applied sciences

MDPI

Article

Modelling the Microphone-Related Timbral Brightness of Recorded Signals

Andy Pearce, Tim Brookes * and Russell Mason

Institute of Sound Recording, Department of Music and Media, University of Surrey, Guildford GU2 7XH, UK; apearce36@googlemail.com (A.P.); r.mason@surrey.ac.uk (R.M.)
* Correspondence: t.brookes@surrey.ac.uk

Abstract: Brightness is one of the most common timbral descriptors used for searching audio databases, and is also the timbral attribute of recorded sound that is most affected by microphone choice, making a brightness prediction model desirable for automatic metadata generation. A model, sensitive to microphone-related as well as source-related brightness, was developed based on a novel combination of the spectral centroid and the ratio of the total magnitude of the signal above 500 Hz to that of the full signal. This model performed well on training data (r = 0.922). Validating it on new data showed a slight gradient error but good linear correlation across source types and overall (r = 0.955). On both training and validation data, the new model out-performed metrics previously used for brightness prediction.

Keywords: audio characterisation; audio indexing; audio search; auditory perception; machine learning; music information retrieval; psychoacoustics; sound quality; sound recording; timbre

Citation: Pearce, A.; Brookes, T.; Mason, R. Modelling the Microphone-Related Timbral Brightness of Recorded Signals. *Appl. Sci.* **2021**, *11*, 6461. https://doi.org/10.3390/app11146461

Academic Editor: Francesc Alías

Received: 4 June 2021
Accepted: 9 July 2021
Published: 13 July 2021

1. Introduction

Audio database searching can be facilitated by metadata relating to the characteristics of each audio excerpt. The ability to generate such metadata automatically can therefore be useful. A key characteristic of sound is its timbre, and brightness is one of the most commonly used timbral descriptors, e.g., it is in the top 3 most-searched timbral attributes on the Freesound audio sample repository [1].

The timbre of a recorded sound will potentially be affected both by the sound source and by the choice of microphone used to record it. Previous work by the authors in [2] identified 31 perceptual attributes of recorded sound that can be affected by microphone choice. The largest impact of microphone choice was found to be on the brightness of the resulting recording. Models of brightness perception have been developed and tested previously, but not specifically in connection with microphone performance. The work documented in this paper aimed to develop a model sensitive to microphone-related, as well as source-related, brightness differences and to evaluate it against existing brightness models found to be useful in other application areas, e.g., differentiating between musical instruments [3–5], between variations of a single instrument [6], between musical performances [7], or between vocal articulations [8].

This paper is split into four sections: Section 2 describes listening tests used to collect listener ratings of perceived brightness; Section 3 introduces previous research into the acoustic correlates of perceived brightness, outlines potential improvements upon these, and develops a linear regression model using the most suitable; Section 4 then presents an evaluation of the brightness model; and finally, in Sections 5 and 6, the model's performance is discussed and conclusions are drawn.

2. Ratings of Brightness

To provide a dataset on which to base a model, listener ratings of brightness must be obtained for stimuli recorded using a variety of microphones. The stimuli used for the

collection of brightness ratings were those described in detail by the authors previously [2]; the stimulus creation method is summarised below.

2.1. Selecting Microphones

Microphones were selected in two stages. Firstly, microphones were identified that represented the extremes in each of seven physical/acoustic factors that are known to affect performance: transduction type, sensitivity, frequency response, directivity pattern, self-noise, diaphragm size, and transient response. This resulted in a shortlist of eight studio microphones. Five independent experts were then asked to say whether use of these microphones would be likely to result in recordings exhibiting a full range of perceptual differences, and to suggest additional microphones if not. Responses indicated that no additional microphones were required. In addition to these eight studio microphones, two MEMS (MicroElectricalMechanical System) microphones with markedly different perceived sonic qualities were chosen; outside of professional recording environments, mobile phones are probably the most common devices used to capture audio and these employ MEMS microphones. The resulting ten microphones were AKG C12, AKG C414 B-XLS, AKG C451, Coles 4038, DPA 4006-TL, Electrovoice RE20, sE 2200a, Shure SM58, Wolfson WM7131, and Knowles SPU0410HR5H.

2.2. Selecting Sources

In previous work by the authors, it was found that musical sources were good at revealing the perceptual differences between microphones (better than vocal sources) [9]. Therefore, only musical sources were considered for this study. Sources were selected to provide a variety of frequency spectra, dynamic ranges and transient components, in order to: (i) be able to reveal the likely effects of the physical/acoustic factors listed in Section 2.1; and (ii) be likely to cover a wide range of source brightnesses. This resulted in five sources being selected: double bass, drums, acoustic guitar, string quartet, and trumpet. On each of these sources, trained musicians played unaccompanied musical phrases lasting 7–12 s.

2.3. Recording Stimuli

The selected microphones were set up in a circular array of 150 mm diameter, as shown in Figure 1, in an ITU-R BS 1116 compliant listening room [10] (which has an acoustic that is dry but not anechoic). All sources (instruments and musicians) were positioned 1.5–2 m from the array, no closer than ten times the array size. This setup has been shown previously by the authors to ensure that the perceived differences between the recordings made using any two of the microphones are predominantly due to the differing characteristics of those microphones, rather than due to their differing positions in the array [9].

All microphones were connected to a Presonus Digimax FS microphone preamplifier feeding an RME Fireface 800 audio interface. Audio was captured with a sample rate of 44.1 kHz and a bit-depth of 24 bits. The MEMS microphones were supplied with 2.7 V power and recorded through the instrument inputs of the preamplifier due to their high output impedance. To equalise differing microphone output levels, pink noise was reproduced at a measured level of 74 dB_{SPL} at the microphone array through a Genelec 1032 loudspeaker positioned 1.5 m directly in front of the array. The input gain on each of the ten preamplifier channels was then adjusted to give a digital signal level of −36 dBFS at the audio interface.

