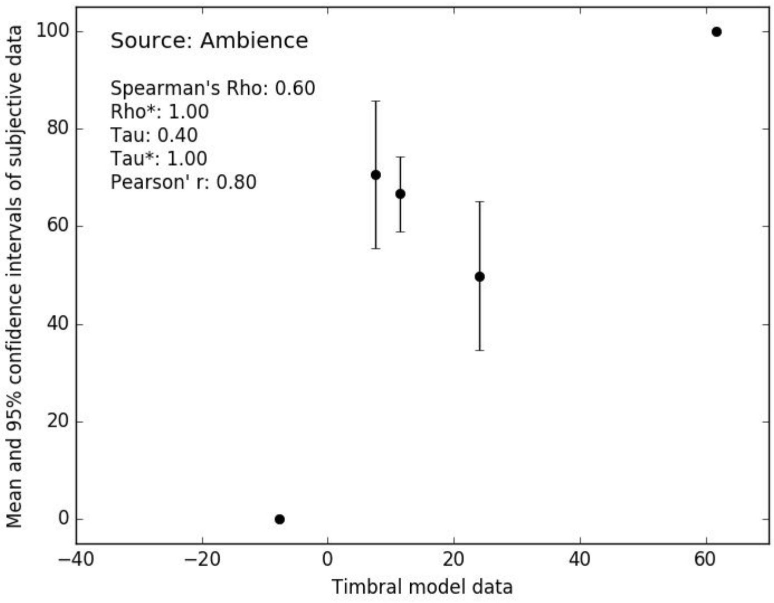


Figure 23 - Evaluation of the brightness model with the ambience source type.

5.1.3 Model performance – Piano

A scatter plot of the brightness model's predictions against the subjective data for the piano source type is shown in Figure 24. This shows good performance of the brightness model (Rho = 0.8); however, the least bright stimulus was not predicted as the least bright.

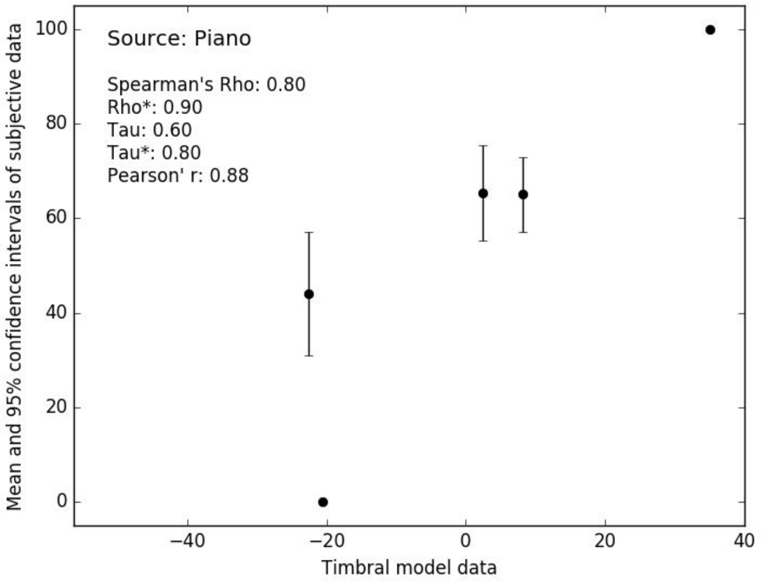


Figure 24 - Evaluation of the brightness model with the piano source type.

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

5.1.4 Model performance – Bell

A scatter plot of the brightness model's predictions against the subjective data for the bell source type is shown in Figure 25. This plot shows that the brightness model works very well, predicting the rank order perfectly (Rho = 1.0) with a fairly linear response (r = 0.89).

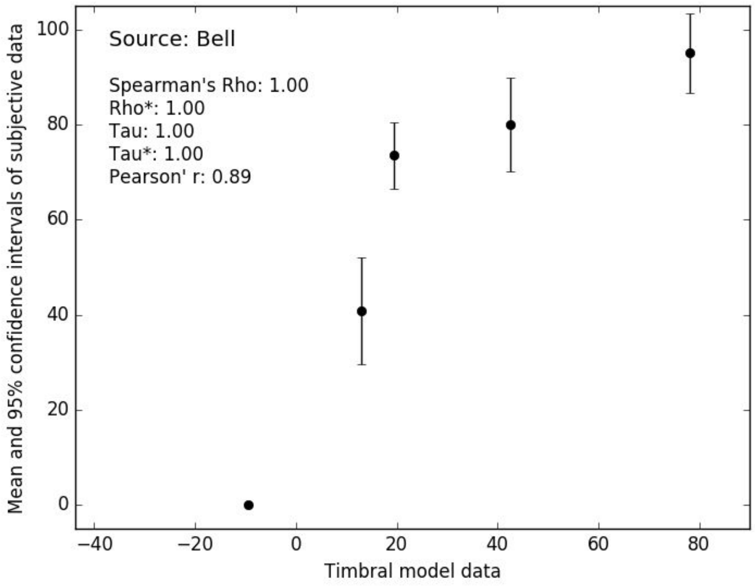


Figure 25 - Evaluation of the brightness model with the bell source type.

5.1.5 Model performance – Voice

A scatter plot of the brightness model's predictions against the subjective data for the voice source type is shown in Figure 26. The model, again, performs well with the exception of predicting the least bright stimulus. If this stimulus was to be predicted correctly, the model would predict the rank order perfectly.

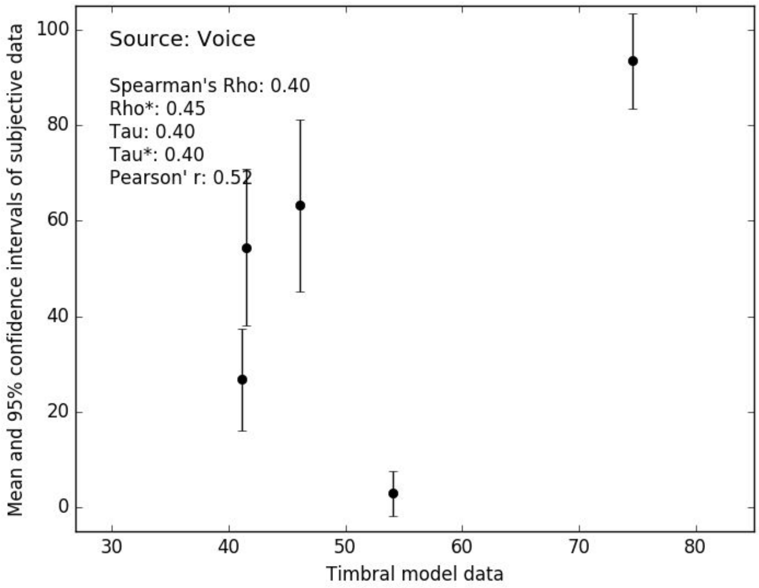


Figure 26 - Evaluation of the brightness model with the voice source type.

5.1.6 Model performance – Swoosh

A scatter plot of the brightness model's predictions against the subjective data for the bell source type is shown in Figure 27. This shows that the brightness model performs very well (Rho = 0.9), almost predicting the rank order exactly. The data points whose rank order was not correctly predicted were close in the subjective scores and confidence intervals overlap. Considering the subjective variance, the model is able to predict the rank order of the data perfectly (Rho* = 1.0).

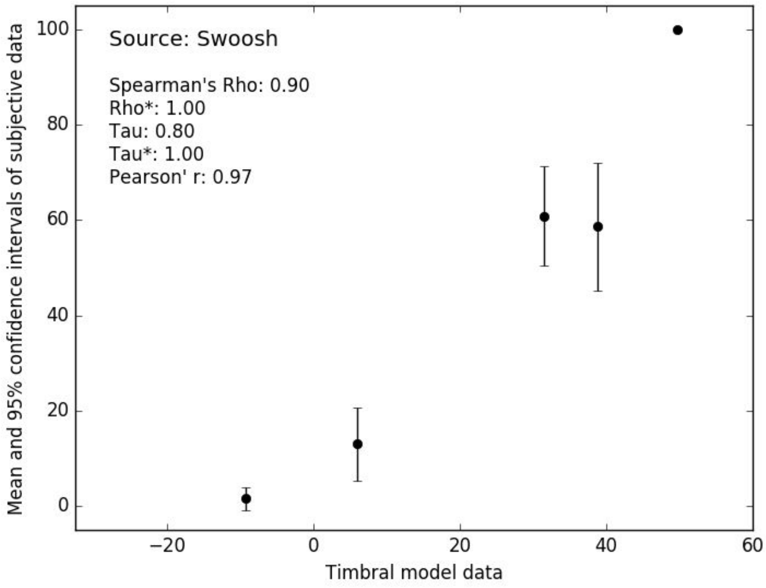


Figure 27 - Evaluation of the brightness model with the swoosh source type.

