MuSA Voice: An Intelligent Platform for Automated, Reflective Vocal Technique Analysis

Suvi Hääräa (0009-0009-1236-3873) (suvi.haeaerae@upf.edu) [1],
Isabelle Oktay (0009-0004-6359-6101) (isabelle.oktay01@estudiant.upf.edu) [2], and
Rafael Ramírez-Meléndez (0000-0002-3294-0764) (rafael.ramirez@upf.edu) [3]

[1,2,3] *Music Technology Group*, **Universitat Pompeu Fabra**, Barcelona, Spain

ABSTRACT

We present **MuSA Voice**, an open, web-based platform for reflective feedback on vocal technique. The system deploys specialised Vocal Technique Classification (VTC) using two large pretrained audio models, CLAP and Whisper, trained separately by singer gender and modulation type (pitch and timbre), within an analysis pipeline that visualises technique usage across a recording. Applying these models to VTC for the first time, we find that Whisper achieves state-of-the-art accuracy and CLAP offers a competitive lightweight alternative, consistent with their strong generalisation to downstream audio tasks. **MuSA** employs descriptive rather than corrective feedback design to support reflection and development of individual expression. We further conduct the first cross-dataset generalisation evaluation of VTC, using GTSinger, which shows that generalisation to real-world singing remains an open challenge. Model performance is evaluated on VocalSet, and system performance is characterised on a real deployment server. **MuSA Voice** integrates these models into a deployed, pedagogically motivated platform, offering a starting point for technique-aware feedback tools for singers.

1. INTRODUCTION

Despite advances in deep learning models for audio, singers remain relatively unsupported by existing performance feedback interfaces. Most tools only address attributes such as pitch, leaving expressive techniques—such as vibrato, straight tone, belting, and breathiness—largely outside the realm of automatic feedback. This limits the human-machine interplay available for singers: rather than engaging in a technique-aware feedback loop (as in Figure 1), learners are confined to signal-level information that does not fully reflect the expressive dimensions central to the voice. However, recent advances in deep learning suggest that this gap may be possible to close. Large pretrained models such as Whisper [1], and CLAP [2] demonstrate strong generalisation to downstream audio tasks [3–7], making automatic detection of high-level vocal techniques increasingly feasible. By developing specialised models for pedagogical voice analysis and integrating them into music performance feedback systems, we can provide singers with the technique-aware interplay that existing tools have yet to offer.

The master-apprentice model (MAM) of weekly individual face-to-face lessons is prominently used in vocal coaching [8], meaning learners spend most of their time practising alone. This can reinforce incorrect habits, as even motivated students can move from deliberate to naive practice without consistent guidance [9–12]. Furthermore, high-quality vocal education can be inaccessible due to its cost. Mobile and automated feedback tools support independent music learning more broadly [13–15], but do not provide feedback on vocal technique, leaving an opening for systems that engage singers at the level of expressive control.

To address this, we present **MuSA Voice** [1] [2], an open web-based singing feedback platform. The singer contributes a raw vocal recording; specialised CLAP- and Whisper-based VTC models analyse pitch- and timbre-modulation techniques at 500ms resolution; and the platform returns interpretable visualisations of technique usage, supporting reflective practice. The contributions of this work are threefold: **(1)** applying Whisper and CLAP to VTC for the first time, including the first cross-dataset generalisation evaluation of VTC using GTSinger [16]; **(2)** evaluating gender- and modulation-type-specific (Pitch Modulation Techniques (PMT) and Timbre Modulation Techniques (TMT)) model specialisation, motivated by its pedagogical relevance and **(3)** integrating these into **MuSA**, extending it with the first openly available vocal technique feedback capability (**MuSA Voice**). We evaluate model performance on VocalSet [17] and characterise system performance on a real deployment server.

[1] **https://analyzer.appskynote.com/musa-voice**
[2] https://youtu.be/DhFouk8CSaE

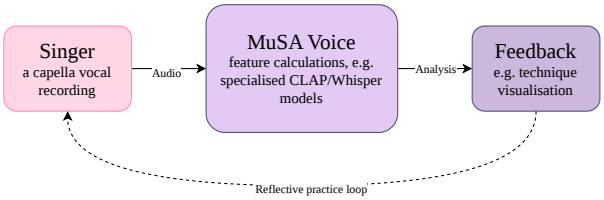


Figure 1. Overview of **MuSA Voice** reflective feedback loop engagement.

