Towards an Acoustic-Semantic Space of Extreme Metal Vocal Styles

Isabella Czedik-Eysenberg[1], Eric Smialek[2], Jan-Peter Herbst[2]

[1] *SinES, Department of Musicology, University of Vienna, Austria*
[2] *Department of Music and Design Arts, University of Huddersfield, United Kingdom*

Background

Growled vocals in extreme metal are characterized by low harmonicity and high roughness and are often associated with expressive traits like "aggressiveness" [18]. Audio features can help classify metal vocals into broad style categories such as low- and high-pitched fry screams, rough singing and clean singing [12][9][7]. Apart from these basic categories, techniques such as inhaled vs. exhaled singing, jaw height, lip rounding, and tongue shape vary the vibrations that resonate along supraglottic structures, shaping extreme metal vocal timbres and contributing to acoustic markers of specific sub-genres. Such markers inform the nuanced semantic descriptions found in music reviews of black and death metal such as *"dark"*, *"haunting"*, *"energetic"* or *"atmospheric"* [15, p. 133].

Despite a widespread, informal recognition of vocal effects specific to individual subgenres, the perceptual organization of these styles has not yet been empirically demonstrated. This study addresses this gap by collecting participant responses and relating them with relevant audio features.

Aims

We aim to provide empirical evidence of how listeners interpret subgenres of extreme metal vocals. We seek to synthesize acoustic and verbal evidence via a semantically meaningful space of verbal associations that is correlated with audio features to explore the following questions:

- What are the main perceptual dimensions of different styles of metal vocals?
- Which audio signal features can be used to describe these dimensions?
- Can specific semantic descriptions such as *"monstrous"*, *"powerful"* or *"atmospheric"* be mapped onto this perceptual space and be related to specific acoustic attributes?

Methods

We extracted short phrases (~ 3–5 seconds duration) from 115 metal vocal tracks. The recordings did not contain any instrumental parts as they were taken from the original multitrack recordings of professional metal tracks provided via a partnership with *Unstoppable Recording Machine*. These excerpts were used in perceptual experiments (see below) and analyzed acoustically by extracting audio features using PRAAT/Parselmouth [4][6], Librosa [11], and Essentia [3].

Experiment 1: Similarity Rating

In order to identify the main perceptual dimensions of different metal vocal styles, a first survey asked 14 subjects to rate a subset of 10 excerpts for pairwise similarity, resulting in 45 comparisons in total. In an online listening test, each pair of excerpts was presented separately, allowing participants to listen to each stimulus as many times as they liked before providing an estimation of similarity on a slider (*"very dissimilar"—"very similar"*). The order of pairs presented was randomized.

We then constructed a similarity matrix by calculating the average similarity score across all participants for each pair of stimuli. This similarity matrix underwent metric multidimensional scaling (MDS) via scikit-learn [13] in Python. MDS is a method to find an n-dimensional spatial configuration of data that preserves pairwise (dis)similarity between examples. Acoustical (dis)similarity in our dataset appears visually as spatial distance between data points. Any resulting disparities between the original similarity values and the distance between two points in the spatial embedding are measured by means of a stress value.

Finally, we used the coordinates of the vocal stimuli along the MDS dimensions to examine correlations between the dimensions of the perceptual similarity space and the extracted audio features.

Experiment 2: Verbal Associations

In a second experiment, vocal excerpts were played to participants from the entire dataset on a self-developed web platform to collect verbal descriptions of the vocals. Participants were presented one excerpt at a time in randomized order and responded by both typing free associations and using preselected tags. The list of preselected tags was compiled based on a previous analysis of descriptive terms used in reviews of black and death metal music [15, p. 121]. Additionally, we added some terms that had often arisen in participants' free-text associations in the beginning phase of the experiment. Figure 1 shows the final list of preselected tags as presented to participants on the web interface.

Figure 1: Preselected tags displayed on the web interface.

In total, 67 people participated in the task, providing 6,073 descriptive adjectives overall (4,493 tags and 1,580 free associations).

We analyzed the verbal data collected in this associations task using two different approaches: (1) a co-occurrence network of the free associations constructed using NetworkX [5] in Python; and (2) a multiple correspondence analysis (MCA) of the preselected tags using FactoMineR [10] in R.