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

5.1.7 Summary

Table 5 below summarises the performance of the brightness model with the subjective data from each of the source types.

Table 5: Summary of the brightness model's performance for each source type

Source type	Spearman's Rho	Rho*	Kendall's Tau	Tau*
Ambience	0.6	1.0	0.4	1.0
Piano	0.8	0.9	0.6	0.8
Bell	1.0	1.0	1.0	1.0
Voice	0.4	0.45	0.4	0.4
Swoosh	0.9	1.0	0.8	1.0

By looking at the performance metrics, it appears that the voice source type was not predicted well. However, inspection of the Figure 26 shows that the poor performance was the result of the stimulus that was rated as the least bright not being predicted as the least bright. For all other source types, the brightness model tends to perform very well.

5.2 Between source types analysis

As with the within source type analysis, the subjective data was first analysed to identify any subjects who made ratings drastically differently from the rest of the group.

5.2.1 Assessment of participants

The inter-subject agreement is shown in Figure 28.

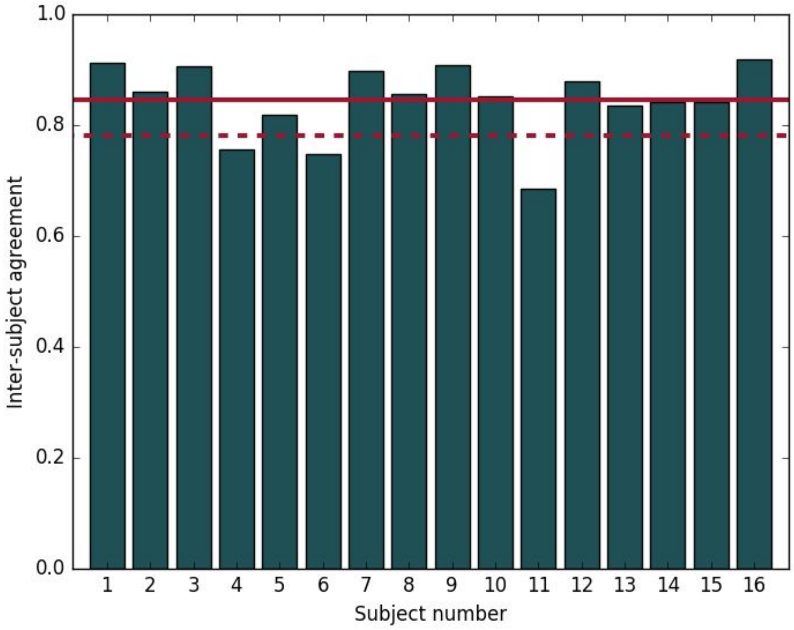


Figure 28 - Inter-subject agreement for the between source type analysis of the brightness attribute.

This shows good agreement between listeners except for subjects 4, 6, and 11. These subjects met the exclusion criteria and were removed from subsequent analysis.

5.2.2 Model assessment

Figure 29 below shows a scatter plot of the mean between source types results against the brightness model's predicted values. As can be seen the model does not perform as well as in the within source type tests. The rank order performance of the model is adequate for a larger dataset, Rho = 0.65, and improves when considering the variation of the subjective data, Rho* = 0.96. The data is also fairly linear, r = 0.69.

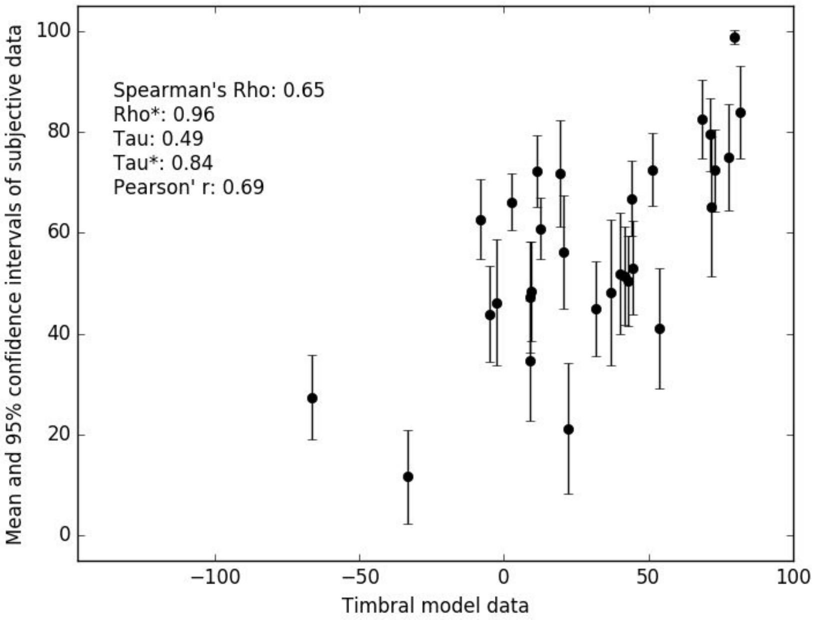


Figure 29 - Evaluation of the brightness model against the between source type dataset.

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

6 Metallic-nature model evaluation

6.1 Within source type analysis

Prior to assessing the performance of the metallic-nature model, the within source type listening test data was assessed to potentially remove any outlier listeners.

6.1.1 Assessment of participants

The inter-subject agreement is shown in Figure 30. This shows very good agreement between subjects, the grand mean of inter-subject agreement of 0.88. However, subjects 10 and 11 met the exclusion criteria and were removed from subsequent analysis.

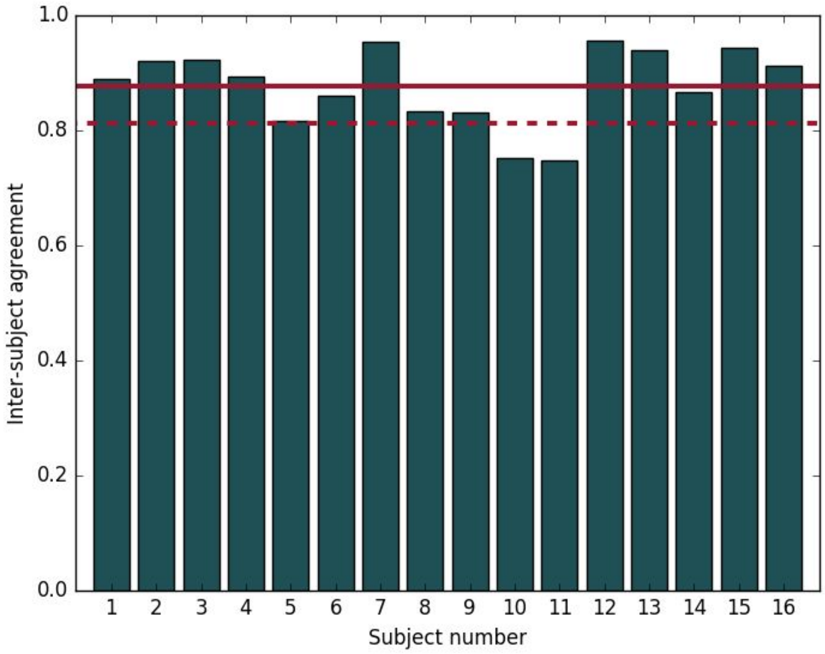


Figure 30 - Inter-subject agreement for the metallic-nature attribute.

6.1.2 Model performance – Clang

Figure 31 shows a scatter plot of the metal-nature model's predicted scores against the subjective data for the clang source type. This plot shows that the metallic-nature model performs very poorly, with negative Rho, Tau, and Pearson's r values.

