Towards Potential Applications of Machine Learning in Computer-Assisted Vocal Training

Antonia Stadler[1], Emilia Parada-Cabaleiro[1,2,3] and Markus Schedl[1,2]

[1] Institute of Computational Perception, Johannes Kepler University Linz, Austria
[2] Human-centered AI Group, Linz Institute of Technology (LIT), Austria
[3] Department of Music Pedagogy, Nuremberg University of Music, Germany
emiliaparada.cabaleiro@hfm-nuernberg.de

Abstract. The usefulness of computer-based tools in supporting singing pedagogy has been demonstrated. With the increasing use of artificial intelligence (AI) in education, machine learning (ML) has been applied in music-pedagogy related tasks too, e. g., singing technique recognition. Research has also shown that comparing ML performance with human perception can elucidate the usability of AI in real-life scenarios. Nevertheless, this assessment is still missing for singing technique recognition. Thus, we comparatively evaluate classification and perceptual results from the identification of singing techniques. Since computer-assisted singing often relays on visual feedback, both an auditory task (recognition from *a capella* singing), and a visual one (recognition from spectrograms) were performed. Responses by 60 humans were compared with ML outcomes. By guaranteeing comparable setups, our results indicate that ML can capture differences in human auditory and visual perception. This opens new horizons in the application of AI-supported learning.

Keywords: AI-supported Education, Singing Techniques, Perception

1 Introduction

Singing techniques, as well as the strategies to teach them, have evolved over the history, in correspondence with chronological and geographical factors influencing music development [1]. Nevertheless, singing pedagogy has been mostly based in oral tradition, which is the reason why the description of how to perform such techniques is, in some cases, vague and imprecise [2]. Due to this, while experienced singers and teachers can naturally evaluate the quality of singing by simply following their intuition [3], this task might be particularly challenging for beginners.

The advantages of using computer-based applications to support teaching and learning have been shown [4]. Within music pedagogy, the use of computer-assisted singing tools, able to enhance singers' awareness, have become of common use in combination with traditional pedagogy [5]. Indeed, some of these tools have shown to be particularly effective in supporting beginners' training [6]. Due to the ubiquity of artificial intelligence (AI), machine learning (ML) methods have also been applied in the automatic assessment of singing quality [7]. Similarly, research on the automatic recognition of specific singing techniques has recently gained popularity [8, 9].

Nevertheless, the development of ML tools to support singing training is still on its infancy, which comes along with not yet well-defined use-cases and prevents a real connection between music pedagogy and the AI field. In this work, we present a preliminary study aimed to pave the way for future research on the use of AI in singing pedagogy. Since it has been shown that assessing how well a ML algorithm performs in comparison to humans can bring light about the utility of AI in real life [10–13], we assess, for the first time, the performance of ML methods in singing technique classification with respect to humans. By evaluating the perceptual ratings of two participant groups (with and without musical expertise) in comparison to ML we aim to: (i) assess how different feature representations perform in comparison to different learners level; and (ii) try to define potential applications of ML in singing education scenarios.

2 Related Work

The use of technology as an auxiliary educational tool has shown to successfully enhance singing pedagogy [14]. This is achieved by integrating acoustic voice analysis in the learning context as well as by using it as a biofeedback for singers' training [15]. Indeed, analysing audio recordings and computer-based feedback are two important elements of up-to-date singing pedagogy [16]. In particular, it has been shown that using visual representations of vocal properties effectively supports learners [5]. For instance, the understanding of phrasing can be enhanced by illustrating vocal pressure [17]. ALBERT [18] and VOXed [19], aiming to promote a more effective singing learning, are tools developed for real-time educational visual feedback. Finally, the use of computer-based tools complementing traditional pedagogy has shown to effectively promote curiosity and motivation [20], two essential aspects for a successful learning.

Within AI, the automatic classification of singing techniques has gained relevance, which lead to the development of dataset such as VocalSet [8] or J-POP [21]. Research on VocalSet showed that features learned from multi-resolution-spectrograms can outperform the original baseline, based on a Convolutional Neural Network (CNN), with a much less sophisticated architecture, i. e., Random Forest [9]. Similarly, a recent work on automatic recognition of paralinguistic singing attributes, e. g., vocal register and vibrato, has confirmed that feeding traditional ML models, such as Support Vector Machine (SVM), with spectrograms is a suitable approach for singing-related tasks [22].

3 Methodology

3.1 Dataset, Preprocessing, and Evaluation Metrics

In this work, we use VocalSet [8], a dataset consisting of 3 560 audio instances (10.1 hours of recordings) produced by 11 male and 9 female singers performing 17 different singing techniques. As in the original baseline, the experiments were performed by considering only 10 singing techniques (1 736 audio instances), i. e., the most relevant

Singing technique	Number of instances	Duration
Belt	205	26.24
Breathy	200	28.00
Inhaled	100	9.95
Lip Trill	202	24.40
Spoken	20	4.06
Straight	361	71.65
Trill	95	18.45
Trillo	100	14.54
Vibrato	255	57.79
Vocal Fry	198	34.10

Table 1: Overview of the samples from VocalSet used in the experiments. For each singing technique, the total number of instance and overall duration in minutes is given.