Recording 5 sources using 10 microphones provided 50 stimuli for rating. This number is in the upper half of the range of numbers of stimuli employed in the brightness-related perceptual studies cited in Section 3.1 (range 16–72, median 32, mean 37). All stimuli are publicly available in the data archive referenced in the Data Availability Statement at the end of this paper.

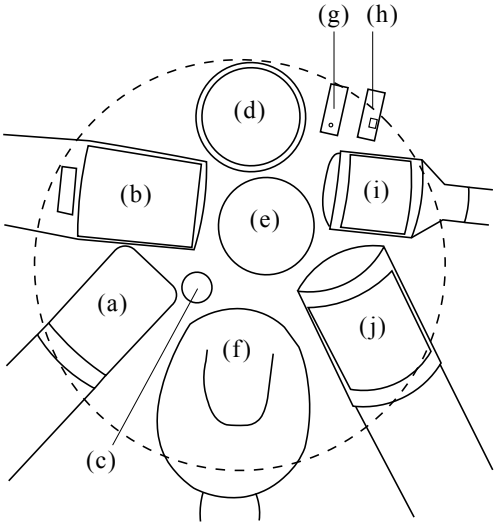

Figure 1. Recording microphone array. (a) AKG C12; (b) AKG C414 B-XLS; (c) DPA 4006-TL; (d) Electrovoice RE20; (e) Shure SM58; (f) Coles 4038; (g) Wolfson WM7131; (h) Knowles SPU0410HR5H; (i) AKG C451; (j) sE 2200a. Dashed line represents 150 mm diameter circle within which all microphone capsules lie.

2.4. Ratings of Brightness

Ratings of brightness were obtained using a two-part multiple-stimulus comparison listening test. In part 1, each page of the listening test presented all ten recordings of a particular source, and listeners were asked to rate the perceived brightness of each recording. Listeners were asked to use the full rating range on each listening test page. This was to minimise scale compression effects and potentially poor discrimination between stimuli that might otherwise have resulted from a particular source having unusually low or unusually high inherent brightness. For each listener, two of the sources were randomly selected to be repeated to assess listener consistency.

In part 2, the most and least bright recordings of each source were presented all on a single test page, and listeners were again asked to rate perceived brightness. Results from a pilot experiment indicated that for each source the Knowles SPU0410HR5H and Coles 4038 microphones produced the most and least bright recordings, respectively. The results of this part of the test will allow the part-one results to be scaled appropriately and combined across sources.

Stimuli were presented diotically over a pair of Sennheiser HD650 headphones driven by a Focusrite Virtual Reference Monitoring (VRM) Box interface, with the VRM feature disabled. All stimuli were loudness matched: for each stimulus, five listeners independently adjusted playback gain until the stimulus loudness matched that of a reference stimulus judged to provide a comfortable listening level; for each stimulus, the mean playback gain (across the five listeners) was then adopted for the listening test. Twenty participants took part in the experiment, all of whom were undergraduate students on the Music and Sound Recording course at the University of Surrey; all had participated in multiple listening tests previously, and all had passed a taught module in technical listening.

2.5. Analysis of Listening Test Results

To screen listeners, for each listener the within-subject consistency (the mean square difference between ratings for each of the two repeated sources) and between-subject agreement (the Pearson's correlation to the mean of all subjects' responses, for each of the sources) were calculated. In both measures, only one subject performed worse than one standard deviation away from the mean consistency/agreement across listeners. This listener was removed from subsequent analysis.

To retain a balanced dataset across all sources, the results from the repeated pages in each subject's tests were discarded. To allow direct comparison of the results across sources, the results from part 2 of the test were used to scale the results from part 1 using a linear transform. The transform was calculated to adjust the part 1 experiment results so that the mean values of the most and least bright stimuli for each source in part 1 equalled the mean values of the identical stimuli in part 2. Prior to the scaling, the part 1 test results spanned a range of zero to 100 for each source type. The scaled results are shown in Figure 2.

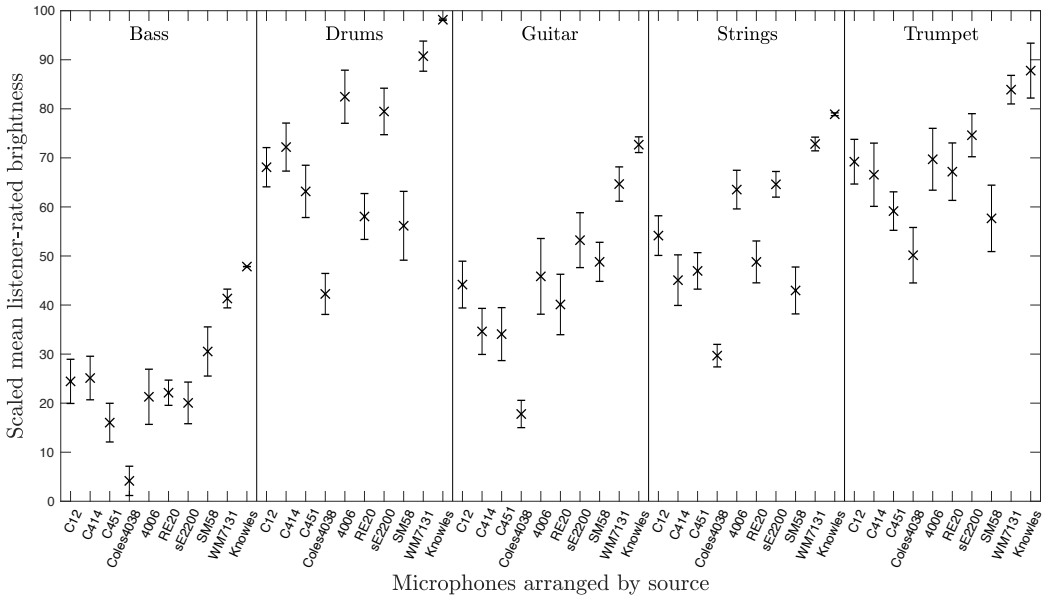

Figure 2. Listener ratings of brightness. Data for each source (from test part 1) have been scaled according to the cross-source ratings of the most and least bright recordings of that source (from test part 2).