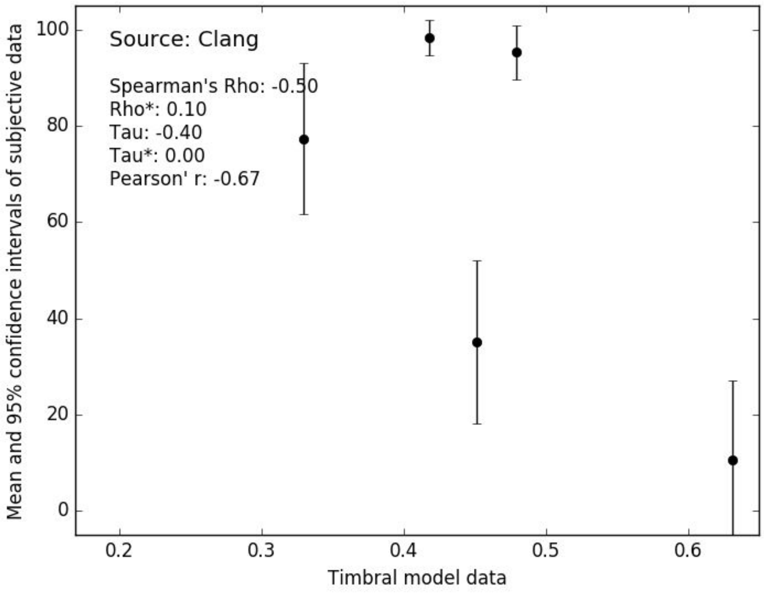


Figure 31 - Evaluation of the metallic-nature model with the clang source type.

6.1.3 Model performance – Footsteps

The values predicted from the metallic-nature model with the footsteps source type are shown in Figure 32 plotted against the subjective results. From this plot, it seems that the metallic-nature model works reasonably well, achieving perfect rank ordering when considering the variation of the subjective results, and predicting the three metallic sounding stimuli higher than the two non-metallic sounding stimuli. However, unlike the other developed timbral models, the metallic-nature model is a logistic regression, as opposed to linear regression. This means that the output of the model is a probability of the stimulus sounding metallic, and as such, scale effects do not exist and should be reviewed when evaluating the model's performance.

Although the model appeared to work well for the footsteps source type, all values from the model are less than 0.5, indicating that the model does not consider any of the stimuli to sound metallic.

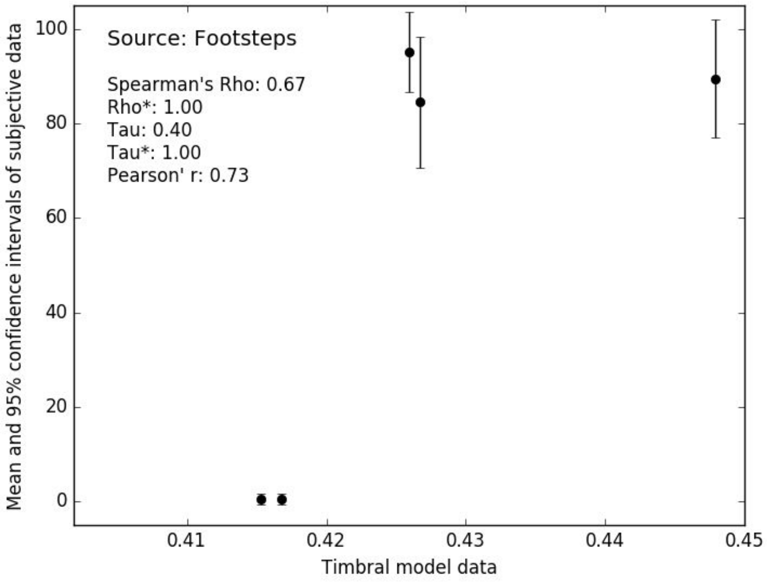


Figure 32 - Evaluation of the metallic-nature model with the footsteps source type.

6.1.4 Model performance – Creak

Figure 33 shows the performance of the metallic-nature model with the creak source type. As can be seen, the model works very poorly, producing almost the reversed rank order of the subjective data.

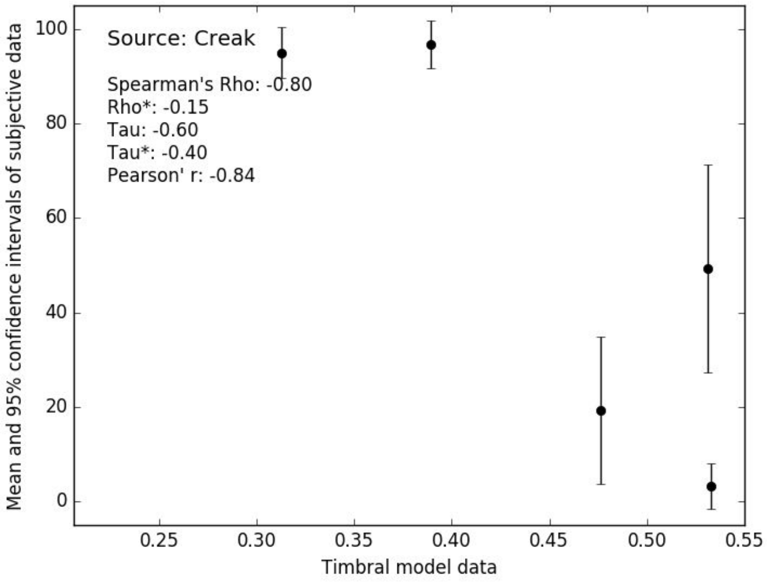


Figure 33 - Evaluation of the metallic-nature model with the creak source type.

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

6.1.5 Model performance – Scrape

Figure 34 shows the performance of the metallic-nature model with the scrape source type. From this, it can be seen that the model reasonably predicts the subjective data. However, it should be noted, once again, that the model only predicted two of the stimuli as having metallic-nature probabilities greater than 0.5, even though subjects were very confident that these sounded metallic (indicated by the small confidence intervals).

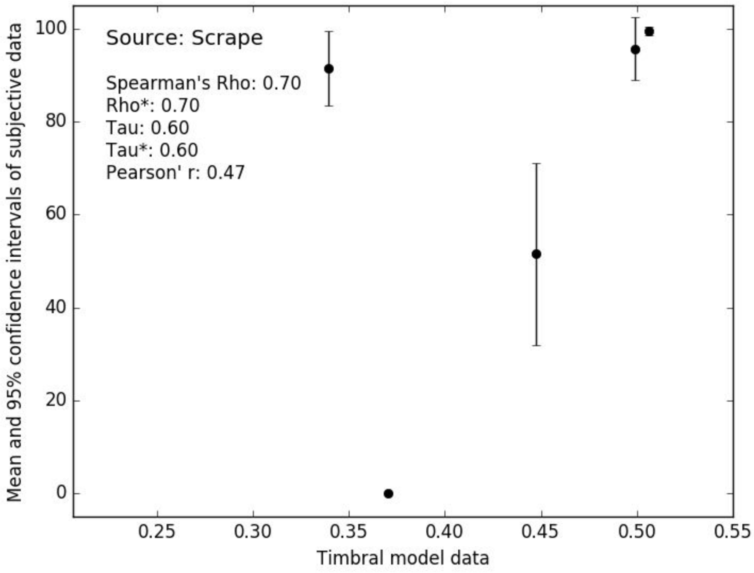


Figure 34 - Evaluation of the metallic-nature model with the scrape source type.

6.1.6 Model performance – Whoosh

Figure 35 shows the performance of the metallic-nature model with the whoosh source type. Again, this shows fairly poor performance of the metallic-nature model. The Kendall's Tau for this model indicates there is no relationship at all between the metallic-nature model and the subjective results (Tau = 0.0).

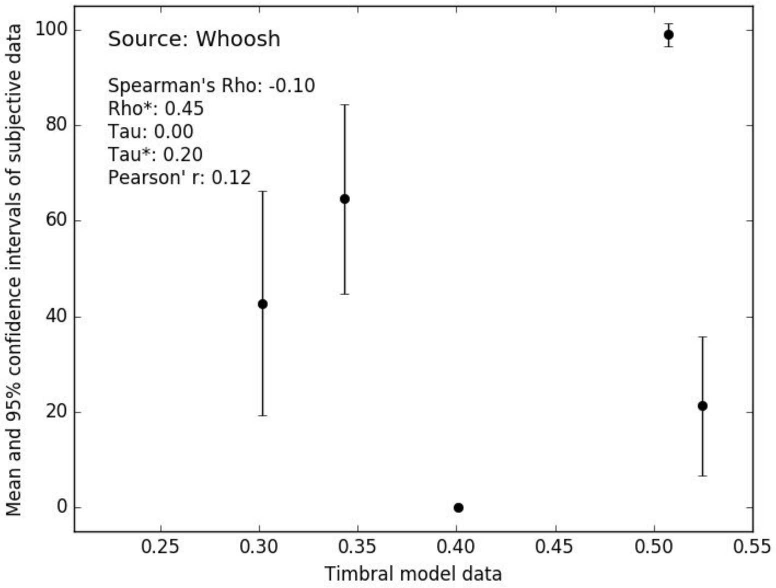


Figure 35 - Evaluation of the metallic-nature model with the whoosh source type.