Copyright: © 2026. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

2. RELATED WORK

2.1 Music Performance Analysis Systems

Music performance analysis (MPA) systems are underused in music education, including singing. Historically, technology has supported tasks like ear training, classroom organisation, and music theory [18], but has not addressed technique-level feedback for singers, a need traditionally met through MAM [19]. Existing MPA tools fall into two categories: real-time performance feedback (RTPF) and non-real-time performance feedback (NRTPF). RTPF platforms (e.g., Yousician [3], VocaBerry [4]) provide feedback during performance, which can compete with the attentional demands of singing itself, which may not be best suited for post-performance reflection [20, 21]. NRTPF avoid this by design, but existing options are either computationally focused (e.g., Sonic Visualiser [5]) or research-oriented (e.g., Praat [6], VoceVista Video [7]). Often they are not pedagogically designed and require high technical literacy [22], limiting their use for self-reflection.

Both NRTFP and RTFP offer limited support for expressive and interpretative skills, which AI-driven systems can support during vocal training [23]. Although recordings can support reflection, learners seldom review them due to the time required to navigate large backlogs [24]. Even in MAM, reflection is often neglected due to time pressure and the perception that reflection is less productive than active playing [25]. This gap in systems that support reflection motivates **MuSA**'s post-performance, descriptive feedback approach.

2.2 Vocal Technique Classification

Vocal Technique Classification (VTC) sits within a broader field of computational singing voice research spanning topics such as singing voice synthesis [26], singing voice detection [27], and automatic lyric transcription [28]. As a classification task, VTC shares the general challenge of limited annotated data common to most ML efforts, compounded by unusually high inter-singer timbral variability. VocalSet [17] established the benchmark for VTC: a 10-way classification task across vibrato, straight, belt, breathy, lip trill, spoken, inhaled singing, trill, trillo, and vocal fry, evaluated using a singer-independent train-test split. Subsequent work has shifted from traditional machine learning towards deep learning [29–31] and self-supervised approaches [32].

Despite this progress, all prior VTC work shares three limitations that constrain both model performance and pedagogical relevance. First, male and female voices are modelled together, despite their distinct acoustic characteristics. Second, the 10-way scheme conflates pitch modulation techniques (PMT: vibrato, straight, trill, trillo) with timbre modulation techniques (TMT: belt, breathy, spoken, inhaled singing, vocal fry), even though these rely on different physical mechanisms and can co-occur in expressive singing. This means a model may predict a co-occurring technique and be penalised as incorrect. Third, no prior work evaluates generalisation beyond VocalSet, leaving open the question of how well these models perform on real-world singing. This work addresses all three.

2.3 Transfer Learning

Transfer learning from large pretrained models lowers training costs and data requirements [33], making it well-suited to VTC given VocalSet's limited size. Whisper [1], a speech recognition model trained on 680,000 hours of labelled audio, has demonstrated strong cross-task generalisation to classification tasks including dysarthric severity classification [4], audio tagging [5], and music genre classification [6]. CLAP [2], which aligns audio and text representations in a joint latent space, generalises to zero-shot and multi-label audio classification [2, 3]. Neither model has been applied to VTC, and no prior VTC work evaluates cross-dataset generalisation or integrates models into a practical feedback tool—gaps that motivate this work.

3. VTC MODELS FOR MUSA VOICE

3.1 Data

Following prior work, we use VocalSet [17], a singing dataset containing approximately 10 hours of monophonic recordings from 11 male and 9 female professional singers across 17 vocal techniques, five vowel sounds, and four performance contexts (scales, arpeggios, long tones, and excerpts). The dataset is imbalanced across both techniques and genders: the largest class, straight, accounts for approximately 25% of the data, while the smallest, spoken, accounts for under 2%; approximately 44% of recordings are female and 56% male. We use the original singer-independent train-validation split [17] without data augmentation, to ensure that test-set singers are unseen during training and allow direct comparison with prior work.

For state-of-the-art (SOTA) comparison, we follow the 10-way classification task established by VocalSet (see Section 2.2). For pedagogically motivated specialisation, we partition the data into PMT- and TMT-specific subsets (see Section 2.2); lip trill is excluded as its production mechanism does not align clearly with either modulation type. We additionally train models on male and female subsets separately. For deployment within **MuSA Voice**, the specialised models are trained using 500ms windows rather than the 3-second windows used in prior work. While shorter windows modestly reduce classification accuracy, they enable finer-grained temporal feedback, allowing the platform to show how technique usage evolves across a recording rather than in coarse 3-second blocks, which is more meaningful for reflective practice.