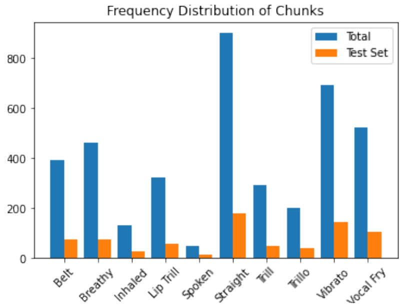


Fig. 1: Distribution of chunks across singing techniques. Besides the total, those used in the user study and as test set in the machine learning experiments, are displayed.

in practice: *Belt*, *Breathy*, *Inhaled*, *Lip Trill*, *Spoken*, *Straight*, *Trill*, *Trillo*, *Vibrato*, and *Vocal Fry*. In Table 1, the frequency distribution of the used audio instances across the singing techniques, as well as their duration in minutes, is indicated.

Following the pre-procesing guidelines used in the baseline of VocalSet [8], the silence at the beginning, middle, and end of the audio files were removed and the instances were split into chunks of approx. 3 seconds length. The distribution of the resulting 3 934 audio chunks across the corresponding singing techniques is displayed in Figure 1 (cf. Total). For the user study and as a test set for the ML experiments, the chunks extracted from the audio instances produced by singers F2, F6, M3, and M11 (i. e., 777), were considered (cf. Test Set in Figure 1). These singers were selected as they produced samples for all the considered techniques.

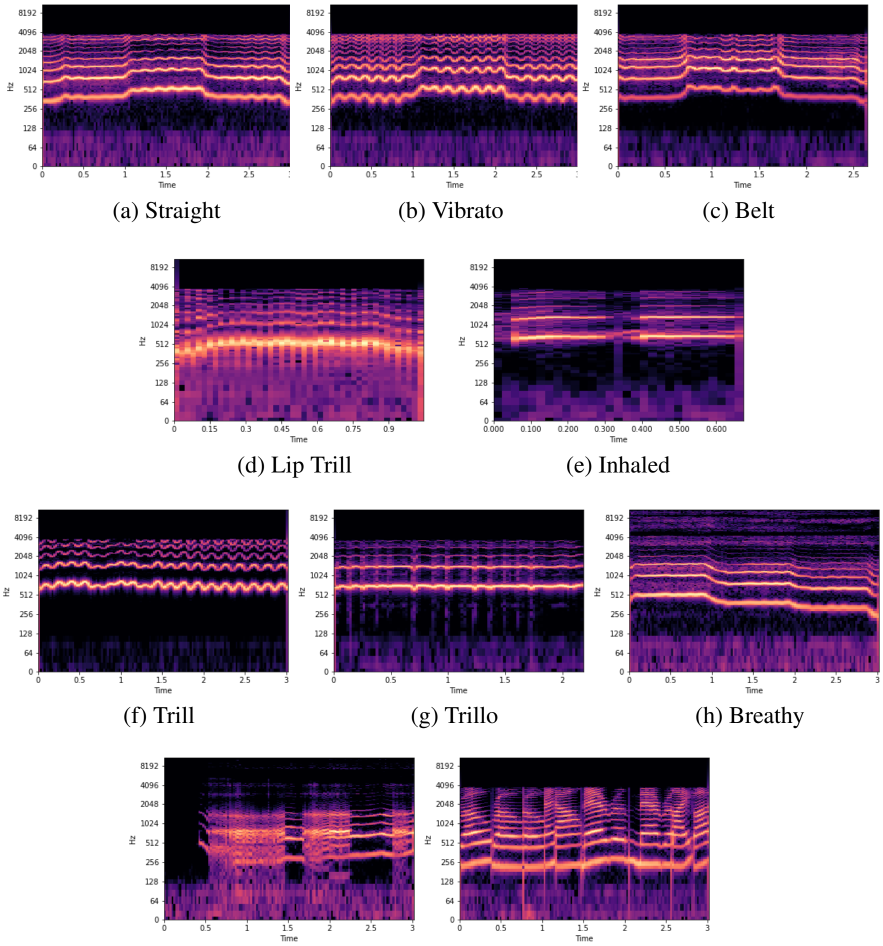
The experimental results, for both the user-based and the ML experiments, will be evaluated in terms of Unweighted Average Recall (UAR), precision, and recall. UAR, also known as Balanced Accuracy, is the recommended metric for datasets with an imbalanced distribution of samples across classes [23]. Besides precision and recall, confusion matrices will be used to interpret confusion patterns amongst classes.

3.2 Singing Techniques

To enable a better interpretation of the results, a brief description of each singing technique (illustrated by a spectrogram generated with Praat, cf. Figure 2), is presented. Since not all the techniques are produced through the same vocalisations in VocalSet, the spectrograms display a variety of them, i. e., arpeggios, long tones, and scales.

The sound produced by the technique *Straight* is natural, without any pressure or ornamentation. This is what we typically refer to as 'normal' singing, with the complete elimination of vibrato [24], which is shown by the horizontal lines in the spectrogram representing the pitch (cf. Figure 2a). In contrast, when singing *Vibrato*, the fundamental frequency and amplitude are intentionally altered by the singer [25], oscillations clearly visible in the spectrogram generated from the same instance (cf. Figure 2b).