6.1.7 Summary

Table 6 below summarises the performance of the metallic-nature model with each of the five source types.

Table 6: Summary of the metallic-nature model's performance for each source type

Source type	Spearman's Rho	Rho*	Kendall's Tau	Tau*
Clang	-0.5	0.1	-0.4	0.0
Footsteps	0.67	1.0	0.4	1.0
Creak	-0.8	-0.15	-0.6	-0.4
Scrape	0.7	0.7	0.6	0.6
Whoosh	-0.1	0.45	0.0	0.2

From this evaluation, it was shown that the metallic-nature model performs very poorly for the majority of the source types. Even for the source types where the performance appeared to be suitable, the metallic-nature model's results covered a small range close to the centre of its scale. Unlike the other timbral models, the metallic-nature model was a logistic regression model, the output from which is a probability that the stimulus sounds metallic. For the source types that did show a reasonable rank order, the probability values were very low and would not indicate that the sources sounded metallic.

6.2 Between source types analysis

As with the within source type analysis, the subjective data was first analysed to identify any subjects who made ratings drastically differently from the rest of the group.

6.2.1 Assessment of participants

The inter-subject agreement is shown in Figure 36.

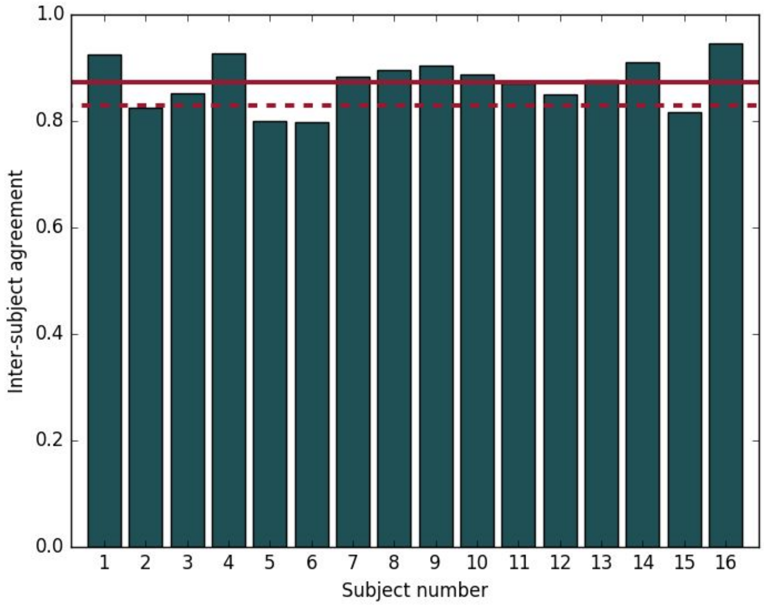


Figure 36 - Inter-subject agreement for the between source analysis of the metallic-nature model.

This shows good agreement between listeners. However, subjects 5 and 6 met the exclusion criteria of having inter-subject agreement scores less than 0.8 and less than one standard deviation below the grand mean. Therefore, subjects 5 and 6 were removed from subsequent analysis.

6.2.2 Model assessment

Figure 37 below shows a scatter plot of the mean between source types results against the metallic-nature model's predicted values. This shows very poor performance of the metallic-nature model. Many of the stimuli that were rated as clearly sounding metallic or non-metallic were not rated correctly by the model.

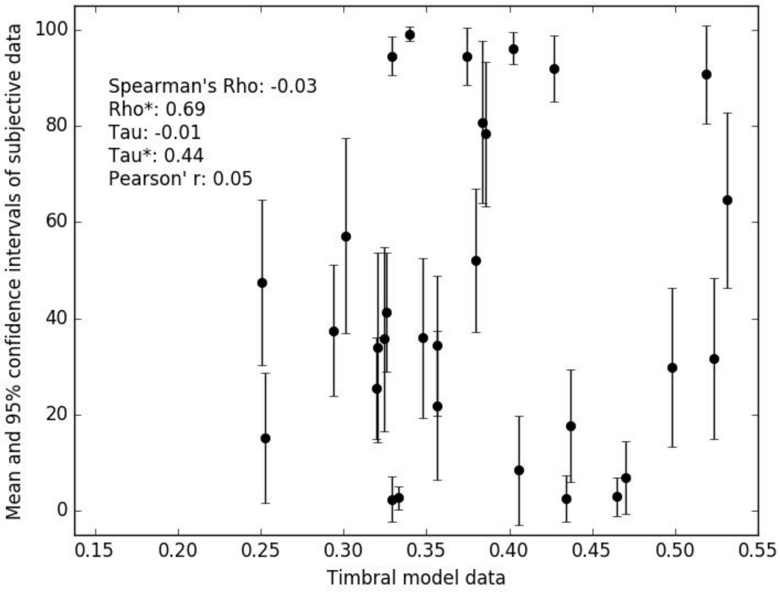


Figure 37 - Evaluation of the metallic-nature model against the between source type dataset.

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

7 Reverb model evaluation

7.1 Within source type analysis

Prior to assessing the performance of the reverb model, the within source type listening test data was assessed to potentially remove any outlier listeners.

7.1.1 Assessment of participants

The inter-subject agreement is shown in Figure 38. This shows very good agreement between subjects, with a grand mean of 0.92. None of the subjects had an agreement of less than 0.8, and thus no subjects met the exclusion criteria.

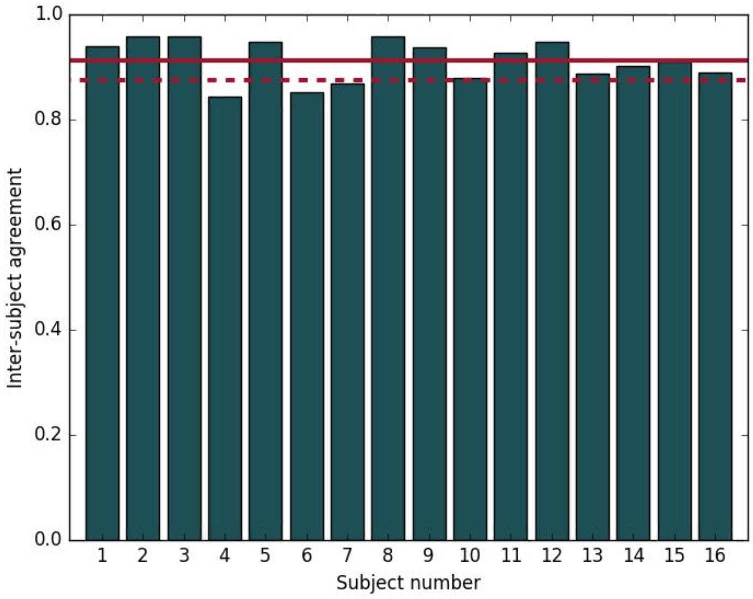


Figure 38 - Inter-subject agreement for the reverb attribute.

7.1.2 Model performance – Snare

The performance of the reverb model against the snare source type is shown in Figure 39. As can be seen, the model does not predict the apparent level of reverb very well (Rho = 0.3). The rank ordering is poor and the stimulus that was consistently rated as having the most reverberation was not modelled as the most reverberant stimulus.

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

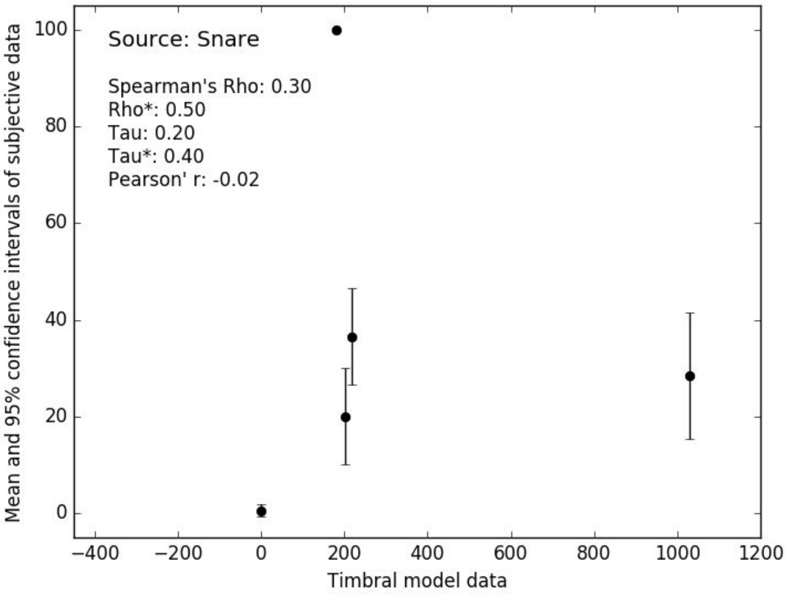


Figure 39 - Evaluation of the reverb model with the snare source type.

