NAVIGATING NOISE: MODELING PERCEPTUAL CORRELATES OF NOISE-RELATED SEMANTIC TIMBRE CATEGORIES WITH AUDIO FEATURES

Lindsey Reymore **Emmanuelle Beauvais-Lacasse** **Bennett Smith** **Stephen McAdams**

Schulich School of Music, McGill University, Canada
lindsey.reymore@mail.mcgill.ca, emmanuelle.beauvais-lacasse@mail.mcgill.ca, bennett.smith@mcgill.ca, stephen.mcadams@mcgill.ca

ABSTRACT

Audio features such as inharmonicity, noisiness, and spectral roll-off have been identified as correlates of "noisy" sounds; however, such features are likely involved in the experience of multiple semantic timbre categories of varied meaning and valence. This paper examines the relationships among audio features and the semantic timbre categories *raspy/grainy/rough*, *harsh/noisy*, and *airy/breathy*.

Participants (n = 153) rated a random subset of 52 stimuli from a set of 156 ~2-second orchestral instrument sounds from varied instrument families, registers, and playing techniques. Stimuli were rated on the three semantic categories of interest and on perceived playing effort and emotional valence. With an updated version of the Timbre Toolbox (R-2021 A), we extracted 44 summary audio features from the stimuli using spectral and harmonic representations. These features were used as input for various models built to predict mean semantic ratings (*raspy/grainy/rough*, *harsh/noisy*, *airy/breathy*) for each sound.

Random Forest models predicting semantic ratings from audio features outperformed Partial Least-Squares Regression models, consistent with previous results suggesting non-linear methods are advantageous in timbre semantic predictions using audio features. In comparing Relative Variable Importance measures from the models among the three semantic categories, results demonstrate that although these related semantic categories are associated in part with overlapping features, they can be differentiated through individual patterns of feature relationships.

1. INTRODUCTION

Several audio features have been identified as correlates of "noisy" sounds, including inharmonicity, spectral flatness, spectral centroid, and spectral roll-off. However, not all types of noise are semantically equal: when timbre categories are nuanced, "noisy" features may be correlates of multiple semantic categories with varied meanings and even varied valence. Through interviews and rating tasks, Reymore and Huron [1] built a 20-dimensional model of musical instrument timbre qualia. Intriguingly, the final model included three timbre dimensions plausibly related to noise components—*shrill/harsh/noisy*, *raspy/grainy* and *airy/breathy*—while a further two dimensions appeared to potentially refer to harmonicity and/or a lack of "noisy" features—*pure/clear* and *focused/compact*. Speculating on correlates of these semantic categories, Reymore and Huron [1] noted that while noise has been often associated with negative valence and high physical exertion as in Wallmark, Iacoboni, Deblieck, and Kendall [2], noise components in breathy timbres, typically measured in speech research with harmonic-to-noise ratio (HNR), may convey a sense of proximity or intimacy that carries positive valence. Thus, a feature such as HNR may be important for multiple semantic categories. Although semantic categories can share acoustic correlates, varying relationship strengths with audio features may create distinctive, perceptible patterns for listeners that are associated with varying semantic content.

The current study examined three semantic categories derived from Reymore and Huron's model: *raspy/grainy/rough*, *harsh/noisy*, and *airy/breathy*. The aim was to determine how these semantic categories may be distinguished based on their relationships with audio features. We used linear and non-linear approaches to model semantic ratings using audio features, with the goal of uncovering distinctive acoustic signatures for each semantic category. Predictors included spectral and harmonic features from a recently updated version of the Timbre Toolbox (R-2021A) [3]. Among these features, several have been associated with noise in previous literature and so were of particular interest for the interpretation of our results (see Section 2.2).

We first describe the methods used in the rating study and in audio feature extraction. These features are then used in models to predict semantic ratings. McAdams et al. [4] used Timbre Toolbox audio features to model affective qualities of timbre using both linear and nonlinear modeling approaches and found that the nonlinear approach was more successful. Accordingly, we compare linear and nonlinear methods for analysis to assess whether this observation holds in a similar dataset. We consider our findings with regard to comparative relative importance

values of features for each of the semantic categories. Our results illustrate in detail relationships between low-level features extracted using MIR techniques and high-level semantic features whose validity and reliability have been established through perceptual studies [1, 5-6].