The least bright stimulus was the bass recorded with the Coles 4038 microphone; the brightest stimulus was the drums recorded with the Knowles SPU0410HR5H microphone. It should be noted that, within the data for each source, several microphones are rated similarly in terms of their brightness, shown by the similar mean values and overlapping confidence intervals. However, the means cover 38–56% of the rating scale within the data for each source, and 94% overall, making the data potentially useful for modelling purposes.

3. Brightness Modelling

With ratings of brightness obtained, a model can be created to predict these ratings. First, a review into the existing correlates of brightness was conducted to identify metrics that may accurately predict perception.

3.1. Correlates of Brightness

In multiple application areas (e.g., comparisons between musical instruments, between variations of a single instrument, or between musical notes/chords), research has found that the spectral centroid (SC) correlates highly with the perception of brightness [3–6]. The SC can be calculated from the discrete frequency representation of an audio signal as

$$C = \frac{\sum_{n=1}^{N} f(n)x(n)}{\sum_{n=1}^{N} x(n)}, \quad (1)$$

where C is the spectral centroid, $f(n)$ is the frequency of the nth bin, and $x(n)$ is the magnitude of the nth bin. For signals that have more high-frequency energy, the SC will be higher, and it will be lower for signals that contain more low-frequency energy.

It has been noted (by a research team that included the second author) that the standard SC metric is calculated with a linearly spaced frequency scale, and suggested that use of a more perceptually meaningful scale, such as mel, Equivalent Rectangular Bandwidth (ERB), or cents, may relate better to perception [11].

Brightness is commonly considered to relate to the high-frequency content of an audio signal [3]. Changes in low-frequency content are therefore unlikely to affect the perception of brightness, yet will affect the SC. One potential improvement to the SC as a brightness predictor would therefore be to calculate it only for signal energy above a particular frequency, f_c, as SC_{fc}. This can be performed as shown in Equation (2), where C_{fc} is the frequency limited spectral centroid and n_c is the frequency bin that most closely matches f_c.

$$C_{fc} = \frac{\sum_{n_c}^{N} f(n)x(n)}{\sum_{n_c}^{N} x(n)} \tag{2}$$

Two ratio-based metrics have been proposed to predict perceived brightness: the ratio of the total magnitude of the high-frequency portion of the signal to that of the low-frequency portion of the signal, as calculated from Equation (3) [7]; and the ratio of the total magnitude of the high-frequency portion of the signal to that of the full signal, as shown in Equation (4) [12]. Several different crossover frequencies, f_c, have been suggested that range from 500 Hz to 3 kHz [7,8].

$$\text{Ratio}_1 = \frac{\sum_{n_c}^{N} x(n)}{\sum_{n=0}^{n_c-1} x(n)}, \tag{3}$$

$$\text{Ratio}_2 = \frac{\sum_{n_c}^{N} x(n)}{\sum_{n=0}^{N} x(n)}. \tag{4}$$

For both the SC and ratio metrics, there are several additional manipulations that may make them more closely relate to the perception of brightness. Frequencies below the low-frequency limit of human hearing are unlikely to affect the perception of brightness but may affect the metrics; therefore, for this study, only frequencies greater than 20 Hz were considered (lower-frequency spectral bands were excluded). Similarly, human hearing is not equally sensitive to all frequencies within the audible range; this study therefore tested the application of A- and C-weighting filters prior to analysis. Finally, it has been shown that high-Q spectral peaks are less perceptible than low-Q peaks [13]; this study therefore trialed spectral smoothing to prevent less perceptible high-Q resonances from overly influencing the metrics.

3.1.1. Metric Parameter Values

Four frequency scales for calculating the SC metrics were considered: Hertz, mel, ERB, and cents (with a 27.5 Hz reference). For the crossover frequencies used to calculate the SC_{fc} and ratio metrics, all values reported in previous research were tested: 500, 750, 900, 1000, 1250, 1500, 2000, and 3000 Hz. The additional value of 6 kHz was also included. Thus, nine crossover frequencies in total were considered.

Each metric was tested with three audio weighting options: unweighted and with A- and C-weighting applied. In terms of spectral smoothing, seven options were tested: no smoothing, octave-band, half-octave, $\frac{1}{3}$-octave, $\frac{1}{12}$-octave, $\frac{1}{24}$-octave, and $\frac{1}{64}$-octave smoothing.

It may also be beneficial to consider the interaction between the various metrics. This can be shown by including the product of the various SC and ratio metrics. Additionally, logarithmic transformations may also improve the correlation of all suggested metrics,

since human perception of frequency and amplitude is roughly logarithmic; all metrics were tested both untransformed and with logarithmic transformation.

3.1.2. Metrics Summary

There were, in essence, four metrics under consideration: two SC types (SC and SC_{fc}) and two ratio types (Ratio_1 and Ratio_2). By combining these metrics, and using multiple options for frequency scale, crossover frequency, audio weighting, spectral smoothing and logarithmic transformation (as detailed in Section 3.1.1), a total of 8484 candidate metrics were created. Table 1 summarises the full complement of metrics and parameter values tested, including combination metrics.

Table 1. Summary of all 8848 candidate metrics for modelling brightness.

