Neural Network-Aided Audio Processing for Automated Vocal Coaching

Neural Network-Aided Audio Processing for Automated Vocal Coaching

Ethan Craigo

Advisors: Demba Ba and Emily Dolan

An undergraduate thesis presented jointly to the School of Engineering and Applied Sciences
and the Faculty of Arts and Sciences in partial fulfillment of a Bachelor of Arts degree
in Computer Science and Music with honors

Harvard University
Cambridge, Massachusetts
March 12, 2019

DEPARTMENT OF MUSIC
HARVARD UNIVERSITY
SPRING 2019

"In submitting this thesis to the Department of Music in partial fulfillment of the requirements for the degree with honors of Bachelor of Arts, I affirm my awareness of the standards of the Harvard College Honor Code."

Name: Ethan Craigo

Signature: Ethan Craigo

The Harvard College Honor Code

Members of the Harvard College community commit themselves to producing academic work of integrity – that is, work that adheres to the scholarly and intellectual standards of accurate attribution of sources, appropriate collection and use of data, and transparent acknowledgement of the contribution of others to their ideas, discoveries, interpretations, and conclusions. Cheating on exams or problem sets, plagiarizing or misrepresenting the ideas or language of someone else as one's own, falsifying data, or any other instance of academic dishonesty violates the standards of our community, as well as the standards of the wider world of learning and affairs.

Contents

List of Figures

List of Tables

Abstract

Research is conducted assessing the feasibility of using neural networks to detect features of the singing voice both idiomatic and unidiomatic to the Western classical singing tradition. This is done in the context of a hypothetical "automated vocal coach" capable of providing singing voice advice independently of any human agent. Data are presented focusing on four criteria of Western classical singing: phonation, laryngeal registration, resonance management, and vibrato. Literature reviews are also presented on all of these criteria to determine their physiology, acoustical properties, and strategies in vocal pedagogy. Networks are binary and ternary classifiers that apply convolution to mel-scaled spectrograms of the isolated singing voice and make judgments based on resulting features. Training results are promising overall but are severely hampered by the absence of large and contrastive datasets illustrating these criteria. Unsurprisingly, factors of singing most visible on a spectrogram are easiest for networks to distinguish. An automated vocal coach does not appear impossible to build, but its constituent networks require much more data collection to be useful in practice.

Acknowledgments

Many thanks are owed to my thesis advisors, Demba Ba of the Electrical Engineering department in the School of Engineering and Applied Sciences and Emily Dolan of the Harvard University Department of Music. Without the help and guidance of Professors Ba and Dolan, this project would not be possible. I would also like to express my gratitude to Bahareh Tolooshams, a Ph.D student in the Electrical Engineering department under the supervision of Professor Ba who provided useful assistance in helping me learn how to train neural networks. I had not taken a class on machine learning before embarking on this project, so this was quite helpful. Furthermore, I am grateful to the team of Barış Bozkurt, Ozan Baysal, and Deniz Yüret for personally supplying me with a subset of their own research data on the human voice. This gesture was instrumental in building the most successful neural network in the whole project.

The basic idea for this thesis came out of another project I completed in the spring of 2017 when I took a psychoacoustics class entitled "Why You Hear What You Hear: The Science of Music and Sound." It was taught by Eric Heller, who was assisted by Dan Tramte. I thank them both for their roles in the genesis of the thesis. Credit also goes to my excellent voice teachers over the years: Annie Leonardi-Merchant, Charles Blandy, and Frank Kelley have collectively taught me most of what I know about the singing voice in practice, and without this knowledge I would not know where to start on this project.

Thanks to my suitemates and friends for putting up with my pre-deadline flakiness and offering help whenever it was needed. Finally, thanks to my parents John and Peggy for being relentlessly supportive of all my obscure interests and pursuits as much as they could throughout my life. Without their love and devotion I would not have ended up here in the first place.

1 Introduction

1.1 Vocal pedagogy and technology

Vocal pedagogy is a difficult research field to understand fully. It is an exceedingly interdisciplinary subject, as its study incorporates aspects of biology, acoustics, music theory, and psychology. Its tenets are ill-defined and disagreed upon often; the lack of a universal certification for vocal instructors means that many contradictory opinions and philosophies flourish between different teachers, even when it comes to the same style of music [1, p. 39]. Moreover, it is a difficult subject in which to gauge success, because students' learning styles vary quite widely and one never knows whether a certain approach will connect with a certain singer. These facts, alongside a popular notion that voice students are best taught subjectively and by example, have tended to discourage active, formal research in vocal pedagogy in the past.

At the same time, vocal pedagogy is a profession and a field that holds the potential to impact the lives of millions [2, p. 404]. There are a great many people in the world seeking careers as vocalists, and more still who would like to pursue singing as a hobby. It is important that they learn how to sing properly; improper vocal technique and vocal overuse contribute to numerous vocal fold injuries, sometimes with long-term effects on health [3]. The number of singers in the world combined with the dangers of improper vocal production demand that research continue in the field of vocal pedagogy, despite its difficulties. As experts argue, it is important for teachers to be versed in this research [4, p. 40], so that they do not make demands of their students that reflect misunderstandings of physiology and run the risk of harming voices [1, pp. 39, 49].

The larger research field of voice science, of which vocal pedagogy is a subset, is widely read by medical and musical professionals alike. Its history is inextricably linked to singing: the very first device enabling doctors to look directly at the vocal folds of living subjects, the laryngoscope, was popularized by an innovative singer and voice teacher named Manuel García in 1854 [5, pp. 27-28]. Over the course of the next century and a half, a number of seminal discoveries were made that solved some of the greatest mysteries of the professional singing voice. The brilliant, ringing quality of opera singers' voices that enables them to easily be heard over orchestras unamplified was revealed to be a high-frequency peak in vocal resonance around 2.4-3.2 kHz known as the "singer's formant" [6], created by very fine adjustments of the larynx and vocal tract [7, pp. 17-18]. The muscles controlling the action of the vocal folds have been identified, and in particular much has been written about the opposing actions of the thyroarytenoid and cricothyroid muscles to control the pitch of phonations [8, 9, 7, 10]. Some researchers have even contrasted different vocal techniques and attempted to formalize the character of operatic singing as opposed to musical theatre singing [11].

In recent decades, the number of papers published in the field of voice science each year has steadily increased [12], most likely spurred on by the greater availability and convenience of sophisticated analytical tools for audio signals. While these tools have been useful to researchers for years now, they are now beginning to move from the laboratory into the voice studio and home, adapted for singers. Software like Estill Voice Training's VoicePrint [13] and VoceVista [14] helps provide useful statistics on audio recordings, while voice synthesizers like Tolvan Data's Madde application can give vocalists acoustical models to which to aspire [7, pp. 96-98]. All of the tools mentioned above require educated interpretation to serve a vocalist properly, though, save for the most basic pitch-matching trainers. In an age where we have become increasingly integrated with technology in our day-to-day lives, it might be appropriate to ask ourselves whether this will always be the case.

Let us suppose, hopefully without controversy, that one of the goals of vocal pedagogy is to ensure that all students seeking to improve their singing voices can receive effective and high-quality vocal instruction. As has been established above, research is a necessary component in moving toward this goal. However, there is a larger obstacle in reaching the goal that analytical research and professional interpretation cannot address. This is that many people around the world have no practical access to high-quality vocal instruction, whether limited by geography, finances, or both. Voice teachers, having a very specialized profession, are often not present in high quantities in rural areas. Furthermore, professional-quality vocal instruction, often required for well-informed suggestions on how to simply sing properly, can easily cost more per month than a generous family cell phone plan in the United States [15, 16, 17].[1] There is no question that vocal instructors deserve a living wage, but particularly for students only looking to pick up singing as a hobby, spending a large portion of one's annual income on vocal lessons is an untenable option if there is no money to spare in the first place. This can discourage people from training their voices.

1.2 Project scope

Given the recent activity in voice science research and analytical tools based upon it, it is natural to wonder how one might use all this new information to help make affordable, scientifically informed vocal training accessible to as many people around the world as possible. In the absence of a professional instructor, an ideal tool for amateur singers to address the problem outlined above might take the form of an "automated vocal coach." This would be a diagnostic application capable of performing detailed analysis on singing voice recordings in short amounts of time, entirely unaided by humans. This analysis would be translated into practical advice for the amateur vocalist. To give some examples with advice taken from the late Oren Brown, vocalists whose tone is judged to be "tense" or "pressed" might be asked to try taking in a bit less air when they inhale before a note [18, pp. 26, 30], and singers struggling with high notes might be asked to imitate a yawning motion as they sing to induce a sympathetic lengthening of their vocal folds [18, p. 54].

Such a diagnostic application would, of course, never be a replacement for a good singing coach. It is of the author's opinion and seems to be conventional wisdom in the field that in addition to supplying diagnostics and physical instructions, vocal coaches often fulfill a limited role as therapists to their pupils. Vocalists' instruments being parts of their bodies, their careers are heavily impacted by lifestyle choices [19, 20] and, as the life and work of Alfred Wolfsohn suggests, psychological conditions [21]. Surveys suggest that singers are largely aware of these things and have a great number of anxieties about their voices [22], and untrained singers often risk their instruments when they wrongly identify the causes of vocal fatigue and injury [20]. It is unrealistic to suspect that software could fix all of these issues and occupy this therapeutic role. Furthermore, researchers and pedagogues agree that good posture plays a large role in proper vocal production [23, 24]. Correcting students' posture for proper singing is something that a teacher can easily do, but would be incredibly difficult to achieve in a piece of software. Instead of replacing vocal instructors, this "automated vocal coach" would be a provisional tool to help anyone around the world learn the basics of how to sing, no matter their background and no matter their means. It could also be of use as a supplement to vocal instruction during rehearsal while instructors are away, or as a means for inexperienced instructors to attune their ears to vocal problems.

The fact that an automated vocal coach does not already exist on the market attests to the difficulty of building it. Products like SINGPRO [25], Vanido [26], and SING&SEE [27] are all available for purchase online, but mainly focus on video lessons, real-time vocal pitch feedback, and visualizations of and statistics on the recorded voice. At the time of this writing, there is no product available for purchase that is autonomously capable of diagnosing more sophisticated issues than intonation, and it is precisely this kind of application that is the focus of this writing. Assuming that one could easily diagnose vocal issues, it

[1] While there are few formally published studies determining the average cost of singing lessons around the United States, this originally researched claim was calculated using several sources. In 2014, a national report on average music lesson cost in different cities in the United States was commissioned by TakeLessons, an online service connecting music teachers with students across the country. The PowerPoint summary of its findings is available online [15]. It determined that the median music lesson costs in the most populous cities in the United States average out to more than $40 per hour of instruction, and that the median music lesson cost in Boston is about $60 per hour of instruction. The report also found that vocal lessons are on average the most expensive type of basic music lesson available to the public. Additionally, using the professional services website Thumbtack.com to search for voice instructors in the Cambridge, Massachusetts area on September 28th, 2018 yielded a number of price quotes for hour-long lessons with different studios, the absolute lowest of which was $56 [16]. Assuming a standard frequency of one hour-long lesson a week, this last quote works out to a monthly cost of $224 for one student.

On September 24, 2018, an article entitled "Best Cellphone Plans 2018" written by Philip Michaels and Stewart Wolpin was published on the technology review website Tom's Guide [17]. It surveyed price quotes for a four-line family phone plan with unlimited data allowance from the four most popular United States cell phone carriers: Verizon, AT&T, Sprint, and T-Mobile. Surveying the article, one finds that all four carriers offer basic pricing for family plans meeting these criteria between $140 and $160 per month. This is demonstrably less than, or in the most extreme case equal to, the median cost of vocal lessons in many locales in the United States. All sources listed in this explanation are formally cited below.

is also difficult to assess the nature of optimal vocal technique objectively and through a computer when so many legitimate disagreements on strategies and philosophies exist between instructors today. However, vocal instructors have also arrived at many broad consensuses about basic aspects of healthy singing [1, pp. 40-42]. Indeed, the basic mechanics of proper vocal production in the context of singing have been established firmly enough to be described matter-of-factly in medical textbooks intended for otolaryngologists in training [3]. This provides hope for at least a broad set of criteria that constitute proper technique.

The focus of this writing is to exploit these professional and scientific consensuses on vocal mechanics in healthy singing to explore the feasibility of building an automated vocal coach at a future date. This exploration focuses on what seem to be the most difficult aspects of constructing such a coach: determining a set of criteria for healthy and proper singing in a particular music idiom, and developing algorithms for recognizing these criteria in audio recordings of singing voices. In the hopes of solidifying these criteria to the greatest possible degree, special attention is paid to the Western classical tradition of operatic singing, the tradition on which the majority of academic research on singing technique has tended to focus in the past [12]. Research is presented below outlining four categories as particularly important in the Western classical tradition: proper phonation, laryngeal registration, resonance management, and proper application of vibrato to notes sung. Notably absent from this list is pitch accuracy or intonation, which can now be very efficiently and accurately determined with a number of real-time pitch detection algorithms that have been developed in decades past and improved upon in recent years [28, 29].

It is challenging to imagine how computers, generally programmed to follow rigidly specified algorithms to meet certain goals, could be able to recognize these criteria in audio recordings that humans have no capability of fully theoretically specifying. For example, while most vocal professionals can recognize what a properly resonant voice sounds like, they cannot name a specific frequency at which a voice's overtones ought to "peak." Most other decision boundaries for what constitutes proper singing, aside from pitch accuracy, are similar in their complexity.

Fortunately, the extremely active field of machine learning in computer science offers us a way around this problem by thinking about it from a human perspective. Well-informed vocal professionals have learned to distinguish ideal techniques from non-ideal ones through a combination of their own rote learning and numerous examples of people they have heard sing throughout their lives. Machine learning algorithms are algorithms with roots in pattern recognition problems, and which employ the use of neural networks. Pursuant to a goal, they oftentimes are employed to make judgments on inputs subject to analyses and adjustable parameters. These algorithms "learn" through many examples, adjusting parameters to refine the accuracy of their classifications according to their goals [30]. A subfield of machine learning, deep learning, has produced algorithms capable of great success in many wildly different applications, such as automatic image recognition, machine-aided language translation, and game-playing [31]. More relevantly, deep learning algorithms are now also being used successfully in audio processing applications, such as digital voice synthesis and the separation of voices from noise in monaural vocal recordings [32, 33]. If all this is possible, it appears a reasonable goal to train neural networks to develop good representations of proper Western classical singing technique in several different dimensions.

Having established criteria for proper Western classical singing and discussed their acoustics, physiology, and pedagogical applications in depth, attempts to train machine learning algorithms to recognize each of these criteria using recordings of amateur and professional singers alike are presented below. Details on neural network architectures as well as training, validation, and test data are detailed for each of these attempts, as well as the ultimate accuracy and testing of the algorithms built through the usage of these networks. Following the presentation of this data, analyses and discussion of the results are presented. Possible impacts, limitations, and future directions of the project are given afterward, with a mind toward building a hypothetical future product. While it will undoubtedly be very difficult to build an automated vocal coach in the near future, this work is an important first step toward this final goal. With more work, this project may help to spread high-quality vocal instruction around the world.

2 Background

This section contains information intended to orient readers in the subjects of Western classical singing and machine learning, or at least the aspects of these subjects that are related to this project. For an introduction to acoustic terminology utilized here, readers are advised to consult Heller (2013) [34].

2.1 Western classical singing

There are a number of disputes in scholarly discourse over what precisely is meant by the terms "Western classical singing," "operatic singing," and "*bel canto*." A full discussion of them would take thousands upon thousands of words. For the sake of convenience and consistency throughout this project, it is best that only theories on the voice that are supported by the existent scientific literature be discussed at length. Therefore, the term "Western classical singing" will, for the rest of this writing, be constrained to the "scientific" school of Western operatic singing arguably founded by Manuel García in the mid-nineteenth century and continued by otolaryngologists and vocal coaches alike in the following years [35, 5]. This school has its critics. The famed writer George Bernard Shaw was known to have argued that "[y]ou can no more sing on physiological principles than you can fence on anatomical principles, paint on optical principles, or compose on acoustical principles" [5]. While a certain degree of singing remains shrouded in artistry and interpretation, perhaps all for the better, it is fortunate for this project that with regard to the basic mechanics of the human voice, Shaw has decidedly and empirically been proven wrong. The scientific method and its applications to the voice will support all further research in this paper, as theories unfounded by empirical evidence are too unsubstantiated to be the base for a piece of software claiming any kind of authority.

2.2 Machine learning for audio signals

2.2.1 Machine learning and artificial neural networks

The term "machine learning" was first coined by Arthur L. Samuel in his 1959 paper entitled "Some Studies in Machine Learning Using the Game of Checkers." In the paper, Samuel defines the term "machine learning" as the phenomenon of digital computers behaving in a way we might term "learning" if we observed it in animals [36]. This definition, although vague and metaphorical, is a useful intuitive concept in understanding machine learning as it is defined today. Machine learning is a subfield of artificial intelligence, meaning that it is focused on programs or agents that receive input from their environments and take various actions based upon this input [37, p. viii]. Artificially intelligent programs are said to *learn* if their performances upon specific tasks are evaluated repeatedly and improve with greater amounts of environmental observations [37, p. 693]. In this process it is implied that programs learn to parse meaningful patterns from the data supplied to them, and act according to the patterns they identify; in this manner pattern recognition and machine learning are said to be different ways of describing the same phenomenon [30, p. vii]. As they separate signal from noise in input and recognize patterns, machine learning algorithms act in a way that directly imitates human learning and characterizes the function of the human brain [38].

In learning information about its environment through data supplied to it, a machine learning agent must be careful not to *overfit* to its data. Overfitting, a term borrowed from statistics, occurs when an agent memorizes so many features of its input data that it cannot accurately generalize the important features of the input data to its environment and does not learn the task at hand [39, 37]. An agent that overfits to its input data can recognize the previous input data very well, but this is not the goal of machine learning and instead amounts to machine memorization. We may term the opposite phenomenon *underfitting*, in which an agent does not learn enough information about its input data to perform well in its environment. Neither overfitting nor underfitting are desirable for machine learning agents; rather, a balance between the two extremes is sought.