3.2 Data Preprocessing

Audio was loaded in mono, resampled to 16 kHz to match the model input requirements of CLAP [2] and Whisper [1], and segmented into 3-second (or 500ms) windows with 25% overlap. CLAP segments were encoded into 512-dimensional embeddings using the `msclap` (2023)

[3] https://yousician.com/
[4] www.vocaberry.org/
[5] www.sonicvisualiser.org/
[6] www.fon.hum.uva.nl/praat/
[7] www.vocevista.com/

Architecture	Test F1	Test Acc	Test B-Acc	Test Top-2	Test Top-3
1DCNN [17]	0.488	0.584	0.484	0.764	0.863
Wav2Vec2.0 [32]	0.561	0.583	0.601	0.748	0.821
WavLM [32]	0.556	0.608	0.579	0.753	0.838
MERT [32]	0.541	0.585	0.585	0.761	0.853
MapMusic2Vec [32]	0.608	0.660	0.621	0.790	0.869
RAT + ALA + cRT [31]	0.692	0.701	0.716	0.876	0.922
CLAP	0.685	*0.704*	0.714	0.874	*0.936*
Whisper	**0.771**	**0.802**	**0.809**	**0.934**	**0.967**

Table 1. VTC performance of prior and proposed models on the VocalSet validation set (3-s window). Prior-model metrics and SOTA from Chen et al. [31]. Bold: best per metric; italic: exceeds prior SOTA.

Specialisation	Test F1	Test Acc	Test B-Acc	Test Top-2	Test Top-3
Whisper - female 10-way	0.936	0.928	0.920	0.980	0.996
Whisper - male 10-way	0.705	0.709	0.715	0.862	0.940
Whisper - mixed PMT	0.854	0.893	0.835	0.975	0.991
Whisper - mixed TMT	0.886	0.884	0.907	0.948	0.985

Table 2. VTC performance of gender or PMT/TMT specialised Whisper models.

encoder. Whisper features were computed as 80-channel log-mel spectrograms (25ms window, 10ms hop) using the Hugging Face `WhisperFeatureExtractor`, consistent with the original pretraining configuration. Class labels were standardised across partitions using shared `LabelEncoders`.

3.3 Model Architectures and Training

CLAP was used as a frozen feature extractor: 512-dimensional embeddings were passed to an MLP classifier comprising a linear layer (512 units, ReLU activation), dropout of 0.3, and a linear output layer, trained with Adam ($lr = 1 \times 10^{-4}$) and cross-entropy loss. Freezing the encoder makes CLAP more computationally lightweight, a key consideration for web deployment. **Whisper base** was fine-tuned end-to-end: encoder outputs were mean-pooled across time and passed to a linear classification layer, with all parameters updated using Adam and cross-entropy loss. The base architecture balances depth and computational cost, and end-to-end fine-tuning enables deeper task-specific adaptation than a frozen encoder allows. Both models were trained for up to 100 epochs with early stopping based on validation accuracy (patience 10) on Google Colab (CLAP: T4 GPU; Whisper: A100 GPU).

3.4 Performance on VocalSet

Following prior work, we report F1-score, accuracy, balanced accuracy, and top-2 and top-3 accuracy, implemented using `scikit-learn`. Balanced accuracy accounts for VocalSet's class imbalance, while F1 captures precision–recall trade-offs across techniques. Top-2 and top-3 accuracies are relevant here, as PMT and TMT techniques can co-occur, and a secondary prediction may reflect a present technique rather than an error.

Table 1 reports 10-way classification performance on the singer-independent VocalSet test split using 3-second windows; prior-model metrics are reproduced from Chen et al. [31]. Whisper outperforms all prior models across every metric (F1: 0.771; B-acc: 0.809), demonstrating that large-scale supervised pretraining on audio transfers effectively to VTC. CLAP matches or exceeds the previous state-of-the-art in overall accuracy and top-3 accuracy, establishing it as a competitive lightweight alternative. These results confirm that pretrained audio models offer an advantage over architectures trained from scratch on VocalSet alone.

Table 2 reports the effect of gender and modulation-type specialisation using Whisper with 3-second windows. Specialising by modulation type improves performance for both PMT (F1: 0.854) and TMT (F1: 0.886), reflecting the benefit of separating techniques that rely on distinct physical mechanisms. Gender-specific training yields a substantial improvement for female voices (F1: 0.936). However, for male voices, it decreases slightly (F1: 0.705).

Model	Test F1	Test Acc	Test B-Acc	Test Top-2	Test Top-3
Female 10-way	0.130	0.369	0.345	0.598	0.824
Male 10-way	0.092	0.308	0.276	0.520	0.630
Female PMT	0.242	0.568	0.573	0.960	0.990
Male PMT	0.353	0.692	0.695	0.944	0.985
Female TMT	0.256	0.568	0.565	0.780	0.915
Male TMT	0.152	0.213	0.281	0.669	0.871

Table 3. Performance of Whisper models when evaluated on GTSinger.

