UNIVERSITÄT AUGSBURG
Universitätsbibliothek

OPEN ACCESS
OPUS AUGSBURG

A machine learning-driven interactive training system for extreme vocal techniques

Johanna Holzinger, Alexander Heimerl, Ruben Schlagowski, Elisabeth André, Silvan Mertes

Angaben zur Veröffentlichung / Publication details:

Holzinger, Johanna, Alexander Heimerl, Ruben Schlagowski, Elisabeth André, and Silvan Mertes. 2024. "A machine learning-driven interactive training system for extreme vocal techniques." In *AM '24: proceedings of the 19th International Audio Mostly Conference: Explorations in Sonic Cultures, September 18-20, 2024, Milan, Italy*, edited by Luca Andrea Ludovico and Davide Andrea Mauro, 348–54. New York, NY: ACM. https://doi.org/10.1145/3678299.3678334.

Nutzungsbedingungen / Terms of use:

CC BY 4.0

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:
CC-BY 4.0: Creative Commons: Namensnennung
Weitere Informationen finden Sie unter: / For more information see:
https://creativecommons.org/licenses/by/4.0/deed.de

A Machine Learning-Driven Interactive Training System for Extreme Vocal Techniques

Johanna Holzinger
johanna.holzinger@student.uni-augsburg.de
University of Augsburg
Augsburg, Germany

Alexander Heimerl
alexander.heimerl@uni-a.de
University of Augsburg
Augsburg, Germany

Ruben Schlagowski
ruben.schlagowski@uni-a.de
University of Augsburg
Augsburg, Germany

Elisabeth André
elisabeth.andre@uni-a.de
University of Augsburg
Augsburg, Germany

Silvan Mertes
silvan.mertes@uni-a.de
University of Augsburg
Augsburg, Germany

ABSTRACT

The scarcity of vocal instructors proficient in extreme vocal techniques and the lack of individualized feedback present challenges for novices learning these techniques. Therefore, this work explores the use of neural networks to provide real-time feedback for extreme vocal techniques within an interactive training system. An Extreme-Vocal dataset created for this purpose served as the basis for training a model capable of classifying False-Cord screams, Fry screams, and a residual class. The neural network achieved an overall accuracy of 0.83. We integrated the model into a user application to enable real-time visualization of classification results. By conducting a first qualitative user study involving 12 participants, we investigated whether interacting with the training system could enhance self-efficacy regarding the correct application of extreme vocal techniques. Our study participants indicated that they found the training system helpful for learning and categorizing extreme vocal techniques.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; **Interaction design**.

KEYWORDS

Extrem Vocals, Learning Application, Interactive Learning, Vocal Techniques

ACM Reference Format:
Johanna Holzinger, Alexander Heimerl, Ruben Schlagowski, Elisabeth André, and Silvan Mertes. 2024. A Machine Learning-Driven Interactive Training System for Extreme Vocal Techniques . In *Audio Mostly 2024 - Explorations in Sonic Cultures (AM '24), September 18–20, 2024, Milan, Italy.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3678299.3678334

This work is licensed under a Creative Commons Attribution International 4.0 License.

AM '24, September 18–20, 2024, Milan, Italy
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0968-5/24/09
https://doi.org/10.1145/3678299.3678334

1 INTRODUCTION

Since its inception, the metal genre has been renowned for its intense sonic characteristics. Attributes of heavy metal, which gained popularity in the 1970s [18], incorporate a deep, distorted, and forceful guitar tone accompanied by rough vocalizations. With the genre's increasing popularity and subsequent diversification of sonic features, it was eventually categorized into subgenres in the 1980s. One such subgenre is extreme metal. Extreme metal's vocal techniques and characteristics differ significantly from those of more traditional styles. While conventional heavy metal vocals maintain pitch clarity, extreme metal emphasizes vocal distortion, resulting in sounds reminiscent of screams [14]. The association with uncontrolled screams may explain the misconception that such distorted vocal techniques harm the vocal cords and larynx. However, this assumption has proven incorrect, provided extreme vocal techniques are executed precisely and controlled [5]. Two prominent vocal techniques in extreme metal are *False-Cord Screams* (sic) and *Fry Screams* [14]. While False-Cord Screams are produced by activating the vestibular folds [14], Fry Screams involve tightly closed vocal folds to produce fry distortion. Many extreme metal vocalists can alter and refine the basic sounds of both techniques by shaping different parts of the vocal tract. However, it's crucial to note that misuse of these techniques can lead to permanent vocal damage [2], underscoring the importance of seeking professional guidance to better understand and properly apply the techniques. However, the limited availability of scientific research in this field often leads to divergent opinions among experts regarding the correct execution of extreme vocal techniques. Furthermore, the application of extreme vocal methods is limited to a niche segment, leading to a shortage of vocal instructors who can teach these specific techniques. The lack of individual feedback from experts can be a significant challenge for newcomers, especially considering the non-intuitive nature of these methods.