2. METHODS

2.1 Semantic Ratings

2.1.1 Participants

Participants (n = 153; F = 95, M = 57, other = 1), were recruited using the internet platform Prolific. Participants were on average 32 years of age (SD = 11.2, range = 18–68); 41 of the 153 self-identified as musicians using the single-question measure from the Ollen Musical Sophistication Index [7]. As identified through Prolific's screening process, all participants were native English speakers. Participants provided informed consent and were compensated for their participation.

2.1.2 Materials

Stimuli consisted of 156 approximately 2-second sound clips of single notes (pitch class C) played by various orchestral instruments, normalized and matched for loudness by the researchers. Stimuli were taken from three sound banks: Vienna Symphonic Library (VSL) [8], McGill University Master Samples (MUMS) [9] and conTimbre [10]. The stimulus set included 42 instruments playing in 5 registers (C2–C6) using both traditional and unconventional playing techniques. The unconventional playing techniques that were selected generated additional noise components (such as bowing a violin at the bridge). In selecting stimuli, we aimed to sample as widely as possible from the semantic space of interest—that is, to sample sounds representing high, moderate, and low ratings on all given categories of interest. To help achieve this goal, the final stimulus selection process was guided by the results of a pilot study (n = 10) of 46 sounds.

Principal Component Analysis of the pilot results, in which sounds were rated on individual rather than grouped terms (e.g., separate ratings were made for "airy" and "breathy") confirmed the appropriateness of grouped terms. Specifically, a three-component PCA model with promax rotation demonstrated strong loadings which aligned with the groupings of terms used in the main study (*raspy/grainy/rough*, *harsh/noisy*, *airy/breathy*).

2.1.3 Procedure

To avoid an overly long experiment, the stimulus set was partitioned into three subsets for each group of three participants, each of whom rated one-third of the stimulus set (52 sounds). This resulted in 51 complete sets of rating data on the entire stimulus set.

Participants rated how applicable each semantic category was to a given stimulus using a continuous sliding scale from 1 (*does not describe at all*) to 7 (*describes extremely well*), where the midpoint was labeled *describes moderately well*. Participants also rated valence (*negative* to *positive*) and perceived playing effort (*little to no exertion* to *high exertion*).

Ratings were made in separate blocks for each scale; participants thus rated their subset of 52 stimuli a total of five times. At the beginning of each trial (except the first in each block), the stimulus was automatically played, and participants could play the stimulus again as many times as desired. The presentation order of the scales and the presentation order of stimuli within each block were randomized. The experiment took approximately 30 minutes to complete.

2.2 Audio Feature Extraction

To investigate relationships between semantic timbre categories and acoustic features, we used an updated version of the Timbre Toolbox [3]. The Timbre Toolbox calculates spectral, temporal, and spectrotemporal acoustic features from an audio signal in Matlab [11]. First, input representations of the signal are computed. Then, both scalar and time-series features are extracted from different input representations. Lastly, the Timbre Toolbox calculates interquartile range (IQR) and median values of time-series features. These values represent the central tendency and variability of the audio features, respectively [12].

For this study, we used the STFT (Short-Time Fourier Transform) and HARM (Harmonic) input representations. The STFT is a spectrotemporal representation obtained using a sliding-window analysis over the audio signal. Then, the amplitude spectrum of the STFT is used as one of the representations to derive the audio features. HARM (sinusoidal harmonic model) is a harmonic representation that uses frame analysis to estimate slowly varying amplitude and frequency of individual harmonics [12].

Each stimulus was analyzed in the Timbre Toolbox to determine the median and IQR of audio features. Several features are derived from both the STFT and HARM representations. Results reported here for these overlapping features were taken from the STFT representation. In total, we used medians and IQRs of 22 features provided by the Timbre Toolbox [3], listed in Table 1.

Several of the features provided by these two representations in the Timbre Toolbox have been previously associated with noise, including inharmonicity, noise energy, noisiness, spectral flatness, HNR, and spectral centroid.

