Machine Learning for Monitoring Vocal Health and Performance of Professional Singers

Reni, S., Jones, S. and Kale, I.

Machine Learning for Monitoring Vocal Health and Performance of Professional Singers

Samuel P. Jones, Saumya Kareem Reni and Izzet Kale
Applied DSP and VLSI Research Group (ADVRG)
School of Computer Science and Engineering, University of Westminster
115 New Cavendish Street, London W1W 6UW, United Kingdom
Email: jonesmondiale@gmail.com, {S.Reni, kalei}@westminster.ac.uk

***Abstract*—This paper gives an insight into interdisciplinary research examining the use of Machine Learning techniques to monitor the vocal health of professional singers. The work reported establishes the viability of using a dataset of audio samples of the human voice to train a convolutional neural network to assess fluctuations in vocal performance of professional singers. Variations in the ease and quality of vocal production are a common experience among those who rely on their voice for a living, and vocal health issues can often prove traumatic and debilitating. Yet the use of data gathering and analysis among professional singers remains rare. The work reported in this study provides a novel basis for a method via which singers, and others who use their voice professionally, can make informed investigations into the potential causes of those fluctuations, and facilitate preventative medical intervention where appropriate.**

Keywords—supervised Machine Learning, convolutional neural networks, image classification, vocal health, digital audio analysis and signal processing, Western classical singing

I. Introduction

The use of biometric data in professional sports has become common practice in recent years. Analysis of these data is used to assess and identify the causes of variations in athletes' performance, and to enable early medical intervention to prevent injury [1]. By contrast, professional singers are reliant on less technologically advanced methods of monitoring and analysing their vocal performance, conditioning, and health. In a survey conducted amongst 28 vocal professionals for this study [2], all respondents reported experiencing variations in the quality or ease of production of their voice, but only one reported having used a software tool or application as an approach to identifying the possible causes of those variations. The scarcity of objective data-driven approaches to vocal performance monitoring can result in a reluctance to discuss vocal health amongst singers and other vocal professionals, and even more harmfully, a disinclination to seek professional medical help [3].

Existing approaches to Machine Learning (ML) research in this area have examined the possibility of using voice data gathered from users to quantify various metrics of vocal performance, and make assessments of vocal health based on monitoring of those metrics [4] - [7]. This study takes an alternative approach, ultimately empowering voice users to rate and classify their own vocal performance in any given recording sample. In this aspect, the task is fundamentally different from clinical diagnosis of pathological conditions, where the clinician is the ultimate arbiter of vocal health; in the context of this study, where a vocal issue may fall short of the threshold of a pathological dysfunction, the true authority on the condition of the voice must ultimately be the singer themselves. This avoids the issue of intra-rater variability identified by Gupta et al. [8].

Taking this approach entails using a single metric assessed from each voice sample, rather than extracting multiple features, and as a result necessarily requires a larger number of files in the training dataset than is typical of existing clinical studies in this area. By definition, the process will involve supervised ML, ultimately with the singers themselves defining the quality of the sound files they generate. Care should be taken that only the quality of vocal performance is assessed, and not the quality of the sound file dependent on other factors such as recording equipment and environment.

This study examines the viability of this task using two datasets of audio samples from two sources. The first dataset was from the open-access Saarbrücker Stimmdatenbank (Saarbrücken Voice Database) [9]. The second dataset was generated by a professional opera singer during recording sessions for the specific purposes of this study. The use of ML in this way to assess the quality of a singer's voice at a given moment is a novel aspect of this study. The study limits itself to a binary classification of voice samples as being characteristic of healthy or unhealthy vocal production (see Fig. 1), in order to establish initially the validity of the concept, specifically in the context of orthodox Western classical singing technique. A trained model achieving high levels of accuracy in this task would suggest that the underlying concept is sound and could provide a basis for more sophisticated future iterations of the software.

Four further sections of this paper follow. The second section describes the methodology, data and pre-processing used in the study. The third section reports the results of the tests carried out, the fourth section discusses those results, and the fifth section presents a conclusion to this study.