7.1.3 Model performance – Atmosphere

The performance of the reverb model with the atmosphere source type is shown in Figure 40. As can be seen, the model performs poorly, with low rank order and very poor correlation (Rho = 0.3). Additionally, the stimulus that was rated as the most reverberant was not modelled as such.

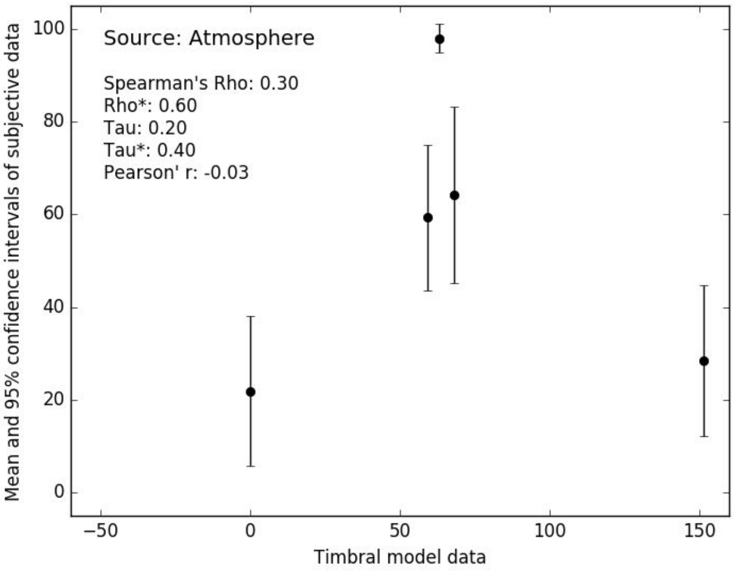


Figure 40 - Evaluation of the reverb model with the atmosphere source type.

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

7.1.4 Model performance – Hit

The performance of the reverb model with the hit source type is shown in Figure 41. From this, it can be seen that only one stimulus was predicted as having a reverberance greater than zero. The timbral reverb model outputs a zero when it is unable to detect a suitable free decay region. This issue needs to be addressed if the model is to predict the apparent level of reverberation more effectively.

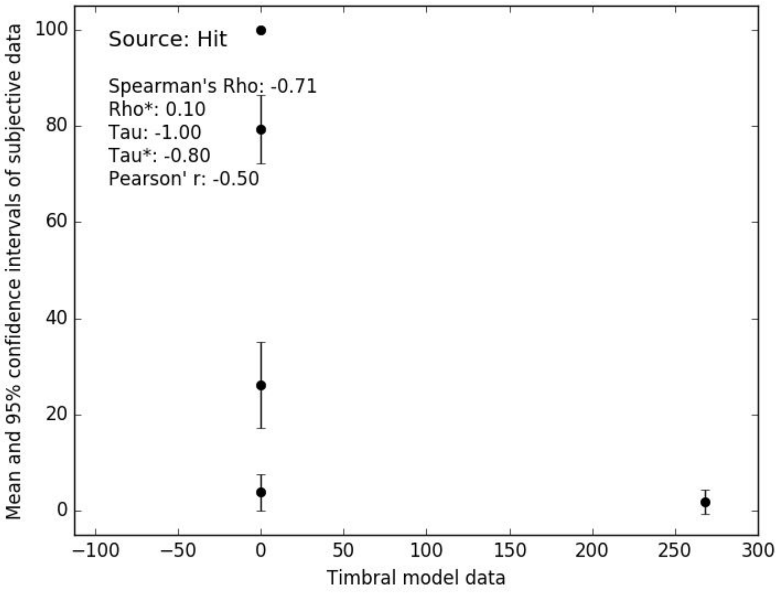


Figure 41 - Evaluation of the reverb model with the hit source type.

7.1.5 Model performance – Footsteps

The performance of the reverb model with the footsteps source type is shown in Figure 42. This, once again, shows several stimuli where the reverb model returned a zero value. Additionally, for the three stimuli where a reverb time was calculated, the rank order was incorrect. One of these stimuli was clearly rated by subjects as the most reverberant, indicated by the lack of confidence intervals, yet the model did not predict this as the most reverberant.

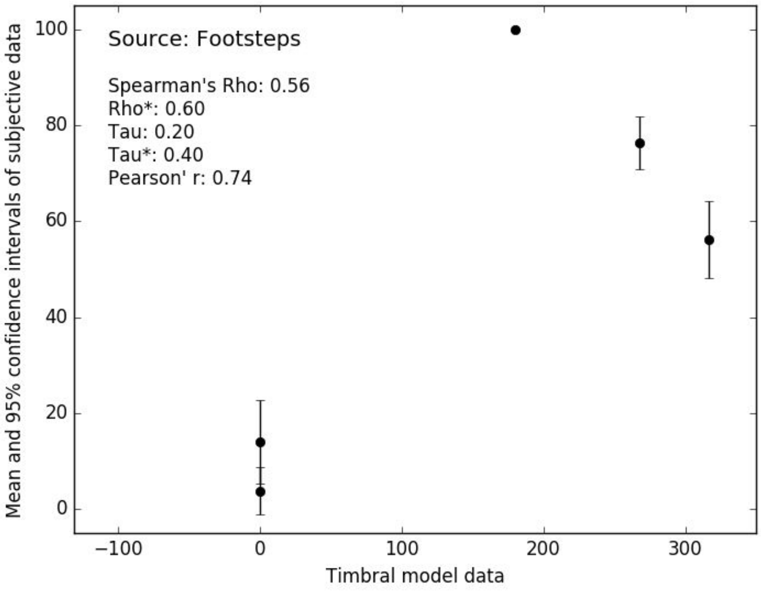


Figure 42 - Evaluation of the reverb model with the footsteps source type.

7.1.6 Model performance – Vocal

Figure 43 shows the performance of the reverb model with the vocal source. This plot shows that the reverb model performs poorly (Rho = 0.1), predicting the reverb of sources rated as having similar reverb times as different, and not correctly predicting that the most reverberant stimulus as the most reverberant.

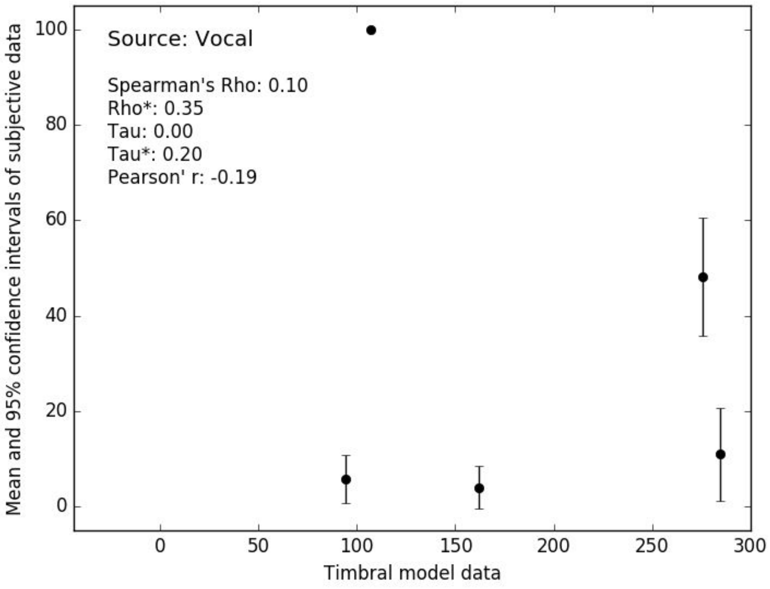


Figure 43 - Evaluation of the reverb model with the vocal source type.

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

7.1.7 Summary

Table 7 below summarises the performance of the reverb model with all five source types.

Table 7: Summary of the reverb model's performance for each source type

Source type	Spearman's Rho	Rho*	Kendall's Tau	Tau*
Snare	0.3	0.5	0.2	0.4
Atmosphere	0.3	0.6	0.2	0.4
Hit	-0.71	0.1	-0.1	-0.8
Footsteps	0.56	0.6	0.2	0.4
Vocal	0.1	0.35	0.0	0.2

Over all source types, the reverb model performs very poorly. There are multiple stimuli where the reverb model is unable to measure a reverb time at all, even though there is clearly a difference in the subjective reverberation. Additionally, for the stimuli where the reverb time can be estimated, the rank order is often wrong.