Audio Features from Timbre Toolbox	
Inharmonicity	Spectral Spread
Noisiness	Spectral Centroid
Noise Energy	Spectral Variation
Harmonic Energy	Spectral Roll-Off
Pitch	Spectral Decrease
Harmonic-to-Noise Ratio	Spectral Skewness
Tristimulus 1	Spectral Flux
Tristimulus 2	Spectral Kurtosis
Tristimulus 3	Spectral Flatness
Harmonic Odd-to-Even Ratio	Spectral Crest
Harmonic Spectral Deviation	Spectral Slope

Table 1. List of Timbre Toolbox audio features*;* median and IQR values were extracted for each feature.

Inharmonicity is the degree to which the frequencies of overtones depart from multiples of the fundamental frequency; inharmonicity within the auditory signal manifests as noise [2], [13]. Noise energy is calculated as the energy of the signal not explained by stable partials [3]. (Note that this definition differs from that of the feature in previous versions of Timbre Toolbox, which calculates noise energy based only on stable harmonic partials [12]). Noisiness refers to the ratio of noise energy to total energy; high noisiness values indicate a signal that is mainly nonharmonic. Spectral flatness, which has also been associated semantically with "noisy" timbres [2], [14], roughly discriminates noise from harmonic content because sinusoidal components produce a peak in the spectrum, whereas white noise produces a flat spectrum [12].

Phonetic processes, such as breathy and creaky voice, may also offer insight into the acoustic correlates that underlie semantic timbre categories. Keating et al. [15] found that different acoustic features or combinations of these features, including HNR, characterized different varieties of creaky voice. HNR is an assessment of the ratio between periodic and non-periodic components comprising a segment of an acoustic signal [15].

Spectral centroid is the center of mass of the power spectrum of an acoustic signal and is related to the perception of brightness [16]. Because Wallmark [14] suggests that increased brightness is associated with the perception of physical exertion, and because we anticipate that our semantic categories of interest will be related to perceived exertion, spectral centroid may be a relevant correlate for one or more categories.

3. ANALYSIS

3.1 Semantic Ratings

All analyses reported in this paper were carried out in R, version 4.0.5 [17]. Cronbach's alpha was calculated among complete sets of ratings using the *alpha* function in the *psych* package [18], where each set of ratings was completed by three participants (see 2.1.3). All alpha values were greater than .9, indicating excellent internal consistency. Mean semantic ratings were distributed over a large portion of the 1–7 rating scale for each category, suggesting that our stimulus set was successful in representing the semantic space of interest. Ranges among mean ratings and Cronbach's alpha values are reported in Table 2.

With Holm corrections implemented by the *corr.test* function in the *psych* package [18],[1] we observed significant Pearson correlations between *harsh/noisy* and *rough/raspy/grainy*, $r(154) = .53$ and between *harsh/noisy* and *airy/breathy*, $r = -.54$. The correlation between *rough/raspy/grainy* and *airy/breathy* was not significant.

Semantic category	Min	Max	Cronbach's α
Raspy/grainy/rough	1.63	6.72	.97
Harsh/noisy	2.12	6.45	.95
Airy/breathy	1.50	5.65	.93

Table 2. Range of mean ratings among stimuli and Cronbach's alpha for each semantic category.

[1] This method is used for all correlations reported in this paper.

3.2 Models

Following McAdams et al. [4], we performed both linear and nonlinear modeling. Scaled and centered values for the audio features from the Timbre Toolbox were used to predict mean semantic ratings; separate models were generated for each of the three semantic categories. The linear method of analysis used was partial least-squares regression (PLSR), a supervised learning algorithm that takes a dimension-reduction approach including a Principal Component Analysis process. Unlike principal component regression, however, PLSR takes both the predictor and outcome variables into account when building the linear model. This kind of statistical approach can handle data that exhibit multicollinearity and thus was appropriate for our dataset. Random forest regression was used as the nonlinear method of analysis. A random forest (RF) is a supervised machine learning algorithm that builds multiple decision trees by randomly selecting observations and specific variables and then averaging the results [19]. Both types of models were built with the *caret* package [20].