Our paper presents a novel approach by exploring the use of neural networks to provide individual real-time feedback for extreme vocal techniques within an interactive training system. The aim is to enable learning of scream techniques without the need for expert guidance. We created a dataset containing audio data and extracted features resembling distinct extreme metal techniques for our endeavor. Using this dataset, we trained a neural network to classify False-Cord screams, Fry screams, and a residual class. We

then integrated this model into a user application, providing real-time visualization of audio recordings' classification results. In a first qualitative user study, we investigated whether this interactive application could enhance the participants' perceived self-efficacy [7] in proficiently executing extreme vocal techniques.

2 EXTREME VOCAL TECHNIQUES

The foundation of all extreme vocal techniques used in metal music lies in vocal distortion [14]. In a broader acoustic context, distortion denotes the alteration of an acoustic signal's waveform [19]. Distorted vocals are achieved by incorporating harmonic overtones and undertones into a fundamental pitch [19]. While distortion in some vocal techniques, such as *Kargyraa*, may exhibit periodic characteristics, the aperiodic distortion prevalent in extreme metal singing renders the identification of specific pitches challenging [14]. The distortion of vocal resonance can be achieved through various methods, each engaging different muscle groups. In the metal genre, Fry screams and False-Cord screams are predominantly utilized [14]. These techniques are not uniformly defined. Therefore, the terms and associated definitions of the techniques in this work are aligned with the following descriptions.

2.1 False-Cord Screams

To produce a False-Cord Scream, the vestibular folds are activated by deliberately introducing strong airflow controlled by the diaphragm, resulting in intensified airflow due to increased subglottal pressure [14]. However, extreme metal vocalists' and vocal coaches' opinions are divided on how False Cords should be activated. While some advocate for an extended sigh or breath distortion technique, others propose deliberate activation methods akin to those found in techniques such as *Kargyraa*. The collision of the vestibular folds generates distortion, allowing spectral energy to spread across a wide frequency range without significant strain on the primary vocal folds [17]. These folds remain in an open position during this technique, minimally vibrating [1]. While their robust tissue structure generally prevents harm to vocal health, excessive airflow can strain the vocal folds, particularly when they are open, reducing air efficiency [1]. Modulating glottal pressure offers control over these flows, potentially reducing air loss by closing the arytenoids [1]. Extreme vocal coaches discern various categories of false-cord screams, delineated by distinctive physiological sensations and vibrational patterns: Type A involves a sensation in the front of the throat and attachment to the base of the tongue, with vibrations occurring in the soft palate and back wall of the throat. Type B primarily exhibits vibrations in the soft palate, potentially differing enough from Type A to be considered a distinct false cord scream subtype. This work, however, does not distinguish between these two types due to the challenge of distinguishing between them across different vocalists.

2.2 Fry Screams

Within the extreme-vocal community, a wide range of viewpoints exists regarding the potential involvement of the vestibular vocal folds alongside the vocal folds in the production of a *Fry Scream*. In this work, the *FryScream* technique involves the active participation solely of the vocal folds in generating distortion, with the vocal folds positioned in a firmly closed state. Simultaneously, subglottal pressure is increased, leading to the emission of distorted sounds due to the escape of air through the tightly stretched vocal folds. Varied perspectives among extreme-vocal coaches exist concerning the correlation between Vocal Fry and Fry Screams. One perspective proposes that Vocal Fry, characterized by closed vocal folds, can be forcefully projected to induce Fry distortion. Conversely, an alternative viewpoint posits that the vocal folds lack sufficient tension during Vocal Fry to elicit Fry distortion. The airflow is projected behind the soft palate and further into the nasal cavities. Adequate tension of the vocal folds and controlled airflow are crucial to prevent potential damage to the vocal folds. Special attention should also be paid to avoiding excessive strain on unrelated muscle groups. Additionally, applying glottal pressure is possible in this technique, although strong airflow is generally not required for generating Fry Screams. Nevertheless, divergent perspectives exist among extreme vocal coaches and practitioners regarding the volume of fry screams, with variability potentially attributed to individual vocalists.