A particularly active branch of machine learning in research today focuses on constructs that even more closely imitate the inner workings of the human brain: artificial neural networks. Neural networks are simply defined as groups of neurons, the basic functional unit in the animal brain, connected together in some way [40]. Artificial neural networks most often occur in the form of digital representations of these biological neural networks. Modern conceptions of neural networks ultimately stem from a 1943 paper published by William McCulloch and Walter Pitts, in which the phenomenon of sympathetic neuron stimulation is distilled into a formal logic upon which mathematical computation may be modeled [41]. Early applications of artificial neural networks focused heavily upon their capacity to model and uncover the mysteries of biological neural networks [30, p. 226]. Within a decade, researchers began to investigate the plausibility of using these sorts of networks for computation and pattern recognition [42, 43]. These early explorations were achieved with the help of the Hebbian theory of neuroplasticity, which postulates a means for neural conditioning: synapses or connections between neurons are strengthened through the constituent neurons' co-activity over time [44, p. 62]. Efforts to apply these early network experiments to practical pattern recognition problems stalled for several decades with computational and algorithmic limitations, but advances in parallel computing and the invention of the backpropagation method by Paul Werbos [45] have since made artificial neural networks a popular topic for the last few decades.

2.2.2 The mechanics of artificial neural networks

It is far beyond the scope of this project to explain all of the mathematical equations and computational concepts involved in the machine learning topics utilized within it. However, an endeavor to explain the basic mechanism of learning with artificial neural networks is probably warranted. Here we will paraphrase Bishop (2006) and Russell and Norvig (2010), two excellent textbooks on the subject that can be consulted for more detail [30, 37]. The most basic unit of an artificial neural network, as it is most commonly implemented today, is the *artificial neuron.* An artificial neuron, often simply called a unit in texts on artificial neural networks, is a digital parallel to the biological neuron. In artificial neural networks, artificial neurons are connected to each other with links in the fashion of edges on a directed graph. Just as in a directed graph, each link or edge carries a weight signifying the type and strength of a connection between two neurons [37, p. 728].

Taking and slightly modifying the terminology in an example from Bishop, suppose we have a simple network with input values x_1 through x_D representing encoded input data, output values y_1 through y_K representing output data, and in between these a "layer" or collection of units z_1 through z_M [30, p. 228]. The units z_i involved in neither input nor output, invisible to an external user, are called *hidden units*, and a layer of hidden units is called a *hidden layer.* Here, from a functional perspective, our goal is to eventually train our hidden units to isolate important features of our input, make some judgment based upon these features, and output data related to this judgment. In this *feed-forward network*, termed as such because information progresses unilaterally forward from the inputs x_i to hidden units z_i to outputs y_i [37, p. 729], each z_j has a link from each input value x_i weighted according to a distinct parameter $w_{i,j}^{(1)}$, and each output y_k has a link from each hidden unit z_j weighted according to a parameter $w_{j,k}^{(2)}$. We may also set up *biases* that occur in the form $w_{0,j}^{(1)}$ and $w_{0,k}^{(2)}$, and respectively assign overall weights to the importance of hidden unit z_j and output value y_k in our network.

Each hidden unit z_i takes in input values according to its weights and then creates a weighted sum of its inputs, applying a certain activation function g to this linear combination a_i to determine its output:

$$z_i = g(a_i) = g\left(\sum_{j=1}^{D} w_{j,i}^{(1)} x_j + w_{0,i}^{(1)}\right) \tag{1}$$

Following this, in a similar fashion, each output unit takes in the output from each of the hidden units according to its own input weights and sums these numbers, applying an output activation function h to the resulting linear combination b_i to determine its own value:

$$y_i = h(b_i) = h\left(\sum_{j=1}^{M} w_{j,i}^{(2)} z_j + w_{0,i}^{(2)}\right) \tag{2}$$

We can thus describe the overall output of our neural network as being dependent entirely on the inputs $\mathbf{x}$ it is being supplied and the weights $\mathbf{w}$ it has been assigned:

$$y_m(\mathbf{x}, \mathbf{w}) = h\left(\sum_{j=1}^{M} w_{j,m}^{(2)} z_j + w_{0,m}^{(2)}\right) = h\left(\sum_{j=1}^{M} w_{j,m}^{(2)} g\left(\sum_{k=1}^{D} w_{k,j}^{(1)} x_k + w_{0,j}^{(1)}\right) + w_{0,m}^{(2)}\right) \tag{3}$$

The precise nature of the activation functions employed in artificial neural network design tend to vary with the goals one attempts to accomplish using networks. They are generally, however, selected to be non-linear functions. Because there are two sets of weights in this model, one from input values to hidden units and another from hidden units to output values, this network is said to have two layers. All of the networks discussed in this project have many more than two layers and some other specialized layer types, but the general principle behind each is analogous to the one just described. The above process from the perspective of a hidden unit is summarized in Figure 1. The circle signifies a hidden unit, and the arrows signify connections between units, inputs, and outputs.

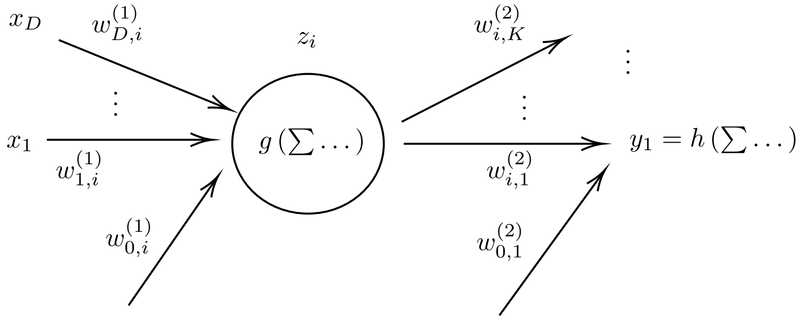


Figure 1: Structure of a hidden unit in an artificial neural network[2]

2.2.3 Learning with neural networks

We now have an understanding of how input is processed in neural networks, but we do not yet have a mechanism for how it learns to recognize patterns in its environment. Abstracting away from the specific inputs given to it, whether or not an artificial neural network performs well on its designated task is entirely based upon its system of weights and biases on connections between units. If its weights and biases are structured properly, a network may perform extremely well. If they are not structured properly, the network will not perform well. Therefore, in order to have a network learn from the data supplied to it over time, we must develop a way of evaluating its performance based upon changes in each of its weights and biases, and adjusting them in a manner that will improve the network's performance.

The key to doing this correctly lies in another way of thinking about neural networks. Neural networks are intended to make different judgments or decisions based on the different inputs supplied to them, with

[2] In this diagram, the circle at the center signifies a hidden unit within our simple neural network whose output is z_i. The arrows leading into this hidden unit are from the initial input layer, with weights marked, as well as the bias applied to this specific unit. Arrows leading away from the hidden unit dignify connections to the output layer with weights similarly marked, as well as biases in the output units. Abbreviated computations within the hidden unit and the output units to which it leads are seen as well.

decision boundaries separating the inputs upon which one judgment is made from the other inputs on which a different judgment is made. This can be conceptualized as a mathematical function mapping inputs to judgments or other outputs based upon those judgments. These functions and decision boundaries might be very complex and difficult to describe, like finding out at which precise point one differentiates a handwritten number "1" from "7." Machine learning with artificial neural networks is a way of approximating the judgments or outputs of these functions with the outputs of our networks, which are complex functions composed of many simpler nonlinear functions. This means that learning with these networks is typically a matter of *nonlinear regression* [37, p. 732], a mathematical topic on which much has been written.

To perform nonlinear regression, we need an *error* or *loss* function E that quantifies the difference between the expected values of the real-world function we are approximating and the values yielded by our network. While the precise specification of E varies with the problem we attempt to solve, it is often calculated according to a specific *training set* of data against which a network's performance is measured. We are always seeking to find a minimum value of $E(\mathbf{w})$ based on our network's weights, but as E is generally quite complex and highly dimensional, we usually settle for finding a local rather than global minimum [30, p. 237]. This is a point at which the *gradient* of E, ∇E, is equal to 0. In the popular method of nonlinear regression known as *gradient descent*, one repeatedly changes weights to take steps in the precise opposite direction of ∇E, after which we know E will be further reduced.

One of the greatest challenges of machine learning throughout the past century was finding a way to efficiently calculate ∇E in a network of arbitrary size, as it is difficult to intuitively know how we ought to structure the weights leading to and from our hidden units [37, p. 733]. Fortunately, the advent of *backpropagation* algorithms allows us to calculate each constituent partial derivative in ∇E quite easily using the chain rule. Let us see an example using our network from above and the function E we just defined:

$$\frac{\partial E}{\partial w_{i,j}^{(1)}} = \frac{\partial E}{\partial a_j} \frac{\partial a_j}{\partial w_{i,j}^{(1)}} \tag{4}$$

The second partial derivative on the right-hand side is easy to compute; looking at the equations above it is clearly equal to x_j. It is the first partial derivative on the right-hand side that is more troublesome. Computing the partial derivative of E with respect to an linear combination for an output b_k is generally easy but depends on the precise formulation of E, and the second partial derivative on the right-hand is also easy to calculate. However, a_j is not an output value. In this case, we first must recursively compute the value of $\frac{\partial E}{\partial b_k}$ for each output combination b_k for which $w_{j,k}^{(2)} > 0$ and then exploit what Bishop (2006) calls the "backpropagation formula" to finally obtain our value [30, pp. 242-244]:

$$\frac{\partial E}{\partial a_j} = \sum_k \frac{\partial E}{\partial b_k} \frac{\partial b_k}{\partial a_j} = g'(a_j) \sum_k \frac{\partial E}{\partial b_k} w_{j,k}^{(2)} \tag{5}$$

In this way, partial derivatives of the error function are calculated first in the output layer of the neural network and then are propagated backwards toward the inputs until all partial derivatives have been calculated. While specific learning procedures will vary among networks, this gives a good general idea of how most networks learn today. Gradient descent using backpropagation, or some version of it, is used to optimize the weights in a network as more data is fed to it over time from the training set. To test the ability of a neural network to perform well on a broader sample of data, oftentimes networks' performance will be measured on a *validation set* of data on which the network has not been trained, and which has been selected for this purpose [37, p. 709]. Based on the output of a network on a validation set, one adjusts its *hyperparameters*. Hyperparameters are parameters of a network that themselves determine the structure of training parameters, such as the network's specific structure or the number of *epochs* or full passes through the training set data over which the network is trained [30, p. 30].

Multiple ways exist to combat overfitting in neural networks, some taking the form of training techniques and others taking the form of specialized layers. One approach, called *early stopping*, is simply the process of stopping the training of neural networks before they are able to memorize too many features and inevitably overfit [30, pp. 261-263]. Stopping learning too early leads to underfitting, but nearly every researcher practicing with neural networks learns to stop training them at some useful point before their validation set performance accuracy decreases. Another approach in building neural networks simply consists of creating networks too simple to memorize all the features of one's data needed for overfitting. Again, this is often useful, but if not done properly this can lead to underfitting.

Two specialized layer-based approaches to combat overfitting employed in the networks in this project are called *dropout* and *batch normalization*. Dropout, first demonstrated in a paper by Hinton et al. in 2012, is often implemented in neural network frameworks as a layer added to a sequential model, but consists of iteratively and randomly nullifying the connections between two successive layers in a network with a specific probability [46]. Batch normalization is a newer technique invented by Ioffe and Szegedy in 2015 that adjusts or normalizes the weights of input connections to a layer, ensuring that all gradients in optimization are smooth and that networks are less likely to become stuck in useless local minima over their loss functions [47, 48]. It, too, is often implemented in machine learning libraries and frameworks as a specialized layer type.

2.2.4 Network architectures for auditory data

There are a great many popular variations to the picture of a generic neural network presented above, each suited to a different task. When researchers construct neural networks programmatically, it is quite inefficient to build one artificial neuron at a time. To build neural networks, most people tend to use frameworks and interfaces such as TensorFlow [49], Torch [50], or Keras [51] optimized for different programming languages. These frameworks tend to operate in terms of pre-built and customizable layers of multiple neurons. Some layers are termed *dense* or *fully connected*, meaning they contain connections from each neuron in the previous layer and to each neuron in the next layer, all with their own weight parameters [31]. Others are recurrent, meaning they may additionally feed back into themselves, and still others add noise to their outputs and drop some of their connections with certain probabilities [31].

Many ways of classifying networks exist, but one particularly important difference is between networks that perform *supervised* and *unsupervised learning*. In supervised learning, a network is supplied with training data inputs labeled with the proper values that should be outputted from them, and is said to be "taught" by the labels; in unsupervised learning no such teachers or labels exist and it is the network's goal to discover some implicit structure in the input data [30, p. 3]. We will be performing only supervised learning here.

In the context of this project, we are interested only in networks that are able to process audio inputs in some format and output judgments upon whether or not the audio supplied to them is of a person singing idiomatically in some aspect of the Western classical tradition. Artificial neural networks that process sensory, as opposed to purely numerical, input have been extant for decades, but most of them have focused on visual stimuli as opposed to audio [52, 53].[3] Convolutional neural networks, which are types of networks mimicking the activity that occurs in animal brains' visual cortexes [31], have been dominant in many key advances in computer vision and image classification in recent years [54, 55, 56, 57]. While several types of networks enjoy active research in computer audition research, including variants on convolutional neural networks [58], autoencoders [59], and recurrent neural networks [60], we will focus only on convolutional neural networks in this project's research. This is because convolutional neural networks have been shown to perform well in a variety of audio classification tasks [61, 62, 58].

[3] This claim is substantiated through searches of the scientific literature available online through two different databases on March 6, 2019, obtaining the number of papers available online focusing on computer vision compared to the number focusing on computer audition. The databases consulted were Google Scholar [52] and Harvard's HOLLIS online library search tool [53]. A database query for "neural network image" yielded 2.61 million results on Google Scholar and 369,000 results on HOLLIS, while a query for "neural network sound" yielded only 1.26 million results on Google Scholar and 123,757 results on HOLLIS. More starkly, a query for the specific term "computer vision" yielded 1.94 million results on Google Scholar and 371,571 results on HOLLIS, while a query for the term "computer audition" only returned 436 results on Google Scholar and 117 results on HOLLIS.

2.2.5 Convolutional neural networks

As stated above, *convolutional neural networks* are inspired by research on the animal visual cortex. In the 1950s and 1960s, David Hubel and Torsten Wiesel published seminal research revealing the structures of cat and monkey visual cortexes [63, 64]. A key finding relevant to the field of machine learning was their discovery that certain neuron columns in the visual cortex map to specific areas in an animal's field of vision that they termed "receptive fields" [63], which are differently sized between different cell groups and often overlap with each other [64]. These columns are excited most by stimuli that exhibit particular patterns of visual characteristics [63]. Some types of cells in the visual cortex, termed "simple," merely fire based on the precise positioning of a stimulus, while other "complex" cells connected to the simple cells detect patterns independent of precise positioning within a receptive field [64].

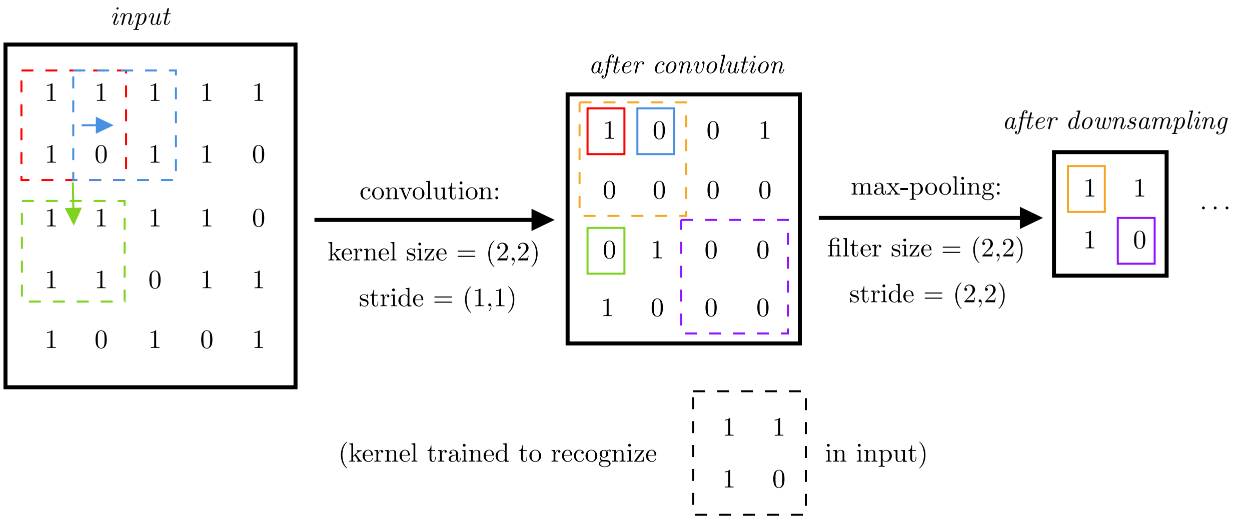
This efficient organization inspired researcher Kunihiko Fukushima to publish a paper in 1980 on a neural network inspired by Hubel and Wiesel's "hierarchy model" of cells recognizing simpler patterns or features repeatedly feeding information into cells recognizing more complex features. In the paper, Fukushima calls his neural network the "neocognitron." It consists of multiple layerings of two types of units: the *S-unit* and the *C-unit* organized into multiple *S-planes* and *C-planes* in *S-layers* and *C-layers* respectively [65].

In this system, all S-units in a single S-plane share identically weighted input connections, although they are displaced from each other by position. Therefore, they are trained to recognize the same pattern at different locations in their input. Multiple S-units from a specific receptive field in an S-plane then feed connections into a single C-unit in the corresponding C-plane in the next C-layer, which is structured to respond uniformly if there is a sufficient output from any S-unit in its field. Following this, the data from each C-plane in this C-layer is connected to each S-unit in a new S-layer, which recognizes higher-level patterns in its own S-planes [65].