Metric	Variables	Total Number
SC	Audio weighting: 3 options Spectral smoothing: 7 options Frequency scale: 4 options Log transform: 2 options	168
SC_{fc}	Audio weighting: 3 options Spectral smoothing: 7 options Frequency scale: 4 options Crossover frequency: 9 options Log transform: 2 options	1512
Ratio_1 & Ratio_2	Ratio type: 2 options Audio weighting: 3 options Spectral smoothing: 7 options Crossover frequency: 9 options Log transform: 2 options	756
SC/SC_{fc} and Ratio Combination	SC type: 2 options Ratio type: 2 options Audio weighting: 3 options Spectral smoothing: 7 options Frequency scale: 4 options Crossover frequency: 9 options Log transform: 2 options	6048

3.2. Initial Modelling of Brightness

The model of brightness was developed in two stages. First, an automated process was used to test each candidate metric against the scaled data from all sources to identify a shortlist of metrics that can predict the overall brightness well. Secondly, these shortlisted metrics were tested against the data separately for each source type to identify the metric that is most suitable for use in a model of brightness.

3.2.1. Automated Metric Selection: Across All Sources

A linear regression model was calculated for each of the candidate metrics, and calibrated to the overall scaled brightness ratings. Linear regression was chosen as the modelling technique since the simplicity of the method reduces the likelihood of producing an overfitted model, therefore producing a model that is more likely to be generalisable. The performance of each candidate model was then assessed in terms of its Pearson's r (correlation), Spearman's rho (rank correlation), and root mean square error (RMSE), shortlisting the metric that performed best for each of these measures.

The highest Pearson's r and lowest RMSE were achieved by the same model, which used one of the SC and Ratio_2 metrics. The highest Spearman's rho was achieved by two models, both of which used one of the SC_{fc} and Ratio_1 metrics. These metrics were therefore shortlisted and the corresponding parameters are summarised in Table 2. Of the

two models that achieved the best Spearman's rho, the difference between the underlying metrics was the logarithmic transformation, with the logarithmic version performing much better in its Pearson's r and RMSE.

Table 2. Best-performing (lowest root mean square error (RMSE), highest Pearson's r, highest Spearman's rho) metrics from the original 8484 candidates.

	Best Pearson's r & RMSE	**Best Spearman's rho**	
Metric	SC & Ratio_2	SC_{fc} & Ratio_1	SC_{fc} & Ratio_1
Audio weighting	None	None	None
Frequency scale	Hertz	Hertz	Hertz
Crossover freq. (f_c)	500 Hz	1250 Hz	1250 Hz
Spectral smoothing	$\frac{1}{3}$-octave	$\frac{1}{3}$-octave	$\frac{1}{3}$-octave
Log transform	Yes	Yes	No
Pearson's r	0.922	0.919	0.646
Spearman's rho	0.902	0.906	0.906
RMSE	8.52	9.928	16.779

3.2.2. Selecting One Metric: Each Source Independently

To select which of the three shortlisted metrics is the most appropriate for a general model of brightness, a linear regression model was created from each metric to the brightness ratings for each source independently. The SC and Ratio_2 model performed better in terms of its Pearson's r, Spearman's rho, and RMSE for each source except for the trumpet, where the model achieved the same Spearman's rho as both SC_{fc} and Ratio_1 models ($\rho = 0.9879$), but slightly poorer Pearson's r and RMSE. From this it was concluded that SC and Ratio_2 metric was more suitable and would produce the best model of brightness.

3.3. Model Refinement

Since the chosen metric is the logarithm of the product of the SC and Ratio_2, the metric can be rewritten as

$$\log_{10}(\text{Ratio}_2 \cdot C) = \log_{10}(\text{Ratio}_2) + \log_{10}(C). \tag{5}$$

Using Equation (5), a multilinear regression modelling approach could be taken to find appropriate weights for both Ratio_2 and SC metrics independently. However, applying this approach did not lead to an improvement in the performance of the model, achieving the same Pearson's r, RMSE, and Spearman's rho to three significant figures. Upon performing a multilinear regression on the standardised values of the SC and Ratio_2 components, the coefficients for each were 46.1 and 42.9, respectively (to three significant figures). Since these coefficients are very similar, the relative contribution of each metric is almost equal. This explains why weighting them independently does not improve the performance of the model. Therefore, the model is a linear regression of the product of SC and Ratio_2.

The finalised brightness can be expressed as in Equation (6). The full model is summarised in Figure 3. Its performance is shown in Figure 4.

$$\text{Brightness} = -95.9388 + 44.5552 \log_{10}(\text{Ratio}_2 \cdot C). \tag{6}$$

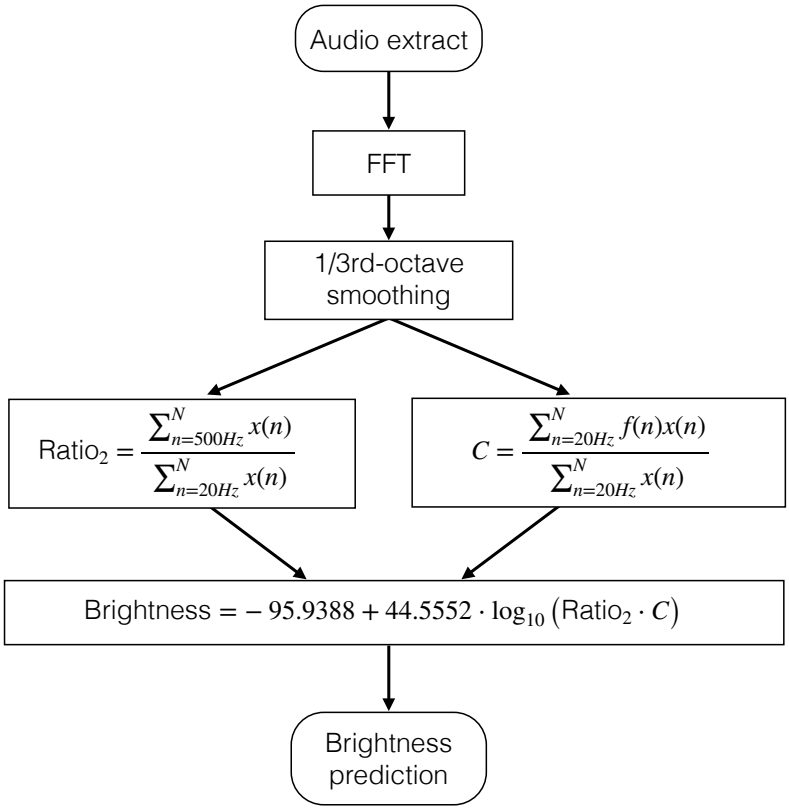

Figure 3. Flow chart of the developed brightness model (FFT = Fast Fourier Transform).