3 RELATED WORK

3.1 Automatic Classification of Singing

Research efforts have delved into different methodological approaches and emphases in exploring singing detection and classification alongside applying real-time feedback on singing to support vocal training. A study by Huang et al. [8] focuses on detecting and localizing screams in urban sound environments, subways, and recognition of screams in noisy or domestic environments. In conjunction with a Support Vector Machine, Mel-Frequency Cepstral Coefficients were employed to classify audio blocks captured by a microphone into discrete categories of scream and non-scream. Pawel et al. [16] present a neural network for automatically recognizing singing quality, categorizing singing voices into nine different quality classes. With an accuracy of over 0.70, the neural network classified the same quality class for the singing voice that experts had assigned before. Additionally, a study by Kalbag et al. [9] addressed the recognition of screaming in heavy metal music. In this study, the authors created a screaming dataset. They further developed various methods for classifying extreme vocal techniques, although no explicit distinction was made between different types of extreme vocal techniques. When classifying singing, screaming, and no voice, a Convolutional Neural Network (CNN) achieved an overall accuracy of 0.87 using Log-Mel spectrograms as input. An approach to classifying singing, high, mid, and low Fry screams, overlaid screams, and no voice was also tested. The highest overall accuracy of 0.45 was achieved using a Support Vector Machine (SVM) receiving VGGish features as input.

Furthermore, Nieto [15] conducted a study performing unsupervised cluster analysis of extreme vocal techniques, including growls (False-Cord screams), Fry screams, and grit, achieving an overall accuracy of 0.92. However, it is noteworthy that the dataset exclusively comprised the voice of a single singer, potentially leading to overfitting of the model to this singer's voice. The mentioned research has limitations in its applicability to extreme vocal techniques in heavy metal music. While they focus on detecting and

assessing singing quality or screams, they do not specifically address the classification or assessment of extreme vocal techniques. Additionally, some of these approaches have limitations, such as relying on datasets exclusively from a single source, potentially leading to overfitting and reduced generalizability of the models.

3.2 Integration of Real-Time Feedback in Vocal Training

An experimental study by Leong et al. [13] investigated the use of real-time visual feedback technologies in vocal training for aspiring music educators. Forty participants were divided into three groups: two experimental groups received real-time feedback software (Sing and See), while the third group served as a control. The experimental groups used the software self-regulatedly over 12 weeks. Tests conducted before and after the usage period showed a significant improvement in vocal sound quality, such as pitch accuracy, among the groups that had used the software. Participants rated the software positively and exhibited an optimistic attitude toward using real-time feedback as a learning method in vocal pedagogy. An article by Fiuza et al. [12] complements existing research regarding the integration of visual feedback in vocal training. It discusses the significance of visual feedback in students' awareness building and presents applications that visualize various aspects of voice production. The contribution provides an approach to promote evidence-based practices in vocal pedagogy and demonstrates the potential of real-time visualization tools for enhancing vocal training. An application by Kumar et al. [11] addresses the challenge of learning singing and the importance of real-time feedback systems for singing students. The authors emphasize that proper singing technique is a coordinated action that integrates various bodily processes. Hence, students could benefit from real-time feedback on their singing. An algorithm for cepstral analysis for pitch determination of speech signals was developed and provided in a user-friendly application, allowing students to hear standard notes and record their performances. However, this approach does not focus on feedback on extreme vocal techniques, where pitch may play a subordinate role.

4 APPROACH

4.1 Extreme Vocal Recognition

The dataset we created for training an Extreme-Vocal Classifier consists of 11 songs and 16 YouTube tutorials demonstrating Fry and False-Cord techniques. Using the annotation software NOVA [6], we divided the audio tracks into three classes: Class 0 represents False-Cord screams, Class 1 stands for Fry screams, and Class 2 corresponds to the residual class. Those classes cover the extreme vocal techniques most commonly used in the metal genre. We conducted extensive online research, including statements from professionals and musicians, to ensure the accuracy of the assignment of distinct vocal techniques. This comprehensive research process entailed scrutiny of various sources, including interviews, live performances, and studio recordings featuring the performers. Subsequently, the categorization was verified by an experienced extreme metal vocalist. To prepare song recordings for real-time classification, we extracted the vocal audio tracks from songs using a neural network (MDX-NET Inst HQ 3) [10]. Subsequently, neural networks were employed using ultimate vocal remover (UVR) [3] to remove existing vocal effects (e.g., reverb or delay) on the isolated vocal track. Background music removal was not conducted for audio from YouTube tutorials, as they had been recorded without background music.

Further, we extracted class labels and associated metadata. The audio track was segmented into smaller samples of precisely 1000ms in duration using this metadata and a sliding window of 200 ms. Subsequently, for each audio segment, 13 Mel Frequency Cepstral Coefficients (MFCCs) were computed at 44 timestamps. Given the disproportionate representation of the residual class within the dataset, a strategic approach was undertaken to address this imbalance. Specifically, we applied undersampling techniques solely to the samples associated with the residual class. This approach focused exclusively on the song samples labeled as belonging to the residual class while preserving the entirety of the tutorials' spoken voice segments. This selective strategy was motivated by the pivotal relevance of the tutorials' verbal content to the training tool's intended use case. The dataset comprises a total of 2,195,429 samples. The final distribution of classes is depicted in Figure 1.