Vibrato is often confused with the technique *Trill*. However, *Vibrato* should sound like one single tone rather than two different ones, which is expected in *Trill* [24]. This is achieved by producing oscillations that do not exceed a semitone beyond the main



(a) Straight (b) Vibrato (c) Belt

(d) Lip Trill (e) Inhaled

(f) Trill (g) Trillo (h) Breathy

(i) Vocal Fry (j) Spoken

Fig. 2: Spectrograms displaying each of the evalauted singing techniques. All of them are generated from samples performed by the female singer F1 producing the vowel ‘a’ except *Spoken*, for which a text is read. The used vocalisations are: arpeggio (*Straight*, *Belt*, *Vibrato*, *Lip Trill*); long tone (*Inhaled*, *Trill*, *Trillo*); scale (*Breathy*, *Vocal Fry*).

tone [26]. On the contrary, *Trill* is perceived as a fluctuation between two clearly distinguished pitches [24]. This can be observed in the spectrogram (cf. Figure 2f), where the regular pitch oscillations are clearly defined contrasting with a dark background which indicates much less presence of upper and lower tones.

Trillo is a singing technique described as a rapid *Trill* similar to the sound of a ‘bleating goat’ [24]. It sounds like a quick repetition of one single note and is produced by larynx movement. In the spectrogram (cf. Figure 2g) it can be observed that the pitch oscillations are much less pronounced than for *Trill*. Another distinguishable property are the pitch breaks visible in the spectrogram, which are due to breaks needed by the singer to catch air when performing this exhausting technique.

In comparison to 'normal' singing, *Belt* is produced through a higher subglottal pressure and by keeping more firm vocal cords adduction, which results in higher sound levels [25, 27]. This technique sounds 'forced', i. e., it is not perceived as relaxed singing but rather uptight. *Belting* is referred to as raising the chest voice above the typical register and implies a higher level of physical effort [28]. This can be observed in the spectrogram by the rather straight and tense pitch lines (cf. Figure 2c).

The technique *Lip Trill*, often used as a warm up exercise, is done by continuously vibrating with the lips while simultaneously maintaining phonation [29]. This technique is the only one where the mouth and lips remain closed, something distinctive in the spectrogram, where there is barely any black background (cf. Figure 2d).

Another characteristic technique is *Inhaled*, as its main feature is that, unlike all the other techniques, the sound is produced using an inspiratory airflow instead of an expiratory one. Therefore, the sound is generated while the singer inhales [30], which can be observed in the spectrogram by less clearly defined pitch lines (cf. Figure 2e).

The technique *Inhaled* sounds, to some extent, similar to the techniques *Breathy* and *Vocal Fry*. In *Breathy*, a low subglottal pressure is combined with a less efficient adduction of the vocal cords [31]. This results in a sound characterised by audible airflow, which is shown in the spectrogram by broader and blurrier pitch lines (cf. Figure 2h). In *Vocal Fry*, characterised by lower subglottal air pressure and transglottal air flow, the vocal folds are shortened, even when frequency increases [32]. This is shown in the spectrogram by diffuse and irregular pitch lines (cf. Figure 2i).

Finally, *Spoken*, in contrast to singing, is the only technique that does not require the control of the pitch. The distinguishing feature visible in the spectrogram is a grid-like pattern (cf. Figure 2j) where the horizontal lines (relatively stable) represent the pitch and the vertical ones (unequally spaced out) correspond to the words' articulation.

3.3 User Study

The user study consists on two experiments performed by different groups: (i) musically trained individuals (task based on auditory perception); (ii) non-musically trained individuals (task based on visual perception). Both experiments were performed through a web-based interface and began with an example (either an audio or an spectrogram) of each singing technique. Then, an explanation of the task, presented as a multiple choice test, was given. For each sample, the participants could choose one singing technique out of the ten given possibilities. 60 volunteers (31 female, 29 male; $\mu = 32.3$ years) participated in the study. Most of them were Austrian (43), the rest were German (14) and Australian (3).[4] They were recruited through the authors' social networks and consent, requested through the interface, was a requirement to take part in the experiment.[5]

In the auditory experiment, the participants were expected to identify the singing techniques by listening to the audio excerpts. Since a trained ear is necessary for this task, in the auditory task only participants with a musical education (9 female, 11 male)

[4] Due to the imbalanced distribution of participants, nationalities' role will not be evaluated.

[5] The procedures used in this study adhere to the tenets of the Declaration of Helsinki. Participants consented the use of their anonymous responses only for research.

took part. Their formal training included choir conductor, singing, and vocal studies. In the visual experiment, the participants were expected to identify the techniques by looking at spectrograms generated from the audio excerpts. Spectrograms were chosen since typically used in singing lessons [16], specially to support beginners [33]. Since for the auditory task a trained ear is needed, the visual task was considered a more suitable alternative for the participants without musical background (22 female, 18 male).