R^2 was computed on the complete dataset using ten-fold cross-validation repeated three times. To obtain Q^2, we applied a further five-fold cross-validation to each model. The observations were divided into five subsets; the model was trained on four out of the five subsets and then predicted the last remaining subset. The subsets were rotated to ensure that the training and prediction steps were applied to every combination of the subsets. Within each of the train-test subsets, models were trained using a 10-fold cross-validation repeated three times. This process also produced the RMSE values that are reported below. Table 3 contains values for R^2, Q^2, and RSME for each model.

Semantic category	Model Type	R^2	Q^2	RMSE
Raspy/grainy/rough	RF	.82	.78	.47
	PLSR	.64	.56	.70
Harsh/noisy	RF	.56	.54	.69
	PLSR	.43	.28	.93
Airy/breathy	RF	.45	.43	.78
	PLSR	.36	.29	.89

Table 3. Average R^2 and RMSE values from PLSR and RF models for all three semantic categories.

3.3 Relative Variable Importance

We calculated Relative Variable Importance (RVI) using the *varImp* function from the *caret* package [20]. RVI values are reported in Tables 4 and 5. RVI for the PLSR is based on the weighted sums of the absolute regression coefficients. The weights are a function of the reduction of the sums of squares across the number of PLS components and are computed separately for each outcome [20]. For the RF, the mean squared error is recorded on the out-of-bag portion of the data and after permuting each predictor variable. Differences between these values are averaged across all trees and normalized by the standard deviation of the differences [21]

	Airy/breathy		*Harsh/noisy*		*Raspy/grainy/rough*	
	Feature	**Value**	**Feature**	**Value**	**Feature**	**Value**
1.	Harm Spec Dev IQR	100.00	Spectral Decrease Med	100.00	HNR Med	100.00
2.	Harm Spec Dev Med	72.27	Spectral Centroid Med	64.79	Noisiness Med	92.05
3.	Spectral Roll-Off Med	68.76	Pitch Med	54.92	Spectral Variation IQR	74.63
4.	Spectral Spread Med	64.50	Spectral Spread Med	50.88	Harmonic Energy Med	73.24
5.	Spectral Centroid Med	62.17	Spectral Roll-Off Med	50.03	Inharmonicity Med	73.17
6.	Spectral Flux IQR	58.61	Spectral Variation IQR	46.63	Pitch Med	72.26
7.	Spectral Slope IQR	58.24	Harm Spec Dev IQR	40.62	Spectral Slope Med	70.63
8.	Tristimulus 1 Med	56.17	Harm Spec Dev Med	39.97	Spectral Crest Med	66.84
9.	Spectral Flux Med	52.86	HNR Med	31.52	Tristimulus 3 Med	63.57
10.	Spectral Skewness Med	48.44	Spectral Decrease IQR	31.36	Spectral Variation Med	59.95

Table 4. Top 10 important variables and their respective relative variable importance values for each semantic category using partial least-squares regression.

	Airy/breathy		*Harsh/noisy*		*Raspy/grainy/rough*	
	Feature	**Value**	**Feature**	**Value**	**Feature**	**Value**
1.	Odd:Even Ratio Med	100.00	Spectral Decrease Med	100.00	HNR Med	100.00
2.	Odd:Even Ratio IQR	72.34	Spectral Spread Med	87.68	Inharmonicity IQR	62.11
3.	Harm Spec Dev IQR	71.73	Spectral Roll-Off Med	72.04	Spectral Variation Med	54.34
4.	Spectral Roll-Off Med	49.60	Spectral Centroid Med	71.21	Noisiness Med	40.84
5.	Spectral Flux IQR	40.98	Spectral Spread IQR	62.49	Spectral Variation IQR	39.43
6.	Spectral Spread Med	30.97	Spectral Variation IQR	37.58	Tristimulus 3 Med	21.00
7.	Spectral Centroid Med	29.13	Spectral Flatness IQR	31.09	Inharmonicity Med	18.98
8.	Spectral Variation IQR	27.13	Spectral Variation Med	26.13	Pitch Med	18.81
9.	Spectral Skewness IQR	19.02	Spectral Flatness Med	24.44	Harmonic Energy Med	5.11
10.	Tristimulus 1 IQR	16.67	Noisiness IQR	24.24	Tristimulus 1 Med	3.13

Table 5. Top 10 important variables and their respective relative variable importance values for each semantic category using random forest regression.