II. Method

A. Dataset 1

The first dataset was taken from the Saarbrücken Voice Database (SVD), an open-access database created and maintained by the Universität des Saarlandes (UdS) [9]. The SVD contains audio samples from male and female subjects across a range of ages, some in good vocal health, and others with a variety of vocal pathologies. A total of 12314 audio samples were downloaded, of which 6131 (3000 male, 3131 female) were pathological samples, and 6183 (2331 male, 3852 female) were healthy samples. Pathological conditions included and their descriptions are given in [2].

Audio files were downloaded as mono .wav files at a bit rate of 800kbps (i.e. around CD-quality). Samples contained recordings of the subject phonating on three vowels (in the International Phonetic Alphabet (IPA) [ɑ: i: u:]) at a constant low, medium or high pitch (presumably relative to each subject's own typical day-to-day speaking pitch).

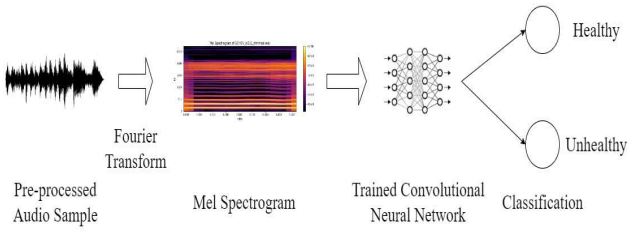

Fig. 1. Process from audio data to classification

Data Pre-processing:

The data acquired needed to be pre-processed into a format suitable for training a convolutional neural network (CNN). The SVD documentation provided by UdS states that any audio files below a minimum quality level have already been removed from the database [10]; for example, all samples included in the database contain distinguishable phonated sound from the outset.

The raw audio files were renamed with a prefix denoting their classification as healthy or pathological, and saved to appropriately labelled folders. Further cleaning and merging was then carried out. Duplicate files as a result of subjects with multiple pathologies were removed, retaining only one copy of the sample file in each case. Files shorter in duration than a threshold value of 500ms were discarded. Human voices can phonate in the Western classical style to frequencies as low as around 60Hz [11], and the consistency of vocal vibrations was assumed to be one factor in vocal quality. Hence a duration threshold of 500ms was set, resulting in a minimum of 30 cycles for this consistency to be present to an assessable degree in the samples. All remaining audio files in the dataset were trimmed to a duration of 500ms. Where the duration was longer than 500ms, the first 500ms of the file was selected, since the clarity of onset of sound is a distinguishing factor of healthy vocal production in classical singing. Finally, a mel-scale spectrogram was generated for each file.

The mel spectrogram has the advantage of separating the resonant frequencies of an audio sample in a manner which reflects the sensitivity of the human ear more closely than a linear scale, allowing for a more well-adjusted visual analysis of the resultant image by the CNN's filters [12]. Via the above steps, 11552 audio files were pre-processed and used to generate 11552 mel spectrograms, 6183 from healthy voice samples and 5369 from pathological voice samples. These formed the dataset which would be used to train, validate and test the CNN. This full dataset was then shuffled. 1600 files (800 healthy and 800 pathological) were first randomly selected and removed from the dataset, to be kept entirely unseen and reserved for further testing. From the remaining files 5600 files (2800 healthy + 2800 pathological) were selected for training, 2400 files (1200 + 1200) for validation, and 1000 (500 + 500) for initial testing.

B. Dataset 2

Whereas datasets taken from the SVD have been used in previous studies of ML, the use of a dataset generated by an individual singer to train a bespoke CNN to classify further voice samples from that singer is a novel area of study. The second dataset was generated during two recording sessions capturing voice samples from a professional classical singer, a 49-year-old male operatic bass-baritone. Over the course of two days, six vowels (IPA [ɑː ɛː iː ɔː uː yː ɜː]) were recorded under home studio conditions with some noise isolation, using a Røde NT-USB external microphone, situated facing the singer at a distance of approximately 50cm, and connected to a 2020 Apple MacBook Air running Audacity for macOS version 2.4.2. Audio files were recorded via a mono channel using Audacity's default settings of a sample rate of 44.1kHz at a depth of 32-bit float, and saved and exported in .wav format with signed 16-bit Pulse Code Modulation (PCM) encoding. Each vowel was recorded at semitone intervals across a full pitch range of two octaves – slightly more on some vowels than others, according to the singer's physical comfort, giving between 72 and 90 samples for each vowel. This range was covered twice for each vowel: in the first instance, the singer generated the sound with a healthy vocal production; the singer then repeated the recordings, this time emulating the effect of vocal fatigue or other vocal pathologies (which by and large cause inflammation of the vocal folds, and consequently a lack of consistent contact between the folds) by use of laryngeal constriction, providing a rough, "unhealthy" sound.