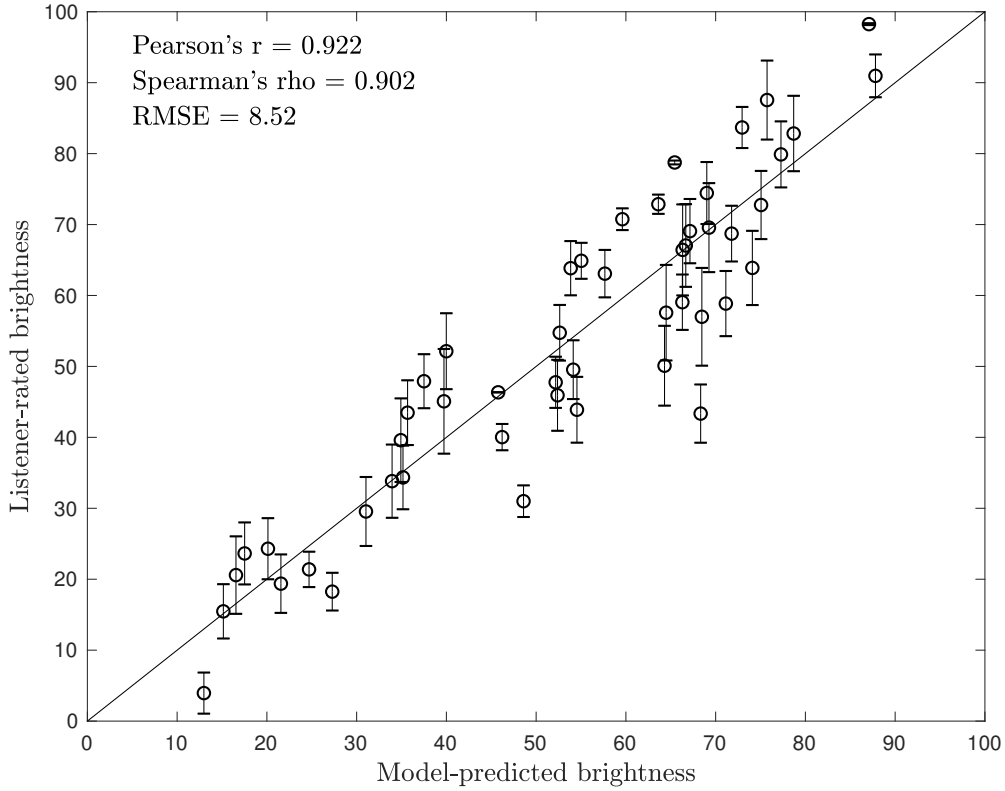

Figure 4. Listeners' brightness ratings against the best-performing model, across all sources.

4. Model Validation

In order to validate the developed brightness model, new stimuli were recorded, using new sources and microphones. Brightness ratings of these new stimuli were obtained and compared against the brightness predicted by the model.

4.1. Creating Validation Stimuli

Validation stimuli were again created as described in Section 2: selecting microphones using a combination of physical/acoustic factors and expert advice; selecting sources best able to reveal the likely effects of those physical/acoustic factors, whilst providing a range of source brightnesses; and recording stimuli with all microphones simultaneously.

4.1.1. Selecting Microphones

Ten studio microphones were selected that weren't used in the generation of the training dataset and that varied in terms of their sensitivity, self-noise, transient response, distortion, diaphragm size, transduction type, directivity, and frequency response: Sony C800; AKG D12; Schoeps CMC-6U with 2H capsule; Countryman B3; Neumann U87 (cardioid); Royer R-121; Sony F730; DPA 4015; Sony ECM670; and Hebden HS 3000 (hypercardioid). Eight independent experts were then asked to suggest microphones representing each extreme of the scale for the perceptual attributes of brightness, noise level, harshness, clarity, and piercing (the five timbral attributes contributing most to perceptual differences between microphones [2]). Analysis of these responses led to the addition of the HS3000 (omnidirectional) and AKG C1000S microphones.

In the microphone set used to generate the brightness model training data, two MEMS microphones were included. For consistency, three new MEMS microphones were included in the set used to generate the validation data: WM7132; WM7138; and WM7331. Two of each of these microphones were included in the validation microphone array: one with the soundhole facing the source (face-on, more typical studio microphone orientation); and a second with the soundhole perpendicular to the source (side-on, more typical MEMS microphone orientation).

4.1.2. Selecting Sources

Sources were selected to best reveal the likely effects of the physical/acoustic factors that differ from one microphone to another, and to provide a wide range of source brightnesses. Additionally, the model of brightness developed in Section 3.2 had a crossover frequency of 500 Hz, and so at least one source was selected that produces energy predominantly below 500 Hz, and another source which produces energy predominantly above 500 Hz, and a broadband source with similar magnitude above and below 500 Hz. This resulted in a list of six sources: banjo, cello, clarinet, glockenspiel, trombone, and ukulele.

Stimuli were recorded exactly as described in Section 2.3. Figure 5 shows the layout of the validation microphone array. All recorded stimuli were loudness matched by a panel of six listeners to produce a comfortable listening level when presented diotically over a pair of Sennheiser HD650 headphones driven by a Focusrite VRM Box interface, with the VRM feature disabled.