4. DISCUSSION

4.1 Relative feature importance profiles for semantic categories

Partial least-squares and random forest models were built to predict mean semantic ratings from extracted audio features. These models were most successful in predicting ratings of *raspy/grainy/rough* (Q^2: RF .78, PLSR .56); models predicting *harsh/noisy* (Q^2: RF .54, PLSR .28) and *airy/breathy* (Q^2: RF .43, PLSR .29) were also moderately successful. McAdams et al. [4] found that a nonlinear approach produced better models than a linear approach when modeling affective qualities using Timbre Toolbox audio features. We also observed an advantage for the nonlinear method, as random forest models consistently yielded higher Q^2 values and lower RMSE values than the linear PLSR models. Because random forest regression offered the more successful models, the current discussion of results focuses on the random forest models unless otherwise noted.

HNR median was the most important variable in predicting ratings of *raspy/grainy/rough*. Spectral decrease median was the most important feature for predicting *harsh/noisy*, and harmonic odd-to-even ratio median was the most important feature for *airy/breathy*. Spectral variation IQR was in the top ten important features predictive of ratings for all three semantic categories.

Patterns of variable importance were distinct for each semantic category. Particularly among the RF models, features ranking especially high in relative importance were often unique to one of the three semantic categories, though some important features were overlapping between categories. This suggests that specific combinations of features may be important for the perception of varying semantic information.

One method of comparing feature importance among the three semantic categories is to choose a minimum importance value in order to define what constitutes a "relevant" feature. Relevant features—i.e., the features exceeding that threshold for each category—can then be compared across models. For example, spectral variation IQR is the only feature with an RVI value over 25 for all three semantic categories, suggesting that it is at least moderately relevant for models of all three categories.

In this manner, we can identify which features are uniquely "relevant" to each semantic category, where "relevant" is defined by the researcher as referring to features with RVI greater than a given value. A threshold of 25 was set for the purpose of this analysis, based on the distribution of RVI values and tractability for discussion. Definitions of "relevance" in similar interpretations can be defined with respect to the goals of the interpretation.

With this definition in mind, uniquely relevant features for *raspy/grainy/rough* include the HNR median, inharmonicity IQR, and noisiness median. Of these features,

Pearson correlations (r) demonstrate that median HNR and median noisiness were strongly negatively correlated in the dataset, $r(154) = -.96$, $p < .001$.