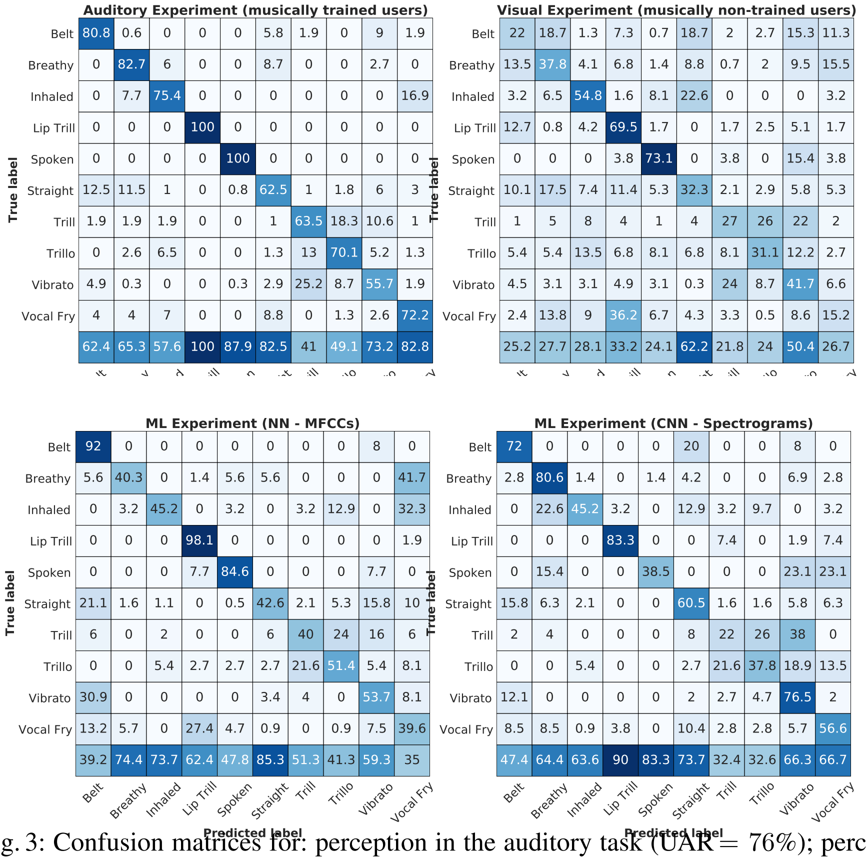
In order to avoid fatigue, the 777 excerpts were randomly distributed across the participants. For the auditory task, this was made in a way that each would annotate between 75 and 80 audio chunks. Since we expect the evaluation of spectrograms to required more time than assessing audio samples, in order to preserve the reliability of the experiment, for the visual task each participant would annotate between 37 and 41 images. In both experiments, in order to prevent individual biases, each sample was evaluated by two different participants, which lead to 1 554 annotations per task.

We are aware that assessing two user groups (experts and non-experts), makes the setups not comparable within the user study. However, the final goal of this study is to make a one-to-one comparison between perception (auditory as well as visual) and ML. In addition, in base of the principle that learning should be tailored to individuals capabilities [34] (which are not the same for musically trained users and non-trained ones) we believe that considering the same task for both user-groups would heavily penalise the non-trained group. Thus, to perform a fair comparison of trained and non-trained users with the ML algorithms, two different perceptual experiments were performed.

3.4 Machine Learning Setup

Following previous works on singing classification [8, 22], both traditional models and neural-based were implemented. Due to space limitations, the results for the traditional models (outperformed by the neural ones) will not be reported. A Neural Network (NN) and a Convolutional Neural Network (CNN) were implemented in the tensorflow framework. The NN, presenting eight layers, Relu as activation function, and categorical crossentropy as loss function, was trained for 40 epochs. The CNN was implemented as in the VocalSet baseline [8], i. e., consisted of seven convolutional layers, seven max pooling layers, learning rate of 0.001, a momentum of 0.6, and categorical crossentropy as loss function. It was trained for 30 epochs.

Two type of features were considered: Mel-Frequency Cepstral Coefficients (MFCCs) and spectrograms. They were chosen as suitable representations according to state-of-the-art literature [35] and their corresponding outcomes will be compared with the auditory and visual perceptual results, respectively. The features were extracted from the audio files (sampling rate: 44100 Hz) with default parameters of the librosa package: fft-size of 2048; frame size of 93 ms; and frame step of 23 ms. For the MFCCs, the first 20 coefficients were extracted. As already mentioned, the 777 excerpts produced by the singers F2, F6, M3, and M11 were used as test set and the remaining 3 157 excerpts as training set. By this guaranteeing a comparable setup w. r. t. the user study, where only the 777 excerpts were assessed.



Auditory Experiment (musically trained users)

	Belt	Breathy	Inhaled	Lip Trill	Spoken	Straight	Trill	Trillo	Vibrato	Vocal Fry
Belt	80.8	0.6	0	0	0	5.8	1.9	0	9	1.9
Breathy	0	82.7	6	0	0	8.7	0	0	2.7	0
Inhaled	0	7.7	75.4	0	0	0	0	0	0	16.9
Lip Trill	0	0	0	100	0	0	0	0	0	0
Spoken	0	0	0	0	100	0	0	0	0	0
Straight	12.5	11.5	1	0	0.8	62.5	1	1.8	6	3
Trill	1.9	1.9	1.9	0	0	1	63.5	18.3	10.6	1
Trillo	0	2.6	6.5	0	0	1.3	13	70.1	5.2	1.3
Vibrato	4.9	0.3	0	0	0.3	2.9	25.2	8.7	55.7	1.9
Vocal Fry	4	4	7	0	0	8.8	0	1.3	2.6	72.2
	62.4	65.3	57.6	100	87.9	82.5	41	49.1	73.2	82.8