Architecture	Specialisation	Test F1	Test Acc	Test B-Acc	Test Top-2	Test Top-3
Whisper	female-PMT	0.682	0.814	0.657	0.935	0.977
CLAP	female-PMT	0.679	0.803	0.667	0.928	0.984
Whisper	male-PMT	0.624	0.695	0.598	0.897	0.961
CLAP	male-PMT	0.594	0.656	0.569	0.890	0.968
Whisper	female-TMT	0.787	0.821	0.775	0.930	0.969
CLAP	female-TMT	0.773	0.794	0.773	0.936	0.981
Whisper	male-TMT	0.704	0.694	0.733	0.818	0.929
CLAP	male-TMT	0.700	0.676	0.728	0.877	0.967

Table 4. Performance of specialised CLAP and Whisper models (500ms window) on VocalSet validation data, filtered by singer gender while preserving the original split.

3.5 Generalisation to GTSinger

To assess whether the models extend beyond VocalSet's controlled conditions, we evaluate the Whisper models on GTSinger [16], which contains a subset of the same vocal techniques (vibrato, straight, breathy, not-breathy) recorded as full musical excerpts across diverse voice types and recording conditions. This is more representative of the real-world audio **MuSA Voice** is designed to receive. Results are reported in Table 3. The drop in performance is substantial across all models, with F1 scores falling to 0.09–0.35, reflecting the distributional shift between VocalSet's controlled vowel-scale recordings and GTSinger's naturalistic excerpts. The differences in the labelled techniques may also impact the results. PMT-specialised models generalise better than TMT models, likely because pitch modulation is less dependent on individual vocal timbre. Female models generally outperform male models, as in previous experiments.

3.6 500ms Models for Deployment

Table 4 reports the performance of gender- and modulation-specialised CLAP and Whisper models trained with 500ms windows on the VocalSet validation set. The 500ms models attain accuracies ranging from 66% to 82%, and PMT-specialised and female models continue to outperform their counterparts. Notably, CLAP performs only moderately below Whisper at this resolution, mak-

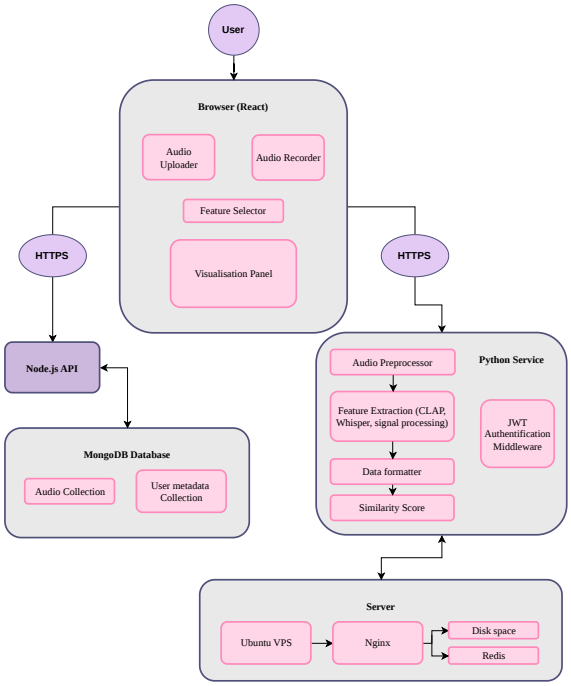


Figure 2. MuSA System Diagram [36]

ing it a viable lightweight alternative for the deployment-constrained server environment described in Section 4. These specialised models are integrated into **MuSA Voice**.

4. MUSA SYSTEM DESIGN

MuSA Voice is deployed on an 8GB RAM, 10GB SSD, 2 vCPU server with 4TB of data transfer capacity. It is designed to work with raw a capella vocal recordings, as the current implementation does not contain models for vocal isolation or removing effects. Audio recordings longer than 30 seconds may not produce output due to current server processing limitations.

The design of **MuSA Voice** is grounded in three pedagogical principles. First, feedback is *descriptive* rather than corrective: model outputs are presented as visualisations of technique distribution across a recording, without labelling performances as right or wrong, encouraging singers to develop their own musical judgement. Second, representations use *musically meaningful* scales rather than raw computational features, reducing the interpretive burden on non-technical users. Third, the system supports *comparison* between a reference and a performance recording, mirroring how a teacher demonstrates a desired sound—with responsibility for interpretation gradually transferred to the learner [34, 35].