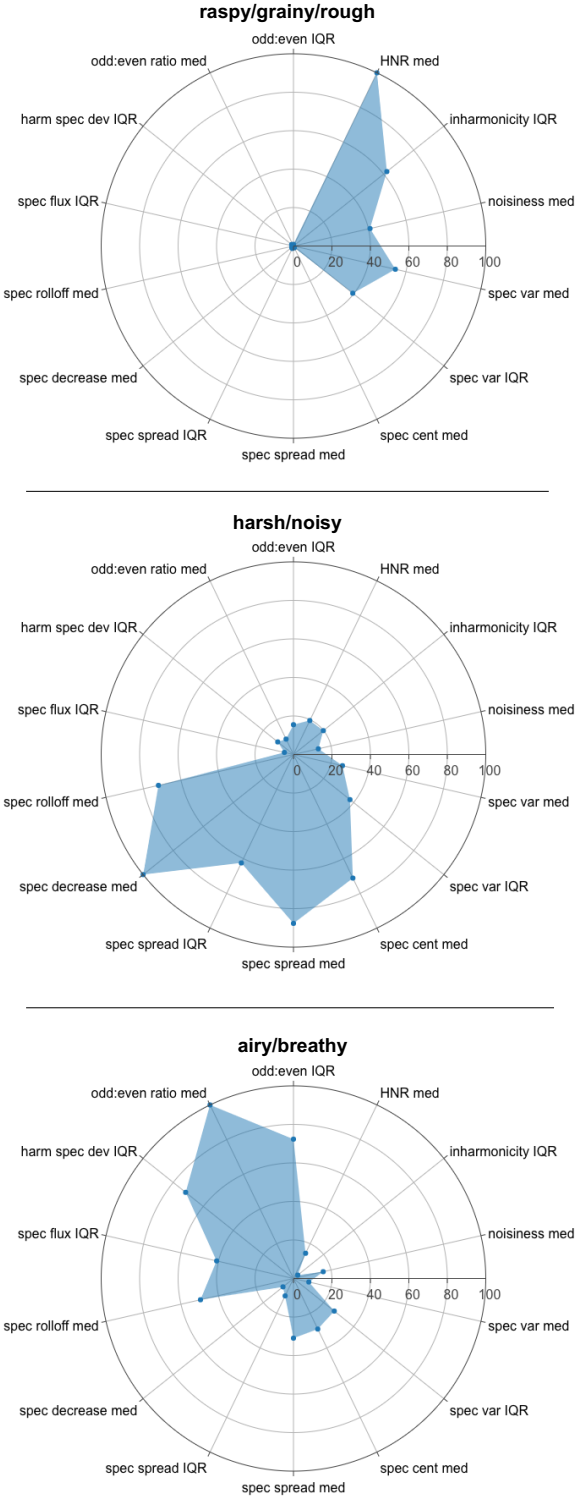

Figure 1. Radar plots of Relative Variable Importance measures for random forest models of each semantic category.

For *harsh/noisy*, uniquely relevant features include the spectral decrease median, spectral spread IQR, and spectral flatness IQR. Unique features for *airy/breathy* include the harmonic odd-to-even ratio (median and IQR, which are intercorrelated at $r = .99$), harmonic spectral deviation IQR, and spectral flux IQR.

The spectral variation median was relevant for both *raspy/grainy/rough* and *harsh/noisy*. *Harsh/noisy* and *airy/breathy* also shared relevant features—spectral roll-off median, spectral spread median, and spectral centroid median; these three features were strongly correlated among our stimuli set (roll-off/spread, $r = .97$; roll-off/centroid, $r = .95$, centroid/spread, $r = .91$).

RVI values for the 14 most important features across semantic categories from the random forest models are illustrated in the radar plots of Figure 1. The radius represents RVI; features are listed in the same order around the circles for all three plots in order to facilitate visual comparisons of semantic categories.

4.2 Noise-related features

In Section 2.2, we reviewed several features which previous literature suggests may be relevant for our semantic categories of interest, including inharmonicity, noisiness, noise energy, spectral flatness, spectral centroid, and HNR. We will now consider these specific features in relation to our semantic rating results.

Inharmonicity IQR was of particular relative importance in both linear and nonlinear models predicting *raspy/grainy/rough*, but neither the IQR nor the median were ranked highly in importance for either of the other semantic categories. Spearman correlations (ρ) suggest a robust monotonic relationship between mean ratings of *raspy/grainy/rough* and inharmonicity IQR, $\rho(154) = .78$; correlation with the median is $\rho = .59$. *Harsh/noisy* values demonstrate a moderate correlation with inharmonicity IQR, $\rho = .40$, but not with the median.

The noisiness median also received high RVI values for both linear and nonlinear models predicting *raspy/grainy/rough*, but this feature (and the corresponding IQR) received relatively low RVI values for the other semantic categories. However, both the noisiness median and noisiness IQR correlated positively with ratings of all three semantic categories. Spearman correlations were strongest for *raspy/grainy/rough*: median, $\rho = .75$; IQR, $\rho = .55$. *Harsh/noisy* demonstrated moderate correlations, median, $\rho = .39$; IQR, $\rho = .38$, and *airy/breathy* was weakly correlated but not significant, median, $\rho = .15$, $p = .06$; IQR, $\rho = .14$, $p = .08$.

Noise energy somewhat unexpectedly was not given particular importance in any of the models and demonstrated relatively weak correlations with semantic ratings.

HNR was the most important contributor to models of *raspy/grainy/rough*. Neither median nor IQR were significantly correlated with *airy/breathy*, but both were correlated with *raspy/grainy/rough*, median, $\rho = -.77$; IQR, $\rho = .40$, and *harsh/noisy*, median, $\rho = -.42$; IQR, $\rho = .35$, where higher ratings in these categories were associated with a lower HNR median but a higher HNR IQR. Given this feature's use in speech research in relation to breathy voice, the lack of significant correlation with *airy/breathy* was

surprising. Because higher HNR is also associated with higher ratings for *raspy/grainy/rough* and *harsh/noisy*, one explanation for this may be that our dataset contained many stimuli with high HNR that were very *raspy/grainy/rough* and/or very *harsh/noisy* but not *airy/breathy*. While the HNR-*airy/breathy* correlation was not significant, it was in the anticipated direction, $\rho = -.11$, $p = .17$.