Visual Experiment (musically non-trained users)

	Belt	Breathy	Inhaled	Lip Trill	Spoken	Straight	Trill	Trillo	Vibrato	Vocal Fry
Belt	22	18.7	1.3	7.3	0.7	18.7	2	2.7	15.3	11.3
Breathy	13.5	37.8	4.1	6.8	1.4	8.8	0.7	2	9.5	15.5
Inhaled	3.2	6.5	54.8	1.6	8.1	22.6	0	0	0	3.2
Lip Trill	12.7	0.8	4.2	69.5	1.7	0	1.7	2.5	5.1	1.7
Spoken	0	0	0	3.8	73.1	0	3.8	0	15.4	3.8
Straight	10.1	17.5	7.4	11.4	5.3	32.3	2.1	2.9	5.8	5.3
Trill	1	5	8	4	1	4	27	26	22	2
Trillo	5.4	5.4	13.5	6.8	8.1	6.8	8.1	31.1	12.2	2.7
Vibrato	4.5	3.1	3.1	4.9	3.1	0.3	24	8.7	41.7	6.6
Vocal Fry	2.4	13.8	9	36.2	6.7	4.3	3.3	0.5	8.6	15.2
	25.2	27.7	28.1	33.2	24.1	62.2	21.8	24	50.4	26.7

ML Experiment (NN - MFCCs)

	Belt	Breathy	Inhaled	Lip Trill	Spoken	Straight	Trill	Trillo	Vibrato	Vocal Fry
Belt	92	0	0	0	0	0	0	0	8	0
Breathy	5.6	40.3	0	1.4	5.6	5.6	0	0	0	41.7
Inhaled	0	3.2	45.2	0	3.2	0	3.2	12.9	0	32.3
Lip Trill	0	0	0	98.1	0	0	0	0	0	1.9
Spoken	0	0	0	7.7	84.6	0	0	0	7.7	0
Straight	21.1	1.6	1.1	0	0.5	42.6	2.1	5.3	15.8	10
Trill	6	0	2	0	0	6	40	24	16	6
Trillo	0	0	5.4	2.7	2.7	2.7	21.6	51.4	5.4	8.1
Vibrato	30.9	0	0	0	0	3.4	4	0	53.7	8.1
Vocal Fry	13.2	5.7	0	27.4	4.7	0.9	0	0.9	7.5	39.6
	39.2	74.4	73.7	62.4	47.8	85.3	51.3	41.3	59.3	35

ML Experiment (CNN - Spectrograms)

	Belt	Breathy	Inhaled	Lip Trill	Spoken	Straight	Trill	Trillo	Vibrato	Vocal Fry
Belt	72	0	0	0	0	20	0	0	8	0
Breathy	2.8	80.6	1.4	0	1.4	4.2	0	0	6.9	2.8
Inhaled	0	22.6	45.2	3.2	0	12.9	3.2	9.7	0	3.2
Lip Trill	0	0	0	83.3	0	0	7.4	0	1.9	7.4
Spoken	0	15.4	0	0	38.5	0	0	0	23.1	23.1
Straight	15.8	6.3	2.1	0	0	60.5	1.6	1.6	5.8	6.3
Trill	2	4	0	0	0	8	22	26	38	0
Trillo	0	0	5.4	0	0	2.7	21.6	37.8	18.9	13.5
Vibrato	12.1	0	0	0	0	2	2.7	4.7	76.5	2
Vocal Fry	8.5	8.5	0.9	3.8	0	10.4	2.8	2.8	5.7	56.6
	47.4	64.4	63.6	90	83.3	73.7	32.4	32.6	66.3	66.7

Fig. 3: Confusion matrices for: perception in the auditory task (UAR = 76%); perception in the visual task (UAR = 41%); classification from a Neural Network (NN) fed with MFCCs (UAR = 59%); and classification from a CNN fed with Spectrograms (UAR = 57%). Darker cells indicate higher values (%); rows encode real labels. Recalls are given in the diagonal; precisions are shown in the last row of each matrix. Note that the UAR is an overall measure computed from the whole confusion matrix.

4 Results

4.1 User Study

As expected, the experimental outcomes show a higher performance from the musically trained participants: UAR = 76% for the auditory task w. r. t. to a UAR = 41% for the visual one. In Figure 3 the confusion matrices for both experiments are displayed. The higher recall and precision achieved by musically trained users is shown for all the techniques, which is displayed by a well defined diagonal and a darker precision row for the auditory results. The confusion between singing techniques experienced by users without musical training is shown by the spread of responses across the matrix as well as by the lower precision (cf. light colour of the last row) for the visual results.

Remarkable results are shown for the techniques *Lip Trill* and *Spoken*, recognised with the highest recall in both experiments: in the auditory, both techniques achieved

100% recall; in the visual experiment, they achieved 69.5% and 73.1%, respectively. Indeed, these two techniques are particularly distinctive w. r. t. the others, which make them more easily recognisable. As mentioned in Section 3.2, from an auditory point of view, *Lip Trill* is the only technique produced with a closed mouth and *Spoken* is the only one for which the pitch is not controlled. Although these aspects are visible in the spectrograms, it is important to note the low precision achieved for both techniques in the visual task: 33.2% and 24.1%, respectively; which indicates that despite their characteristics, these techniques are often wrongly chosen by the non-experts group.