4.1 Back-End Architecture

A core component of **MuSA** is the Python service responsible for feature extraction, caching, and user token generation (see Figure 2). Feature extraction is implemented as a standalone Python service due to model customisation needs and the audio length limitations of JavaScript libraries such as `Essentia.js`. Models can be swapped by adjusting configuration files. `Flask` and `Gunicorn`

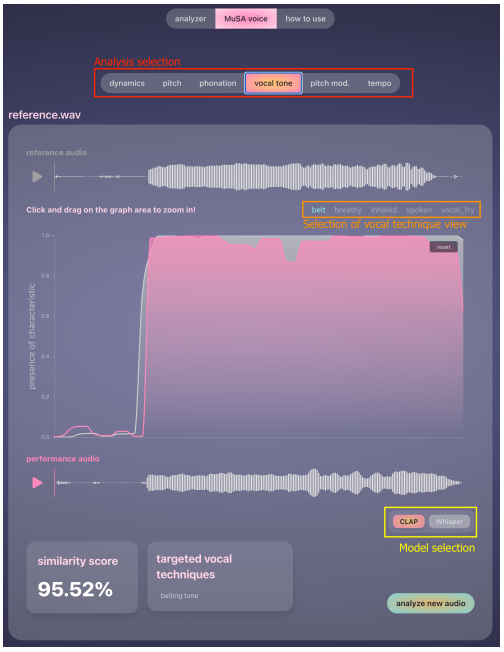


Figure 3. MuSA Interface

handle routing and deployment. Redis manages short-term caching and user session data via JWT tokens, minimising redundant computation and supporting concurrent users. The legacy `Node.js` service manages MongoDB communication from the front end for user and audio metadata.

4.2 Front-End User-Interface

The **MuSA** UI is built with `React` and `TailwindCSS`, with `D3.js` rendering feature visualisations (see Figures 2 and 3). To reduce cognitive load, a tab-based interface isolates individual analytical dimensions, guiding attention toward specific aspects of technique. Pitch is mapped to a piano-roll background, tempo to a BPM y-axis. For VTC-specific features—vocal tone and pitch modulation—the graph displays model activations for any selected class across the full recording, allowing users to pinpoint where a technique is detected. A reference recording can be uploaded alongside the performance recording, and a similarity score summarises overall correspondence without indicating whether greater similarity is desirable. The singer, not the system, determines whether convergence toward the reference reflects their goals.

4.3 User Flow

The full user flow is illustrated in Figure 4. The workflow involves three main steps: (1) uploading or recording both a performance and a reference audio (uf.3), (2) selecting a feature to analyse (uf.5), and (3) exploring the resulting visualisations and similarity score (uf.6). On login, users complete a short survey about their recording conditions and technique focus to help calibrate future analyses. The Python service then processes the audio, extracting the selected features; results are cached in Redis, feature data stored in MongoDB, and audio files saved locally. Users can play back recordings, zoom into specific passages, and inspect extracted feature data in the visualisation panel.

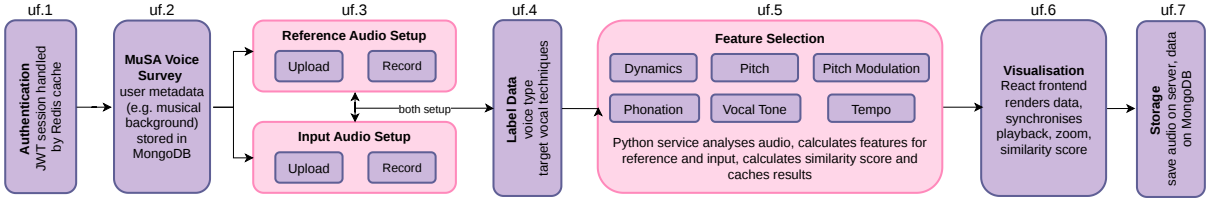


Figure 4. MuSA System user flow

Feature / Model	Front-end Request-Response Time (s)	Back-End Model Time (s)	CPU Avg (%)	RAM ↑ Baseline (%)
Dynamics	1.27	0.48	42	0.05
Pitch	4.86	4.40	51	0.04
Phonation	2.13	1.04	59	0.00
Vocal Tone (CLAP)	92.32 (combined)	23.37	89	3.86
Vocal Tone (Whisper)	92.32 (combined)	67.52	99	3.46
Pitch Modulation (CLAP)	93.04 (combined)	22.98	90	2.30
Pitch Modulation (Whisper)	93.04 (combined)	68.65	99	3.95
Tempo	1.92	0.39	14	0.02

Table 5. Inference latency and computational load for each feature in the **MuSA** server. Front-end time includes preprocessing, model execution, and network transfer. Back-end time reflects isolated model inference. "Combined" denotes features sharing the Whisper/CLAP pipeline calls.