We utilized the resulting dataset to train an Extreme Vocal Classifier employing a recurrent neural network architecture. The structure of the Extreme Vocal Classifier comprises an input layer with dimensions (44, 13), followed by two recurrent (SimpleRNN) layers, each with 64 units, two Dense layers with 64 and 32 units, respectively, a Dropout layer with a dropout rate of 0.1, and an output layer with three units corresponding to the classification classes. The choice of recurrent layers, specifically SimpleRNN, was made to account for the sequential nature of the audio data and to capture potential short-term contextual dependencies within the audio data. The network architecture is depicted in Figure 2. We conducted an initial split into training and testing data for every performer in the dataset to perform a Leave-One-Out-Cross-Validation (LOOCV) later. In this process, the data of a single performer served as the basis for testing. The training dataset was further divided into training and validation data using a split factor of 0.15 to facilitate unbiased performance monitoring of the model during training. Subsequently, we trained multiple models with training data excluding one single performer over a maximum of 10 epochs with a batch size of 32 and a learning rate of 0.00001. Following this procedure, we used every vocal performer for the training process. We also utilized an Early Stopping mechanism to prohibit training if no improvement or a deterioration of the model's performance was observed in a subsequent epoch. In such cases, we restored the optimal weights of the model to prevent potential overfitting due to excessive training epochs.

4.2 Interactive Learning Environment

The developed training system facilitates audio data acquisition through microphone input, enabling users to engage in real-time monitoring and feedback mechanisms pertaining to extreme vocal techniques. The system monitors the volume of the recorded audio signal. Once the recorded signal reaches a length of 1000 milliseconds and a certain volume threshold is exceeded, the neural network initiates the prediction of the employed vocal technique. The prediction updates occur rapidly, every 40 milliseconds, which

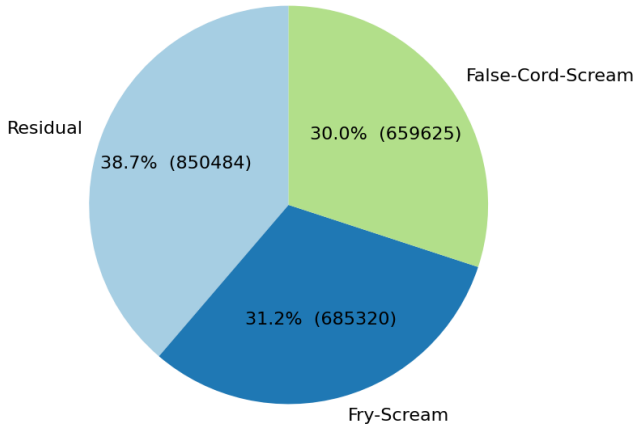

Figure 1: Class distribution of the Extreme Vocal Techniques dataset.

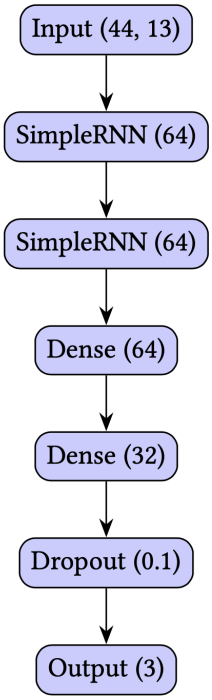

Figure 2: Model Architecture of the Extreme Vocal Classifier.

can lead to volatile fluctuations. A buffering mechanism is employed to accumulate a series of five predictions for clearer feedback and to increase user-friendliness. This approach ensures a more stable and consistent presentation of predictive outcomes. Following the accumulation of five predictions – more specifically, the class probabilities for all three classes – a cyclic removal process is enacted, whereby the oldest prediction is purged from the buffer to accommodate subsequent predictions, thus maintaining a constant buffer size. Subsequently, per class, an arithmetic mean is computed over

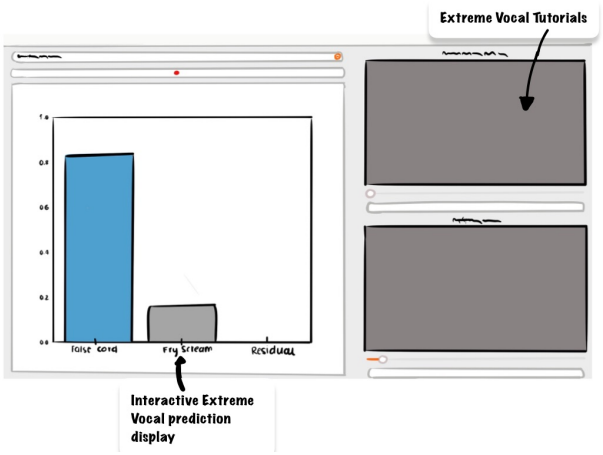

Figure 3: Schematic of the interactive real-time training system for extreme vocal techniques.