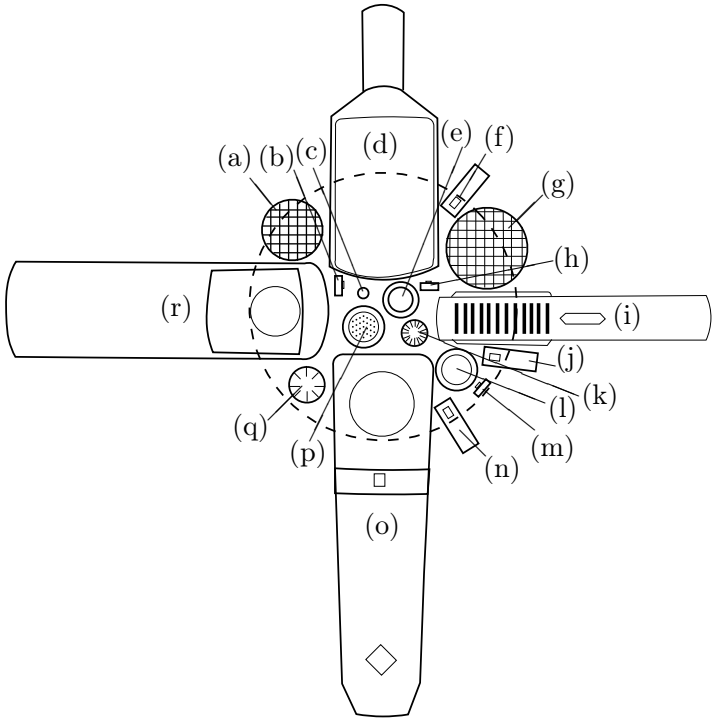

Figure 5. Validation microphone array. (a) AKG C1000S; (b) Cirrus WM7331 (side on); (c) Countryman B3; (d) AKG D12; (e) Schoeps CMC6U with 2H capsule; (f) Cirrus WM7331 (face on); (g) Sony F730; (h) Cirrus WM7132 (side on); (i) Royer R-121; (j) Cirrus WM7132 (face on); (k) Sony ECM670; (l) Hebden HS3000 (omni); (m) WM7138 (side on); (n) WM7138 (face on); (o) Neumann U87 (cardioid); (p) Hebden HS3000 (hyper-cardioid); (q) DPA 4015; (r) Sony C800.

To encourage similar scale usage here to that in the previous gathering of brightness ratings, the most and least bright stimuli overall from the training dataset (drums recorded with the Knowles SPU0410HR5H and bass recorded with the Coles 4038, respectively) were included in the part-two scaling experiment. Since these scale anchors formed part of the training dataset, the associated responses were excluded from the dataset used for validation.

4.2. Validation Ratings of Brightness

For each source, a pilot listening test was used to identify ten stimuli covering the full range of brightnesses. Three subjects were asked to rate the brightness of all eighteen stimuli per source. From the analysis of these results, only the most and least bright MEMS microphones were retained for the main experiment: the Cirrus WM7138 (face-on) and Cirrus WM7331 (side-on), respectively. Eight studio microphones were selected by identifying those rated most and least bright, as well as six others that spanned the range of brightness ratings, removing those whose ratings of brightness were similar to others. Those selected were: AKG D12; Sony C800; Countryman B3; Hebden HS3000 (hyper-cardioid); Neumann U87 (cardioid); Sony ECM670; and Sony F730.

Using the same two-part procedure as described in Section 2.4, eighteen listeners then rated the brightness of each retained stimulus.

4.3. Validation of Model

Figure 6 shows the brightness predicted by the model against the mean listener ratings. The model achieves a Pearson's r of 0.955, but the line of best fit (grey dashed line) shows a different gradient to the ideal line (solid back), leading to under-prediction for high-brightness stimuli. This is most likely due to inconsistency in scale usage between the training and validation tests (the inclusion of just two stimuli common to both might not

have been sufficient). For the purposes of audio database searching, however, a gradient mismatch between prediction and perception is unlikely to be of any consequence: if brightness metadata are used to search for sounds with a specific brightness then, since there is no standard unit of brightness perception, the search would require selection of a reference sound with the desired brightness, and the metadata would allow accurate identification of the other sounds having the most similar brightnesses; if brightness metadata are instead used to order search results, then the high value of Spearman's rho indicates that this also will be achieved accurately.