7.2 Between source types analysis

As with the within source type analysis, the subjective data was first analysed to identify any subjects who made ratings drastically differently from the rest of the group.

7.2.1 Assessment of participants

The inter-subject agreement is shown in Figure 44.

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

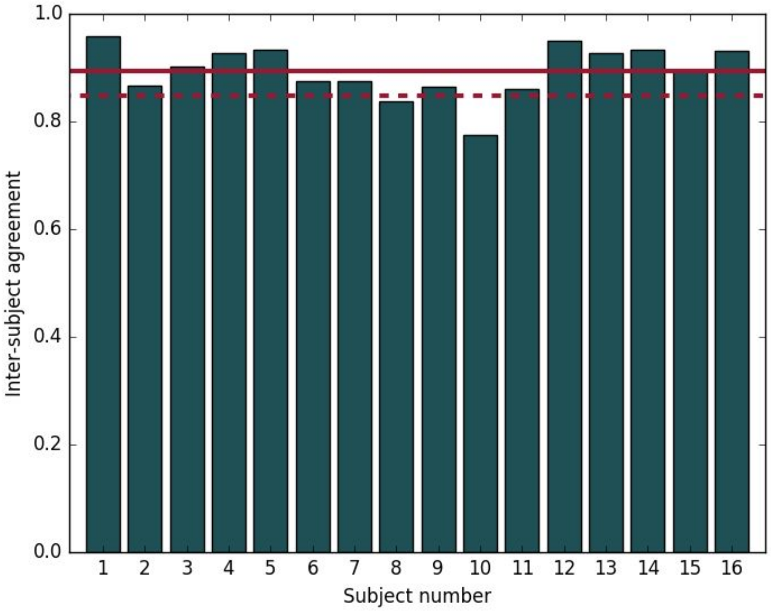


Figure 44 - Inter-subject agreement for the between source type analysis of the reverb attribute.

This shows good agreement between listeners. However, subject 10 met the exclusion criteria and was removed from subsequent analysis.

7.2.2 Model assessment

Figure 45 below shows a scatter plot of the mean between source types results against the reverb model's predicted values. This shows extremely poor performance of the reverb model. The model has no correlation with the subjective results (*r* = -0.01) and many of the stimuli are predicted as having 0 reverb time, due to the algorithm's inability to detect any suitable free decay regions. Additionally, the Spearman's Rho is extremely low and negative (Rho = -0.12) and does not significantly improve when considering the variance (Rho* = 0.56). The stimulus that is clearly rated as the most reverberant (consistently rated as the most reverberant by all subjects, indicated by no confidence intervals), was identified as having a reverb time of 0.0.

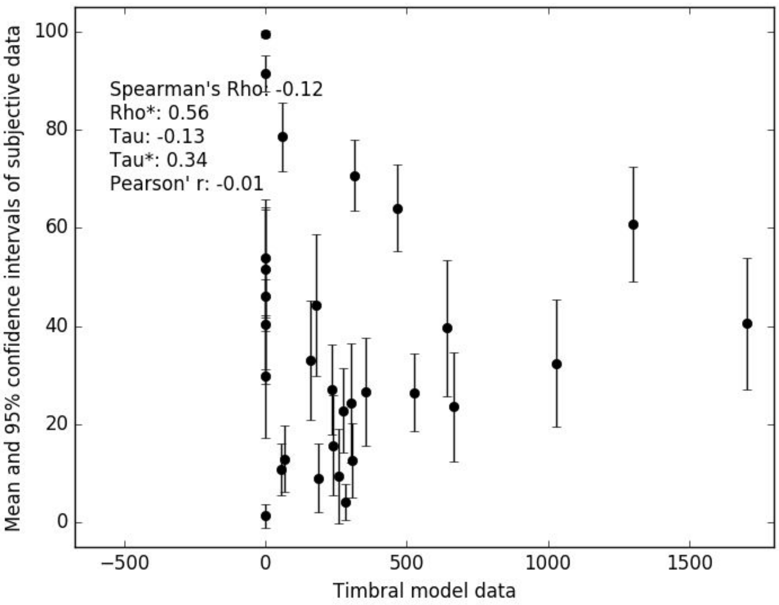


Figure 45 - Evaluation of the reverb model against the between source type dataset.

8 Roughness model evaluation

8.1 Within source type analysis

Prior to assessing the performance of the roughness model, the within source type listening test data was assessed to potentially remove any outlier listeners.

8.1.1 Assessment of participants

The inter-subject agreement is shown in Figure 46. This shows very good agreement between subjects, the grand mean of the inter-subject agreement of 0.86. Only subject 5 met the rejection criteria, having an agreement less than 0.8 and an agreement below one standard deviation less than the grand mean, and was therefore removed from subsequent analysis.

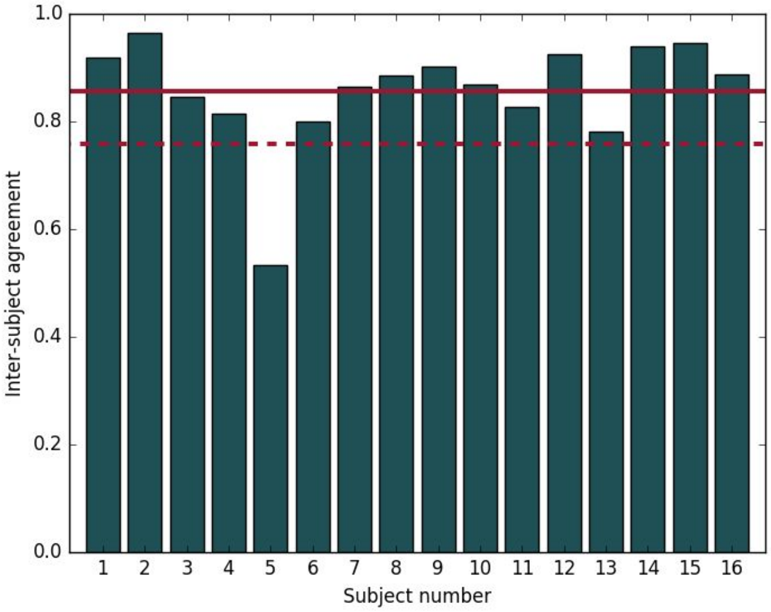


Figure 46 - inter-subject agreement for the roughness attribute.

8.1.2 Model performance – Bell

The performance of the roughness model against the bell data is shown in Figure 47. This shows that the model is able to correctly predict the rank order of the stimuli (Rho = 1.0). However, the results appear to be nonlinearly related, with a clustering of model predictions close to zero.

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

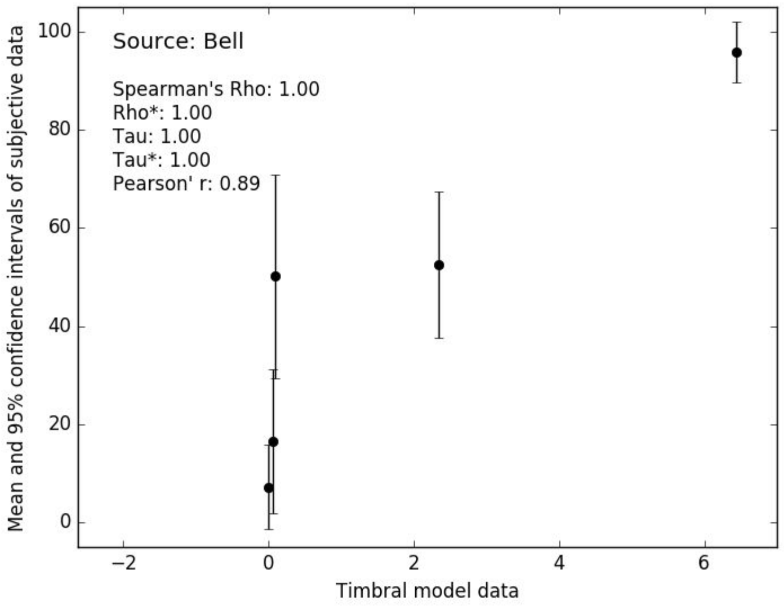


Figure 47 - Evaluation of the roughness model with the bell source type.

8.1.3 Model performance – Synth

Figure 48 shows the performance of the roughness against the synth source type. As can be seen from this plot, the model seems to predict the subjective data very well. However, there is, once again, this nonlinear trend between the subjective results and the model's output.