the ensemble of class probabilities within the buffer, providing a cohesive and representative summary of predictive trends. This mean value is visually depicted within the interface, as illustrated in Figure 3 (left), where the predominant neural network classification is color-coded to enhance discernibility. As such, the interface provides real-time visualization of analysis outcomes, allowing users to observe changes as they happen. Additionally, short tutorials on Fry screams, and False-Cord screams (see Figure 3: right) were integrated to enhance the learning process for beginners, offering foundational knowledge on these techniques. This enables the system to be effectively utilized even without prior knowledge.

5 EVALUATION

5.1 Computational Evaluation

5.1.1 Methodology. To evaluate the recurrent Extreme-Vocal-Classifier, a Leave-One-Out-Cross-Validation (LOOCV) was employed to assess the model's generalization ability and reduce potential biases in the evaluation [20]. Additionally, LOOCV aims to ensure that the model appropriately classifies the extreme vocal techniques and does not merely identify the performer.

5.1.2 Results. The precision values for False-Cord-Scream, Fry-Scream, and Other classes are 0.63, 0.64, and 0.95, respectively. Correspondingly, the recall values for False-Cord-Scream, Fry-Scream, and Other classes are 0.72, 0.65, and 0.91. The F1-scores for False-Cord-Scream, Fry-Scream, and residual classes are 0.67, 0.65, and 0.93, with an overall accuracy of 0.83. Detailed classification metrics are presented in Table 1. Notably, the confusion matrix in Figure 4 highlights that the False-Cord and Fry Scream classes exhibit higher confusion. In contrast, the residual class demonstrates less frequent misclassifications with extreme-vocal techniques.

5.2 User Study

We conducted a user study to investigate the impact the proposed learning application has on self-efficacy regarding the correct application of extreme vocal techniques.

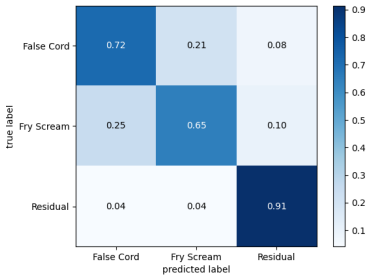

Figure 4: Confusion matrix from the assessment of the Extreme Vocal Classifier (normalized values)

Table 1: Classification report of the LOOCV of the Extreme-Vocal-Classifier.

	Precision	Recall	F1-score
False Cord Scream	0.63	0.72	0.67
Fry Scream	0.64	0.65	0.65
Residual	0.95	0.91	0.93
Accuracy			0.83
Macro Avg	0.74	0.76	0.75
Weighted Avg	0.84	0.83	0.84

5.2.1 Study Setup. Firstly, we assessed participants' prior knowledge in the field of extreme vocal techniques as well as other singing techniques. If needed, participants' previous experiences were further specified. Subsequently, we measured participants' self-efficacy regarding their ability to apply extreme vocal techniques correctly. Therefore, we used a variation of a single item (via Likert Scale) proposed by Bernacki et al. [4]. Specifically, we chose to formulate the item as *I feel able to apply extreme vocal techniques correctly.*. This assessment was repeated after participants had viewed two short tutorials on Fry screams and False-Cord screams integrated into the learning application. Active application of the techniques while watching the tutorials was not explicitly prohibited. Participants performed various vocal exercises unrelated to harsh vocal techniques based on a 5-minute vocal warm-up video to stimulate blood circulation in the vocal cords and laryngeal area. Subsequently, we initiated an interactive interaction with the learning application. Participants were free to decide how and for how long they interacted with the application. In the final quantitative assessment, we measured participants' self-efficacy regarding their ability to correctly apply extreme vocal techniques once again after completing the interactive phase. In addition, we asked participants to express both positive and negative aspects of the learning application and provide free-form feedback (i.e., open feedback).

All participants engaged with the learning tool under standardized technical conditions to mitigate discrepancies in participant interaction and study results resulting from technical disparities. Specifically, each participant utilized the same dynamic microphone (Shure SM58) to record their voice. Before the interaction, microphone input levels were individually calibrated to accommodate each participant's vocal volume. This calibration process aimed to optimize the program's sensitivity to sound, reducing the likelihood of interference frequencies or clipping at the digital audio input/interface. The study setup is shown in Figure 5.