Data pre-processing and augmentation then followed. Each sample was edited so that it contained distinguishable phonated sound from the outset, and was a minimum of 500ms in duration. In total, 573 healthy and 552 (human-emulated) unhealthy samples were generated during these sessions. This dataset of audio files was then augmented via six processes. Firstly, the volume of each file was adjusted. Four sets of adjustments were made: increases of 25% and 50%, and decreases of 25% and 50% of the original volume level, providing some basic simulation of variations in recording conditions. A fifth augmented set of files was produced by adding white noise at 0.5% of the maximum noise level to the background of each audio file. Finally, a sixth augmented set was produced with a 125ms echo effect being added to each audio file. These six augmentations produced an augmented dataset of 3438 healthy and 3312 emulated unhealthy samples. When added to the original dataset, this produced a full dataset of 4011 healthy and 3864 emulated unhealthy samples, giving 7875 files in all. As with Dataset 1, this full dataset was shuffled, and a selection made of 5000 files for training (2500 healthy + 2500 emulated unhealthy), 1000 for validation (500 + 500), 800 for initial testing (400 + 400), with 800 files (400 + 400) being set aside for the purposes of further unseen testing.

C. Neural Network (NN) Training

Image analysis of mel spectrograms by means of a trained CNN was selected as a suitable method of classifying the audio samples. Western classical singers' distinctive voices are understood to be characterised by the presence of a "singer's formant" – an unusually resonant frequency in the voice which enables the singer to be heard unamplified in large spaces [13]. The presence and strength of this formant in the frequency spectrum is likely to be a significant factor in how a "healthy" sound is identified (contrast Fig. 2 and Fig. 3, for example), and so analysis of distinct frequencies in each sample is necessary. In essence, the task being demanded of the CNN is to analyse audio samples in a manner as similar as possible to a human listener. Thus, the separation of frequencies in mel spectrograms in a manner which models the response of the human ear to audio signals [12] makes them a very strong contender for analysis of audio samples in this context.