Beyond the expected performance differences between listeners' groups, a prominent confusion pattern is common in both experiments, i. e., samples from *Vibrato* are wrongly identified as *Trill*. In both tasks, the amount of misclassifications is nearly half of the correctly identified samples. For the auditory experiment, 25.5% misclassifications vs. 55.7% correct hits; for the visual one, 24% misclassifications vs. 41.7% correct hits. The confusion pattern is also shown in the opposite direction, i. e., *Trill* instances are wrongly identified as *Vibrato*, a result consistent with previous research showing that *Trill* might be similar to *Vibrato* performed with an 'exaggerated extent' [36]. The described confusion pattern involves *Trillo* as well, i. e., *Trill* and *Trillo* are misclassified not only as *Vibrato*, but also amongst themselves. Indeed, the three techniques are similar, since produced by modulating the fundamental frequency (cf. Section 3.2).

Finally, a prominent confusion is displayed for the visual experiment, i. e., *Vocal Fry* is wrongly identified as *Lip Trill*. The percentage of misclassifications exceeds by far the amount of correctly identified instances: 36.2% vs 15.2%. The pattern is not shown for the auditory experiment, which suggest that this type of confusion relates to similarities in the spectrograms difficultly disentangled without audio information.

4.2 Machine Learning

Amongst the evaluated algorithms and feature sets, the best performing model was the NN fed with MFCCs (UAR = 59%) followed by the CNN fed with spectrograms (UAR = 57%). Confirming the results shown in both perceptual experiments, *Lip Trill*, and to some extent *Spoken*, are also the two techniques best recognised by the model fed with MFCCs: 98.1% and 84.6% of recall, respectively; cf. diagonal in Figure 3 (NN - MFCCs). This was also shown for the model fed with spectrograms concerning *Lip Trill*, achieving the highest recall (83.8%), but not for *Spoken*, reaching only 38.5% recall; cf. Figure 3 (CNN - Spectrograms). It is important to note, that despite the low recall for *Spoken*, the precision for this technique is lower for the NN than for the CNN, which indicates that the promising recall is only due to the high confusion attracted by the class; the same is displayed for the visual experiment but not for the auditory one.

The results from the model trained with MFCCs show that except for *Belt* (recall = 92%), all other techniques achieved a considerably lower recall: 39.6%≤recall≤53.7%. *Belt* was also well recognised in the auditory experiment but not in the visual one, which suggests that acoustic properties characteristic of this technique, recognisable by ear, can be better captured by specific acoustic features such as MFCCs than by spectrograms. In fact, this is to some extent confirmed by the lower recall for *Belt* achieved by the CNN trained with spectrograms, i. e., 72%.

As shown in the user study, the most prominent confusion pattern displayed by the ML results is between *Trill*, *Trillo*, and *Vibrato*. This is clearly shown by the misclassification of *Trill* instances as *Trillo*: 24% and 26% for the model trained with MFCCs and spectrograms, respectively; as well as those misclassified as *Vibrato*: 16% and 38%, respectively. However, unlike in the user study, this confusion is not displayed in the opposite direction for the ML task, i. e., almost no instances of *Vibrato* are wrongly classified as neither *Trill* nor *Trillo*, misclassifications $\leq 4.7\%$ for both models.

Interestingly, *Vibrato* is particularly well classified by the CNN, i. e., the model trained with the spectrograms (76.5%). This is also shown, to some extent, by the non-trained user participating in the visual task, for whom this technique is identified as the fourth best (41.7%). Differently, in the auditory study, *Vibrato* was the technique worse recognised (55.7%), and also for the NN (model trained with MFCCs), *Vibrato* was by far worse classified than for the CNN (53.7% vs 76.5%). This suggest that spectrograms are more suitable than acoustic features for characterising *Vibrato*'s properties, something observable both perceptually and from a computational point of view.

Finally, another prominent confusion pattern shown by the model trained with MFCCs is given by the high percentage of *Breathy* and *Inhaled* samples wrongly classified as *Vocal Fry*: 41.7% and 32.3%, respectively. This is partially mirrored by the results from the user study. A major confusion of *Inhaled* towards *Vocal Fry* is shown in the auditory task (16.9%); while a major confusion of *Breathy* towards *Vocal Fry* is shown in the visual experiments (15.5%). However, this confusion pattern is not shown for the model trained with spectrograms, for which the misclassification is shown between *Breathy* and *Inhaled* themselves: 22.6% of *Inhaled* samples are wrongly classified as *Breathy*. This suggests that training a ML model with acoustic features such as MFCCs might enable to artificially mirror, and even amplify, perceptual patterns shown by humans assessing different modalities. Something not possible when using spectrograms.

5 AI in Singing Education: Future Directions

Within the e-learning context, the most obvious use-case for a system able to recognise singing techniques is to provide feedback during students' training. For instance, since the singing technique *Breathy*, sometimes also referred to as *Rough*, is not desired in most genres [22], the ML-based application would first detect *Breathy* singing and subsequently suggest exercises to prevent it. Our comparative results confirm previous works on human vs. machine speech identification [13], indicating that the most predominant perception patterns shown by humans can be mirrored by ML. Nevertheless, while our models outperform non trained users, they are still less accurate than musically trained individuals. This indicates that standard ML architectures (as those used in this study) could be useful in providing feedback to beginners; however, more sophisticated models should be developed to meaningfully support advanced learners.