Results

Similarity Space

The mean similarity matrix obtained from Experiment 1 forms the basis for a perceptual similarity space computed using MDS. To determine an appropriate number of dimensions, we first evaluated the resulting stress value as a function of the number of dimensions used (see Figure 2).

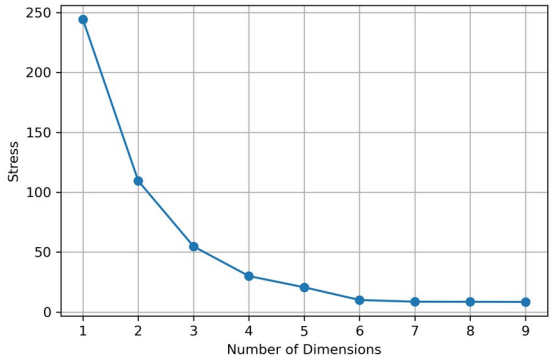

Figure 2: Scree plot showing the stress value of the MDS solution in relation to the number of dimensions used.

Based on the elbow criterion, we decided on a three-dimensional configuration, which is visualized in Figure 3.

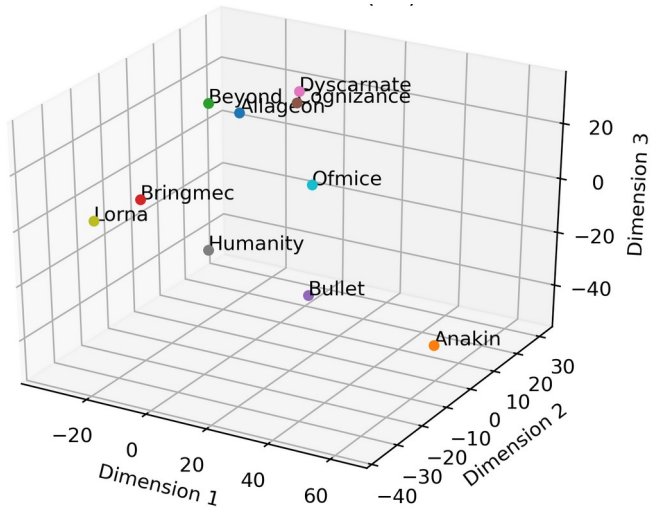

Figure 3: Visualization of the acoustic stimuli within the three-dimensional similarity space resulting from MDS.

The correlations between the resulting dimensions and audio features extracted from the stimuli show that the first major axis contrasts harmonic vs. inharmonic vocals (*Harmonic-to-Noise Ratio [HNR]*: r = 0.837, p = 0.005; *Spectral Complexity*: r = -0.959, p < 0.001). The second perceptual dimension shows no significant linear correlations with the extracted sound features but is nevertheless consequential and may correspond with higher-order categories beyond the scope of this study and briefly discussed in the conclusion below. Lastly, the third dimension relates to the position of the higher formants (e.g. *F2*: r = -0.855, p = 0.003).

Co-occurrence Network of Free Associations

From the free verbal associations, we constructed a semantic network computationally through co-occurrence analysis. We considered words (nodes) to be co-occurring (edges) if they were used—by any participant—to refer to the same stimulus. To avoid overly idiosyncratic associations, we only included in the graph co-occurrences that appeared in at least three stimuli.

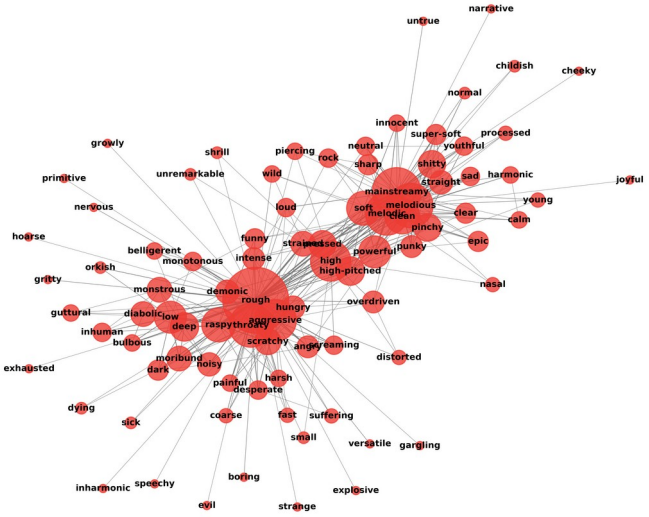

Figure 4: Co-occurrence network of free associations

Overall, the verbal associations within the semantic network can be divided into two main communities that reveal a dichotomy between *"rough"/"raspy"* and *"clean"/"melodious"* vocals. The network also shows subcommunities of more fine-grained descriptions.