5. SYSTEM EVALUATION

Responsiveness and computational load were evaluated using a 13-second mono soprano vocal recording captured via the **MuSA Analyzer** for voice. The same audio file was used for all tests to ensure consistency. Each feature extraction pipeline was executed 10 times, and we report the averaged front-end request–response latency, back-end model inference time, CPU utilisation, and RAM usage in Table 5. Front-end request–response latency includes preprocessing, model inference, feature computation, and all communication between the front end and the Python back-end service.

Lightweight feature calculations such as dynamics, phonation, tempo, and pitch are completed in under 5 seconds end-to-end, with back-end inference times typically below one second. These features impose minimal computational overhead, remaining below 60% CPU and 0.1% RAM above baseline, making them well-suited for interactive use. The combined vocal tone and pitch modulation analyses depend on Whisper and CLAP, resulting in an end-to-end latency of 92–93 seconds. Whisper is the main computational bottleneck, requiring about 67 seconds for back-end inference and consuming more computational resources than CLAP. CLAP, by comparison, is lighter, completing inference in approximately 23 seconds with an average CPU load of approximately 90%.

6. DISCUSSION

The results highlight both the potential and current limitations of applying large pretrained audio models to VTC in a musical context. Whisper and CLAP both perform competitively on VocalSet, confirming that large-scale pretraining transfers effectively to VTC. However, the substantial drop on GTSinger reveals that models trained on controlled, vowel-scale recordings struggle to generalise to naturalistic singing. It should be noted that in this test, GTSinger labels are applied at the excerpt level rather than the phoneme level, so some segments may carry a technique label where the technique is not consistently active. The results should be interpreted as indicative of the domain gap rather than as a precise performance benchmark. A similar issue exists in VocalSet. Listening reveals that clip-level labels sometimes cover only partial instances of a technique, which likely contributes to the difficulty of generalising. More fine-grained annotation of both training and evaluation data is a prerequisite for more robust VTC, and by extension, more reliable feedback.

Despite these limitations, the cross-dataset results are directionally informative: PMT models generalise better than TMT models, consistent with pitch modulation being less dependent on individual vocal timbre. The asymmetry in gender splits—improved model performance for female voices, and reduced for male ones—likely reflects how the female subset of VocalSet is dominated by sopranos, consequently having less timbral variability. The male singers span more diverse voice types [17].

The specialisation experiments more broadly suggest that decomposing vocal technique into structurally meaningful dimensions improves both performance and interpretability. Presenting singers with technique-specific signals—rather than a single 10-way prediction—gives them something they can engage with selectively, focusing on the dimensions most relevant to their current practice goals. In expressive musical domains, classification accuracy alone is an insufficient basis for feedback. What matters more is whether the output is consistent and interpretable enough to support reflection. **MuSA Voice** addresses this by visualising technique activations across a recording rather than issuing correctness judgements, reframing VTC as a form of computationally mediated listening rather than replacing musical judgement. Initial evidence on how users engage with this form of feedback in practice is reported separately in [37].

Finally, the system evaluation exposes a practical accuracy–latency trade-off. Whisper offers higher predictive performance but is computationally heavy for the current server, while CLAP provides a more responsive alternative at modest accuracy cost. Lighter Whisper variants and formal user studies comparing both representations will be beneficial future work.

7. CONCLUSION AND FUTURE WORK

This work presented **MuSA Voice**, an open platform integrating specialised CLAP and Whisper models for descriptive, technique-aware feedback in singing. Our find-

ings confirm that large pretrained audio models transfer effectively to VTC, consistent with prior work on downstream audio classification tasks, with Whisper achieving state-of-the-art performance and CLAP offering a competitive lightweight alternative suited to constrained deployment. The primary contribution of this work is the integration of these models into **MuSA Voice**, externalising aspects of vocal technique to support reflective practice. Cross-dataset and real-world generalisation remains an open challenge, motivating **MuSA**'s descriptive feedback design as a deliberate response to current limitations in models.

Future work should target cross-dataset robustness by means of collecting and annotating a broader range of naturalistic, non-studio recordings with a wider range of vocal ranges, timbres, and singing styles. Another direction for future work would be incremental, background feature extraction during recording, which could reduce post-performance latency and enable the development of real-time performance feedback. Furthermore, we are investigating whether synthetic voice data can supplement real recordings for training in vocal tasks, given the scarcity of annotated, diverse vocal technique data. While the pilot user study evaluating singers' engagement with MuSA Voice's feedback is reported separately in [37], formal, larger-scale studies of learning outcomes remain future work.

Acknowledgments

This work was supported in part by groups funding the Music and Machine Learning Lab at the UPF MTG, namely the Maria de Maeztu Units of Excellence Programme CEX2021-001195-M funded by MCIU (MCIU/AEI/10.13039/501100011033), the IMPA project PID2023-152250OB-I00 funded by MCIU (MCIU/AEI/10.13039/501100011033/FEDER, UE), and the Musitopia project funded by the Fundació Barcelona Music Lab.