Spectral flatness figured in *harsh/noisy* models but was not highly important for any of the three categories. Spearman correlations suggest positive monotonic relationships between spectral flatness (both median and IQR) with *rough/raspy/grainy*, median, $\rho = .34$; IQR, $\rho = .39$, and *harsh/noisy*, median, $\rho = .49$; IQR, $\rho = .50$. These relationships with *airy/breathy* were both weaker and in the opposite direction, median, $\rho = -.26$; IQR, $\rho = -.22$.

Spectral centroid, the correlate for semantic brightness, figured as relatively important in models for *harsh/noisy* and *airy/breathy*, but not for *raspy/grainy/rough*. The median correlated positively with ratings of *harsh/noisy*, $\rho = .58$, and negatively with *airy/breathy*, $\rho = -.45$. The IQR correlated positively with both *harsh/noisy*, $\rho = .43$, and *raspy/grainy/rough*, $\rho = .36$.

In summary, among our features of interest, we found that inharmonicity IQR, noisiness median, and HNR median seemed to be most specifically associated with *raspy/grainy/rough*, although some moderate relationships among these features can also be identified with *harsh/noisy*. Spectral flatness was weakly to moderately correlated to all three categories but did not figure prominently in models. Spectral centroid was primarily associated with *harsh/noisy* and *airy/breathy*. Both median and IQR for noisiness were correlated positively with all three categories, whereas for roll-off, flatness, and centroid, correlations for *raspy/grainy/rough* and *harsh/noisy* were in the opposite direction than those for *airy/breathy*.

4.3 Variance in valence and perceived exertion associated with semantic categories

Rating results demonstrated that the three semantic categories varied in perceived valence and playing exertion. While ratings of *raspy/grainy/rough* and *harsh/noisy* were negatively correlated with valence, $r(154) = -.90$ and $r = -.61$, respectively, ratings of *airy/breathy* were positively correlated with valence, $r = .31$. *Raspy/grainy/rough* and *harsh/noisy* were also moderately correlated with increased exertion, $r = .50$ and $r = .46$, respectively; however, ratings of *airy/breathy* did not correlate significantly with perceived playing exertion. Thus, we can consider *rough/raspy/grainy* to be associated strongly with negative valence and moderately with perceived exertion. *Harsh/noisy* is moderately associated with negative valence and perceived exertion, and *airy/breathy* is moderately associated with positive valence but not associated with exertion.

These descriptive statistics demonstrate how two semantic categories with relatively similar perceptual relationships to emotional valence and exertion may be differentiated by patterns of relationships with audio features; for example, the most important predictors of *raspy/grainy/rough* include HNR, inharmonicity, and noisiness, while the most important predictors of *harsh/noisy* include spectral decrease, spread, roll-off, and centroid. We can also see that categories with differing relationships to emotional valence and perceived exertion may both have relevant relationships with a given feature. Such overlapping relationships may be in either opposite directions—for example, *harsh/noisy* is positively associated with spectral roll-off median and spectral flatness median—or in the same direction—for example, noisiness median and HNR median.

5. CONCLUSION

This research examined associations between spectral and harmonic audio features and the timbre semantic categories *raspy/grainy/rough*, *harsh/noisy*, and *airy/breathy*. We collected semantic ratings from 153 participants for 156 orchestral instrument sounds varying in register, instrument family, and playing technique. Ratings confirmed that the three semantic categories were distinct, and that categories differed in their relationships with exertion and valence.

We built partial least-squares and random forest models predicting mean semantic ratings for each category. Across the three categories, nonlinear random forest regression models outperformed linear partial least-squares regression models. The spectral and harmonic features used in this paper were most successful for predicting *rough/raspy/grainy*, followed by *harsh/noisy*. Models were least successful in predicting ratings for *airy/breathy*.