Multiple Correspondence Analysis of Verbal Tags

Starting with the selected tags, we conducted a multiple correspondence analysis (MCA) using a binary matrix (stimuli x tags), indicating whether particular tags occurred for particular stimuli.

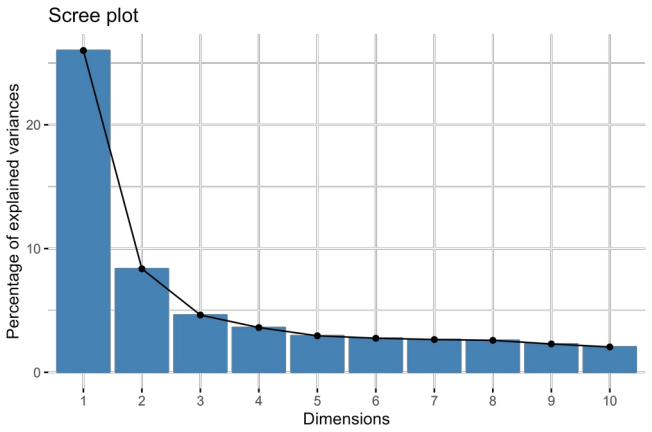

Figure 5: Scree plot showing the percentage of explained variance per dimension of the MCA.

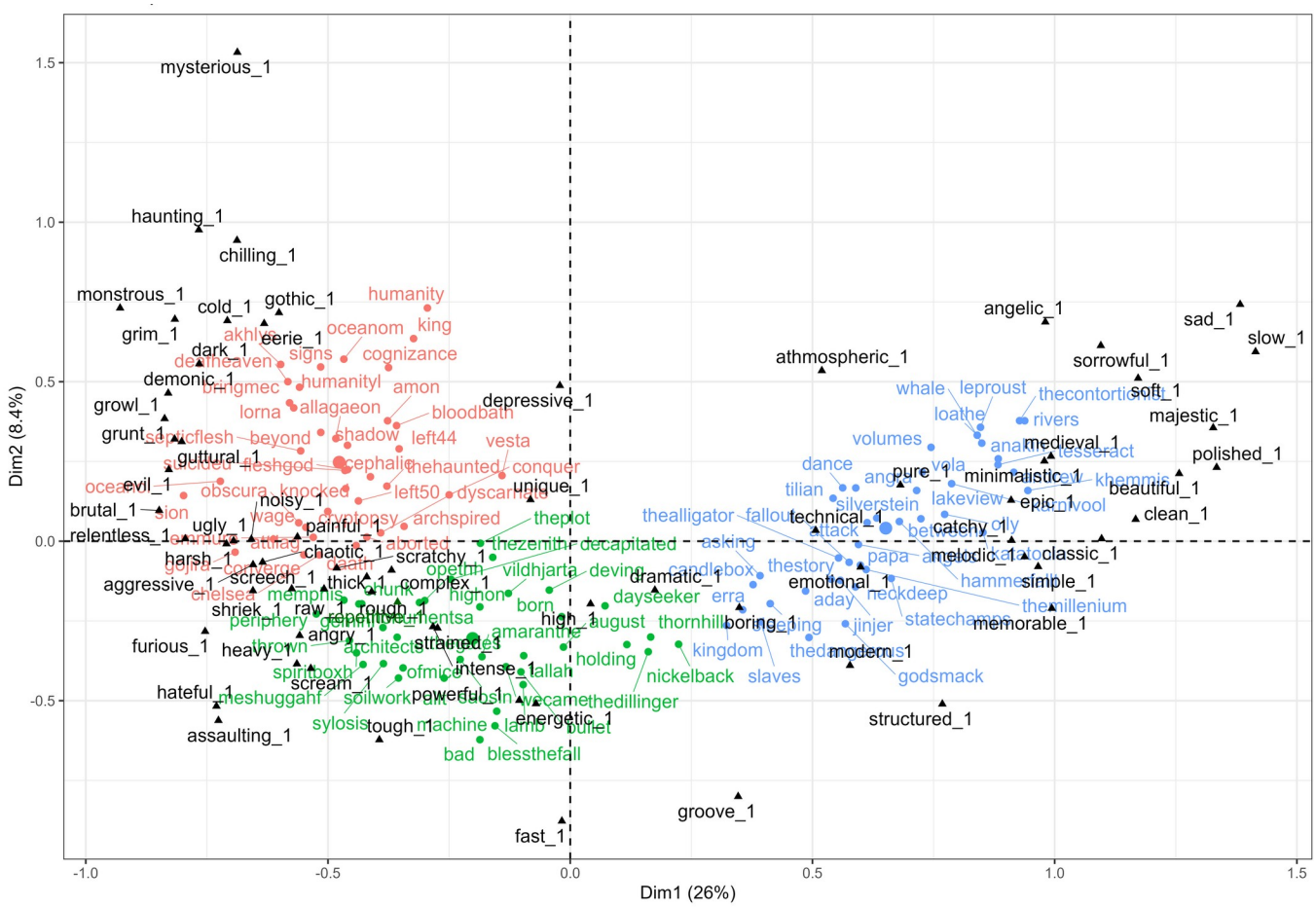

Figure 6: Biplot showing the MCA results. Verbal tags are marked with black triangles and vocal excerpts in three clusters differentiated by colour.

Based on the scree plot (Figure 5), we opted for a two-dimensional configuration, with the first dimension explaining 26% of the variance and the second dimension explaining 8.4%. Figure 6 displays the biplot, visualizing verbal associations (black triangles) as well as the acoustic stimuli (coloured circles) projected onto the space of verbal descriptions. The different colours indicate the results of hierarchical clustering using Ward's method, which resulted in three clusters roughly corresponding to clean vocals (blue), moderately rough vocals akin to emotive speech (green) and growled or whispered very inharmonic vocals (red).

Acoustically, the first dimension of descriptions according to the MCA, shows a very strong correlation with specific audio signal descriptors defined further below (Table 1 displays the four strongest observed correlations). These acoustical features relate to perceptual impressions of *inharmonicity*, *noisiness*, and *roughness*.

Table 1: Strongest audio feature correlations for the first dimension of the MCA.

Audio Feature	Pearson correlation	
	r	p
Harmonic-to-Noise Ratio	0.932	< 0.001
Shimmer	-0.929	< 0.001
Spectral Contrast (400–800 Hz)	0.916	< 0.001
Sensory Dissonance	-0.825	< 0.001