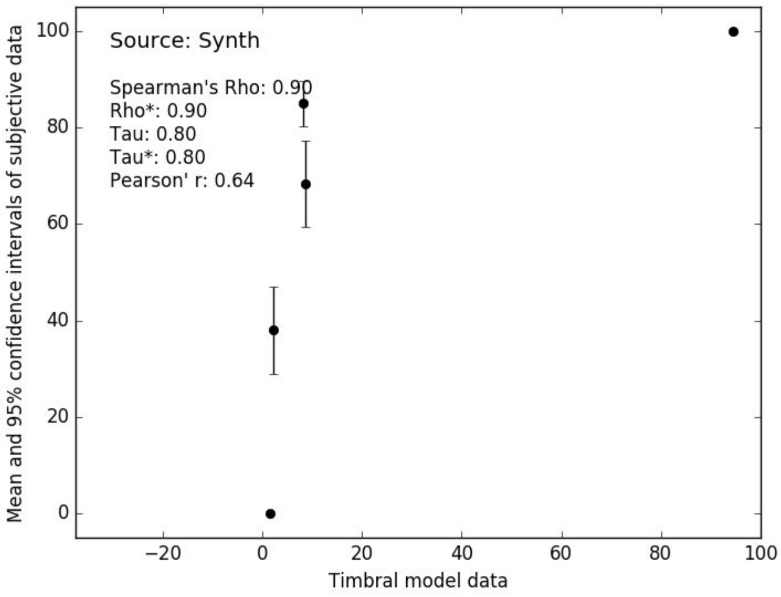


Figure 48 - Evaluation of the roughness model with the synth source type.

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

8.1.4 Model performance – Engine

Figure 49 shows the performance of the roughness model with the engine source type. This shows a slightly worse performance than the previous source types. However, the performance is still reasonable, having Rho* of 0.8.

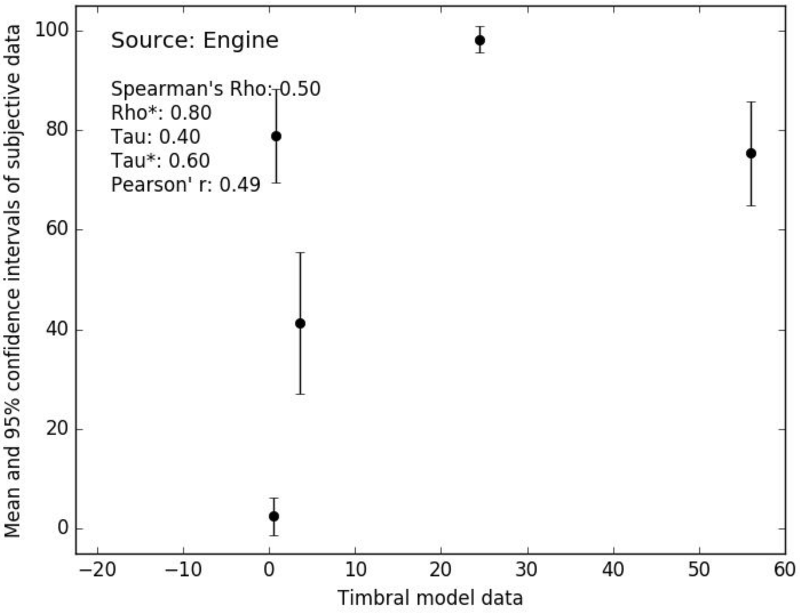


Figure 49 - Evaluation of the roughness model with the engine source type.

8.1.5 Model performance – Alarm

The performance of the roughness model with the alarm source type is shown in Figure 50. Again, the model performs well, almost correctly predicting the rank order of the data, with perfect rank order being attained when considering the variation of the subjective data (Rho = 0.9, Rho* = 1.0).

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

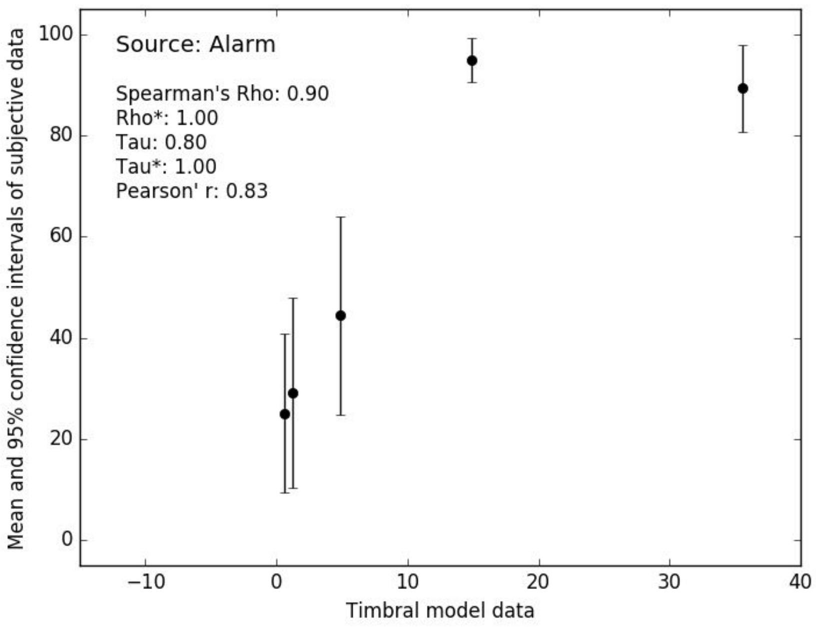


Figure 50 - Evaluation of the roughness model with the alarm source type.

8.1.6 Model performance – Guitar

The performance of the roughness model with the guitar source type is shown in Figure 51. This shows good performance of the model, with a fairly linear response with the exception of one stimulus (*r* = 0.88, Rho = 0.9).

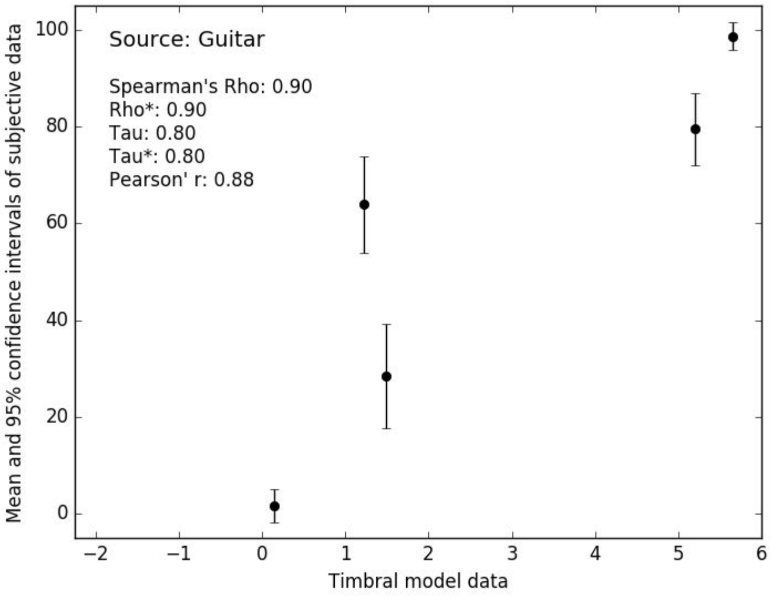


Figure 51 - Evaluation of the roughness model with the guitar source type.

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

8.1.7 Summary

Table 8 below summarises the performance of the roughness model with all five source types.

Table 8: Summary of the roughness model's performance for each source type

Source type	Spearman's Rho	Rho*	Kendall's Tau	Tau*
Bell	1.0	1.0	1.0	1.0
Synth	0.9	0.9	0.8	0.8
Engine	0.5	0.8	0.4	0.6
Alarm	0.9	1.0	0.8	1.0
Guitar	0.9	0.9	0.8	0.8

The model of roughness performs well on all source types. The worst performing source type was the engine sounds. This may have been due to the interpretation of roughness for engine sounds, whereby temporal artefacts may have an effect on the perceived level of roughness. Many of the source types seem to indicate that there is a nonlinear relationship between the subjective data and the model's predictions of roughness.

8.2 Between source types analysis

As with the within source type analysis, the subjective data was first analysed to identify any subjects who made ratings drastically differently from the rest of the group.

8.2.1 Assessment of participants

The inter-subject agreement is shown in Figure 52.

