Distorted Realities:
Classifying Extreme Vocals Between Harmonics and Noise
A Machine–Human Evaluation of Vocal Confusion Patterns

UNIVERSITÄT ZU KÖLN

Xuhong Qiu[1], Emilia Parada Cabaleiro[2]

[1]University of Cologne, [2]Nuremberg University of Music

Abstract

Extreme vocal techniques like growling and shrieking define the intensity of metal and related genres, yet remain underrepresented in both Music Information Retrieval (MIR) and musicology due to limited fine-grained vocal categorisation. Existing systems oversimplify these styles, hindering accurate tagging and retrieval. To assess the potential of machine learning as a reliable alternative to human judgment, this study compares human and machine performance in classifying such vocals. Based on the Extreme Metal Vocals Dataset (EMVD), it includes a listening test with 158 expert participants. Perceptual results achieve 76.2% Unweighted Average Recall (UAR), while a Support Vector Machine (SVM) trained on ComParE features reached 90% UAR on average using leave-3-singer-out cross-validation. These findings highlight the feasibility of automatic vocal annotation and show how MIR methods could help organise and analyse underrepresented genres, promoting diversity in musicological research.

Research Objectives

1. **Investigate confusion patterns** in human perception of extreme vocal styles and the acoustic features involved.
2. **Compare human and machine errors** to uncover shared sensitivities in vocal style classification.

Taxonomy

1. Low / Mid / High Groups
Based on interviews with metal vocalists, we applied a perceptual pitch-based classification (Low/Mid/High) using STFT-derived maximum frequencies. A Welch's ANOVA confirmed significant differences across groups ($F(2, 58.41) = 37.3$, $p < .001$), with Games-Howell post hoc tests showing all pairwise comparisons were significant ($p < .001$), indicating a robust and reliable trend in max_frequency values.

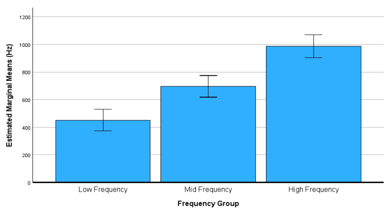


Figure 1: Estimated marginal means of max_frequency for different groups. Error bars show 95% confidence intervals.

2. Vocal Techniques / Vocal Effects
This study used a four-class vocal technique taxonomy from the EMVD dataset for machine learning. Vocal effects and other styles were excluded.

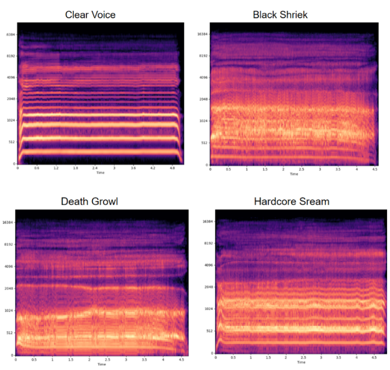


Figure 2. Log-Mel spectrograms of the four investigated vocal techniques performed by Singer ID 10 (male) with 'Mid_a' (vowel 'a' sung at a mid-frequency pitch). In the Y-axis and X-axis, frequency (in Hz) and time (in seconds) are given, respectively.

Methods

Human Perception

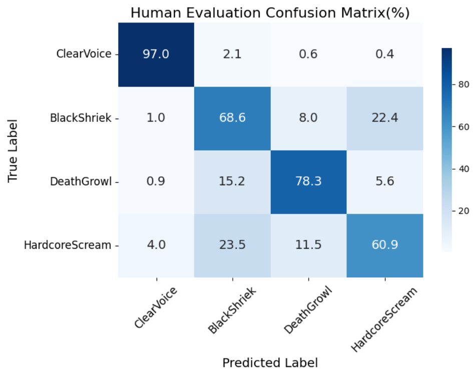
Platform: SoSciSurvey
Participants: 158 expert listeners (*M* = 29.5, *SD* = 8.1; 23 Females)
Recruitment: Metal scenes & online
Experience: 63% fans, 18% performers; 42% >10y
Task: Forced-choice (20 lyric clips, 15s each)
Stimuli: 71 clips (rated 2) from 4 techniques
Design: Demo + test; genres hidden; optional feedback via code
Evaluation: UAR, precision, recall