Harmonic-to-Noise Ratio (or Harmonicity) represents the energy ratio between periodic and non-periodic components within the signal, while *Shimmer* quantifies variations in signal amplitude. It is calculated as the average absolute difference in peak amplitude between consecutive periods, normalized by the average amplitude. Both measures are typically used to study voice pathologies, with higher noise and shimmer being related to hoarseness and breathiness [17]. Similarly, *Spectral Contrast* describes the difference between the energy in the peak and valley of the spectrum, with low contrast values corresponding to a noisier signal. *Sensory Dissonance* corresponds to perceptual roughness and is determined using dissonance curves [14, p. 556].

Both MCA dimensions correlate with their respective MDS dimensions (Dim1: $r = 0.901$, $p < 0.001$, Dim2: $r = -0.813$, $p = 0.004$). In both the MDS and the MCA analyses, the second dimension demonstrates a much less clear relationship with audio features than the first dimension.

Table 2: Strongest audio feature correlations for the second dimension of the MCA.

Audio Feature	Pearson correlation	
	r	p
Valence (Predictive Model)	-0.468	< 0.001
Arousal (Predictive Model)	-0.449	< 0.001
Minimum Frequency (5%)	-0.391	< 0.001
Formant 1	-0.263	0.004

The strongest relationships are found with audio models used to predict the perceived emotional qualities of *valence*

and *arousal* (Table 2). To predict these qualities, we used the AudioSet-VGGish regression models [1] available in Essentia [3]. Besides these high-level features, additional correlations that are comparatively weak can be observed. These include a minimum frequency, below which 5% of the signal energy is found, as well as the frequency of the first formant.