Figure 5: Technical setup for conducting the qualitative user study.

5.2.2 Participants. We acquired twelve participants (10 male, two female, mean age 30.83 years, SD 5.70 years) for the study. There was no specific selection of participants based on their prior knowledge of extreme vocal techniques. However, most participants were active hobby musicians and/or fans of extreme music. Six participants reported no prior knowledge of extreme vocal or other singing techniques. One participant reported being experienced in other singing techniques but not extreme vocal techniques. Three participants reported experience in extreme vocal techniques but not other singing techniques, and two had experience in both areas.

5.2.3 Results. We conducted explorative statistical tests to assess whether the interaction with the developed training system can increase participants' perceived ability to apply extreme vocal techniques. Specifically, we examined whether there was a significant increase in the measured self-efficacy scores after watching the tutorials and after the interaction.

A Shapiro-Wilk test revealed a non-normal distribution of the differences between post-tutorial and post-interaction values ($p = 0.011$, $W = 0.808$). Consequently, the non-parametric Wilcoxon signed-rank test was employed. We did not find significant differences here ($W = 7.50$, $p = 0.556$), indicating that self-efficacy did not significantly increase during the interaction with the application. Descriptive statistics further characterized the data distribution, showing identical statistical values for post-tutorial and post-interaction measurements (mean = 2.58, median = 3.00, std. dev. = 0.996, respectively). Overall, these findings suggest that interaction with the training program, as implemented in the study, did not significantly enhance participants' self-efficacy.

Open feedback analysis. Our open questions explored participants' perceptions and evaluations of an interactive application designed to train them in extreme vocal techniques. We structured our qualitative analysis around thematic categories derived from participants' responses. The extracted results provide insights into participants' experiences and perspectives on the training application:

(1) *Classification Accuracy:*
Participants, with their diverse range of expertise in extreme vocal techniques, conveyed favorable feedback regarding the accuracy of sound classification and found the visual display beneficial in elucidating applied techniques. However, categorizing particularly deep False-Cord-Screams as the residual class was identified as a limitation of the Extreme-Vocal Classifier by several participants. Individuals with limited prior knowledge expressed uncertainties regarding the correspondence between the display and the produced sound.
(2) *Feedback and Display:* Participants commended the simplicity and utility of the visual feedback, perceiving the system as a promising means for reviewing and enhancing their applied techniques. However, they noted the necessity for longer vocalizations to ensure an adequate display, presenting a challenge especially for novice extreme-vocalists. Suggestions were made to refine the display for the residual class or gradually diminish its prominence to mitigate frustrations. Furthermore, recommendations were put forward for integrating natural language feedback and providing supplementary resources on specific techniques, which were identified as potential areas for refinement.
(3) *Insights into Extreme Vocal Techniques:* Participants conveyed that the application enhanced self-reflection on their proficiency in extreme vocal techniques, enabling them to identify strengths and weaknesses across various techniques. This is highly interesting, as self-reflection and self-efficacy might have substantial effects on each other - enhanced self-reflection might be the reason that self-efficacy did not significantly improve. However, the participants acknowledged the application's elucidation of the difficulty levels associated with certain techniques while expressing enjoyment in engaging with the system.

In summary, our qualitative findings indicate favorable responses from participants regarding the utility and effectiveness of the vocal training application, alongside constructive recommendations for its enhanced utilization.

6 DISCUSSION

6.1 Automatic Classification of Extreme Vocal Techniques

The current investigation introduces an Extreme Vocal Classifier, showcasing an overall accuracy of 0.83. Notably, it exhibits strong discriminatory capability between Extreme Vocal classes and the residual class. However, within the subset of Extreme Vocal techniques, the classifier demonstrates a limitation in discerning between Fry-Screams and False-Cord Screams. This may be attributed to spectrogram similarities, particularly in high-pitched False-Cord Screams and Fry-Screams featuring identical fundamental pitch, as depicted in Figure 6.

Another complicating factor in classifying extreme vocal techniques is the diverse range of manipulative options inherent in base screams and their potential overlays. In highly manipulated screams, distinguishing between False-Cord and Fry Screams can be challenging as the underlying technique may become obscured by strong overlays. Subclasses that account for this variability could enhance classification accuracy.

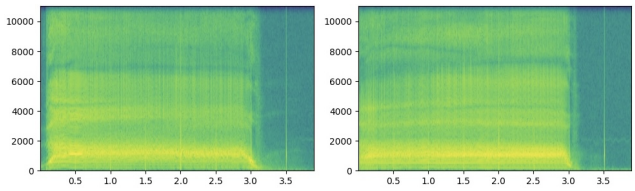

Figure 6: Left: Spectrogram of a high Fry scream; Right: Spectrogram of a high False-Cord scream. The fundamental frequency is identical for both techniques.