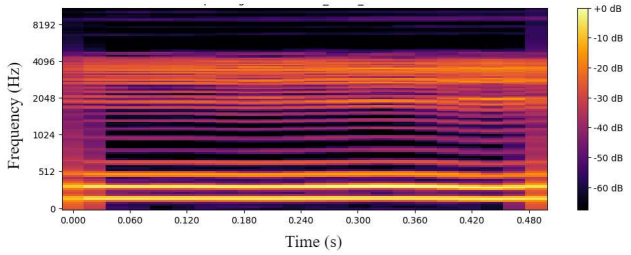

Fig. 2. Mel spectrogram from healthy voice sample

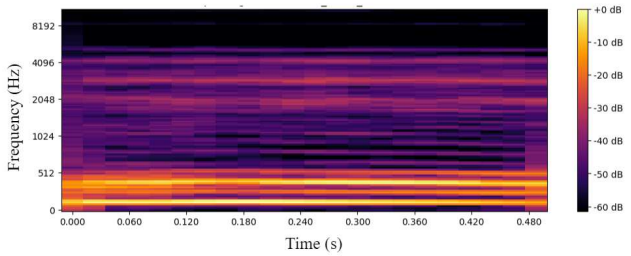

Fig. 3. Mel spectrogram from unhealthy voice sample

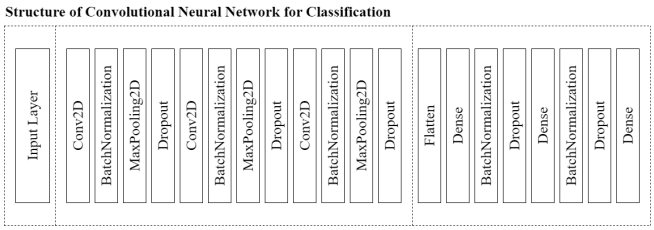

Fig. 4. Architecture of Convolutional Neural Network

Fig. 4 presents the CNN architecture. Note that since this is a binary classification task, the output layer is a fully connected single output dense layer. The process of selecting the most appropriate architecture involved a trade-off between keeping the system as computationally efficient as possible, while still enabling the trained model to achieve high levels of accuracy when classifying unseen data, particularly data from sources beyond that from which the training data was taken.

III. RESULTS

All processes were carried out on an MSI Summit E14 laptop using an 11th generation Intel Core i7-1185G7 processor running at 3GHz and 15.7GB of usable RAM. The Python libraries and versions used for various functions can be seen in Table 1.

TABLE 1. Implementation Platform -Versions and Libraries

Python Library	Version	Function
librosa	0.10.0.post2	Audio manipulation
matplotlib	3.7.2	Results visualisation
numpy	1.23.5	Mathematical functions
scikit-learn	1.3.0	Results analysis
tensorflow	2.13.0	NN architecture

A. Speed of Computation

Using the hardware and software as described, typical execution times were 22.3s per 100 files for pre-processing operations, and 0.5s per 100 files for classification. Given that each file was 500ms in duration, this would suggest a typical execution time of 22.8s for full processing and analysis of 50s of audio data. Consequently, real-time analysis of a continuous audio signal would appear to be a realistic prospect if required.

B. NN Model Trained on Dataset 1 Tested on Dataset 1 files

Using the files from Dataset 1 to train the CNN over 10 epochs, a training accuracy of 0.9927 was reached, with a training loss of 0.0245, alongside a flawless validation accuracy of 1.000 and a validation loss of 0.0003 (see Fig. 5). This trained model was then tested on selections of the 800 files reserved for unseen testing. 100 of these files were randomly selected, and this was repeated over 100 cycles, giving 10000 tests in all. The accuracy achieved in this unseen test was 100%, with the results containing 4913 true positives and 5087 true negatives. These levels of accuracy were typical of those achieved in many previous training runs using various other combinations of files from the dataset. The highest levels of accuracy and lowest levels of loss were reached within 6-8 epochs (Fig. 5), and so a lesser number of epochs may well be a more computationally efficient choice.

C. NN Model Trained on Dataset 2 Tested on Dataset 2 Files

Using the files from Dataset 2 to train the CNN over 10 epochs, a training accuracy of 0.9920 was reached, with a training loss of 0.0234, alongside a validation accuracy of 0.9990 and a validation loss of 0.0035 (see Fig. 6). This trained model was then tested on selections of the 800 files reserved for unseen testing. 100 of these files were randomly selected, and this was repeated over 10 cycles, giving 1000 tests in all. The accuracy achieved in this unseen test was 100%, with the results containing 494 true positives and 506 true negatives. Maximum accuracy and minimum loss were achieved within 4 epochs (Fig. 6), and so again a lesser number of epochs may well be a more computationally efficient choice, as well as helping to avoid potentially overfitting the model to the training data [14].