Importantly, connections from input data and C-layers to S-layers are modifiable, and one trains the neocognitron by modifying these connections' weights. Connections from S-layers to C-layers are evenly distributed and unmodifiable, so as to ensure that pattern recognition occurs independent of position [65]. To summarize the function of these two types of layers, S-layers recognize multiple patterns in each of their S-planes, and C-layers recognize each of the patterns in the preceding S-layer independent of position while simplifying the information to be fed into the next S-layer.[4]

Fukushima's structure is essentially the core of the structure of a modern convolutional neural network, although modern networks use much updated terminology. In modern parlance, the weights that define the operation of an S-plane on input data constitute a *kernel* that is trained to recognize a specific feature and applied in overlapping positions across the input space in a process conventionally called *convolution*, hence the name of the network type [66]. Thus Fukushima's S-layers are now termed *convolutional layers*. C-layers eliminate position-dependent information about the patterns identified in the S-layers leading to them, reducing the amount of data fed into the following layer. Because of this behavior they are called *downsampling* or pooling layers [66], depending on the specific algorithms they employ. The processes of convolution and downsampling are visualized in Figure 2. A major method of training convolutional neural networks is, as in other neural network types, backpropagation, with Yann LeCun et al. (1989) first demonstrating this method on a handwritten digit classifier [67].

[4] To accurately understand these concepts, it helps to visualize them with images. The figures in the Fukushima paper are helpful with this [65], as is Figure 2.



input

1	1	1	1	1
1	0	1	1	0
1	1	1	1	0
1	1	0	1	1
1	0	1	0	1

after convolution

1	0	0	1
0	0	0	0
0	1	0	0
1	0	0	0

after downsampling

1	1
1	0

(kernel trained to recognize

1	1
1	0

in input)

Figure 2: Mechanics of a simple convolutional neural network[5]

Unfortunately, there is occasionally some confusion between the terms "filter" and "kernel" in articles about convolutional neural networks: some researchers use the terms interchangeably, while the two words mean different things to others. In the interest of disambiguation, throughout this project the word "kernel" will be used to describe a set of weights in a convolutional layer learning a specific feature, while "filter" will refer to the specific type of downsampling accomplished in a downsampling layer. Accordingly, "kernel size" will refer to the receptive field size of a unit in a convolutional layer, while "filter size" will refer to the receptive field size of a unit in a downsampling layer. The degree to which receptive fields overlap in convolutional and downsampling layers' units is determined by a hyperparameter called *stride*, which summarizes the horizontal and vertical distance between adjacent units' receptive fields. Oftentimes convolutional layers will have a stride of (1, 1), meaning that their kernels' receptive fields move horizontally and vertically by just 1 unit between neighboring neurons. Downsampling layers often have strides equal to their filter sizes, to ensure that they are efficiently degrading information. To avoid the resolution loss caused by kernel sizes, convolutional layers will sometimes artificially "pad" their inputs with empty information in a process known as *zero padding* before performing a convolution operation; this is often used in networks containing many layers [66]. Following conventional terminology used in the Keras machine learning framework, convolutional layers used in this project employing zero padding will be referred to as "padded," while convolutional layers without zero padding will be labeled "valid."

The modern architecture of a convolutional neural network consists of a number of layers sequentially stacked upon one another. Input data is fed into a series of alternating convolutional and downsampling layers, in order to learn increasingly high-level features about the data. An activation function is applied to the output of each convolutional layer, as with other layer types in neural networks. The resulting downsampled data is then passed through several traditional fully connected layers to make judgments about features recognized

[5] In this diagram the first solid black square represents the units in the input layer of a very simple convolutional neural network. Here the convolutional layer that follows it consists of a single kernel whose weights have been trained to recognize the feature shown in the dashed square below the three layer outputs. Each unit in the convolutional layer outputs a 1 if it recognizes its trained feature in its receptive field and a 0 otherwise. The size of each convolutional unit's receptive field, or the kernel size, is two-by-two, while the horizontal and vertical distance between adjacent convolutional units' receptive fields, or stride, is one-by-one. The second black square represents the outputs of each unit in the convolutional layer. One can clearly see the effects of applying the kernel to the input layer by looking at some convolutional outputs' corresponding inputs in the dashed squares within the input layer. By the arrows within the input layer, one can also see how kernels are often visualized to "slide" across input fields vertically and horizontally, looking for the same feature at each step.

The convolutional outputs are then fed into units within a downsampling layer. The weights in the downsampling layer are fixed and implement the max-pooling downsampling algorithm, outputting the maximum of all the elements in their receptive fields. Here the receptive field size or filter size is also two-by-two, but the stride is also two-by-two, so the inputs to units in this layer do not overlap. Similarly to the above one can see the effects of the downsampling outputs by looking at some downsampling outputs' corresponding inputs in the dashed squares within the convolutional layer.

and render the data meaningful to humans before finally being outputted [66]. The downsampling of data enables very high-level features to be learned about images and other input data without the time and memory costs needed for a basic fully-connected network to perform the same work [54, 55]. Given time and computational power constraints, this makes convolutional neural networks a natural fit for this project. Now having a practical understanding of machine learning as it relates to this project, we may proceed ahead to the project's implementation.

3 Methodology and literature review

This section contains information on the organization and structure of this project and its constituent neural network experiments. Firstly, a description of the criteria of idiomatic Western classical singing chosen for this project, as well as a justification for the inclusion of each criterion, is offered. Outlines for simple classification algorithms are presented for each criterion, for experimentation in this project. Following this, descriptions of training data gathered for the networks are given, as well as specifications of transformations performed on datasets and the general structures of the networks operating on the data. Finally, an overall procedure for the study of each of these criteria is outlined.

3.1 Criteria research

3.1.1 Sources

It is an inevitably imperfect exercise to try to distill the central pillars of Western classical singing into only a few different categories. As is mentioned in the section above, an inarguable definition of the term is difficult to reach in the first place due to disagreements on which locations and time ranges to include in one's survey of the genre. This thesis, as has been established, focuses upon the newer school of Western classical singing that unifies some of the tenets and aesthetic preferences of the past with the scientific and anatomical knowledge of more recent years. Here we are concerned primarily with the improvement of individual voices in a solo setting, and it is no accident that this choice makes the majority of voice science research in the past century readily available for us to use [12].

In selecting the criteria characterizing idiomatic Western classical singing for neural networks to learn, it was important to consult literature by respected figures in the field of vocal pedagogy, ranging over multiple eras. This literature needed to be general and instructional in nature, properly weighing the importance of different aspects and techniques of Western classical singing against each other in a teaching context. Three respected texts, published by different authors in three different centuries, were selected as sources for finding these criteria.

The first of these is a work by the great teacher Manuel García, whose foundational work unifying vocal pedagogy and anatomy began in the 1830s and reached into the first years of the twentieth century [5]. García himself was surrounded by singing for the entire duration of his life: while he abandoned his own performing career before reaching the age of thirty, his father Manuel García the Senior originated the Count Almaviva role in Rossini's famous *Il barbiere di Siviglia*. His sisters were the famous Maria Malibran and Pauline Viardot. Over the course of his life, García became well-known for the fame of his pupils, including the "Swedish Nightingale" Jenny Lind and Mathilde Marchesi. He held esteemed positions at the Paris Conservatory and Royal Academy of Music in London for many years [5].

García pioneered the procedure of laryngoscopy to test his own theories about the production of the voice [5]. Such was the rigor of his studies of the larynx that he received an honorary M.D. from the University of Königsberg in 1862 [35]. His *magnum opus*, in which he details his philosophy on how to sing from basic physical actions to sophisticated textual interpretation, went through several different editions and translations throughout the mid-nineteenth century. This thesis consults an edition published in Boston in the 1870s entitled *Garcia's New Treatise on the Art of Singing* [68]. The portions of this work relevant to this project focus heavily on descriptions of the physical production of singing. The work is especially useful because it is steeped in the vocabulary in use at the time that many great works defining the Western classical tradition of voice were being composed and performed.

The second work consulted is the book *Discover Your Voice*, written by Oren Brown and published in 1996 [18]. Brown, another long-lived baritone whose teaching career spanned many decades, honed his teaching skills working as a voice therapy lecturer and voice therapist in his own right at the Washington University School of Medicine in St. Louis [69]. Working alongside a number of otolaryngologists as a professional from a very different background, he gained an intimate understanding of each muscle involved in vocal production, both within and external to the larynx. He learned to treat speaking voices as well as singing voices [18,

p. ix], and developed a reputation for prolonging the professional careers of singers in multiple styles in the decades that followed [69]. He served on the voice faculty at the Juilliard School for nineteen years. Among his most famous students were the *heldentenor* James King, jazz singer Buddy Greco, and baritone Bo Skovhus. *Discover Your Voice*, a respected text in its field written at the end of Brown's life, emphasizes healthy and effortless voice production. Containing material on which beginners and professionals alike can work, it also has tips for other voice instructors and choral conductors. It functions as a more fleshed-out, modern version of García's text, with a better understanding of the science behind the voice.

Finally, the third book surveyed in the selection of these criteria is *Practical Vocal Acoustics*, published in 2013 [7]. It was written by Kenneth Bozeman, the current Frank C. Shattuck Professor of Music at Lawrence University. Bozeman is an experienced pedagogue and tenor who is extremely active in the field of voice science, chairing the editorial board of the *Journal of Singing* in which many of the articles cited in this thesis were published. His students have gone on to sing with many esteemed opera companies around the world [70]. Bozeman's research often works with the voice from an analytical, acoustic perspective, and focuses less on the anatomical origins of the voice. *Practical Vocal Acoustics* is no exception to this, and is written in a way that makes it quite clear which techniques are and are not idiomatic in Western classical singing. Bozeman uses terminology that is particularly useful for analyzing and classifying audio recordings of the human voice.

3.1.2 Selections

Studying the above three sources closely, a list of four essential aspects of idiomatic Western classical singing was produced. Criteria were determined selectively and purposefully at a rather low level, although a project with more time and resources might include more high-level criteria such as proper agility and phrasing. These criteria were selected to be at an intermediate difficulty level for classification, with decision boundaries complex enough that a single deterministic algorithm might have trouble classifying them properly but not difficult enough that no clear pathway to classification existed with current technologies.

The criteria selected for this project, as mentioned in the introduction, are as listed: proper phonation, laryngeal registration, resonance management, and proper application of vibrato to notes sung. These vary in difficulty of execution and also vary in difficulty of classification from an external perspective. Each criterion selected is discussed below, with a short description of how a simplistic classifier might work on it.

3.1.3 Proper phonation

It is a fact agreed upon in medical and voice research literature, as well as all three main sources consulted for finding these criteria, that all manifestations of the human voice are driven by the breath [71, 72, 68, 18, 7]. The vast majority of human communication is driven by exhalation from the lungs as regulated by the diaphragm and intercostal and abdominal muscles. In voiced phonation, the column of air expelled from the lungs passes through the partially adducted vocal folds in the larynx. The opening between the vocal folds is known as the *glottis*. The folds vibrate at a certain frequency depending upon both the *subglottal pressure* of air being driven through them and their size and tension at the time of phonation [72]. Being able to control this pressure over time, as well as the flow of air across the vocal folds, is a phenomenon known to voice teachers and students as "breath support" [71] or "breath management" [18, pp. 34-35]. This is very important in singing, being fundamental to its basic function.

Managing the regularity of subglottal pressure and glottal airflow throughout a passage through this phenomenon of "breath support" seems a difficult aspect of singing for neural networks to be able to classify well. This is because gauging breath support over time requires a network to take long samples of audio into account at once, meaning that such a network would need a great deal of training material to perform well and would also be quite computationally intensive. Additionally, the character of sound in irregularly supported singing is not immediately distinguished from that of uniformly well-supported singing with any widely available audio signal transformations. Perhaps this problem can be solved by focusing on breath support in just Western classical singing.

Manuel García writes that one of the common sources of "unsteadiness" during singing, a phenomenon which he views negatively, is irregular motion of the glottis [68, p. 33]. He advocates for the forced contraction of the glottis to keep it somewhat steady against the air flowing through it to avoid a "veiled" or "extremely dull" sound, evident when too much air escapes [68, p. 7]. This would seem to correlate with the "weakness and tremulousness" [68, p. 7] that García notes as a result of the unsteady glottis and low subglottal pressure. Conversely, Oren Brown writes on the opposition between the compression upon the lungs and the pressure against the glottis and cautions against a "strained expansion" of the lungs in singing for proper tone production [18, pp. 30, 34], which would naturally create a subglottal pressure in excess of that required for the desired tone.

From these cautions against imbalances of opposing pulmonary and subglottal pressures, we can see that there is a specific character to the subglottal pressure and airflow used in skilled Western classical singing, as distinguished from other genres of singing. Kenneth Bozeman, whose background is in this style of singing, makes this very clear in his writings. He creates a "phonation equation" relating breath (thoracic) pressure and airflow to glottal resistance, and harmonically distinguishes between three phonation modes: pressed, breathy, and flow. In pressed phonation the glottal resistance is too high and the resulting timbre is "metallic," in breathy phonation the resistance is too low and the timbre is "fluty" and "airy," and in flow phonation the glottal resistance is at an acceptable medium level and the timbre is "balanced" [7, pp. 5-6]. This is a common classification system derived from Johan Sundberg in his 1987 book *The Science of the Singing Voice*, with the addition of a neutral mode in which both subglottal pressure and airflow across the glottis are low [73].

In publishing a dataset on computationally detecting phonation modes, Rhodes and Crawford (2012) affirm that baritones in the Western classical tradition are taught to remain in the flow phonation mode whenever they sing. [74]. They also characterize breathy phonation as characteristic of some jazz and pop records, and pressed phonation as characterized by James Brown's record "I Feel Good" [74]. From all this research, we can gather that the Western classical tradition clearly values a proportional balance between the airflow rate across the glottis and the subglottal pressure. Breathy and pressed phonation modes appear to be unidiomatic, while neutral and flow phonation modes appear to be acceptable.

Given that the most obvious way to distinguish between phonation modes is the one described by Bozeman, a suitable goal for a phonation classifier in Western classical singing for amateur vocalists might be to distinguish between his precise modes of phonation. Singers' phonation ought to constantly be balanced in this style, so in order to yield useful advice in the context of an automated vocal coach, this phonation classifier would have to distinguish between airy phonation, pressed phonation, and flow or neutral phonation. This is the goal for the classifier built to satisfy this criterion.

3.1.4 Laryngeal registration

All three texts surveyed for these five criteria also discuss an idea of "registration" across one's vocal range [68, 18, 7]. Registration is an idea that all three authors seem to acknowledge as somewhat nebulous. There is widespread agreement in the field of voice science that the concept of registers of the voice is one that is highly contentious and often confusing, due to widespread disparities in precise classifications and names and a lack of understanding of the physiology behind them [75, 76]. The difficulty in understanding vocal registers lies chiefly in the fact that they originate from and are physically confined to the vibration of the vocal folds during phonation. Since vocal folds are very small pieces of tissue that generally do not exceed 3 centimeters in length, and their vibratory activity happens at frequencies too high for a normal camera [72], it is difficult to study their precise modes of vibration.

The generally accepted definition of a register in the human voice was first put forward by Manuel García in the very text studied closely in this paper, and affirmed and refined by voice researchers in later years. A register is a zone or series of frequencies emitted by the human voice having a similarity to each other in sound and produced by a similar physical mechanism [68, 75, 77, 76]. Registers are also, perhaps confusingly, called *voices* in the context of vocal pedagogy, reflecting the perception that these regions in vocal production are so different from each other that they may as well be different voices. Some researchers claim that registers often have either no or very little overlap in terms of the frequencies that belong to them [75]. However, many vocal pedagogues allow for overlaps between registers that enable different modes of singing on various pitches, and spend a great deal of time discussing ways that singers can “mix” or “blend” adjacent registers so that the characters of their voices are somehow in between the timbres produced by each constituent register in the “mix” [68, 78]. Barring certain medical conditions affecting the larynx, everyone has a similar if not identical set of vocal registers, but these registers occur in different bands of frequencies in different people [75, 76]. Singing in different registers can cause different areas of the body to sympathetically resonate with the frequencies produced, although this does not mean that the frequencies are produced in these areas [68, p. 5].

There are at least 107 names that have been used by vocal pedagogues and voice scientists to refer to different registers [75]. There is even widespread disagreement on the number of registers affirmed to exist in the scientific literature, although papers reviewed by the author have tended to refer to between three and five registers in their discourse. Some papers tend to only define registers located in frequency ranges that are often used in singing, while others define registers across the entire range of phonation. None of the three sources primarily consulted in determining criteria for Western classical singing agreed on terminology for registration.

García, who sometimes uses the word “voice” to refer to registers, defines three registers in his text and alludes to a fourth register in another location. From lowest to highest frequency bands, these are the chest voice, the *falsetto* register, and the head register or head voice, with the so-called *voix-mixte* or mixed voice located between chest and *falsetto* registers [68, pp. 6, 7]. The chest and head voices are termed as such because of the areas of the body in which one feels resonance when one sings in them. He claims that the falsetto and head registers are located in the same positions for male and female singers [68, p. 6], a claim at odds with most successive literature on the subject [75, 18, 78]. According to him, proper registration on various frequencies differs widely between voice types. While basses and baritones ought to stay almost entirely within their chest voices, tenors benefit from “blending” some falsetto tone into their upper notes [68, p.9]. Alto voices ought to use chest and falsetto registers, while mezzo-sopranos should remain mostly in the falsetto register and sopranos in the falsetto and head registers [68, pp. 8-9].