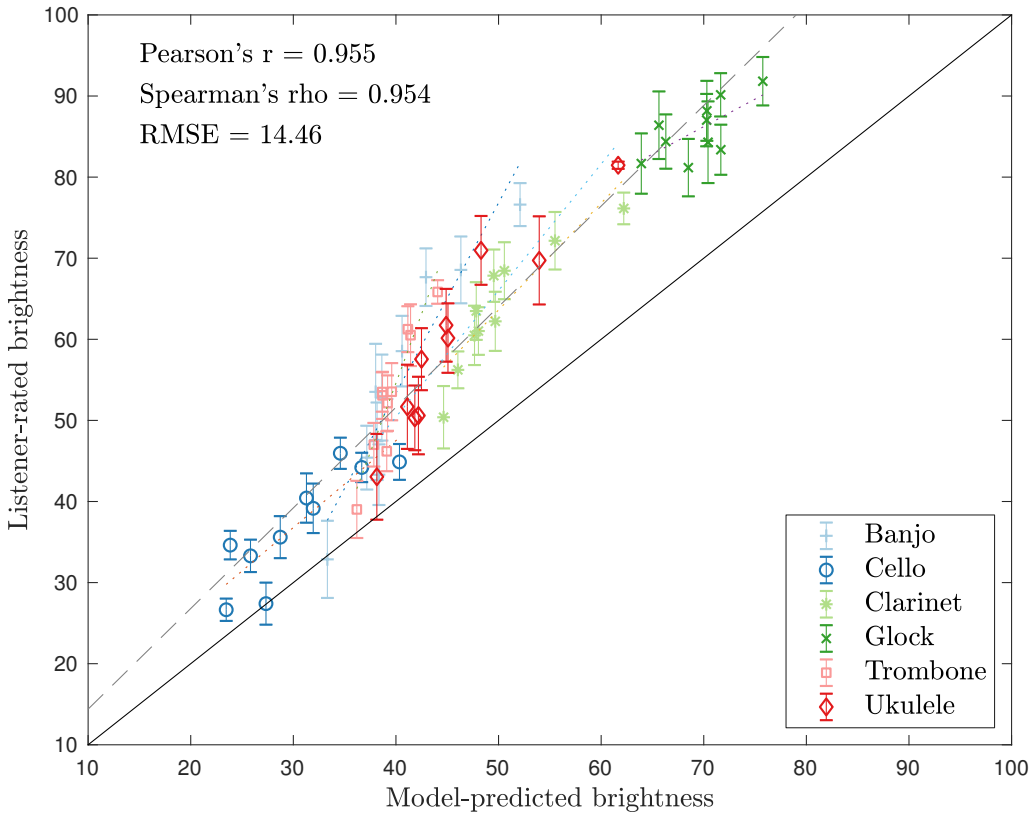

Figure 6. Validation brightness ratings against the model of brightness, colours indicating each source type. Coloured dotted lines represent the line of best fit for each source, and the grey dashed line is the overall line of best fit (Colour online).

In Section 3.3, it was shown that the contributions of the *SC* and Ratio_2 were similar by performing a multilinear regression with the standardised metrics. Performing the same analysis with the validation data produces coefficients of 12.172 and 6.815, respectively. This indicates that the *SC* accounts for more of the variance within the validation dataset than the Ratio_2 metric. However, neither constituent metric accounts for the vast majority of the variance.

4.4. Comparison to Existing Models

Table 3 compares the performance of the developed model against models built on each of the three existing metrics introduced in Section 3.1: spectral centroid (*SC*), Ratio_1, and Ratio_2. For each of these three metrics, a linear regression model was trained with the training dataset, and tested against the validation dataset. The crossover frequencies for the ratio metrics were selected from those listed in Section 3.1.1; choosing the frequencies that gave the highest Pearson's r and lowest RMSE.

Since logarithmic transformations were tested for each metric during the development of the new model, the logarithmic transformations of the *SC* and ratio metrics were also calculated for comparison.

Table 3. Comparing the performance of all brightness models.

Model	Pearson's r	Spearman's rho	RMSE
	Tested against training data		
New model	0.922	0.902	8.52
SC	0.760	0.789	14.30
log *SC*	0.811	0.789	12.87
Ratio_1 (1 kHz crossover)	0.734	0.867	14.93
log Ratio_1 (2 kHz crossover)	0.907	0.887	9.28
Ratio_2 (2 kHz crossover)	0.888	0.887	10.11
log Ratio_2 (3 kHz crossover)	0.900	0.892	9.59
	Tested against validation data		
New model	0.955	0.954	14.46
SC	0.830	0.801	19.33
log *SC*	0.864	0.801	24.00
Ratio_1 (1 kHz crossover)	0.713	0.901	17.73
log Ratio_1 (2 kHz crossover)	0.862	0.822	23.62
Ratio_2 (2 kHz crossover)	0.851	0.838	24.05
log Ratio_2 (3 kHz crossover)	0.798	0.787	34.58

For both the training and validation datasets, the new model out-performs models built on all existing metrics of brightness, in all performance measures.

5. Discussion

Much of the literature regarding the perception of brightness suggested that the spectral centroid is a suitable metric which correlates with perceived brightness [3–6]. Literature has also suggested that ratio models are suitable for modelling perceived brightness [7,12]. However, the results from this paper show that a combination of the *SC* and Ratio_2 produces a better model. The reasons for this are unclear but the authors suggest that *SC* is perhaps too sensitive a measure of spectral balance (being affected by, for example, movement of a spectral peak from a bright-sounding 8 kHz to an equally bright-sounding 7 kHz) whereas Ratio_2 is perhaps too coarse (being entirely unaffected by spectral changes that are either wholly below or wholly above 500 Hz). (Adding energy above 20 kHz will lead our model (and any other based on the same *SC* and/or ratio metrics) to over-predict brightness. If significant energy variations above 20 kHz are considered likely, and a high enough sample rate is employed to represent them, then the problem can be fixed simply by setting the upper limits of the ratio and *SC* sums to 20 kHz.)

The proposed model out-performed the existing metrics, as well as their logarithmic transformations, with all performance measures, shown in Table 3. Whilst the logarithmic transformation improved the correlation of the *SC* metric for both the training and validation dataset, a logarithmic transformation did not always improve the correlation of the ratio metrics. This is shown by the lower Pearson's r for the logarithmic models with the validation data. Interestingly, although the *SC* metric is far more commonly associated with brightness in the literature, the ratio metrics tended to produce higher correlation to the subjective brightness ratings.

The best crossover frequency, f_c, for the new model was identified as 500 Hz. This is within the range of crossover frequencies suggested by Lartillot and Toiviainen [12].

Adding a A- or C-weighting filters, or using a perceptual frequency scale to calculate the *SC*, did not improve the performance of the model. This was unexpected since these perceptual transformations were proposed to better relate metrics to subjective impressions. It may be that the logarithmic transformation applied to the selected metrics was sufficient to simulate the nonlinearities of the human hearing system.

The most appropriate level of spectral smoothing was found to be $\frac{1}{3}$-octave, implying that the stimuli contained some degree of high-Q peaks/troughs in their frequency spectra that were adversely affecting the relationship between the metrics and perceived level of

brightness. Although the $\frac{1}{3}$-octave smoothing has likely removed these, applying a wider smoothing (half-octave or octave-band width) may have resulted in a loss of detail in the signal, and thus reduced the predictive ability of the metric.

Detailed analysis of the exact impact of each parameter on brightness prediction accuracy is outside the scope of this paper but could make an interesting follow-up study.