8. REFERENCES

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022, arXiv. [Online]. Available: https://arxiv.org/abs/2212.04356

[2] Y. Wu, K. Chen, T. Zhang, Y. Hui, M. Nezhurina, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," 2022, arXiv. [Online]. Available: https://arxiv.org/abs/2211.06687

[3] S. Braun and H. Gamper, "Multi-label audio classification with a noisy zero-shot teacher," in *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2024, pp. 240–244.

[4] S. Rathod, M. Charola, A. Vora, Y. Jogi, and H. A. Patil, "Whisper features for dysarthric severity-level classification," in *Proceedings of Interspeech 2023*, 2023, pp. 1523–1527.

[5] Y. Gong, S. Khurana, L. Karlinsky, and J. Glass, "Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers," in *INTERSPEECH 2023*, ser. interspeech_2023. ISCA, Aug. 2023. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2023-2193

[6] B. Shi, H. Wang, C. Lu, and M. Zhao, "Whispar: Transferring pre-trained audio models to fine-grained classification via prompt and adapter," *Knowledge-Based Systems*, vol. 300, p. 112242, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705124008761

[7] R. Ma, A. Liusie, M. J. F. Gales, and K. M. Knill, "Investigating the emergent audio classification ability of asr foundation models," 2023, arXiv. [Online]. Available: https://arxiv.org/abs/2311.09363

[8] I. M. Hanken, *The role and significance of masterclasses in creative learning*. Oxford University Press, Dec. 2017, vol. 1. [Online]. Available: https://academic.oup.com/book/11108/chapter/159534829

[9] G. F. Welch, "Variability of practice and knowledge of results as factors in learning to sing in tune," *Bulletin of the Council for Research in Music Education*, pp. 238–247, 1985.

[10] G. E. McPherson, , and J. M. Renwick, "A Longitudinal Study of Self-regulation in Children's Musical Practice," *Music Education Research*, vol. 3, no. 2, pp. 169–186, Sep. 2001, publisher: Routledge. [Online]. Available: https://doi.org/10.1080/14613800120089232

[11] S. Hallam, T. Rinta, M. Varvarigou, A. Creech, I. Papageorgi, T. Gomes, and J. Lanipekun, "The development of practising strategies in young people," *Psychology of Music*, vol. 40, no. 5, pp. 652–680, Sep. 2012, publisher: SAGE Publications Ltd. [Online]. Available: https://doi.org/10.1177/0305735612443868

[12] K. A. Ericsson and K. W. Harwell, "Deliberate Practice and Proposed Limits on the Effects of Practice on the Acquisition of Expert Performance: Why the Original Definition Matters and Recommendations for Future Research," *Frontiers in Psychology*, vol. 10, 2019. [Online]. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.02396

[13] M. Goncharova and I. Gorbunova, "Mobile technologies in the process of teaching music theory," *Propósitos y Representaciones*, vol. 8, no. SPE3, p. e705, sep. 2020. [Online]. Available: https://revistas.usil.edu.pe/index.php/pyr/article/view/705

[14] P. J. Yu and M. Z. Xiong, "Remote vocal singing course design based on embedded system and

internet of things," *Mobile Information Systems*, vol. 2022, no. 1, p. 8712081, 2022. [Online]. Available: https://doi.org/10.1155/2022/8712081

[15] Y. Rui, "Simulation of e-learning in vocal network teaching experience system based on intelligent internet of things technology," *Entertainment Computing*, vol. 50, p. 100711, 2024. [Online]. Available: https://doi.org/10.1016/j.entcom.2024.100711

[16] Y. Zhang, C. Pan, W. Guo, R. Li, Z. Zhu, J. Wang, W. Xu, J. Lu, Z. Hong, C. Wang, L. Zhang, J. He, Z. Jiang, Y. Chen, C. Yang, J. Zhou, X. Cheng, and Z. Zhao, "Gtsinger: A global multi-technique singing corpus with realistic music scores for all singing tasks," 2024, arXiv. [Online]. Available: https://doi.org/10.48550/arXiv.2409.13832

[17] J. Wilkins, P. Seetharaman, A. Wahl, and B. Pardo, "Vocalset: A singing voice dataset," in *19th International Society for Music Information Retrieval Conference*, 2018, pp. 468–474. [Online]. Available: https://doi.org/10.5281/zenodo.1203819

[18] J. Mroziak and J. Bowman, "Music tpack in higher education: Educating the educators," *Handbook of technological pedagogical content knowledge (TPACK) for educators*, pp. 285–296, 2016.