Brown’s treatment of registers differs wildly from García’s. Citing work in the 1980s done by Harry Hollein and his associates, he prefers to dispose with misleading names altogether and numbers his registers 1 through 4 [18, p. 51]. Register 1 may be known by such names as pulse, creak, and vocal fry, and is the lowest part of the human voice; it is not ordinarily used for singing, and strangely enough was found by Hollein to be placed at a similar average level between men and women [75, 18]. Register 2, also known as modal, chest, normal, and heavy, is the main register used for both speaking and singing, and is the register already well-developed in most beginning singers [18, p. 54]. Register 3, also termed falsetto, light, head, and loft [75], is used more often in singing than in speaking. Insofar as it is referred to as “false” or “fake,” it is quite underused in men’s voices and often referred to as absent in women’s [18, pp. 55-56], although in the latter case it is very much present unlike what many have classically claimed [79]. Register 4, sometimes called flute or whistle, seems only to be found in some women and children’s voices, though some men can learn to exercise it as well. It feels fairly effortless to use when it is developed, and is associated with coloratura lines and the highest notes in high female voices [18, pp. 58-60]. This classification system is a fairly good representation of the current academic classification of registers.

Brown asserts that males and in particular tenors must learn to develop their falsetto voices to facilitate ease with their higher ranges as they learn to blend their registers. He states that women generally have less trouble accessing their falsetto or head-voice range than their chest voice in singing, although they generally speak in their chest voices; they too must learn to blend the registers [18, pp. 57]. He spends much time talking about register blending between registers 2 and 3, stating that the goal for every voice is to create a seamless range from low to high [18, pp. 50-51, 56, 61]. Much like García, he alludes to a register 2A often termed head, mid, middle, or upper that is located in between registers 2 and 3 and must be developed in beginning singers [18, pp. 51-52]. Above all, in training register blending, it is important not to overuse any part of one’s voice and keep the throat feeling as good as possible [18, p. 63].

Bozeman's approach to registration is simple, but not rigidly categorical. Laryngeal registration is not the focus of his book, so he does not spend much time discussing it, but when he does, he refuses to talk in terms of absolute registers. Vocalizations inevitably fall somewhere on a scale between "heady" and "chesty" for him. "Chestier" registrations in "vibrational mode one" involve shorter and thicker vocal folds cutting the airflow more abruptly and have stronger, "brassier" higher harmonics, while "headier" registrations in "vibrational mode two" involve a smaller portion of the vocal folds coming into contact with the air more gently and are "flutier" and more sinusoidal, having less higher harmonics [7, pp. 6-7]. In a more dedicated later section of the book, Bozeman discusses the ideal of having a "smooth" perceptual register transition between vibrational modes one and two by shifting greater engagement from thyroarytenoid to cricothyroid muscles as one moves up through one's range. The range in which this occurs is known as the *zona di passaggio* in some classical schools [7, pp. 93-95]. Perfectly smooth register changes in actual laryngeal function rarely occur in practice, but one can smooth one's register changes through practicing with a semi-occluded vocal tract. It is important, Bozeman says, that the larynx not be allowed to rise in the transition from vibrational mode one to mode two [7, p. 94-95].

As mentioned above, the science is not yet completely settled on how registers are physically formed [76]. However, a few significant results have been noted by researchers in recent years. An experiment involving female subjects found greater thyroarytenoid muscle engagement and vocal fold adduction in lower registers than higher ones [76], and another study involving male subjects was found to support previous claims that the period in a single oscillation of the vocal folds during which the folds are open is greater in the modal register than the falsetto register [77]. These conclusions support the physical characterization of the falsetto and chest registers given in Miller and Schutte (2005) [78] and supported in Bozeman's book [7, pp. 6-7]. Loudness seems to also play a role in the perception of registration, as described in Li and Yiu (2006). The researchers noted that falsetto production seemed to begin at lower frequencies during quieter phonations in the general literature, a result which their own conclusions supported. This may be because softer, breathier tones naturally favor falsetto-like vibrations of the vocal folds, while louder phonations require a more concerted stretching of the vocal folds to reach this vibration mode [80]. Still other research by Miller and Schutte on female voices argues that there is a large role that the vocal tract plays in tuning and damping harmonics to "blend" registers [78].

In the context of usefulness for vocal pedagogy in the Western classical school of singing, it indeed seems most important from studying all these texts that students learn to "mix" or "blend" their registers well. While both "brassy" and "fluty" tones are accepted as valid in this musical idiom, vocalists need to be able to transition perceptively smoothly between these types of tones to be able to sing pieces with wide vocal ranges. In any discriminatory software agent of this sort, it is extremely important that we not train any neural networks to reject the sounds of certain voices outright simply for being divergent from a perceived norm. Therefore, it seems to make the most sense for an automated vocal coach to test only for smoothness in register transitions.

An ideal automated vocal coach trained in laryngeal registration might be able to make at least a three-way distinction between voice recordings it analyzes. Singers could be making ideally smooth register transitions between different ranges in their voices. They could also err in this category in one of two ways. They could either attempt to bring "flutier" and headier tones down to frequencies that were too low for a smooth transition, or they could bring "brassier" and chestier tones up to frequencies too high to be sustainable. The natural disparity between the characters of people's voices might make fair classification of these characters impossible, though. It might be far simpler to watch for the *results* of abrupt register changes induced by insufficiently smooth registration. A network drawing a binary distinction might watch for sudden changes in vibration mode, looking for abrupt changes in the strength of upper harmonic partials and sound pressure. Due to time and resource considerations, it is the latter type of network that is to be studied in this project, differentiating *broken* and *smooth* registration.

3.1.5 Resonance management

"Resonance management" has a great deal less history and controversy in the scientific literature than the two preceding subjects. It is, rather, a convenient term for strategies of resonance within the canon of

Western classical singing that have largely to do with the actions of the vocal folds. In the context of this project, it is a term that refers to the presence or absence of the phenomenon known to most voice scientists as the “singer’s formant.” The singer’s formant is, as mentioned in the introduction to this writing, a band of frequencies between 2.4 and 3.2 kHz that is characteristically an area of high sound pressure and resonance in the trained Western classical voice [6, 7]. In Italian opera it is also known as *squillo* [81], although the two terms are not completely equivalent. It is a “ring” that enables the professional Western classical voice to be heard easily above an orchestra [82], and acoustically results from the clustering of formants 3, 4, and 5 in the voice [83]. It also lies within the highest range of frequency sensitivity in the human ear, between about 2 and 5 kHz [84, 7, 85]. It may be specifically attuned to the sound of the Western classical orchestra in Western classical singing; other operatic traditions seem to have different singer’s formants more attuned to their specific orchestras [86].

The science surrounding the production and phenomenon of the singer’s formant is more clearly settled than with many other topics in voice science. Its existence was first noted in scientific literature in Wilmer Bartholomew’s explorations of “good voice-quality” in male singers in 1934 as a “high formant” usually present in the voice between 2.4 and 3.2 kHz, although higher in voices judged to be of lower quality; its prominence relative to other frequencies in the voice was judged to be a factor of quality in the Western classical tradition [87]. Undoubtedly, in addition to being called *squillo*, it is a component of *chiaroscuro* voice [7, pp. 17-18], judged by Italian pedagogues over the centuries to be an essential part of any well-trained operatic voice [88].

Chiaroscuro, a combination of the Italian words for “bright” and “dark,” refers to the proper balance between various areas of resonance of the human voice: both low and high resonances ought to be in proportion to each other in this tradition of singing [88]. This balance is an ultimate goal of learning to sing in the Western classical tradition, but as Bartholomew notes, this high formant *is* an important factor in subjective voice quality [87], and generally most people who can sing at all can produce lower harmonics with little difficulty. Therefore, while the perceived quality of the “ring” has been shown to be affected by the relative proportions of lower partials and other factors like proper phonation mode [82], we will focus here on simply the higher frequencies implicit here in the term “singer’s formant.” This is done both for the sake of simplicity and network accuracy.

Johan Sundberg first coined the term “singing formant” in 1974 [6]. Speaking acoustically, it is often defined as a “cluster” of multiple formants in the voice placed on top of each other or in close succession [7, pp. 17-18]. These formants are often interpreted as the third, fourth, and fifth formants exhibited in a spectrogram of the voice proceeding from low to high frequencies, often termed F_3, F_4, and F_5 [82, 83, 81]. The effect of diminishing their volume in a recording of an isolated professional voice, as done by the author a number of times in the course of building the dataset for this criterion, has quite a marked effect on one’s perception of the voice: it makes the same notes sound as though they are being sung by an entirely different, less well-trained singer. Physiologically these frequencies tend to be accentuated not as the result of any changes in the vocal tract above the larynx, but are instead due to reflections in sound around the edges of the larynx and in between the larynx and epiglottis, particularly in the laryngeal ventricle and piriform sinuses [87, 6, 89]. The singer’s formant may be accentuated by a lowered larynx, as well as an open throat and constricted aryepiglottic sphincter [6, 90, 81, 7].

There is a lively debate in the research community over which voice types possess a singer’s formant as strictly defined above and which voice types do not possess this. It is almost universally agreed, ever since the first paper put forward by Bartholomew, that trained male voices have a particularly strong singer’s formant, and also that the singer’s formant is less accentuated in female voices than male voices [87, 6, 82, 89]. Some, however, assert that due to the lowered volume of the formant in the typical frequency range one would expect and the occasional presence of two separate formant peaks in this location in soprano voices, sopranos do not cluster their formants and thus cannot be said to have a “singer’s formant” [91, 83]. Other studies contradict this and report a singer’s formant in soprano voices [92]. Regardless of the precise terminology one uses for it, though, all the figures in these studies show that soprano voices do have a peak or peaks in frequency in the broad range in which singer’s formants often occur. Therefore, in this research we will not distinguish between the phenomena reported in soprano and other voice types and refer to all the formant peaks as “singer’s formants” for the sake of simplicity.

Manuel García's research on this topic was surprisingly prescient. In his *Art of Singing*, he discusses "timbres of singing" at length, defining them as general characters of the voice that can be perceived when a voice is in any of its different registers [68, p. 6]. Timbres are produced by interactions between the larynx, pharynx, and other parts of the vocal tract, and are all classified as being in one of two characters: open or clear and closed or sombre [68, p. 6]. This would seem to parallel the *chiaro* and *oscuro* components of the voice discussed earlier by Italian pedagogues and mentioned above. Like those pedagogues, García stresses a balance between the two characters of the voice, negatively describing situations in which one character or the other is excessively exaggerated [68, p. 6]. Strikingly, García mentions the role of "tendons" and folds surrounding the glottis contributing to the formations of various timbres, as well as the lowering of the larynx [68, p. 8]. However, he does not make a normative judgment on any constant timbre one ought to assume in singing.

Oren Brown, in his chapter on resonance and power of the voice, discusses the acoustic and physiological origins of the singer's formant, noting its origins in the lowered larynx and widened pharynx as well as its power to make a voice heard over an orchestra [18, p. 80]. Uniquely in these texts, he suggests that students learn to accomplish this first by relaxing the tongue, allowing the larynx to assume its lowered position and facilitating the formant's presence [18, pp. 80-81]. He also emphasizes that the development of resonance takes time and concerted training, noting that training allows the laryngeal muscles to develop greater resistances to air flowing through the larynx and therefore produces a greater range of control over resonance [18, p. 83]. He is careful to tell any student of voice not to try to force the development of robust resonance too early, and stresses singing with a connected flow of air [18, pp. 83-84]. Bozeman effectively summarizes most of the scientific revelations on the acoustics and physiology of the singer's formant that we have discussed above, not focusing on advice to the singer; therefore, a summary of his claims here would be redundant [7, pp. 17-18].

From the context of an automated vocal coach, it is fairly clear to see how one would classify voices with regard to the level of the singer's formant. In the most basic binary classifier, it is either present in the voice to an appreciable extent and is therefore idiomatic, or it is absent from the voice and therefore suggests unidiomatic Western classical singing. Checking the level of the singer's formant is also quite clear acoustically; one looks to see if there is a high amount of sound pressure in the 2.4-3.2 kHz frequency band relative to the fundamental frequency of one's voice, and makes sure that this phenomenon is not simply due to "noise" or uniform energy across the frequency spectrum. The challenge, then, is determining just what the precise decision boundary for idiomatic Western classical singing would be, given many examples of professional and amateur singers. This is a perfect task for a neural network, and one that will be explored in a later section.

3.1.6 Proper vibrato

Vibrato is a feature of singing emblematic of, and fairly uniform across, most Western classical singing as it is performed today. It is an iconic feature of opera in pop culture, as exhibited by the Western cultural trope of the soprano singing a quivering, sustained high note on stage in a Viking hat. In its broadest sense, vibrato as it is understood in the West is defined as some sort of periodic fluctuation of a fundamental pitch in a note played or sung by a musician [93]. Western classical singing is fairly distinct from other styles of singing around the world in both the presence and uniformity of vibrato within the style: the minority of other musical styles such as Peking opera around the world that use vibrato tend to use different types of it for different characterizations [94, 95]. Today vibrato is so universal in many types of Western classical music that many pedagogues have taken to claiming, without much supportive evidence, that singing without it is detrimental to vocal health [96].

The general speed of vibrato in Western classical singing is between 5 and 7 Hz, and the general depth of the fluctuations within vibrato is within two semitones in either direction from the central pitch of a note [93, 94]. Anything outside of the usual range is perceived as unidiomatic. When vibrato is not present or has insufficient depth, the phenomenon produced is known as "straight-tone" singing [94]; when the fluctuation of pitch in vibrato is too wide, it calls to mind older singers who have not been able to keep their voice agile in age [93]. Vibrato slower than 5 Hz tends to be perceived as quite slow, and vibrato faster than 8 Hz as too "nervous" [93].

In the same paper in which he published his revolutionary results on the singer's formant, Bartholomew (1934) also discussed normative ideas on how vibrato ought to sound in well-trained voices: he described the ideal vibrato as consisting of a constant oscillation in pitch and timbre of about 6 or 7 Hz, the depth and speed of which ought to remain constant [87]. Crucially, he emphasizes how "disagreeable" unidiomatic vibrato sounds to the classically trained ear, and at the same time remarks that it is the result of the coordination of a great many muscles that is difficult to consciously control in many people [87]. This latter claim resonates with this author's own anecdotal experience, having worked with many singers who complain about the character of their vibrato as judged against this normative standard and feel powerless to change it.

Today we have no studies claiming a full understanding of all the muscles involved in vibrato production. It appears that there are simply too many factors in vibrato occurring in a very small area of the body at once to accurately pinpoint precisely what happens. Some studies have, however, highlighted cricothyroid muscle involvement and neural and auditory feedback into the larynx as two physiological factors implicated in the process [97, 98]. Still others have suggested abdominal muscles are primarily responsible for vibrato activity [97], among myriad other points of view. A study in 1993 also utilized electromyography to report muscle involvement in sustained vibrato in four sopranos and found at least some level of activity in anterior suprahyoid, extralaryngeal, massetter, and perioral muscles [99]. At any rate, vibrato appears to be a phenomenon largely recognized and reinforced instead of specifically and purposefully enacted, with many claiming that it arises naturally out of healthy, relaxed singing [100, 101]. A hypothesis that natural nerve impulses involved in pitch regulation produce vibrato, as put forth by a number of researchers and pedagogues [18, 98], would tend to naturally conform with this view.

Some sources take care to distinguish between idiomatic vibrato and unidiomatic "tremor," "tremolo" (here understood not to be of the sort produced by rapid stopping and starting, but a fluctuation in pitch), or "wobble" [87, 100, 18, 98]. This term, as its many names suggest, is quite ill-defined. Bartholomew argues that, while it sounds like a similar phenomenon to vibrato, it occurs as the result of a separate physiological process [87], while Westerman (1938) advises that the term be dispensed with entirely, presumably because it implies a separation that does not exist [100]. Modern research has been kinder to Westerman's point of view, noting common vibratory patterns in the same muscles in both tremor and vibrato [98]. Here we will discard the difference between the two, simply speaking of unidiomatic and idiomatic vibrato.

One complication in discussing vocal vibrato in the context of Western classical music is that vibrato was not used nearly as much centuries ago as it is today. It is not the case, as is sometimes assumed, that contemporaneous performers of early music did not use any vibrato while singing [102, 96]. However, the technique did not enjoy the nearly constant use it undergoes today in much Western classical singing, often being relegated to the place of an ornament or deliberate stylistic choice [103]. Particularly in many English styles, it was dispensed with in the interest of perfect pitch accuracy and agility [102, 103, 18], while French and other traditions seemed to naturally incorporate it more [102, 18]. Writers in the Renaissance period seemed to have a shared distaste for vibrato on sustained notes, but by the eighteenth century and proceeding onward, this practice became more and more accepted [96].

It appears that by the nineteenth century, some traditions had come to accept a more or less continual vibrato from some of their soloists, spurring on the dominance of the practice in the twentieth century and beyond in operatic music [96]. This normative standard led to the promulgation of many radically pro-vibrato points of view that survive today among pedagogues and listeners alike [101, 96]. In an extremely myopic statement exemplifying this point of view, Westerman (1938) claims that it is utterly impossible to sing acceptably without vibrato [100]. There were still some critics of the practice, though, particularly in Anglophone regions. Our old friend from earlier, George Bernard Shaw, was known to excoriate singers exhibiting what he considered to be excessive vibrato, complaining at one point that vibrato was "sweeping through Europe like the influenza" [104].

Even today in the Western classical world there are contexts in which vibrato is somewhat unidiomatic. Many choir directors nowadays actively discourage singing with vibrato, often for the benefit of pitch accuracy and uniformity of timbre [101, 96, 105]. In early music ensembles attempting to stay faithful to contemporaneous practices existing when pieces were written, vibrato is often diminished or absent [102, 96, 105]. Such vibrato-less singing is often termed *straight-tone* [101, 105]. Absent a single directive or consensus in the music world, the general rule followed in other areas of this project is to go with the normative operatic prescription for singing, as it is more widely studied and taught than any other style explored in the project. Therefore, singing with a vibrato as defined by the standards set above in Sundberg (1994) would be the standard taught by the hypothetical automated vocal coach explored here.