Discussion and Conclusion

The three analytical approaches all indicate that *Harmonicity* is the most important perceptual axis for evaluating different styles of metal vocals. Most prominently, both the first dimension of a perceptual similarity space obtained via MDS on similarity ratings as well as the first dimension of an MCA based on verbal associations of metal vocal excerpts show a high correlation with the *Harmonic-to-Noise Ratio (HNR)* extracted from the stimuli ($r = 0.837$ and $r = 0.932$, respectively). The salience of this first dimension, which also relates to the perceptual roughness (*Sensory Dissonance*) of the vocal sounds, might be explained by the distinguished role roughness seems to play in communicating danger in screamed vocalizations [2].

With both the MCA and the MDS, the second dimension does not seem to be as directly related to low-level audio features as the first. However, it may represent higher-level concepts. Semantically, this second dimension constitutes a continuum between two groups of associations that appeared in the MCA: *fast, groove, tough, energetic, assaulting* vs. *mysterious, haunting, chilling, angelic, atmospheric*. This could indicate a broad dichotomy of aesthetic tropes related to *"quotidian human toughness"* vs. the *supernatural*. Smialek [16] argues that this distinction, which can be found in lyrical themes, music videos, and similar paratexts, separates traditional metal genres from more controversial, newer forms like metalcore.

Our findings can be explored interactively through a web application, allowing users to experience them both aurally and visually: https://sinestools.univie.ac.at/emv/space/.

Acknowledgement

This research was generously funded by EU Horizon / the Engineering and Physical Sciences Research Council (EPSRC) in the United Kingdom under grant EP/X02590X/1.

References

[1] Alonso-Jiménez, P., Bogdanov, D., Pons, J., & Serra, X. (2020). Tensorflow audio models in Essentia. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 266–270). IEEE.

[2] Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A. L., & Poeppel, D. (2015). Human screams occupy a privileged niche in the communication soundscape. *Current Biology*, *25*(15), 2051–2056.

[3] Bogdanov, D., Wack, N., Gómez Gutiérrez, E., Gulati, S., Boyer, H., Mayor, O., Roma, G., Salamon, J., Zapata, J. & Serra, X. (2013). Essentia: An audio analysis library for music information retrieval. In *14th Conference of the International Society for Music Information Retrieval (ISMIR)*, pp. 493–498.

[4] Boersma, P. (2001). PRAAT, a system for doing phonetics by computer. *Glot International*, *5*, 341–345.

[5] Hagberg, A., Swart, P., & S Chult, D. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference*, 11–15.

[6] Jadoul, Y., Thompson, B., & De Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, *71*, 1–15.

[7] Kalbag, V., & Lerch, A. (2022). Scream detection in heavy metal music. *arXiv* preprint arXiv:2205.05580.

[9] Kato, K., & Ito, A. (2013). Acoustic features and auditory impressions of death growl and screaming voice. In *Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing* pp. 460–463. IEEE.

[10] Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: an R package for multivariate analysis. *Journal of Statistical Software*, *25*, 1–18.

[11] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E. & Nieto, O. (2015). Librosa: Audio and music signal analysis in Python. In *Proc. of the 14th Python in Science Conference*, Vol. 8, pp. 18–25.

[12] Nieto, O. (2013). Unsupervised clustering of extreme vocal effects. In *Proc. of the 10th International Conf. on Advances in Quantitative Laryngology*, pp. 115–116.

[13] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P, Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825-2830.

[14] Plomp, R., & Levelt, W. J. M. (1965). Tonal consonance and critical bandwidth. *The Journal of the Acoustical Society of America*, *38*(4), 548–560.

[15] Smialek, E. T. (2015). *Genre and expression in extreme metal music, ca. 1990–2015*. PhD Dissertation, McGill University (Canada).

[16] Smialek, E. (2023). Contempt-of-Core: A Reception History of Metalcore Subgenres as Abject Genres. In J.-P. Herbst (Ed.), *The Cambridge Companion to Metal Music* (pp. 281–298). Cambridge University Press.

[17] Teixeira, J. P., Oliveira, C., & Lopes, C. (2013). Vocal acoustic analysis–jitter, shimmer and HNR parameters. *Procedia Technology*, *9*, 1112–1122.

[18] Tsai, C. G., Wang, L. C., Wang, S. F., Shau, Y. W., Hsiao, T. Y., & Auhagen, W. (2010). Aggressiveness of the growl-like timbre: Acoustic characteristics, musical implications, and biomechanical mechanisms. *Music Perception*, *27*(3), 209–222.