[19] H. Gaunt, "Apprenticeship and empowerment: the role of one-to-one lessons," in *Musicians in the Making: Pathways to Creative Performance*, J. Rink, H. Gaunt, and A. Williamon, Eds. Oxford University Press, Dec. 2017, p. 0. [Online]. Available: https://doi.org/10.1093/acprof:oso/9780199346677.003.0003

[20] M. Çorlu, C. Muller, F. Desmet, and M. Leman, "The consequences of additional cognitive load on performing musicians," *Psychology of Music*, vol. 43, no. 4, p. 495–510, Jul. 2015. [Online]. Available: https://journals.sagepub.com/doi/10.1177/0305735613519841

[21] C. Provenzale, F. Di Tommaso, N. Di Stefano, D. Formica, and F. Taffoni, "Real-time visual feedback based on mimus technology reduces bowing errors in beginner violin students," *Sensors*, vol. 24, no. 12, p. 3961, Jun. 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/12/3961

[22] J. Hua and W. Yi, "Seeing the singing voice: The current state of vocal data visualisation," 08 2025, pp. 158–163.

[23] D. Wu, "Utilizing ai and machine learning in vocal training programs: A contemporary approach," *Learning and Instruction*, vol. 102, p. 102278, 2026, early Access Online. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959475225002026

[24] R. Ramirez-Melendez and G. Waddell, "Technology-Enhanced Learning of Performance," in *The Oxford Handbook of Music Performance, Volume 2*, G. E. McPherson, Ed. Oxford University Press, Apr. 2022, p. 0. [Online]. Available: https://doi.org/10.1093/oxfordhb/9780190058869.013.24

[25] C. Dos Santos Silva and H. Marinho, "Self-regulated learning processes of advanced musicians: A prisma review," *Musicae Scientiae*, vol. 29, no. 1, p. 89–108, 3 2025. [Online]. Available: https://journals.sagepub.com/doi/10.1177/10298649241275614

[26] Y.-P. Cho, F.-R. Yang, Y.-C. Chang, C.-T. Cheng, X.-H. Wang, and Y.-W. Liu, "A survey on recent deep learning-driven singing voice synthesis systems," in *2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, 2021, pp. 319–323.

[27] R. Monir, D. Kostrzewa, and D. Mrozek, "Singing voice detection: A survey," *Entropy*, vol. 24, no. 1, 2022. [Online]. Available: https://www.mdpi.com/1099-4300/24/1/114

[28] X. Gu, L. Ou, W. Zeng, J. Zhang, N. Wong, and Y. Wang, "Automatic lyric transcription and automatic music transcription from multimodal singing," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 20, no. 7, May 2024. [Online]. Available: https://doi.org/10.1145/3651310

[29] Y. Yamamoto, J. Nam, and H. Terasawa, "Deformable cnn and imbalance-aware feature learning for singing technique classification," in *Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH)*. INTERSPEECH, 2022, pp. 2778–2782, arXiv. [Online]. Available: http://arxiv.org/abs/2206.12230

[30] Y. Yamamoto, J. Nam, H. Terasawa, and Y. Hiraga, "Investigating time-frequency representations for audio feature extraction in singing technique classification," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. APSIPA, 2021, pp. 890–896. [Online]. Available: http://www.apsipa.org/proceedings/2021/pdfs/0000890.pdf

[31] G.-Y. Chen, Y.-F. Yeh, and V.-W. Soo, "Rat: Radial attention transformer for singing technique recognition," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[32] Y. Yamamoto, "Toward leveraging pre-trained self-supervised frontends for automatic singing voice understanding tasks: Three case studies," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2023*. Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1745–1752.

[33] M. Iman, H. R. Arabnia, and K. Rasheed, "A review of deep transfer learning and recent advancements,"

Technologies, vol. 11, no. 2, 2023. [Online]. Available: https://doi.org/10.3390/technologies11020040

[34] L. S. Vygotsky and M. Cole, *Mind in society: Development of higher psychological processes*. Harvard university press, 1978.

[35] N. Granott, K. W. Fischer, and J. Parziale, "Bridging to the unknown: A transition mechanism in learning and development," *Microdevelopment: Transition processes in development and learning*, pp. 131–156, 2002.

[36] I. Oktay, "Musa: A new tel platform for enhancing self- reflection and musical understanding through saliency analysis of performance recordings," Jul. 2025. [Online]. Available: https://doi.org/10.5281/zenodo.17303802

[37] S. Häärä, I. Oktay, and R. Ramirez-Melendez, "Musa voice: Evaluating deep learning-based vocal technique feedback for self-regulated, reflective singing practice," in *Proceedings of the 18th International Conference on Computer Supported Education - Volume 1: CSME*, INSTICC. SciTePress, 2026, pp. 545–556.