Here, our three main pedagogical sources are conspicuously silent on the practice and teaching of idiomatic vibrato. The word "vibrato" is mentioned by neither García nor Bozeman in their texts, and is only discussed at length by Brown on one page of *Discover Your Voice* [18, p. 96]. Furthermore, there is some evidence that García, valuing steadiness in the voice, may have argued very strenuously against the continuous usage of vibrato in singing. Twice in *The Art of Singing* he discourages the constant usage of what he calls the *tremolo* and defines as a "vacillation" of the voice [68, p. 33, 69], going as far to say that many voices have been "lost" to overuse of vibrato and are thereafter incapable of the art of phrasing [68, p. 69]. García was quite frustrated when a natural tremor came into his voice with age, describing it as "execrable" and "an abomination" [5]. Brown certainly appears to value vibrato more positively than García, but cautions against a "wobble" characteristic of older voices and produced by excessive, persistent breath pressure over time [18, p. 96]. He does not explain mechanisms for producing vibrato in voices that do not naturally exhibit it [18, p. 96].

This all begs the question of how, given that we have very little to work with in our model texts, we are to teach proper vibrato with an automated vocal coach if we learn to recognize it well computationally. Some directions have emerged in the research world, some quite recently. Mitchell and Kenny (2004) present research supporting the open throat as conducive to the production of vibrato in the voice [106]. In a very expansive and informative thesis written partially on pedagogical applications of vibrato and straight-tone singing phenomena, Danya Katok (2016) argues that this dimension of singing is largely regulated by subglottal pressure levels, with higher levels of subglottal pressure conducive to vibrato and low subglottal pressure levels producing straight-tone singing [105]. Finally, it is this author's view based upon his own observations and interactions with teachers that vibrato tends to occur at least in his own voice spontaneously with greater engagements of abdominal and extralaryngeal muscles. Perhaps exercises could be devised to have students learn to isolate these muscles and consciously engage or disengage them when vibrato or straight-tone singing is required, eventually developing enough muscle memory that conscious thought about these actions would become unnecessary.

To summarize, while there is a relative dearth of material on acceptable and unacceptable vibrato in the three main sources consulted for this research and a disagreement on views of vibrato in the singing world, vibrato itself is so very characteristic of the Western classical style of singing that it would be an unacceptable omission to leave it out of any hypothetical automated vocal coach. Vibrato tends to exist along a one-dimensional perceptual scale in Western singing, with straight-tone singing on one side of the scale, unacceptably slow, fast, or wide vibrato on the other, and idiomatic vibrato levels somewhere in the middle. A simple neural network could focus on small intervals of time in voice recordings fed into it and perform this ternary classification on each slice of audio. The eventual output would outline idiomatic and unidiomatic areas of vibrato in these recordings.

3.2 Datasets

The definition of machine learning agents given in the section above can be characterized as agents with the power to learn from their environments. Without input from its environment, a machine learning agent cannot learn and is limited only to what is hard-coded into it. Furthermore, agents cannot spontaneously learn information that is external to or not implicit in its environment. Therefore, proper supply of information to these agents is absolutely fundamental to their proper functioning. In the context of neural networks, this information takes the form of pieces of data fed into a network, and the environment is the whole set of data on which the network operates. If one seeks, as is the goal of this project, to build a network with applications to the real world or a subset of it, one must provide that network with a dataset that is as reflective of the cases the network will encounter in the real world as possible. For a classification network, this means that data for each label in a dataset ought to be diverse enough to encompass suitable occurrences of the label in reality, although they ought not to be so broad or numerous that they cause a network to overestimate their occurrence in realistic data.

In the context of this project, data takes the form of recordings of the human voice singing in various styles, ideally with as little background noise or musical accompaniment as possible to ensure that our networks are not prevented from learning important details and features of the singing voice. Just as with any scientific experimentation, when main variables or labels that one is attempting to study differ between trials or pieces of data, it is best to control for all variables not being studied by keeping them constant. This is to ensure maximum probability that our networks correctly isolate and learn to recognize a contrast between the appropriately labeled data. Therefore, when data are available for a single voice singing in ways that can be classified under multiple labels, they are represented as much as possible in our datasets. At the same time, the goal of this project is to build a product that could be used to help people with almost any voice or phonatory pattern learn to sing properly according to the Western classical tradition. Given this, we also want to represent as many different voice types and parts as is humanly possible in our datasets.

With unlimited high-fidelity data which are broad and contrastive, as well as unlimited computational power, it is this author's view that nearly any feature of the real world could be learned to an appreciable degree by a neural network. Unfortunately, there are a great number of practical limitations to the data that can be used in this project in terms of scope and feasibility. One limitation is that, as mentioned above in note 3, the subfield of research on computer audition is not as active as the field dealing with computer vision. Therefore, datasets focusing on images and visual data are present in the world in high quantities now that machine learning has been a very active field of research for many decades, but publicly available audio datasets are relatively few and far between for people looking to work on projects like this one.[6]

In recent years, there have been inroads into this issue through the publication of some now well-known large labeled public datasets, such as the Free Music Archive, the Million Song Dataset, and LibriSpeech [108]. Fortunately for the public, most of these datasets tend to be quite general and to serve as benchmarks for testing different audio classification architectures. Unfortunately for this project, there are very few large datasets available focusing on the isolated human singing voice. Some exist and have been utilized in this project, but right now there is no single dataset large or regularized enough for any of the classification tasks attempted in this project to be performed by any neural network with human-like accuracy. The available datasets are often small, do not encompass all sounds and phonemes utilized in common examples of singing, or do not contain enough examples of unidiomatic singing, or some combination of the preceding factors.

One could endeavor to address this problem by building multiple high-quality datasets by one's self, but this would be an extremely time-consuming endeavor involving recording or gathering high-quality voice samples and painstakingly labeling each datum within them. This is an eventual goal of this project as it extends beyond the scope of this thesis, but was not feasible within time constraints over the past year. To address the above problems as well as was possible, it was necessary for this project's implementation to build larger datasets out of existing smaller public ones, and in some cases necessary to construct somewhat low-quality labeled datasets by one's self. Data augmentation was occasionally performed in the form of manually editing existing pieces of data to produce new, differently classified ones or adding some maximally contrastive, high-quality recordings of the author's voice to data.

While there are not a great many singing-voice datasets in the research world that have been made available publicly, there was enough aggregate material published to make this project just barely feasible. One can say without any hesitation that this project would not have been possible for a single undergraduate student three years ago. The first dataset acquired by the author in the research for this project was the publicly available AudioSet, first published in 2017 by researchers at the Sound and Video Understanding teams in

[6] Many large companies that produce or sell voice-activated digital assistants such as Google, Amazon, Microsoft, and Apple are thought to have amassed extremely vast quantities of data for the purposes of speech recognition that are privately held, and add to their collections of data every day through the operations of their own products [107]. Since these companies usually do not disclose statistical information or even estimates about the amounts of data that have been used to train their proprietary inventions, we have no way of knowing the size gap between the amount of data publicly available and labeled in datasets for academic purposes and the amount privately available to corporations. Due to the fact that one does not often see people making their own speech recognition systems by themselves today, though, one has good reason to suspect that it is large.

the research department of Google [109].

AudioSet consists of short audio samples of millions of YouTube videos falling into different categories of sounds. Each sound indexed in the dataset, far too large to download as a standalone file, is tagged with the YouTube video from which it comes and the timecode of the audio sample. It is particularly useful for locating videos of amateur singers, which was occasionally necessary. The downside of this dataset is that it is difficult to find high-fidelity recordings on YouTube of the singing voice in an isolated environment; very often videos tagged as "a cappella" are not so or are of groups of multiple people singing *a cappella*. Even though it often required searching through minutes upon minutes of video to find good samples, this dataset was useful a number of times in this project.

The most broadly valuable dataset in this project was VocalSet, announced in a paper in the International Society for Music Information Retrieval Conference of 2018 and published by Julia Wilkins et al. in the same year [110]. It contains 10.1 hours of recordings of multiple professional operatic singers in multiple voice parts trained in the Western classical style demonstrating a number of different voice techniques [110], most of which are idiomatic to this style and some others which are not. These techniques are demonstrated through single long tones, scales, arpeggi, and song excerpts. Each of the four datasets built for this project used recordings from this dataset extensively, as it is for the most part correctly hand-labeled and its constituent recordings are of extremly high quality. The one major limitation of the dataset is that it only contains recordings of twenty different voices and does not contain any recordings of contralto voice types, but it is a very time-consuming undertaking to ask a person to record minutes upon minutes of high-quality singing-voice audio in the first place, so this is quite understandable.

The Singing Voice Dataset, presented by Dawn Black et al. in 2014, contains a number of recordings made by singers of Chinese descent, most of them performing traditional Chinese opera arias [111]. Many of these singers are extremely well-trained professionals exhibiting vocal styles that are precisely cultivated but quite unidiomatic to the Western classical style of singing, which makes their recordings useful for training networks. While this dataset does not contain recordings of many different people either, many of its recordings are multiple minutes long, meaning that it is a very rich and well-defined representation of the Chinese operatic tradition of singing. One limitation of this dataset for this project is its labels; the dataset helpfully contains song excerpts similar to those a person would be singing into an automated vocal coach, but was constructed for learning the differences between songs expressing positive and negative emotions in Chinese opera. Its labels are therefore irrelevant to this project, and this audio must be manually sampled at length to determine useful labels for it.

The Phonation Modes Dataset, published by Proutskova et al. in 2013, is part of a study conducted by the same team of researchers in the same year focused on a similar goal to this project [112]. These researchers were seeking to classify voices exhibiting the exact phonation modes specified above in the methodology section, and their dataset features recordings that contrast these phonation modes very well. It consists of second-long recordings of varying vowels sung in breathy, pressed, resonant, and neutral modes. Proutskova et al. were able to distinguish between these modes with an average classification accuracy of over 60% when processing their input data properly [112]. Classification tasks and accuracies were calculated separately for different vowels, however, and there is no intentional inclusion of consonants in any of the recordings of the human voice such as is present in the songs people would be singing to an automated vocal coach. Therefore, the phonation mode classifier utilized in this project used this data heavily, but augmented it with song excerpts in each of the phonation modes for each label to be classified. Moreover, since both flow and neutral phonation modes are idiomatic and utilized in the Western classical singing tradition, there are twice as many data points available from this dataset that would be labeled as "balanced" phonation in the context of this project.

Finally, a subset of the dataset employed in Bozkurt et al.'s 2017 paper "A Dataset and Baseline System for Singing Voice Assessment" was obtained by the author with the permission of the researchers [113]. Uniquely in this list of datasets, it contains well-cultivated recordings made exclusively of amateur singers. In these recordings, these singers attempt to repeat short melodic passages played to them on a piano as part of an audition for a musical conservatory. This was quite useful in training networks on voices without very developed resonance. The data was applied to other classification tasks, too, as it was representative of the types of singing that would be used in the actual application of an automated vocal coach in the real world. The author is very grateful to these researchers for their gracious gesture in supplying the data for use in this project.

3.3 Audio representations

It was mentioned above that we would be training convolutional neural networks in this project to classify singing voice recordings, but to say only this is not specific enough for one to have an idea of how said networks would be trained. Convolutional neural networks can operate on audio, as it turns out, in multiple different input formats. The most common audio input format encountered in research literature for these types of networks is the spectrogram [114, 115, 58, 116], though some convolutional models endeavor to operate directly on preprocessed one-dimensional audio signals [33]. We will proceed forward with the former option, as it is much more well-explored and has seen successful results for audio classification in the papers cited above.

Operating as we are on recordings that can be as short as one second and as long as several minutes, depending on the format of the datasets from which they originated, it would be both ridiculous and computationally infeasible to try to pass entire recordings into convolutional neural networks one at a time. Neural networks are known to operate best on inputs with standardized sizes; this is because standardizing the sizes of one's inputs ensures that input size cannot be viewed by networks as a contrastive feature of one's data. Therefore, we must endeavor to operate on short "slices" of audio instead.

We cannot use slices of audio that are longer than one second, which is the length of the shortest recordings in the datasets in this project. Using audio slices that are much shorter than a second, though, would also be a mistake, as some of our networks operate on features such as vibrato whose correct classifications require watching frequencies and resonance over time. Therefore we have chosen a standard slice length of one second for each data point inputted into each of our networks. Processing audio into these slices before feeding it into the network gives us the additional advantage of being able to potentially generate hundreds of data points from just one source recording.

It is true that hundreds of data points extracted from recordings of a single person's voice will never be as good to have as hundreds of data points from hundreds of peoples' voices. However, spectrograms taken from the same recording often differ from each other quite widely while remaining in the same label for classification as each other, which makes them useful for us. These time slices do not overlap in the experiments in this thesis, as it was determined that overlapping slices would increase the time it took to train models without adding many unique features to datasets. The criteria analyzed in this project are either more or less constant phenomena of the voice or extremely short and localized events; neither type of phenomenon has a high probability of suffering from being split over two slices of data. Regardless, experimentation using overlapping time slices of audio is a good subject for further research.

In order to amplify amounts of data being fed into convolutional neural networks and prevent overfitting, some researchers perform common image transformations upon their input images to the networks, including stretching, shearing, and rotation operations in a process known formally as *data amplification* [117]. This would be utterly disastrous in most cases for the data being fed into these networks. Many of the features involved in processing input data in this project are either dependent on specific time sequences or frequency patterns, and are meaningful only if understood along specific time and frequency scales and axes. Rotation of a spectrogram by 90 degrees, for example, would invert time and frequency axes and render the sound it represents unrecognizable if played back to the human ear alongside the unrotated spectrogram's sound. For this reason we do not do this; the only data augmentation technique used at all in this project is the filtering of specific frequency ranges in audio signals.

The specific resolution of the images along both time and frequency axes depends on the specific feature being learned, but never, ever varies within a single network's input data, as this too would render the data meaningless for the same reasons as above. The frequencies represented in the spectrograms are selected and scaled judiciously. Most commonly the information in the images relevant for classification is scaled from 0 to 6000 Hz, so we use these values at the lower and upper extremes of the frequency axis of each spectrogram

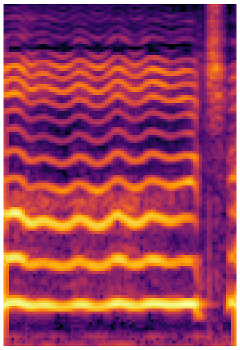


Figure 3: A typical spectrogram image used to train the networks in this project, presented exactly as it is fed to the networks. The X-axis represents time progressing forward from left to right, while the Y-axis represents mel-scaled frequency bands, the lowest at the bottom and the highest at the top of the image.

in three out of the four criteria analyzed. The time scale of the images is most often 512 audio samples per vertical spectrogram frame; this scaling is also used in three out of the four criteria in this project.

These spectrograms are all mel-spaced, meaning their frequency axes are scaled according to the mel scale. The mel scale, developed by Stevens et al. in 1937, scales frequencies according to their relative distance from each other as they are perceived by the human ear [118]. Mel spacing in spectrograms has seen much success in the field of machine learning focused on convolutional neural network processing of audio signals [114, 119, 58], and for this reason it is employed in this project. The spectrograms are produced from audio by a Python script utilizing the "librosa" library for audio processing and the Matplotlib library for image production. They are pseudo-colored according to the 'inferno' color map provided in Matplotlib; this is by arbitrary convention. While most studies using spectrograms for processing with convolutional neural networks operate on grayscale images, studies do exist in which networks operate on pseudocolored spectrograms [116]. Furthermore, limited experimentation using grayscale images of the same sound events as training sets for the models utilized in this project yielded no appreciable difference in training or validation accuracy or typical used case performance for any network. An example of a typical spectrogram image used in a neural network is shown in Figure 3.

3.4 Network architectures and parameters

All neural networks employed in this project, as mentioned above multiple times, are multi-layered convolutional networks consisting at their cores of one or several instances of one or more two-dimensional convolutional layers followed by a type of downsampling layer called *max-pooling* that performs rather well on various networks [120]. Sometimes input is followed by Gaussian-noise distortion before being fed into the first convolutional layer of the network. The shapes of kernels in convolutional layers differ widely between criteria being tested; sometimes more generalized square kernels are used in succession and other times more vertically or horizontally stretched filters are used to recognize more frequency or time-dependent features of spectrograms as they presented themselves in criterion analysis. Occasionally, dropout and batch normalization are used to regularize layer weights and prevent overfitting to the data comprising the small data sets used in this project. When batch normalization is used following various layers, it is applied before the activation function of the various layers, as visualized in the network architecture tables below. These layers' biases are removed, as is common practice when one applies batch normalization to a layer in a neural network [47]. Following the stacks of convolution and max-pooling layers, the output of the final max-pooling layer is fed into a stack of two or three fully connected layers with or without dropout layers between them, and the final fully connected layer outputs the classifications made by the networks.

The activation functions utilized in these neural networks are rectified linear units (ReLUs) and softmax and sigmoid functions. ReLUs are employed after all hidden layers in the networks; they were first demonstrated to perform well on hidden layers by Hinton and Nair (2010) and have been dominant in hidden layer activation functions ever since [121, 122]. Sigmoid and softmax functions are used as the activations for the final layers of binary and multi-label classification networks respectively; they have each also been established as dominant activation functions for final classifications in these respective network types [122]. The adaptive learning rate or optimization function in each neural network is the Adam stochastic optimization algorithm, first demonstrated by Kingma and Ba in 2015 and subsequently revolutionary in the machine learning field [123].[7] The loss function utilized in each network is respectively binary cross-entropy in binary classification tasks and categorical cross-entropy in multi-label classification tasks; these functions are standard in these respective network types [30, p. 236]. Occasionally the loss function output over the network's training will be specially *weighted* according to different *class weights* in order to normalize for imbalanced datasets or cause the networks to pay more attention to classifying data of specific labels correctly.

The datasets built for each classifier utilized in this project were, for the purposes of training the networks, separated into training and validation sets. Depending on the specific network, proportions between the sets were determined such that the number of validation-set images used to train the networks per epoch was approximately one-third to one-fourth of the size of these images in the training set. These proportions can be seen in the tables in the following sections of the paper detailing training parameter sets. All neural networks in this project were built and trained using the Keras framework for Python, an excellently built and maintained machine learning framework that combines bleeding-edge features sourced from recent research and high computational power with simple and easy-to-learn code syntax [51]. The Keras framework was built on top of an existing installation of TensorFlow, a more low-level machine learning library built by Google [49]. All audio and image processing and training was accomplished on a mid-2014 series MacBook Pro. The MacBook could handle these training tasks rather well. The longest network training times were about thirty minutes, although the small size of these data sets necessitated simple architectures to prevent overfitting to the training data, so having sufficient computation power was not a grave concern of the author's.