The proposed model correlates well with the validation data, but there was a gradient difference in the predicted brightness, under-predicting the validation brightness ratings for high-brightness stimuli. It was suggested that this may have been due to inconsistency in scale usage between the training and validation tests. The training dataset ranged from 0 to 100, whereas the validation dataset ranged only from 26.7 to 91.8. It has been noted that prediction models typically express some form of offset or different gradient in the results [14], so this is not completely unexpected.

6. Conclusions

Automatic generation of metadata relating to timbre can facilitate searching of audio databases, and brightness is one of the most commonly used timbral descriptors; it is also the timbral attribute most affected by microphone choice. Models have been developed previously to predict the differing brightnesses of recorded sounds that result from differences between sound sources. This study developed a model intended, additionally, to account for the influence of the microphones used to make the recordings. This model is the product of the *SC* and Ratio_2 (the ratio of the total magnitude of the high-frequency portion of the signal to that of the full signal), with $\frac{1}{3}$-octave smoothing, 500 Hz crossover frequency, and a logarithmic transformation. The model can be summarised with the flowchart shown in Figure 3.

With a training dataset of 50 stimuli, comprising brightness ratings of five sources recorded with ten microphones, the developed model achieved a Pearson's r of 0.922, Spearman's rho of 0.902, and RMSE of 8.52. This performs better than all of the previously proposed metrics that relate to brightness (when calibrated on this dataset). Validating the model on a different dataset, composed of brightness ratings of six different sources recorded with ten different microphones, the proposed model achieved a Pearson's r of 0.955. On this validation dataset the model, again, performed better than all previously proposed metrics.

Author Contributions: Conceptualization, A.P., T.B. and R.M.; methodology, A.P., T.B. and R.M.; software, A.P.; validation, A.P.; formal analysis, A.P.; investigation, A.P.; resources, A.P., T.B. and R.M.; data curation, A.P., T.B. and R.M.; writing—original draft preparation, A.P.; writing—review and editing, A.P., T.B. and R.M.; visualization, A.P. and T.B.; supervision, T.B. and R.M.; project administration, T.B. and R.M.; funding acquisition, T.B. and R.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was primarily funded by the UK's Engineering & Physical Sciences Research Council (EPSRC) and Cirrus Logic. Additionally, some of the work was conducted as part of the AudioCommons research project and received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 688382.

Institutional Review Board Statement: The University of Surrey's Self-Assessment for Governance and Ethics—Human and Data Research (SAGE-HDR) process was completed and indicated that this study did not require formal review by an ethics committee.

Informed Consent Statement: Informed consent was obtained from all subjects involved in this study.

Data Availability Statement: The data underlying the findings presented in this paper are available from doi:10.5281/zenodo.322747. Further project information can be found at https://iosr.uk/micQuality (accessed on 4 September 2017) and at https://iosr.uk/AudioCommons (accessed on 31 January 2019).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Pearce, A.; Brookes, T.; Mason, R. Timbral attributes for sound effect library searching. In Proceedings of the AES Conference on Semantic Audio, Erlangen, Germany, 22–24 June 2017.
2. Pearce, A.; Brookes, T.; Mason, R.; Dewhirst, M. Eliciting the most prominent perceived differences between microphones. *J. Acoust. Soc. Am.* **2016**, *139*, 2970–2981. [CrossRef]
3. Schubert, E.; Wolfe, J. Does Timbral Brightness Scale with Frequency and Spectral Centroid? *Acta Acoust. United Acoust.* **2006**, *92*, 820–825.
4. Schubert, E.; Wolfe, J.; Tarnopolsky, A. Spectral centroid and timbre in complex, multiple instrumental textures. In Proceedings of the 8th International Conference on Music Perception and Cognition, Chicago, IL, USA, 3–7 August 2004; pp. 654 – 657.
5. Grey, J.; Gordon, G. Perceptual effects of spectral modifications on musical timbres. *J. Acoust. Soc. Am.* **1978**, *63*, 1493–1500. [CrossRef]
6. Poirson, E.; Petiot, J.; Gilbert, J. Study of the brightness of trumpet tones. *J. Acoust. Soc. Am.* **2005**, *118*, 2656–2666. [CrossRef]
7. Juslin, P. Cue utilization in communication of emotion in music performance: relating performance to perception. *J. Exp. Psychol. Hum. Percept. Perform.* **2000**, *26*, 1797–1813. [CrossRef]
8. Laukka, P.; Juslin, P.; Bresin, R. A dimensional approach to vocal expression of emotion. *Cogn. Emot.* **2005**, *19*, 633–653. [CrossRef]
9. Pearce, A.; Brookes, T.; Dewhirst, M. Validation of experimental methods to record stimuli for microphone comparisons. In Proceedings of the 139th Audio Engineering Society Convention, New York, NY, USA, 29 October–1 November 2015.
10. ITU-R BS.1116. *ITU-R BS.1116-3 Methods for the Subjective Assessment of Small Impairments in Audio Systems*; Recommendation BS.1116-3; International Telecommunication Union: Geneva, Switzerland, 2015.
11. Hermes, K.; Brookes, T.; Hummersone, C. The Harmonic Centroid as a Predictor of String Instrument Timbral Clarity. In Proceedings of the 140th Audio Engineering Society Convention, Paris, France, 4–7 June 2016.
12. Lartillot, O.; Toiviainen, P. MIR in Matlab (II): A Toolbox for Musical Feature Extraction From Audio. In Proceedings of the 8th International Conference on Music Information Retrieval, Vienna, Austria, 23–27 September 2007.
13. Olive, S. The Preservation of Timbre: Microphones, Loudspeakers, Sound Sources and Acoustical Spaces. In Proceedings of the 8th International Conference of the Audio Engineering Society, Washington, DC, USA, 3–6 May 1990; paper 8-018.
14. ITU-T. *ITU-T P.1401 Methods, Metrics and Procedures for Statistical Evaluation, Qualification and Comparison of Objective Quality Prediction Models*; Recommendation P.1401; International Telecommunication Union: Geneva, Switzerland, 2012.