In comparing Relative Variable Importance measures from the models among the three semantic categories, results demonstrate that although these semantic categories are associated in part with overlapping features, they can be differentiated through individual patterns of feature relationships. Among plausibly noise-related features, we observed that inharmonicity IQR, noisiness, and HNR were in general related strongly to *raspy/grainy/rough* and moderately to *harsh/noisy*. Spectral roll-off, flatness, and centroid demonstrated moderate relations to *harsh/noisy* and *airy/breathy*, but in opposite directions. Finally, the directions of associations with HNR and noisiness were the same across all three semantic categories but varied in strength.

These results contribute to efforts to bridge understandings of timbre in MIR and music cognition by clarifying the relationships between low-level audio features and nuanced semantic categories generated from perceptual studies. The methods presented here may be used to build feature profiles of other semantic categories beyond those related to noise. Furthermore, our findings may be useful in timbre synthesis, in that they can help guide the creation of sounds with specific semantic content. Such applications to synthesis may be especially relevant to audio branding and electroacoustic composition.

6. REFERENCES

[1] L. Reymore and D. Huron, "Using auditory imagery tasks to map the cognitive linguistic dimensions of musical instrument timbre qualia," *Psychomusicology: Music, Mind, and Brain*, vol. 30, no. 3, pp. 124–144, June 2020.

[2] Z. Wallmark, M. Iacaboni, C. Deblieck, and R. Kendall, "Embodied listening and timbre: Perceptual, acoustical, and neural correlates," *Music Perception*, vol. 35, no. 3, pp. 332–363, 2018.

[3] S. Kazazis, et al. Timbre Toolbox R-2021A. [Manuscript in preparation.]

[4] S. McAdams, C. Douglas, and N. N. Vempala, "Perception and modeling of affective qualities of musical instrument sounds across pitch registers," *Frontiers in Psychology*, vol. 8, no. 1, June 2017.

[5] L. Reymore, "Characterizing prototypical musical instrument timbres with Timbre Trait Profiles," In *Musicae Scientiae*, vol. 27, no. 1, April 2021.

[6] L. Reymore, "Variations in timbre qualia with register and dynamics in the oboe and French horn," In *Empirical Musicology Review*, In Press.

[7] J. Ollen, "A criterion-related validity test of selected indicators of musical sophistication using expert ratings," Ph.D. dissertation, Music, OSU, Columbus, OH, 2006.

[8] Vienna Symphonic Library GMbH, "Vienna Symphonic Library." 2011. Available: http://vsl.co.at

[9] F. Opolko and J. Wapnick, "McGill University Master Samples (3CDs)." 1987. McGill University.

[10] T. Hummel, "conTimbre Database Project." 2012. Available: http://www.contimbre.com

[11] MATLAB 2021a, The MathWorks, Inc., Natick, Massachusetts, United States.

[12] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, "The Timbre Toolbox: Extracting audio descriptors from musical signals," *The Journal of the Acoustical Society of America*, vol. 130, no. 5, pp. 2902–2916, November 2011.

[13] M. Caetano, C. Saitis, and K. Siedenberg, "Audio content descriptors of timbre," in *Timbre: Acoustic, Perception, and Cognition*, 2019, pp. 297–333.

[14] Z.T. Wallmark, "Appraising timbre: embodiment and affect at the threshold of music and noise," Ph.D. dissertation, Musicology, UCLA, Los Angeles, CA, 2014.

[15] P. Keating, M. Garellek, and J. Kreiman, "Acoustic properties of different kinds of creaky voice," in *ICHPhS*, Glasgow, UK, 2015, pp. 2–7.

[16] E. Schubert and J. Wolfe, "Does timbral brightness scale with frequency and spectral centroid?," *Acta Acustica United with Acustica*, vol. 92, no. 5, pp. 820–825, September/October 2006.

[17] R: A language and environment for statistical computing. (2021). [Online]. Available: https://www.R-project.org/

[18] W. Revelle. psych: Procedures for Personality and Psychological Research. (2020). [Online]. Available: https://CRAN.R-project.org/package=psych Version = 2.1.3

[19] G. Biau and E. Scornett, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.

[20] M. Kuhn. The Caret Package. (2012) [Online]. Available: http://cranrproject.org/web/packages/caret/caret.pdf

[21] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. *R News*, vol. 2, no. 3, pp. 18–22.