3.5 Procedure

In the following four sections, each outlining different experiments, a general framework for procedure is followed. Having established the type of network we will be training and the goals we seek to achieve in the research above, we will first describe the nature of each dataset used for classification. A description of the network architectures used, or architectures used, in the case in which multiple audio architectures were found feasible for the criterion being studied, will then be provided. Following this, the performance statistics for the networks created will be supplied, and the implications of these results on the feasibility of a hypothetical automated vocal coach will be briefly discussed.

One more discussion is necessary before proceeding to the results section of the paper. Most studies on machine learning classification are focused primarily on improving an accuracy rate of classification with regard to a dataset. This makes sense if, as discussed above, a dataset is representative of the practical environment in which a classifier is designed to operate. However, this is not necessarily always the case in this project. As we will see, oftentimes the datasets with which we must necessarily work in this project are quite small, noisy, and of poor contrastive quality. Therefore, the performances of these networks, while they may be quite high in test and validation accuracies, may be absolutely abysmal on real-world data. Oftentimes in training these networks, this was noted to in fact be the case.

The question, then, comes to mind of how we are to systematically and rigorously assess the performances of these final networks on data similar to what an automated vocal coach might be required to handle. This is, in fact, our primary concern due to the poor quality of the datasets used in training. To handle this concern, for each criterion explored in this project, an extremely small test set symbolizing a typical use case, and termed as such in the figures to be presented ahead, was synthesized by this author. The use case sets consist of short recordings of the author's own voice demonstrating beginning phrases from a well-known Italian aria, either Tosti's "La serenata" or Giordani's "Caro mio ben," frequently employed in vocal pedagogy. The author took pains to make each recording differ only from the others in the characteristic criterion explored

[7] As of March 2019, it has been cited over 19,000 times since publication according to Google Scholar [52].

by a particular dataset. While he wishes that he had more time to prepare larger test sets for each network in this project using multiple voices, he is confident that each use case is extremely accurate to the criterion contrast that is the goal of each network's classifications.

The following, therefore, cannot be stressed enough: when results are presented below in each section listing the accuracies of networks on training, validation, and use case sets, the results being presented are the results with the best use case performance rather than the results with the highest training or validation set accuracy. The author would much rather present results that look worse on paper and may be the result of underfitting that perform well on these use cases than results that look better according to traditional metrics but yield useless results in the real world. To most accurately illustrate the performances of networks on these use cases, best-case confusion matrices are presented illustrating the networks' classifications of the use case data in the results for each criterion. Having noted all this, results and a brief discussion on how they apply to the concept of this project's hypothetical automated vocal coach are presented below.

4 Phonation

Results and analysis of neural networks for studying and discerning proper phonation from audio recordings of amateur and professional singers are presented here.

4.1 Implementation

The goal of this phonation classifier is to distinguish between so-called breathy, balanced, and pressed phonation modes, "balanced" being a term used by the author to collectively refer to neutral and flow phonation modes outlined in voice science research. These different phonation modes, when qualitatively analyzed in spectrograms, seem to have relatively constant distinguishing features over large time scales. The most salient visual feature of breathy singing to the author is the broad-spectrum noise exhibited across all voice formants in a spectrogram that is likely the result of air escaping from the throat as one sings. Pressed singing tends to have very strong harmonics extending up through relatively high frequency ranges, hence the "bright" tone noted by Bozeman (2013) [7, pp. 5-6]. Balanced singing is characterized by the lack of either of these two features in a spectrogram. Given all this information, it makes sense for convolutional network architectures to apply small square or balanced rectangular kernels to input to learn low-level features of data, and then apply vertical filters to the outputs of these layers to learn high-level features of these phonation modes over broad frequency spectra.

Architectures built for two different network types are presented. These architectures differ quite a bit from each other, but are fundamentally founded upon processing different input image formats. Unlike other criteria in this project, the frequency bounds on spectrogram inputs to these networks were determined to be from 0 to 10 kHz, as opposed to from 0 to 6 kHz. This choice was made because many of the higher harmonics characteristic of pressed phonation modes in singing occur above 6 kHz. The architecture of one network, known hereafter as the small-image phonation network, operates on images with 128 mel-spaced frequency bands in this range. This number of bands, the same number as in all other criteria in this project, necessarily degraded the precise frequency information in these images compared to images scaled from 0 to 6 kHz. To address this issue, a second architecture was constructed for another *large-image* phonation network, which operated on spectrograms with 256 mel-spaced frequency bands. The best use-case results of each network architecture are compared and contrasted below.

The dataset presented to each of these two networks was the same. It consisted largely of a subset of Proutskova et al. (2013)'s phonation modes dataset, in which neutral and flow phonation modes were lumped together into one single "balanced" category. To these data were added basic recordings from VocalSet of professional vocalists singing scales and arpeggi in "breathy," "vibrato" and "straight-tone," and "belt" modes of singing for each respective phonation mode explored here; spectrograms of these recordings revealed similar respective contrastive features to those in the phonation modes dataset. The major problem with these data so far was that they consisted entirely of vowel sounds and would likely make the network useless for typical singing recordings. Therefore, isolated-vocal song clips sourced from VocalSet for the balanced phonation mode and from popular songs on YouTube for the breathy and pressed phonation modes were added to finish this dataset.[8] The precise balances of all these components had to be edited during training.

Architectures and training parameters for both the small-image and large-image networks are shown below. In architecture tables throughout this paper, layers are shown from greatest to least proximity to input data from top to bottom. In other words, the top layer receives spectrogram input data and the bottom layer outputs the final judgments of the networks.

[8] For the curious, the VocalSet recordings for the balanced phonation mode were of excerpts of all the three songs present in the dataset: "Caro mio ben," "Dona nobis pacem," and "Row, Row, Row Your Boat." The songs presented as examples of breathy singing were "Good for You" by Selena Gomez, "Issues" by Julia Michaels, "Poor Boy" by Nick Drake, "Don't Know Why" by Norah Jones, and "Love Yourself" by Justin Bieber. The songs presented as examples of pressed singing were "I Feel Good" by James Brown, "Rolling in the Deep" by Adele, "Run to the Hills" by Iron Maiden, "Chandelier" by Sia, and "Black Hole Sun" by Soundgarden.

Layer	**Parameters**		
Conv2D padded	Number of kernels: 16	Kernel size: (1, 3)	Activation: ReLU
MaxPooling2D	Filter size: (1, 2)		
Conv2D valid	Number of kernels: 32	Kernel size: (128, 3)	Activation: ReLU
MaxPooling2D	Filter size: (1, 2)		
Dropout	Probability: 0.2		
Flatten			
Dense	Units: 128	Activation: ReLU	
Dense	Units: 128	Activation: ReLU	
Dense	Units: 3	Activation: Softmax	

Table 1: Architecture of small-image phonation network

Layer	**Parameters**		
Conv2D padded	Number of kernels: 8	Kernel size: (3, 1)	Activation: ReLU
Conv2D padded	Number of kernels: 8	Kernel size: (1, 3)	Activation: ReLU
Conv2D padded	Number of kernels: 16	Kernel size: (3, 3)	Activation: ReLU
MaxPooling2D	Filter size: (2, 2)		
Conv2D valid	Number of filters: 32	Kernel size: (128, 3)	Activation: ReLU
MaxPooling2D	Filter size: (1, 4)		
Dropout	Probability: 0.2		
Flatten			
Dense	Units: 128	Activation: ReLU	
Dense	Units: 128	Activation: ReLU	
Dense	Units: 3	Activation: Softmax	

Table 2: Architecture of large-image phonation network

Parameter name	**Small-image network**	**Large-image network**
Optimizer type	Adam	Adam
Loss type	Categorical crossentropy	Categorical crossentropy
Class weights	1 : 1 : 1	1.68 : 1 : 1.19
Image dimensions	(128, 87)	(256, 87)
Training samples per epoch	3000	3000
Validation samples per epoch	760	760

Table 3: Training parameters for both phonation networks

4.2 Results

Best use-case results for the small-image and large-image networks and confusion matrices on the use case sets for both networks are presented here.

Metric	**Small-image network**	**Large-image network**
Number of epochs trained	11	12
Training set accuracy	93.82%	93.47%
Validation set accuracy	71.22%	72.81%
Use case accuracy	55.56%	51.85%

Table 4: Training results for both phonation networks

Here we can see that both networks, despite having quite divergent architectures, performed in manners very similar to each other in terms of number of epochs trained, training set accuracy, and validation set accuracy. The training set accuracy is much higher than the validation set accuracy in both networks. This seems to indicate overfitting, but strangely, the validation set accuracy improved throughout the training of these networks up to this point, so of the types of overfitting to have it is a more acceptable one. The large-image network appears to have only a slightly worse performance on the use case set than the small-image classifier if one looks at overall accuracy, but a study of the confusion matrices on use cases for both networks reveals the imperfection of this statistic.

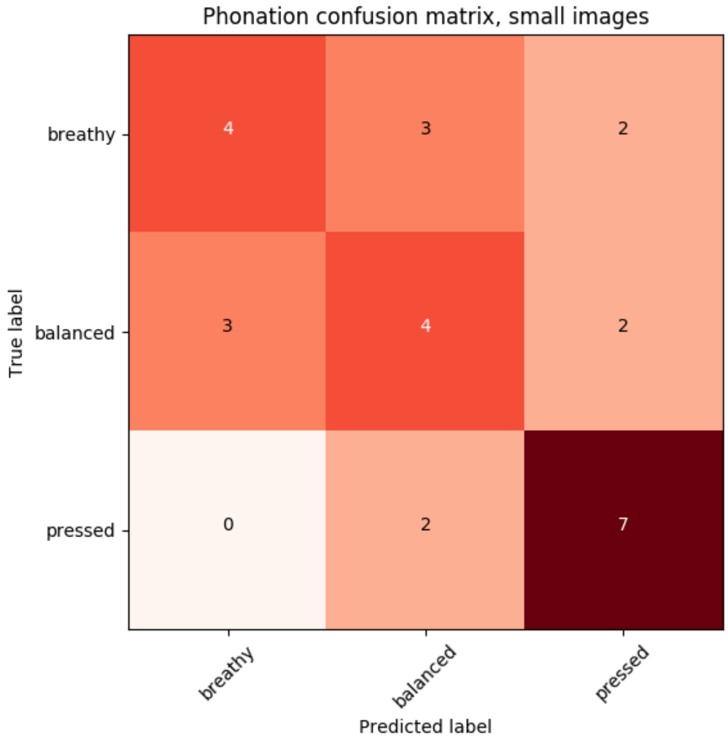


Figure 4: Confusion matrix for small-image network on use case

As can be seen in figures 1 and 2, the small-image network's performance on the use case set is much better than the large-image network's performance on this set. While the small-image classifier shows at least a slight preference for the right label in each phonation mode classified, the large-image classifier appears to have absolutely no facility in recognizing breathy singing and almost unilaterally classifies it as balanced singing, recognizing a pretty stable two-way distinction between breathy and balanced singing and pressed singing.

4.3 Discussion

Training these networks and tweaking hyperparameters for them was difficult. The dataset was just small enough that sufficiently complex networks would easily overfit the data and provide useless results, but noisy enough and outlining a distinction abstract enough that a network that was too simple would fail to learn anything about the data at all over time. Adding noise and regularization to large networks to combat overfitting persistently led to the elimination of breathy singing as a realistic label for classification in the use case set. A probable hypothesis for this could be that the dataset for the classifier is already quite noisy, and adding more noise to the input at any stage of processing completely eradicates the noise-based distinction between breathy and balanced phonation modes.

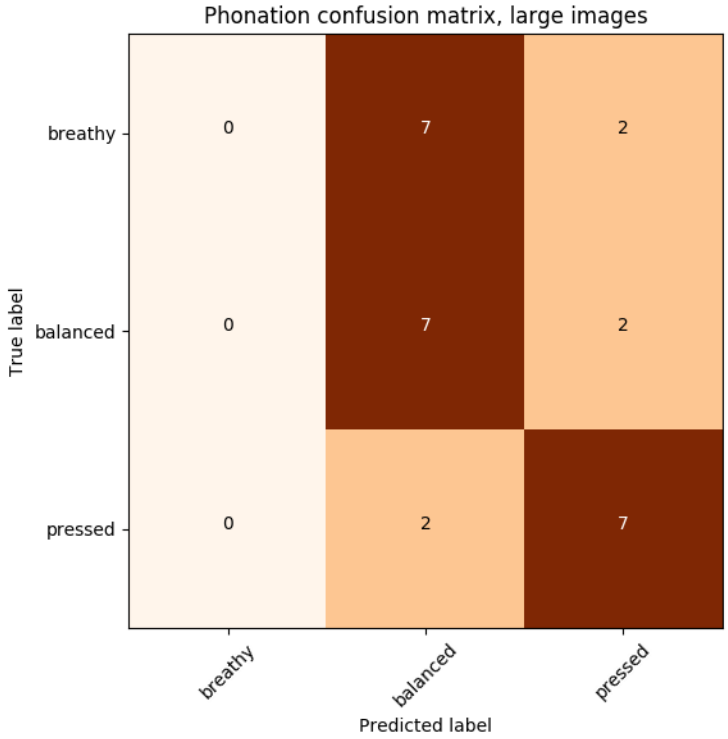


Figure 5: Confusion matrix for large-image network on use case

Getting better results for this classifier therefore probably requires a better dataset with cleaner data, one that does not contain samples of audio sourced from lossy videos on YouTube with lots of background noise out of desperation. Earlier versions of architectures for this criterion were plagued by the “pressed” label being overgeneralized to apply to the “balanced” use case data, and even the “breathy” data in some cases. This problem was effectively removed by inserting the VocalSet recordings of multiple singers that differed *only* in different phonatory modes, featuring the same singers singing the same notes on the same vowels. This type of data regularization and variable elimination is broadly effective in cases like this.

It is difficult to determine whether the results of these phonation matrices tell us much at all about the proper or improper functioning of these phonation mode classifiers on real-world data, due to the fact that the classifiers’ confidences in each of the labels here were sometimes low. However, the shapes of the confusion matrices for each network are somewhat encouraging. Each network could ultimately be useful in its own way in an automated vocal coach. Since the networks operate on small slices of larger audio recordings, results on slices of amateur vocalists singing could be averaged out to provide the vocalists with an idea of the dominant phonation mode engaged in their singing. This could eliminate the problems seen above in relatively insignificant errors in the confusion matrices: going by just the dominant classifications of each label in the small-image network, the network performed perfectly on the use case set. This is even more true for the more solid distinction between balanced and pressed phonation modes in the large-image network.

The author would not recommend for either network to be packaged in an automated vocal coach today, but by itself, the small-image network certainly performs more accurately in a typical use case than the large-image network and would be the superior candidate for inclusion. The large-image network appearing to accurately perform a binary classification might be useful in a final product. If it were combined with another effective binary classifier able to tell the difference between breathy singing on one hand and balanced and pressed on the other, a three-way classification could be created from the joint operation of the two networks.

5 Laryngeal Registration

Results and analysis of neural networks for studying and discerning proper laryngeal registration from audio recordings of amateur and professional singers are presented here.

5.1 Implementation

The goal of this registration classifier is to distinguish singing with smooth registration from register breaks, sometimes known as "voice cracks" in common parlance [124], in a voice recording. This is a binary classifier, as opposed to the ternary classifier built in the previous section, and so is subject to less possible sources of classification error than the network above. There are a number of complications in the construction of these networks and the data fed into them that lead to many, many practical problems in the construction of this network, however.

Firstly, as opposed to in the other criteria outlining idiomatic Western classical singing, errors in registration occur in transitions between notes. They are therefore rather atomic when it comes to the time scale, exhibiting large effects across the frequency spectrum but only for a very limited amount of time. This fact stands in opposition to the other criteria, which outline more or less continuous modes of singing. Small errors in time-segmented recordings in the other network types can be ignored and the general characterization of a voice recording can still be correct. This is emphatically *not* the case for this network. One would have to consciously try to make one's voice crack as much as possible in order for a majority of time slices of a voice recording to be correctly classified as broken registration.

Moreover, due to the large stigma placed on voice cracks in normative Western society [125],[9] false negatives and false positives are quite unacceptable in the operation of an automated vocal coach in this area. Since the voice crack is an easily identifiable phenomenon whose occurrence is widely recognized throughout normative American society, people generally know when their voices have cracked during singing without needing to be told this. False positives might thus be able to be overlooked somewhat, as they could not cause much harm other than making a user lose faith in the accuracy of the software. However, false negatives are actively harmful, as the singer is left knowing they have made an error in smooth registration without being provided any useful advice in how to avoid this problem.

A third problem with network architectures here is that voice cracks do typically present themselves as quite apparent events in a spectrogram, but while the difference is quite apparent to the ear, it is incredibly difficult to distinguish them from different types of phonatory and consonant onsets in images. The distinctions the author can make between all these things occur in very small patterns in relative harmonic energies. To pick up on this distinction correctly and avoid overgeneralizing the "broken registration" label, therefore, it is likely that a very large network would need to be trained on a very large, high-quality dataset contrasting all of these events with proper labeling.

This brings us to the final complication implicit in this classification problem: a large dataset containing a "voice crack" class such as would be necessary for the proper operation of this classifier simply does not exist in the world today. This is a topic so very specialized that hardly anyone would have need to create it. In an attempt to produce this, the author took it upon himself to spend hours finding videos containing audio of voice cracks with relatively low background noise on YouTube. Once the audio for each video was downloaded, it was edited into a supercut containing only audio immediately surrounding each isolated voice crack within the original video.