Our experimental outcomes also show that ML can capture confusion patterns coming from different perceptual modalities. This type of parallelism might be particularly informative when integrated in a XAI system, i. e., a ML systems which besides giving a prediction, is also able to provide a human-understandable reasoning justifying it. Thus, an XAI assistant could propose specific warm-up exercises depending on the singers' voice [37], subsequently assess whether the performed technique match the target, and finally illustrate (either visually or acoustically, depending on which feature representation is more informative), the predicted class (performed by the student) with respect to the target one (performed by a professional singer of the system's database).

Similarly, in base of our results, an XAI assistant could also highlight the most prominent confusion patterns shown for both perception and classification, i. e., the confusion between *Trill*, *Trillo*, and *Vibrato*. By displaying not only a visual (qualitative) representation but also precision (quantitative) measures achieved by the model, learners might gain a more objective understanding of the similarities between techniques, something that beyond being perceived, can also be measured. At the same time, this would also illustrate real challenges in distinguishing amongst some techniques, which would encourage a more constructive learning experience. We believe that the use of intelligent systems as the one just described, specially when including an XAI component, would promote in first place exploration, motivated by the curiosity of interacting with the XAI assistant. Furthermore, another important expected outcome is to encourage the students to carefully evaluate their own performance, both visually and acoustically, which would lead to the development of self-reflective and critical skills.

Needless to say that such a system, in particular considering that the current results are way below human proficiency, would be expected to be used as a complementary tool to traditional teaching, i. e., supporting the student (specially during individual learning), but used under the close supervision of the teacher. Indeed, a full development of the system, including an user interface as well as a usability assessment in a real pedagogical scenario, is still to be done and constitutes one of our future priorities. In this process, a continuous monitoring from singing educators, critically assessing the potential of the system in complementing their own practice, is essential.

Finally, beyond supporting vocal training, the recognition of specific singing techniques in a song might also enable the classification of a given piece into a musical style or genre. For instance, the use of the *Belting* technique, particularly for women, is typically used in pop genre [38] while *Vibrato* is a strong indicator of operatic singing style [39]. The application of this technology in the context of automatic genre classification is clearly relevant for music recommendation systems [40]. Similarly, an efficient singing detection system could also be utilised for an e-learning application aimed to support students' understanding of musical genres in relationship to singing styles.

6 Conclusions

We presented a comparative assessment of humans' and ML performance in singing technique recognition. Our study shows that some confusion patterns typical of perception are mirrored by ML, which highlights the potential of supporting education with AI to illustrate (and further understand) perceptual processed. Our results also indicate that ML can capture patterns displayed by different perceptual cues: auditory and visual. This suggests that AI could be of interest to enhance learning through different perceptual modalities. The presented results seem to encourage further research on the application of XAI in singing pedagogy, which could promote students' reflective and critical skills, by this enhancing the outcomes of a student-centered learning process.

References

1. White, B.D.: Singing Techniques and Vocal Pedagogy. University of Surrey Press, Surrey, UK (1985)
2. Stark, J.: Bel canto: A history of vocal pedagogy. University of Toronto Press, Toronto, Canada (1999)
3. Nakano, T., Goto, M., Hiraga, Y.: Subjective evaluation of common singing skills using the rank ordering method. In: Proceedings of the International Conference on Music Perception and Cognition, Bologna, Italy (2006) 1507–1512
4. Haßler, B., Major, L., Hennessy, S.: Tablet use in schools: A critical review of the evidence for learning outcomes. Journal of Computer Assisted Learning **32**(2) (2016) 139–156
5. Lã, F.M., Fiuza, M.B.: Real-time visual feedback in singing pedagogy: Current trends and future directions. Applied Sciences **12**(21) (2022) 10781
6. Wilson, P.H., Thorpe, C.W., Callaghan, J.: Looking at singing: Does real-time visual feedback improve the way we learn to sing. In: Proceedings of the Asia-Pacific Society for the Cognitive Sciences of Music Conference, Seoul, South Korea (2005) 4–6
7. Gupta, C., Li, H., Wang, Y.: Automatic leaderboard: Evaluation of singing quality without a standard reference. IEEE/ACM Transactions on Audio, Speech, and Language Processing **28** (2019) 13–26
8. Wilkins, J., Seetharaman, P., Wahl, A., Pardo, B.: Vocalset: A singing voice dataset. In: Proceedings of the International Society for Music Information Retrieval Conference, Paris, France (2018) 468–474
9. Yamamoto, Y., Nam, J., Terasawa, H., Hiraga, Y.: Investigating time-frequency representations for audio feature extraction in singing technique classification. In: Proceedings of the IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Tokyo, Japan (2021) 890–896
10. Burkhardt, F., Brückl, M., Schuller, B.: Age classification: Comparison of human vs machine performance in prompted and spontaneous speech. In: Proceedings of Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung, Magdeburg, Germany (2021) 35–42
11. Koemans, J.: Man vs Machine: Comparing cross-lingual Automatic and Human Emotion Recognition in Background Noise. Master Thesis, Radboud University (2020)
12. Parada-Cabaleiro, E., Schmitt, M., Batliner, A., Hantke, S., Costantini, G., Scherer, K., Schuller, B.: Identifying emotions in opera singing: Implications of adverse acoustic conditions. In: Proceedings of the International Society for Music Information Retrieval Conference, Paris, France (2018) 376–382
13. Parada-Cabaleiro, E., Batliner, A., Schmitt, M., Schedl, M., Costantini, G., Schuller, B.: Perception and classification of emotions in nonsense speech: Humans versus machines. PLoS ONE **18**(1) (2023) e0281079
14. McCoy, S.: Singing pedagogy in the twenty-first century: A look toward the future. In Harrison, S.D., O'Bryan, J., eds.: Teaching singing in the 21st century. Springer, New York, NY, USA (2014) 13–20
15. Miller, D.G.: Resonance in singing: Voice building through acoustic feedback. Inside view press, Gahanna, OH, USA (2008)
16. Lã, F.M.: Teaching singing and technology. In Basa, K.S., ed.: Aspects of singing II: Unity in understanding - Diversity in aesthetics. VoxHumana, Nürnberg, Germany (2012) 88–109
17. Friberg, A., Bresin, R., Sundberg, J.: Overview of the kth rule system for musical performance. Advances in Cognitive Psychology **2**(2) (2006) 145
18. Rossiter, D., Howard, D.M.: Albert: a real-time visual feedback computer tool for professional vocal development. Journal of voice: official journal of the Voice Foundation **10**(4) (1996) 321–336