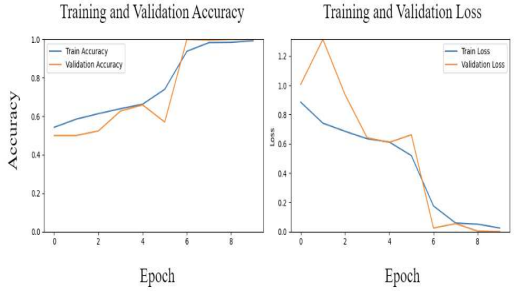

Fig. 5. Training and validation results for Dataset 1

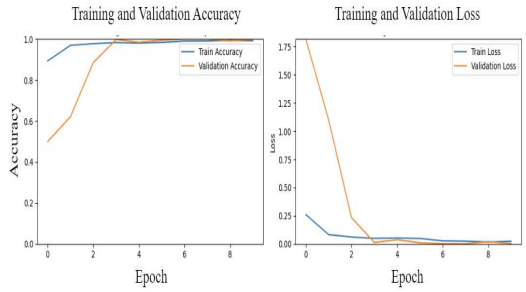

Fig. 6. Training and validation results for Dataset 2

D. NN Model Trained on Dataset 1 Tested on Dataset 2 Files

This experiment involved the NN model that was trained on files from Dataset 1 being tested on audio files from Dataset 2. The objective was to expose the model to a completely new set of files from an entirely different source. In the first set of tests, the testing files were selected only from the original unaugmented audio files gathered for Dataset 2, thus carrying out an assessment of a set of files from a single source. A set of 100 cycles of 100 tests was carried out, giving 10000 tests in all. An overall accuracy of 96.90% was returned, with 4884 true positives, 4806 true negatives, and 310 false positives. Another set of tests was carried out, this time using the full Dataset 2 including the augmented audio files, in order to simulate a set of files taken from multiple sources. 10000 tests returned an accuracy of 99.58%, with 4950 true positives, 5008 true negatives, and 42 false positives.

IV. Discussion

A. Legitimacy and Scope

Given that assessment of vocal quality in singing or speaking is usually considered a highly subjective matter, it might reasonably be asked whether objective, binary categorisations can be applied to audio samples of the human voice. In a short preliminary survey, a small group of professionals with extensive experience of working in classical and operatic vocal music was asked to apply just such a categorisation to a set of six audio recordings of a variety of vowels being sung at a variety of constant pitches, played in a randomised order. These recordings were generated by a professional classical singer: in three of the recordings, the singer used an orthodox, "healthy" classical technique; in the other three, the singer emulated the effect of an "unhealthy" vocal condition in a similar manner to that used in the generation of Dataset 2. The results of the assessment were near-unanimous, with three of the samples being identified by all six participants as "healthy", two being identified by five of the participants as "unhealthy", with one responding "don't know". Within its limited scope, the results of this survey suggest that in the context of classical operatic singing, there is a clear consensus as to what constitutes "healthy" or "unhealthy" vocal production.

It can therefore reasonably be expected that the fundamental concept of classifying vocal production on a binary basis is valid, at least within the context of orthodox Western classical operatic vocal technique. This study limited itself to examination of this area of professional vocal production; the question of whether the conclusions are transferable to other styles and conventions of music and speech remains open for further investigation. Future research might explore the applicability of this approach to other musical genres or even everyday speech patterns, where the line between "healthy" and "unhealthy" vocal production may be less distinct.

B. Critical Evaluation

The high levels of accuracy achieved in many of these tests should be considered in the context of the straightforward nature of the procedure being carried out by the neural network. A binary healthy/unhealthy classification is as simple a task as can be imagined, and accuracies in excess of 99% are not atypical in studies of this kind (see for example Al-Nasheri et al. [7]). These high accuracy levels provide a potential starting point for more sophisticated classification regimes. With a large training dataset, there is a constant danger of producing an overfitted trained model, which will tend to produce lower accuracies when tested on unseen data from different sources. Hence it is notable that the model trained on Dataset 1 was capable of classifying to such a high degree of accuracy sets of unseen files taken from Dataset 2 - an entirely different source and produced under different conditions from the training dataset. This bodes well for the future prospects of designing, building, and training neural networks to classify a diversity of real voices in a range of real-world situations.

From an ethical and artistic standpoint, it should also be emphasised persistently that the singer providing the data for the training of the model should be in control of that data, the prime beneficiary of the model and its application, and most importantly, the sole arbiter of what constitutes ideal and below-ideal quality when it comes to assessing and classifying the performance of their own voice.