Around 800 isolated voice cracks produced around 800 pieces of very noisy, very diverse data attempting to collectively illustrate the anatomy of the voice crack in both male and female voices.[10] The "smooth registration" portion of the dataset used for this criterion was compiled from a variety of voice recordings from VocalSet, Bozkurt et al.'s Singing Voice Assessment dataset, some of the YouTube clips of pop songs taken from the previous criterion's dataset, and a few recordings of the author's own voice providing isolated

[9] A small anecdote in the otherwise quite unrelated article cited above and the wide presence of "voice crack compilation" videos online for people to ridicule are evidence to this claim.

recordings of specific vowels during smooth and broken laryngeal registration. This variety of sources was intended to match the wide variety of source recordings labeled as "broken registration."

The problem of dataset size could not be solved within the time constraints of this project, but some other problems were addressed to some extent. The architectures of two networks presented below have slightly different structures and interestingly distinct classification accuracies. Both exploit rather complex architectures compared to the classifiers developed for the other criteria in this project. The two networks are henceforth called "architecture 1" and "architecture 2." They consist largely of different regularization techniques applied to cascading groupings of 3-by-3 kernels in convolutional layers paired with 2-by-2 max pooling layers, and they have identical training parameters. This is a good network architecture for learning relatively high-level features about input data which is both time and frequency-dependent. The images sent into both networks differ from images used for other singing criteria in this paper in that they contain only 256 samples per spectrogram frame; they are therefore twice as wide as the images used in other classification networks here to allow registration breaks to be visualized in higher time resolution. The architectures for both networks, as well as their training parameters, are summarized in the tables below.

Layer	**Parameters**		
GaussianNoise	Standard deviation: 0.1		
Conv2D padded	Number of kernels: 8	Kernel size: (3, 3)	Activation: none
BatchNormalization	default parameters		
Activation	Type: ReLU		
MaxPooling	Filter size: (2, 2)		
Conv2D padded	Number of kernels: 8	Kernel size: (3, 3)	Activation: none
BatchNormalization	default parameters		
Activation	Type: ReLU		
MaxPooling	Filter size: (2, 2)		
Conv2D padded	Number of kernels: 16	Kernel size: (3, 3)	Activation: none
BatchNormalization	default parameters		
Activation	Type: ReLU		
MaxPooling	Filter size: (2, 2)		
Conv2D padded	Number of kernels: 32	Kernel size: (3, 3)	Activation: none
BatchNormalization	default parameters		
Activation	Type: ReLU		
MaxPooling	Filter size: (2, 2)		
Flatten			
Dense	Units: 128	Activation: ReLU	
Dropout	Probability: 0.5		
Dense	Units: 1	Activation: Sigmoid	

Table 5: Architecture 1 for registration network

[10] For those curious, the YouTube data consisted of twenty videos. Some of these videos were of the aforementioned "voice crack compilations" available in great numbers online, while others were similarly vindictive "tenor voice crack" videos focusing on operatic styles of music relevant to this project. Others were of notably bad *American Idol* auditions the author remembered watching, while still others consisted of people singing poorly on purpose for extended periods of time. Most of the other videos consisted of examples of yodeling, a singing style in which register breaks are not only allowed, but exploited artistically for highly pleasing results [126]. Some of these latter videos were found using AudioSet.

Layer	Parameters		
GaussianNoise	Standard deviation: 0.1		
Conv2D padded	Number of kernels: 8	Kernel size: (3, 3)	Activation: ReLU
MaxPooling	Filter size: (2, 2)		
Dropout	Probability: 0.5		
Conv2D padded	Number of kernels: 8	Kernel size: (3, 3)	Activation: ReLU
MaxPooling	Filter size: (2, 2)		
Dropout	Probability: 0.5		
Conv2D padded	Number of kernels: 16	Kernel size: (3, 3)	Activation: ReLU
MaxPooling	Filter size: (2, 2)		
Dropout	Probability: 0.5		
Conv2D padded	Number of kernels: 32	Kernel size: (3, 3)	Activation: ReLU
MaxPooling	Filter size: (2, 2)		
Dropout	Probability: 0.5		
Flatten			
Dense	Units: 128	Activation: ReLU	
Dropout	Probability: 0.5		
Dense	Units: 1	Activation: Sigmoid	

Table 6: Architecture 2 for registration network

Parameter name	Both architectures
Optimizer type	Adam
Loss type	Binary crossentropy
Class weights	1 : 1
Image dimensions	(128, 173)
Training samples per epoch	1846
Validation samples per epoch	461

Table 7: Training parameters for both registration network architectures

5.2 Results

Best use-case results for vocal registration networks with both architectures are presented here, alongside confusion matrices for both networks.

Metric name	Architecture 1	Architecture 2
Number of epochs trained	1	8
Training set accuracy	66.20%	71.43%
Validation set accuracy	89.78%	81.63%
Use case accuracy	73.68%	68.42%
Normalized use case accuracy	62.86%	75.72%

Table 8: Training results for both registration network architectures

Neither of these networks perform exceptionally well on training, validation, or use case set accuracies. Moreover, the fact that the network with architecture 1 was only allowed one epoch of training on its dataset causes a natural suspicion that its superior use case accuracy is just a fluke of meaningless underfitting to a bad dataset. It is natural that the network with architecture 2 would have slightly better training set accuracy, because more epochs of training had elapsed at the time of its preservation. Comparing validation set accuracies suggests the occurrence of overfitting in architecture 2.

Something here must be noted in use case accuracy. Since the use case recording for broken registration does not consist entirely of voice cracks, the frequency of voice cracks in both use case recordings is much lower than the frequency of smooth registration in the use cases. Therefore, to provide good, interpretable results in vocal registration classification, one must weight the networks' classifications of broken registration samples more heavily than their classifications of smooth registration samples. In confusion matrices, after all, it is really the probabilities of label classifications that matter more than the precise numbers of data points classified correctly and incorrectly. This operation of data normalization shows, as displayed in the confusion matrices in Figures 6 and 7, that although the network with architecture 1 appears initially to have better use case accuracy, architecture 2 yields more useful results in classifying the use case set of recordings.

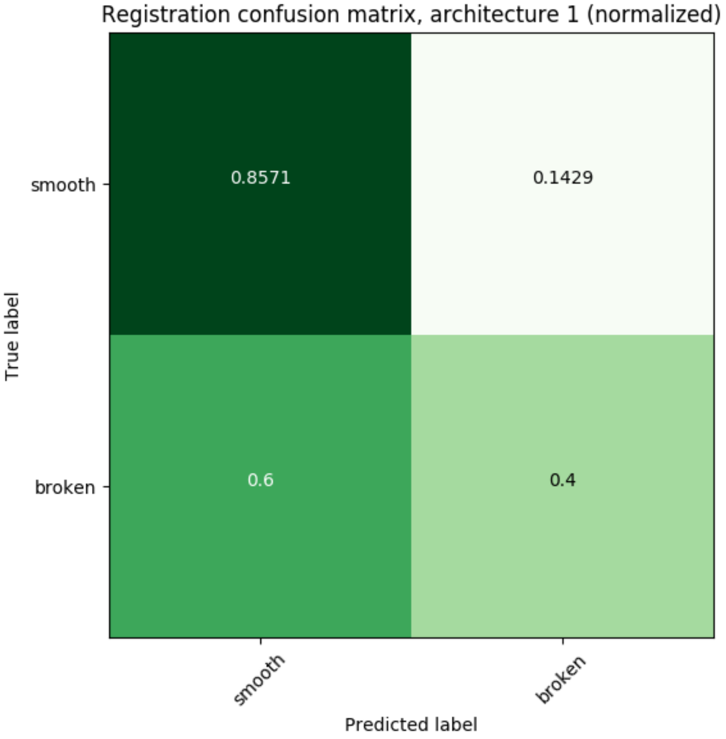


Figure 6: Confusion matrix for registration network on use case, architecture 1

Figure 8 shows the typical results of training a neural network on this dataset, illustrating how its training, validation, and use case set accuracies fluctuate over time. One can see that, as different epochs of training elapse, the training set accuracy can improve to arbitrarily high quantities, with only moderate losses in the validation set accuracy. The model quickly loses its relevance to use case set classification, though, over time. When training and validation set accuracies can improve over time, but test set accuracies cannot improve, it is indicative that the dataset on which one is training one's classification model is of very poor quality and does not accurately reflect realistic environmental scenarios. As one brief, positive note, though, the normalized confusion matrix for the network with architecture 2 displays a relatively good facility of the network in classifying both labels. Given the low confidence in many of the predictions made here, this is probably a fluke of training likely not to generalize out to larger sets of typical use case data. However, this fact is at least worth noting.

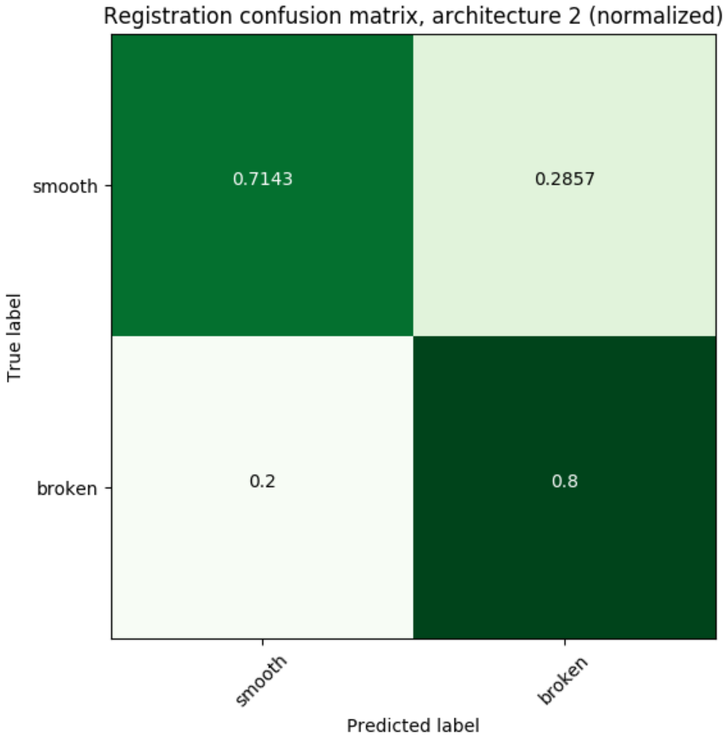


Figure 7: Confusion matrix for registration network on use case, architecture 2

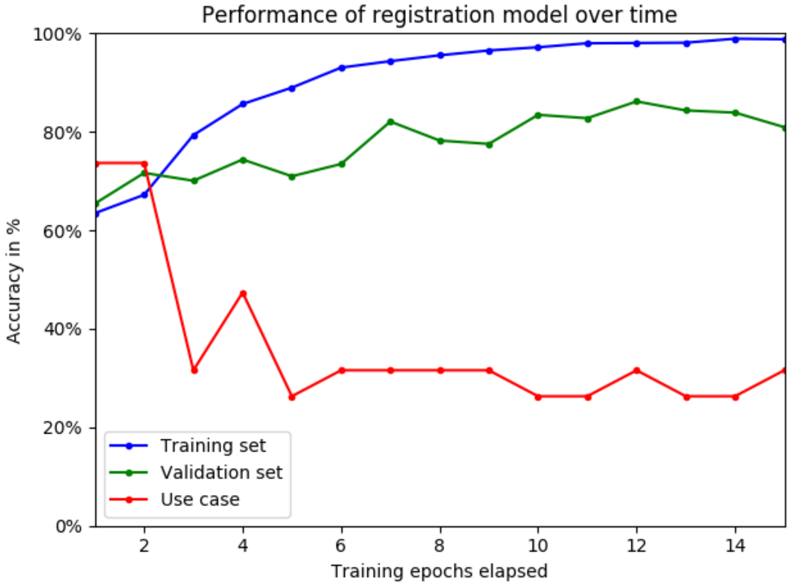


Figure 8: Accuracy of typical registration network over 20 epochs of training

5.3 Discussion

These results are overall rather discouraging for the feasibility of these types of networks to perform well on these types of data, but the dataset provided in training the networks is so small, so noisy, and so poorly assembled that perhaps this could simply instead be the result of a bad dataset. Perhaps both obstacles contribute in their own way to the general failure of these networks in distinguishing between smooth and broken laryngeal registration. It is difficult to conclusively say anything for sure about these results due to the large number of potential pitfalls in the training process and data supply. Many manipulations on the dataset employed here and architectures used in training networks on this dataset provided even worse results than the ones presented here.

Having normalized the confusion matrices and reasoned above that false positives are better than false negatives in one of these classifiers, the network with architecture 2 appears to be the superior network for inclusion in a potential automated vocal coach than the false-negative-biased network with architecture 1. This is not to say that either network appears ready for deployment in this hypothetical application, though. Ideally, other network types would need to be trained on much larger and better datasets until they yielded acceptable results for inclusion in an automated vocal coach. Perhaps other learning paradigms could be explored, operating on different representations of audio data. There are many potential approaches to improving these results, but they are outside the scope of this project.

6 Resonance Management

Results and analysis of neural networks for studying and discerning proper resonance management from audio recordings of amateur and professional singers are presented here.

6.1 Implementation

The hypothetical resonance network outlined in the implementation section of this paper is a binary classifier that distinguishes between recordings of the singing voice that exhibit the "singer's formant," or *squillo*, and recordings that do not exhibit this phenomenon. As opposed to the criteria outlined in the sections above, this criterion is a relatively easy classification to learn even using a small dataset. It is a consistent feature of singing across large spans of time, unlike the results of improper laryngeal registration. Not only that, it is a very easy feature of the human voice to see in a spectrogram. One simply looks for a bright area in the right frequency band in a recording to detect its presence or absence.

The dataset selected for this binary classification of resonance was a combination of VocalSet and Bozkurt et al.'s Singing Voice Assessment dataset. VocalSet was used to supply the network being trained with samples of professional singers with levels of the singer's formant idiomatic to the Western classical style, while the Singing Voice Assessment dataset was used to supply the network with audio clips of amateur singers without idiomatic resonance. Some recordings of vocalists singing Chinese opera in Black et al.'s Singing Voice Dataset were added to the Singing Voice Assessment recordings under the "no-squillo" label to help balance the classes of the training data, which were at first heavily biased towards the "squillo" label. Only the vocalists in the Singing Voice Dataset determined to have low levels of a Western singer's formant were included here. The dataset assembled was somewhat small, but random sampling of its data suggested that it was of high quality.

To assist in the maximum facility of classification, and since the data for the "squillo" label was still a bit more prevalent than for the "no-squillo" label after this, some minor data transformation via filtering was performed upon some of the VocalSet data, effectively removing the squillo from the vocalists, and the resulting audio was added to the "no-squillo" data. The use case set was prepared using this technique as well, since it was rather difficult for the author of this paper to sing in a similar manner without some level of squillo in his voice. This edited data did not sound artificial; rather, it sounded like recordings of real singers with somewhat ineffective strategies of voice resonance from the standpoint of the Western classical tradition. The best use-case network architecture is presented below. It is a very simple network attempting to learn an extremely simple feature of spectrogram representations of recordings of the singing voice on a small dataset. As a result it employs high amounts of regularization on its layers, including applying Gaussian noise to its initial input images and employing both batch normalization and dropout on its hidden layers.

The architectures and training parameters of this network are presented in detail in tables 9 and 10. It operates by first learning low-level horizontal features of the input data to accentuate areas of high resonance within the spectrograms, and then it performs a vertical convolution on the resulting data, finding the exact frequency bands of high resonance areas within a recording of the voice before making its final judgments.

6.2 Results

This network only needed to be trained for six epochs to attain high classification accuracy on training, validation, and typical use case sets, as shown in table 11 and figure 9. It is the only one of the network architectures displayed in this paper to perform perfectly on a use case that might be common in the context of an automated vocal coach. In fact, it was highly confident in all its correct predictions made on the typical use case recordings. Its practical performance could almost certainly be improved with the inclusion of a

Layer	**Parameters**		
GaussianNoise	Standard deviation: 0.01		
Conv2D padded	Number of kernels: 8	Kernel size: (1, 3)	Activation: none
BatchNormalization	default parameters		
Activation	Type: ReLU		
MaxPooling2D	Filter size: (1, 2)		
Dropout	Probability: 0.5		
Conv2D valid	Number of kernels: 16	Kernel size: (128, 3)	Activation: none
BatchNormalization	default parameters		
Activation	Type: ReLU		
MaxPooling2D	Filter size: (1, 8)		
Dropout	Probability: 0.5		
Flatten			
Dense	Units: 128	Activation: ReLU	
Dropout	Probability: 0.5		
Dense	Units: 1	Activation: Sigmoid	

Table 9: Architecture of resonance management network

Parameter name	**Value**
Optimizer type	Adam
Loss type	Binary crossentropy
Class weights	1 : 1
Image dimensions	(128, 87)
Training samples per epoch	2500
Validation samples per epoch	750

Table 10: Training parameters of resonance management network

larger training dataset and more robust test set data, but from all of the statistics available to this author, the training of this network appears to have been an unreserved successs.

Metric name	**Value**
Number of epochs trained	6
Training set accuracy	92.96%
Validation set accuracy	92.57%
Use case accuracy	100.00%

Table 11: Training results for resonance management network

6.3 Discussion

This is the only one of the criteria explored in this project whose contrastive data were able to conclusively be classified correctly by a trained network through experimentation. More statistical testing would need to be done to ensure that this network would be accurate on a variety of different voices to warrant its unmodified inclusion in an automated vocal coach. Particularly interesting would be to see if it performs well in distinguishing typical use case recordings of trained and untrained sopranos, the existence of the singer's formant in whom is debated as discussed in a previous section. Even without this exploration, this is undoubtedly the best candidate network explored here. Moreover, the small size of this network (772 kB) and its simplicity make it lightweight and quick to operate. For all these reasons and more, it is an ideal candidate for practical use.