19. Welch, G.F., Howard, D.M., Himonides, E., Brereton, J.: Real-time feedback in the singing studio: An innovatory action-research project using new voice technology. Music Education Research **7**(2) (2005) 225–249
20. Stavropoulou, S., Georgaki, A., Moschos, F.: The effectiveness of visual feedback singing vocal technology in greek elementary school. In: Proceedings of the International Computing Music Conference, Athens, Greece (2014) 1786–1792
21. Yamamoto, Y., Nam, J., Terasawa, H.: Analysis and detection of singing techniques in repertoires of j-pop solo singers. In: Proceedings of the International Society for Music Information Retrieval Conference, Bangaluru, India (2022) 384–391
22. Xu, Y., Wang, W., Cui, H., Xu, M., Li, M.: 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 (1) (2022) 1–16
23. Bekkar, M., Djemaa, H.K., Alitouche, T.: Evaluation measures for models assessment over imbalanced data sets. Journal of Information Engineering and Applications **3** (2013) 27–39
24. Isherwood, N.: Vocal vibrato: New directions. Journal of Singing **65**(3) (2009) 271
25. Kob, M.: Physical Modeling of the Singing Voice. PhD thesis, Bibliothek der RWTH Aachen (2002)
26. Sangiorgi, T., Manfredi, C., Bruscaglioni, P.: Objective analysis of the singing voice as a training aid. Logopedics Phoniatrics Vocology **30** (2005) 136–146
27. Sundberg, J., Thalén, M.: Respiratory and acoustical differences between belt and neutral style of singing. Journal of Voice **29**(4) (2015) 418–425
28. LeBorgne, W.D., Lee, L., Stemple, J.C., Bush, H.: Perceptual findings on the Broadway belt voice. Journal of Voice **24**(6) (2010) 678–689
29. Gaskill, C.S., Erickson, M.L.: The effect of a voiced lip trill on estimated glottal closed quotient. Journal of Voice **22**(6) (2008) 634–643
30. Vanhecke, F., Moerman, M., Desmet, F., Six, J., Daemers, K., Raes, G., Leman, M.: Acoustical properties in inhaling singing: A case-study. Physics in Medicine **3** (2017) 9–15
31. Proutskova, P., Rhodes, C., Crawford, T., Wiggins, G.: Breathy, resonant, pressed–automatic detection of phonation mode from audio recordings of singing. Journal of New Music Research **42**(2) (2013) 171–186
32. Appleman, R., Bunch, M.: Application of vocal fry to the training of singers. Journal of Singing **62**(1) (2005) 53–9
33. Hoppe, D., Sadakata, M., Desain, P.: Development of real-time visual feedback assistance in singing training: A review. Journal of Computer Assisted Learning **22**(4) (2006) 308–316
34. Schleicher, A.: Educating learners for their future, not our past. ECNU Review of Education **1**(1) (2018) 58–75
35. Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.Y., Sainath, T.: Deep learning for audio signal processing. IEEE Journal of Selected Topics in Signal Processing **13**(2) (2019) 206–219
36. Sundberg, J.: Acoustic and psychoacoustic aspects of vocal vibrato. Vibrato (1995) 35–62
37. Elliot, N., Sundberg, J., Gramming, P.: What happens during vocal warm-up? Journal of Voice **9**(1) (1995) 37–44
38. Spivey, N.: Music theater singing... let's talk. Part 2: Examining the debate on belting. Journal of Singing **64**(5) (2008) 607–614
39. Howes, P., Callaghan, J., Davis, P., Kenny, D., Thorpe, W.: The relationship between measured vibrato characteristics and perception in western operatic singing. Journal of Voice **18**(2) (2004) 216–230
40. Schedl, M., Knees, P., McFee, B., Bogdanov, D., Kaminskas, M.: Music recommender systems. Recommender Systems Handbook (2015) 453–492