It is worth noting that the dataset utilized may need more diversity, as we primarily included performers whose techniques were clearly identifiable and excluded overlaid screams like Tunnel-Throat techniques. Beyond False-Cord and Fry Screams, other scream types such as Hybrid Screams, Epiglottal Growling, and Arytenoid-Cartilage Screams remain unaddressed by the current classifier, although less prevalent in the metal genre. The neural networks we used for voice track extraction from songs may have introduced distortions and complicated classification. Still, this could not be avoided due to the limited availability of YouTube tutorial data. Moreover, the exclusive utilization of Mel Frequency Cepstral Coefficients (MFCCs) without exploring alternative feature sets leaves room for potential improvements. While Nieto's unsupervised clustering [15] achieved higher overall accuracy, albeit on data from a single vocalist, the developed Extreme Vocal Classifier achieved solid results across multiple performers in Leave-One-Out Cross-Validation (LOOCV). Furthermore, this study successfully expanded upon Kalbag et al.'s scream classification [9] to differentiate between False-Cord and Fry Screams.

6.2 Interactive Training System with Real-time Feedback on Extreme Vocal Techniques

The quantitative analysis conducted within the scope of this study does not support the hypothesis of increased self-efficacy regarding the accurate application of extreme vocal techniques. However, we emphasize that this finding does not inherently signify the ineffectiveness of the developed training system. Instead, several methodological constraints in executing the user study may be pivotal in interpreting these outcomes. The limited sample size comprising twelve individuals may have impacted the study's statistical power. Moreover, recruiting participants in the niche domain of extreme vocal techniques presents challenges, inherently restricting participant numbers.

Furthermore, it is essential to acknowledge that self-efficacy assessment may be subject to the participants' prior knowledge levels, with novice participants potentially predisposed to overstating their self-efficacy. This tendency can be attributed to the misconception that techniques appear simpler in tutorials than in actual execution. Hence, self-efficacy levels may remain stagnant or decrease due to such influential factors following brief, one-time interactions. For instance, some participants indicated that the training system highlighted the substantial difficulty of these techniques, potentially indicating initial overestimation. Another critically considered aspect is the limitation of interaction to a single session, which is

likely insufficient to effect significant change in self-efficacy. The intensive physical exertion from the techniques may require a longer and more intensive interaction, but this exceeds the temporal and logistical limits of the present work.

Conversely, the qualitative analysis underscores a positive sentiment regarding feedback on the system. Participants lauded the utility of the intuitive and unobtrusive interface, facilitating technique categorization and control. Nonetheless, shortcomings were identified, notably concerning the classification adequacy of the model and the absence of classifications for screams not encompassed in the dataset. Additionally, participants expressed a desire for supplementary natural language feedback to streamline the learning process, given the current absence of improvement suggestions for screams. Despite providing real-time feedback, users still rely on independent exploration of techniques.

Collectively, these discussions imply that a more comprehensive investigation, encompassing a larger participant cohort and an extended interaction duration, may be indispensable to meaningfully gauge the training system's impact on self-efficacy. Nevertheless, the qualitative evaluation of the conducted user study continues to underscore the positive ramifications of integrating real-time feedback into vocal pedagogy for extreme vocal techniques.

7 CONCLUSION

This work involved creating an Extreme-Vocal dataset containing songs and YouTube tutorials demonstrating two extreme vocal techniques: Fry Screams and False-Cord Screams. MFCCs were extracted from this dataset to train a recurrent neural network for the real-time classification of both techniques besides a residual class. Evaluation using Leave-One-Out Cross-Validation yielded a classification accuracy of 0.83, showcasing the classifier's proficiency in distinguishing extreme vocal techniques from the residual class.

By integrating the trained neural network into a user application, we developed a system that can give aspiring extreme vocalists real-time feedback on their techniques by visualizing predictions on audio data captured via a microphone. Using this application, we conducted a qualitative study with standardized technical conditions to assess the application's impact on the singers' perceived self-efficacy. Despite our first quantitative analysis of self-efficacy scores not yielding significant results, participants exhibited a positive inclination towards the application, indicating an affirmative reception of the overall approach. Nonetheless, areas for improvement remain, particularly in user experience and classification accuracy. Overall, this work underscores the potential and challenges of machine learning-based training systems for extreme vocal techniques, offering insights for future research in vocal pedagogy and real-time feedback applications. Future research endeavors may seek to augment the classification accuracy of the Extreme-Vocal Classifier by expanding the dataset, refining class annotations, and exploring alternative feature sets. Long-term studies with a larger sample size could provide a more comprehensive understanding of the application's impact on self-efficacy. Moreover, additional resources and natural language feedback mechanisms could be incorporated to enhance UX and usability and facilitate the learning experience. Based on positive user feedback and solid classification results, we are confident that further research will yield AI-based tools that can improve extreme vocalists' safety and technical proficiency.