Machine Classification

Classifier: Support Vector Machine (SVM, linear kernel, scikit-learn)
Feature set: ComParE (Eyben et al., 2010. 6373 features via openSMILE)
Feature selection: SelectKBest (ANOVA F-test); k ∈ [100–1000]
Hyperparameter tuning: Grid search (C × k) (C ∈ [1-0.000001])
Cross-validation: Leave-3-singers-out; speaker-disjoint
Test set: Matches 71 human-labeled samples

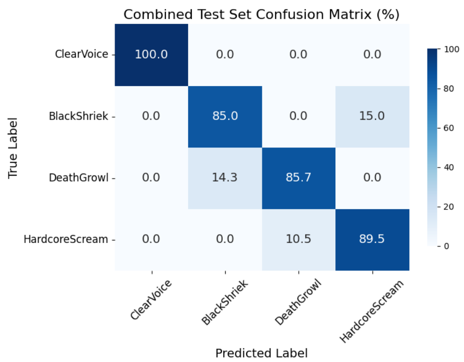
Results

Human Perception



UAR: 76.2%.
Clear Voice (97%) and Death Growl (80%) had highest **recall**.
Black Shriek (69%) and Hardcore Scream (61%) were frequently confused, especially with each other (23.5% Hardcore Scream→Black Shriek, 15.2% Death Growl→Black Shriek).
Black Shriek had lowest **precision** (63%) and may have served as a fallback label.
Results suggest that perceptual categorization was influenced by both acoustic features and contextual cues.

Mashine Learning



UAR: 90.7% (±3.9).
Clear Voice classified with 100% **precision/recall**.
Black Shriek (85%) was misclassified as Hardcore Scream (15%); Death Growl (86%) as Black Shriek (14%).
ML confusion patterns closely matched human results—except Hardcore Scream→Black Shriek, which ML avoided. High consistency and precision across categories indicate ML outperforms human perception in both accuracy and reliability.

Conclusion

This study compared human and machine classification of extreme vocal techniques using the EMVD dataset. SVM models outperformed human listeners (UAR: 90% vs. 76.2%), especially for difficult categories like Black Shriek and Hardcore Scream. Findings highlight the potential of ML for vocal annotation and the need for more inclusive, perceptually-informed classification systems.

We sincerely thank all participants and vocalists who contributed to this study.

References

This poster was first presented at the SummerSoc 2025 (Crete, Greece).

Erbe, M. (2014). By demons be driven? Scanning "monstrous" voices. In E. J. Abbey & C. Helb (Eds.), *Hardcore, punk, and other junk: Aggressive sounds in contemporary music* (pp. 51–72). Lexington Books.
Sakakibara, K., Fuks, L., Imagawa, H., & others. (2004). Growl voice in ethnic and pop styles. *Proceedings of the International Symposium on Musical Acoustics (ISMA 2004)*, Nara, Japan.
Tailleur, M., Pinquier, J., Millot, L., Vogel, C., & Lagrange, M. (2024). Emvd Dataset: a Dataset of Extreme Vocal Distortion Techniques Used in Heavy Metal. *2024 International Conference on Content-Based Multimedia Indexing (CBMI)*, 1-5.
Eyben, F., Wöllmer, M., & Schuller, B. (2010). openSMILE – The Munich versatile and fast open-source audio feature extractor. *Proceedings of the 9th ACM International Conference on Multimedia (MM 2010).* https://doi.org/10.1145/1873951.1874246
Stadler, A., Parada-Cabaleiro, E., & Schedl, M. (2023). Towards potential applications of machine learning in computer-assisted vocal training. *Proceedings of the 16th International Symposium on Computer Music Multidisciplinary Research (CMMR),* Tokyo, Japan.
Xu, Y., Wang, W., Cui, H., Xu, M., & Li, M. (2022). Paralinguistic singing attribute recognition using supervised machine learning for describing the classical tenor solo singing voice in vocal pedagogy. *EURASIP journal on audio, speech, and music processing, 2022(1),* 8. https://doi.org/10.1186/s13636-022-00240-z