V. Conclusion

This paper describes the feasibility of developing an audio signal classification tool that distinguishes a healthy vocal sample from an unhealthy one. The model developed was subjected to training using large datasets from two different sources, containing healthy voice signals and signals of different vocal pathologies, with promising results. As a preliminary study, this interdisciplinary research has the potential to have a significant impact on the field of audio signal processing, facilitating healthy and unhealthy voice classification for early identification and diagnosis of voice problems, and thereby opening new possibilities in vocal healthcare, rehabilitation, and care of the professional voice. Some further investigation would be desirable as to the reproducibility of the results returned by this study, particularly looking at the range of files included in the testing datasets, and also the inclusion of files with a greater diversity, for example in types of voice, augmentation, and methods of collection. Audio samples collected under circumstances which more closely model real-world performance situations might also present insightful challenges to the trained models, leading to further more robust and versatile implementations of this initial concept. The high levels of accuracy achieved in the binary classification suggests that the underlying novel concept of the study, using Machine Learning to assess the quality of a singer's voice at a given moment, is valid, and could form the basis for further development of the concept.

REFERENCES

[1] M. Bates, "The rise of biometrics in sport," *IEEE Pulse*, Available: https://www.embs.org/pulse/articles/the-rise-of-biometrics-in-sports/, 29 June 2020.

[2] S. P. Jones, "Professional vocal performance monitoring software," pp. 16-25, unpublished.

[3] N. Y. K Li-Jessen and C. Jones, "Keeping injured voices hush-hush: why professional singers and actors often don't seek treatment for vocal illness," *The Conversation*, Available: https://theconversation.com/keeping-injured-voices-hush-hush-why-professional-singers-and-actors-often-dont-seek-treatment-for-vocal-illness-183330, 13 June 2022.

[4] Z. Lei, L. Martignetti, C. Ridgway, S. Peacock, J. T. Sakata and N. Y. K. Li-Jessen, "Wearable neck surface accelerometers for occupational vocal health monitoring: instrument and analysis validation study," *JMIR Formative Research*, vol. 6, no. 8, Available: https://formative.jmir.org/2022/8/e39789/, August 2022.

[5] Z. Dankovičová, D. Sovák, P. Drotár and L. Vokorokos, "Machine learning approach to dysphonia detection," *Applied Sciences*, vol. 8, no. 10, Available: https://www.mdpi.com/351524, October 2018, 1927.

[6] Voice Clinical Systems, "What is OnlineLab?," Voice Clinical Systems, Available: https://voiceclinicalsystems.com/en/app-onlinelab/, 2022.

[7] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, K. H. Malki, T. A. Mesallam and M. Farahat Ibrahim, "Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions," *IEEE Access*, vol. 6, Available: https://ieeexplore.ieee.org/document/7906604, April 2017, pp. 6961-6974.

[8] R. Gupta, T. Chaspari, J. Kim, N. Kumar, D. Bone and S. Narayanan, "Pathological speech processing: state-of-the-art, current challenges, and future directions," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP*), Shanghai, China, 2016, pp. 6470-6474, doi: 10.1109/ICASSP.2016.7472923.

[9] Universität des Saarlandes, Saarbrücker Stimmdatenbank, Universität des Saarlandes, Available: https://stimmdb.coli.uni-saarland.de/index.php4.

[10] Universität des Saarlandes, Saarbruecken Voice Database Handbook, Universität des Saarlandes, Available: https://www.stimmdatenbank.coli.uni-saarland.de/help_en.php4, 2007.

[11] P. Clarós, a. Z. Sobolewska, A. Doménech-Clarós-Pujol, C. Pujol and A. Clarós, "CT-based morphometric analysis of professional opera singers' vocal folds," *Journal of Voice*, vol. 33, no. 4, pp. 583.e1-583.e8, July 2019.

[12] S. S. Stevens, J. Volkmann and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185-190, 1937.

[13] S. H. Lee, H. J. Kwon, H. J. Choi, N. H. Lee and S. M. Jin, "The singer's formant and speaker's ring resonance: a long-term average spectrum analysis," *Clinical and Experimental Otorhinolaryngology*, vol. 1, no. 2, pp. 92-96, June 2008.

[14] P. Baheit, "What is overfitting in Machine Learning," *V7 Labs*, Available: https://www.v7labs.com/blog/overfitting, 1 December 2021.