ACKNOWLEDGMENTS

This paper was funded by the DFG through the Leibniz award of Elisabeth André (AN 559/10-1).

REFERENCES

[1] Mathias Aaen, Julian McGlashan, Noor Christoph, and Cathrine Sadolin. 2021. Extreme vocal effects distortion, growl, grunt, rattle, and creaking as measured by electroglottography and acoustics in 32 healthy professional singers. *Journal of Voice* (2021).

[2] Mathias Aaen, Cathrine Sadolin, Anna White, Reza Nouraei, and Julian McGlashan. 2022. Extreme Vocals—A Retrospective Longitudinal study of Vocal Health in 20 Professional Singers Performing and Teaching Rough Vocal Effects. *Journal of voice* (2022).

[3] Anjok07 and other contributors. 2023. *Gui for a Vocal Remover that Uses Deep Neural Networks.* https://github.com/Anjok07/ultimatevocalremovergui

[4] Matthew L Bernacki, Timothy J Nokes-Malach, and Vincent Aleven. 2015. Examining self-efficacy during learning: Variability and relations to behavior, performance, and learning. *Metacognition and Learning* 10 (2015), 99–117.

[5] Mauro Barro Fiuza and Marta Assumpção de Andrada e Silva. 2018. Can singing with rasp be a healthy practice? *Distúrb Comun, São Paulo* 30, 4 (2018), 802–808.

[6] Alexander Heimerl, Tobias Baur, Florian Lingenfelser, Johannes Wagner, and Elisabeth André. 2019. NOVA - A tool for eXplainable Cooperative Machine Learning. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 109–115. https://doi.org/10.1109/ACII.2019.8925519

[7] Peter A Heslin and Ute-Christine Klehe. 2006. Self-efficacy. *Encyclopedia Of Industrial/Organizational Psychology, SG Rogelberg, ed* 2 (2006), 705–708.

[8] Weimin Huang, Tuan Kiang Chiew, Haizhou Li, Tian Shiang Kok, and Jit Biswas. 2010. Scream detection for home applications. In *2010 5th IEEE Conference on Industrial Electronics and Applications*. IEEE, 2115–2120.

[9] Vedant Kalbag and Alexander Lerch. 2022. Scream detection in heavy metal music. *arXiv preprint arXiv:2205.05580* (2022).

[10] Minseok Kim, Woosung Choi, Jaehwa Chung, Daewon Lee, and Soonyoung Jung. 2021. KUIELab-MDX-Net: A two-stream neural network for music demixing. *arXiv preprint arXiv:2111.12203* (2021).

[11] Arvind Kumar, Mahesh Chandra, and Shubham Agarwal. 2016. GUI implementation of real time feedback system to learn singing. In *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. IEEE, 1051–1054.

[12] Filipa MB Lã and Mauro B Fiuza. 2022. Real-time visual feedback in singing pedagogy: current trends and future directions. *Applied Sciences* 12, 21 (2022), 10781.

[13] Samuel Leong and Lee Cheng. 2014. Effects of real-time visual feedback on pre-service teachers' singing. *Journal of Computer Assisted Learning* 30, 3 (2014), 285–296.

[14] Susanna Mesiä and Paolo Ribaldini. 2015. Heavy metal vocals. A terminology compendium. *Karjalainen, Toni-Matti/Kärki, Kimi, Modern Heavy Metal: Markets, Practices and Culture, Helsinki: Aalto University* (2015), 383–392.

[15] Oriol Nieto. 2013. Unsupervised clustering of extreme vocal effects. In *Proc. 10th Int. Conf. Advances in Quantitative Laryngology*. 115.

[16] Żwan Paweł. 2008. Automatic singing quality recognition employing artificial neural networks. *Archives of Acoustics* 33, 1 (2008), 65–71.

[17] Eric Smialek, Philippe Depalle, and David Brackett. 2012. A spectrographic analysis of vocal techniques in extreme metal for musicological analysis. In *ICMC*.

[18] Will Straw. 1984. Characterizing rock music cultures: The case of heavy metal. *Canadian University Music Review* 5, 5 (1984), 104–122.

[19] Glenn D. White and Gary J. Louie. 2005. *The Audio Dictionary* (third ed.). University of Washington Press. 114 pages.

[20] Tzu-Tsung Wong. 2015. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern recognition* 48, 9 (2015), 2839–2846.