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

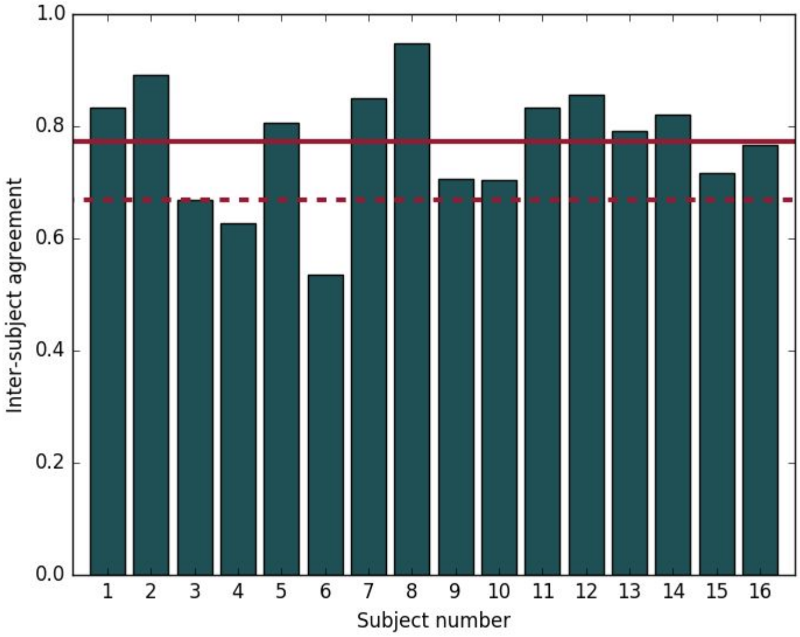


Figure 52 - Inter-subject agreement for the between source type analysis of the roughness attribute.

This shows reasonable agreement between subjects, with a grand mean of 0.77. However, subjects 3, 4, and 6 met the exclusion criteria of having an agreement of less than 0.8 and less than one standard deviation below the grand mean. These three subjects were removed from subsequent analysis.

8.2.2 Model assessment

Figure 53 below shows a scatter plot of the mean between source types results against the roughness model's predicted values. Between source types, the roughness model shows much worse performance. There is still a weak positive correlation (r = 0.6), and reasonable rank order (Rho = 0.60). the rank order improves considerably when considering the variance of the subjective data (Rho* = 0.95). However, this improvement may be due to the wide confidence intervals of the subjective data. This indicates that subjects were inconsistent in their ratings of certain stimuli.

It can also be seen that the model of roughness predicts multiple stimuli close to zero. This may be due to inadequacies in the implementation of the algorithm, or that there are missing features within the model.

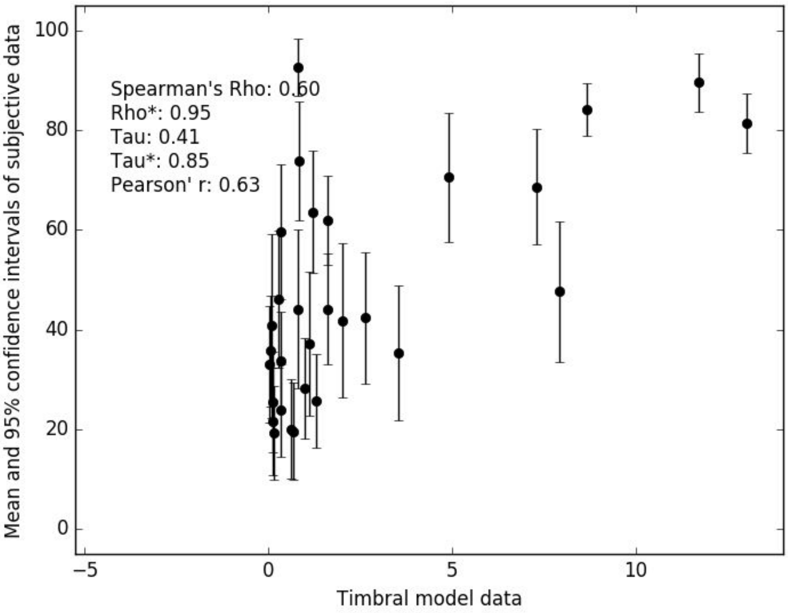


Figure 53 - Evaluation of the roughness model against the between source type dataset.

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

9 Conclusion

In this deliverable, a suitable corpus of audio stimuli was compiled from freesound based on the typical uses of the six timbral attributes that have been modelled: hardness, depth, brightness, metallic-nature, reverb, and roughness. Listening tests were conducted on these stimuli to obtain subjective ratings in each of these six timbral attributes. These subjective ratings were then used to evaluate the performance of each model.

The performance of all models against their respective source types is shown in Table 9, and Table 10 shows the performance of each model for the between source type analysis. The tables are colour coded to show a good performance (Spearman's rho ≥ 0.8), reasonable performance (Spearman's rho of 0.5 - 0.8), and poor performance (Spearman's rho < 0.5).

Table 9: Summary of the performance of all timbral models with their respective source types.

Timbral model	Source type	Spearman's Rho	Rho*	Kendall's Tau	Tau*
Hardness	Kick	0.5	0.75	0.4	0.6
	Piano	0.2	1.0	0.2	1.0
	Cymbal (Hit)	0.9	1.0	0.8	1.0
	Snare	0.2	1.0	0.2	0.8
	Guitar	0.9	1.0	0.8	1.0
Depth	Bass	0.5	0.95	0.4	0.8
	Kick	1.0	1.0	1.0	1.0
	Drum loop	0.9	1.0	0.8	1.0
	Voice	0.8	0.8	0.6	0.6
	Impact	0.7	0.9	0.6	0.8
Brightness	Ambience	0.6	1.0	0.4	1.0
	Piano	0.8	0.9	0.6	0.8
	Bell	1.0	1.0	1.0	1.0
	Voice	0.4	0.45	0.4	0.4
	Swoosh	0.9	1.0	0.8	1.0
Metallic-nature	Clang	-0.5	0.1	-0.4	0.0
	Footsteps	0.67	1.0	0.4	1.0

	Creak	-0.8	-0.15	-0.6	-0.4
	Scrape	0.7	0.7	0.6	0.6
	Whoosh	-0.1	0.45	0.0	0.2
Reverb	Snare	0.3	0.5	0.2	0.4
	Atmosphere	0.3	0.6	0.2	0.4
	Hit	-0.71	0.1	-0.1	-0.8
	Footsteps	0.56	0.6	0.2	0.4
	Vocal	0.1	0.35	0.0	0.2
Roughness	Bell	1.0	1.0	1.0	1.0
	Synth	0.9	0.9	0.8	0.8
	Engine	0.5	0.8	0.4	0.6
	Alarm	0.9	1.0	0.8	1.0
	Guitar	0.9	0.9	0.8	0.8

Table 10: Summary of the performance of all timbral models with the between source types analysis.

Timbral model	Spearman’s Rho	Rho*	Kendall’s Tau	Tau*
Hardness	0.75	0.98	0.52	0.91
Depth	0.72	0.94	0.53	0.81
Brightness	0.65	0.96	0.49	0.84
Metallic-nature	-0.03	0.69	-0.01	0.44
Reverb	-0.12	0.56	-0.13	0.34
Roughness	0.6	0.95	0.41	0.85

Comparing the output from each model to the subjective data collected revealed that the brightness and roughness models performed well for the majority of their source types. The depth and hardness models were shown to perform well on a few source types, but poorly on others. The reverb and metallic-nature models, however, performed poorly on all source types.

The results also revealed that subjects tended to disagree in their ratings of the hardness and depth attributes more so than the others, shown by the lower grand means of the inter-subject agreement scores for these attributes. This may have been due to different interpretations of the attributes, yet additional analysis of the hardness data did not reveal independent groups of ratings. Secondly, subjects tended to have less inter-subject agreement when rating across programme items rather than rating within a single programme item.

The information presented in this deliverable will be used to further guide the development of timbral models with Task 5.2, leading to Deliverable D5.6: Second prototype of timbral characterisation tool for semantically annotating non-musical content. In this follow-up task, the evaluation of these models will be considered, and appropriate action will be taken to either improve upon the performance of the models, or to develop additional timbral models.

10 References

Evans, J. 1996: ‘Straightforward statistics for the behavioural sciences’, Pacific Grove: Brooks/Cole Pub. Co.

Francombe, J., 2014: Perceptual Evaluation of Audio-on-Audio Interference in a Personal Sound Zone System, PhD thesis, University of Surrey, Guildford, Surrey, UK.

Pearce, A., 2017: ‘Perceived differences between microphones’, PhD Thesis, University of Surrey, Guildford, Surrey, UK.

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

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