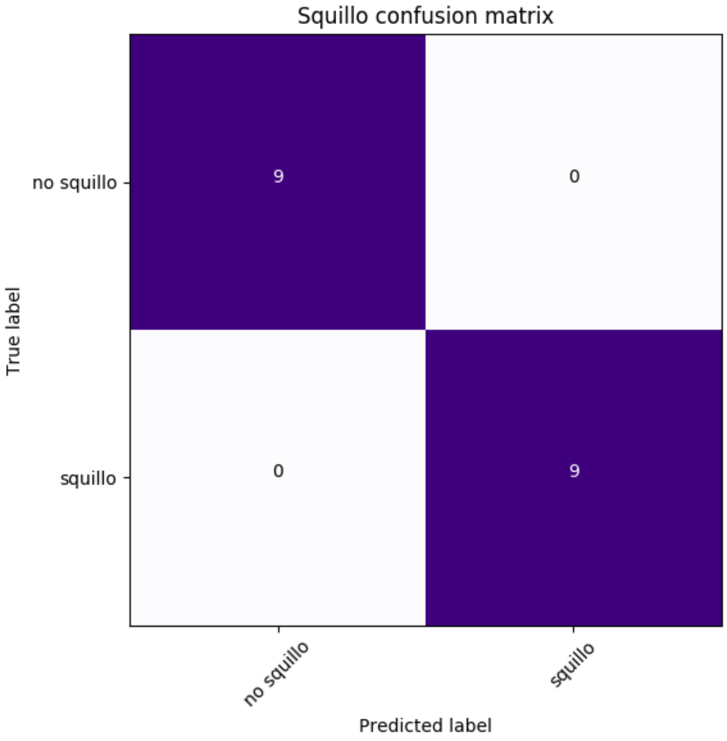


Figure 9: Confusion matrix for resonance management network on use case. As can clearly be seen, there was not much confusion here.

7 Vibrato

Results and analysis of neural networks for studying and discerning proper vibrato management from audio recordings of amateur and professional singers are presented here.

7.1 Implementation

In the section of this paper devoted to research in vibrato, a hypothetical simple vibrato classifier was described capable of making distinctions between straight-tone singing, idiomatic vibrato singing, and unidiomatic vibrato singing in the Western classical tradition. The dataset selected for this ternary classification of vibrato styles was a combination of VocalSet and Black et al.'s Singing Voice Dataset. Different segments of VocalSet were used for demonstrating singing with straight-tone and idiomatic vibrato styles. In the absence of many datasets of singers with wide and slow vibratos singing in the Western style, the wide and slow vibratos in the Chinese opera recordings were fed to the network as examples of unidiomatic, "heavy" vibratos. This was the largest dataset explored in this project. There is only one lamentable fact about this dataset, and it is that recordings of isolated vocalists singing with unacceptably fast vibratos, surprisingly difficult to find online, are not included within it. Overlooking this issue, though, the quality of the dataset is rather high relative to some others in the project.

Classification of this criterion appeared to be more complex than with resonance management above, but a good deal easier than either of the two criteria preceding that. Vibrato is a phenomenon that has both time and frequency dependencies; a regular, uniform oscillation of the pitch and timbre of the voice appearing as a wiggling line on a spectrogram is necessary to identify it. In ideal Western classical singing, these wiggling lines occur with high regularity throughout the frequency spectrum. The wiggling lines of unidiomatic vibrato are of inconsistent and/or excessive depth or speeds that are too slow and/or fast. In clear-cut cases of straight-tone singing, they are absent from a spectrogram entirely.

An architecture and training parameters for the best use-case network explored in this project are presented below in tables 12 and 13 respectively. This is quite a simple network, consisting of two convolutional and two max-pooling layers applied directly to two fully connected layers before classifications are ultimately made. The shape of the kernels of the first convolutional filter is key to understanding how the network operates. These kernels convolve on rectangular, somewhat horizontal portions of data whose frequency bands are roughly of the same width as the widest vibrato one is likely to encounter in classification. This data is compressed in the horizontal axis before being fed to a layer whose kernels learn more high-level data about the distribution of vibrato lines throughout frequency ranges in the spectrogram, and the outputs of *this* convolutional layer are heavily vertically compressed before being fed into the network's first fully connected layer.

Layer	**Parameters**		
Conv2D valid	Number of kernels: 16	Kernel size: (3, 7)	Activation: ReLU
MaxPooling2D	Filter size: (2, 4)		
Conv2D valid	Number of kernels: 32	Kernel size: (3, 3)	Activation: ReLU
MaxPooling2D	Filter size: (8, 4)		
Flatten			
Dense	Units: 128	Activation: ReLU	
Dense	Units: 3	Activation: Softmax	

Table 12: Architecture of vibrato network

Parameter name	**Value**
Optimizer type	Adam
Loss type	Categorical crossentropy
Class weights	1.33 : 1.65 : 1
Image dimensions	(128, 87)
Training samples per epoch	3000
Validation samples per epoch	1000

Table 13: Training parameters of vibrato network

7.2 Results

The results of this network on training, validation, and use-case data are presented in table 14. This network took more epochs to train than any other network displayed in this project. It exhibited somewhat mixed results on the use case set, although its training and validation set accuracy were both somewhat high. Greater detail of this network's performance on the use case set is summarized in the confusion matrix shown in figure 10.

Metric name	**Value**
Number of epochs trained	19
Training set accuracy	90.80%
Validation set accuracy	88.14%
Use case accuracy	57.69%

Table 14: Training results for vibrato network

In the confusion matrix presented in figure 10, one sees that this network identifies straight-tone singing with relatively high probability, although it strangely identifies some parts of straight-tone singing as unidiomatic

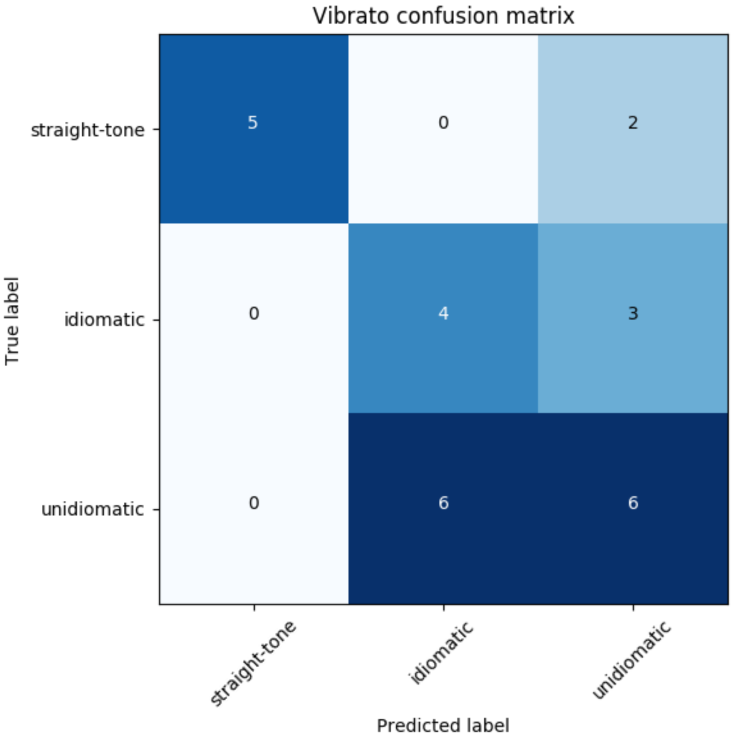


Figure 10: Confusion matrix for vibrato network on use case

vibrato. Its classification performance between recordings identified as exhibiting idiomatic and unidiomatic vibrato, however, is little better than random chance. Just as in the confusion matrix for the large-image phonation network discussed above, we seem to have produced a good binary classifier with a ternary dataset.

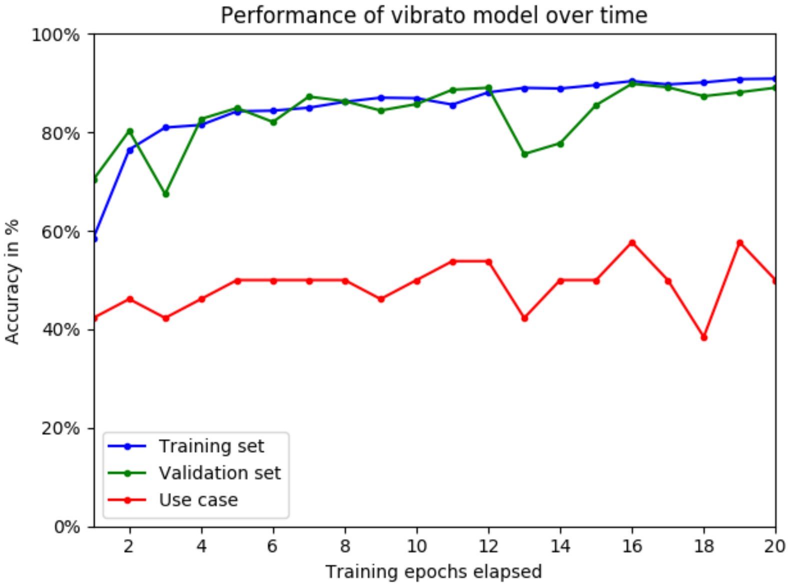


Figure 11: Accuracy of best vibrato network over 20 epochs of training

As a contrast with the graph provided in the registration experiments in this paper, another graph detailing performance of the vibrato network over time on training, accuracy, and test datasets is presented in figure 11. This graph effectively shows that while the use case accuracy of this network very much lags behind both training and validation set accuracy, there is a correlation between the contours of validation set and use case set accuracies over time. This correlation is very surprising, given that the data in training and validation sets are not from a common source and, other than the vibrato strategies they exhibit, are not related to each other in any way. The results of this graph suggest that the classifications that this network is making are valid and result from a dataset that is accurate to real-world use cases.

7.3 Discussion

While results of the confusion matrix in figure 10 are not especially encouraging, the results shown in figure 11 suggest that the network's performance can probably be improved upon through further experimentation. This could come in the form of training the existing network over the course of more epochs; it appears that even at the end of the graph in figure 11, the training and validation set performances are still trending upward over time. Due to the correlation between the validation and use case set accuracies, this probably also means that use case accuracy could improve with time as well. However, the constant displacement between use case and validation set accuracies suggest that perhaps the differential classification of idiomatic and unidiomatic vibrato will remain a problem for the network regardless.

To solve this problem, one could try to apply regularization to the network or experiment with its architecture. However, many different architectures have already been tried for classifying this criterion with worse use-case accuracies in all of these attempts, and any attempt at regularization in this network has tended to greatly decrease all three levels of accuracy for reasons unknown. Failing all these attempts to improve use case set performance, one strategy remains: to improve the quality and size of the dataset used in training this network, which can always be done with more time investment. Hopefully this could be done at least partially by adding unidiomatic vibrato that is too fast to the dataset to complete its breadth, as mentioned above.

Not exhibiting a stable binary distinction between idiomatic and unidiomatic vibrato, this network is not yet suitable for inclusion in an automated vocal coach. As is reasoned in the section for the small-image phonation mode network above, though, it could be part of a system used in a final product if it were made to develop a highly accurate binary distinction at least between straight-tone and idiomatic/unidiomatic singing and were used in conjunction with a network that accurately distinguished between idiomatic and unidiomatic vibrato types. This is also a long way off and would probably require a better dataset, but it could relieve some of the pressures on a single network to make an efficient three-way distinction.

8 Conclusions

The results of this highly interdisciplinary research can be examined according to several different metrics. It is true that the results of this research do not lay very good solid groundwork for the concrete construction of a real-life automated vocal coach. Of the four criteria characteristic of Western classical singing analyzed in this project, only one network in one area was judged to perform well in distinguishing between idiomatic and unidiomatic singing. The distinction this successful network learned was a basic one of looking at relative sound pressure in a particular frequency area, so it was not surprising that it worked well. Some other classifiers appeared to be on the edge of learning to properly label typical use-case data, but were not quite capable of meeting their classification goals. These limited suggest that future research might have some promise in these areas.

Performances in the three less successful categories were hampered severely by small, noisy, and irregularly structured datasets cobbled together by necessity when it was found infeasible to build better datasets. These datasets impaired networks by misrepresenting real-world data. Accurately measuring the performance of convolutional neural networks on singing-voice spectrograms in idiomatic Western classical singing classification, absent any other confounding factors, requires rigorously assembled and large datasets. Ideal datasets would contain high-fidelity audio to which noise could manually be added later; each of their accurately balanced labels would contain a diverse but accurate representation of many different voices singing many different songs in similarly diverse styles. Putting just one of these datasets together would be an extremely time-intensive endeavor, but would probably be necessary to warrant extensive further research into this precise topic. It seems likely, though, that the outputs of convolutional neural networks can only be as good as the input fed into them, and so criteria that are difficult to estimate visually from spectrograms might still suffer from inaccuracies and overgeneralizations in neural network classification. Register breaks may particularly belong to this class of problematic criteria for convolutional neural networks.

Other limitations on networks' performance in this research may be due to restrictive factors imposed upon the project in terms of the general types of networks used to classify audio, as well as the specific audio representation types fed to these networks. Aside from convolutional networks, recurrent neural networks are often used in processing audio. Complex recurrent neural networks are sometimes employed directly atop audio signals [60, 59]. One could try even other combinations of network and representation type, such as one-dimensional convolutional networks applied directly to audio signals or recurrent networks applied to one-dimensional outputs of convolutional networks on spectrograms. As music and singing are both quite time-dependent phenomena, it makes sense that there might be a performance benefit from using recurrent networks on more time-dependent criteria. It is also possible that performance benefits may be enjoyed by feeding only grayscale spectrograms into the same architectures outlined above or using time-overlapping frames of audio spectrogram data in building datasets and changing nothing else; these results seem unlikely but are technically possible.

While this project does not come close to building an automated vocal coach, it succeeds more in providing a foundation upon which similar research can be based in the future. As the first published project to the author's knowledge demonstrating applications of machine learning to vocal pedagogy, it contains a unique synthesis of qualitative data about both topics that provides insights and promising directions for future research. For instance, it provides novel intuitions on the proper shaping of convolutional neural network kernels for the processing of certain features of the human voice. While the specific networks have not yet been trained very well, there is evidence particularly in vibrato and phonation mode classification that the network structures used here show promise in future classification experiments on larger and more robust datasets. This is exemplified by data such as the graph in figure 11 that correlates validation set and typical use case performances that improve over time, motivating further research with better input data.

Supposing one can build networks in the future that accurately distinguish idiomatic and unidiomatic Western classical singing according to these criteria, a number of future steps would be possible for the project. To build an automated vocal coach that is powered by machine learning and serves a broadly useful purpose to amateur vocalists, many more criteria must be capable of being classified by machines. These include, but are not limited to, many types of management of the timbral components of the acoustic singing voice unrelated to the singer's formant, incremental laryngeal registration shifts as discussed briefly above, and identification of various note articulations such as *legato*, *marcato*, and *staccato* notes. They could even include higher-level features of music such as the accurate pronunciation of consonants and vowels of different languages being sung, identification of volume variations in the voice over time, and even extremely high-level features such as emotional expression in singing. The types of future experiments are only limited by one's imagination.

Once enough criteria are properly classified by portable or server-based neural networks to justify inclusion in a final product, a great number of surveys of vocal scientists and voice pedagogues must be conducted in conjunction with a large review of all the scientific literature published on effective singing voice pedagogical techniques. Some useful advice for voice students struggling with areas of discomfort or uncertainty in executing idiomatic singing in the criteria presented in this project are discussed above, but they are only a start to bridging the massive gap between recognizing features of the voice and learning to provide advice on improving it. It is only once this research is conducted and digested that useful recommendations can be made to students based off of the classifications made by this application, and a final product can finally be presented. The work ahead is thorny and difficult in many ways, but the data already published are encouraging and suggest many future research directions. It is this author's great hope to return to this project with more time and resources to spare on it in future years, maybe at a time when greater advances in computer audition have been achieved. With lots of hard work and some amount of luck, perhaps this project can play a part in teaching the world to sing.

References

[1] A. Naseth, “Constructing the Voice: Present and Future Considerations of Vocal Pedagogy,” *Choral Journal*, vol. 53, no. 2, pp. 39–49, 2012.

[2] S. McCoy, “Too Many Singers?” *Journal of Singing – The Official Journal of the National Association of Teachers of Singing*, vol. 73, no. 4, pp. 403–405, 2017.

[3] J. Ongkasuwan and M. S. Courey, “Care of the professional voice,” in *Bailey’s Head and Neck Surgery: Otolaryngology*, 5th ed., J. T. Johnson and C. A. Rosen, Eds. Lippincott Williams & Wilkins, 2014, ch. 72.

[4] P. Kiesgen, “Voice Pedagogy: ”Well, Vocal Pedagogy Is All Subjective Anyway, Isn’t It?”,” *Journal of Singing – The Official Journal of the National Association of Teachers of Singing*, vol. 62, no. 1, pp. 41–44, 2005.

[5] T. Radomski, “Manuel García (1805-1906): A Bicentenary Reflection,” *Australian Voice*, vol. 11, pp. 25–41, 2005.

[6] J. Sundberg, “Articulatory interpretation of the “singing formant”,” *The Journal of the Acoustical Society of America*, vol. 55, no. 4, pp. 838–844, 1974.

[7] K. W. Bozeman, *Practical Vocal Acoustics: Pedagogic Applications for Teachers and Singers.* Hillsdale, New York: Pendragon Press, 2013.

[8] T. Shipp and R. E. McGlone, “Laryngeal Dynamics Associated with Voice Frequency Change,” *Journal of Speech and Hearing Research*, vol. 14, no. 4, pp. 761–768, 1971.

[9] I. R. Titze, E. S. Luschei, and M. Hirano, “Role of the Thyroarytenoid Muscle in Regulation of Fundamental Frequency,” *Journal of Voice*, vol. 3, no. 3, pp. 213–224, 1989.

[10] D. M. Hull, “Thyroarytenoid and cricothyroid muscular activity in vocal register control,” Ph.D. dissertation, University of Iowa, 2013.

[11] E. Björkner, “Musical Theater and Opera Singing-Why So Different? A Study of Subglottal Pressure, Voice Source, and Formant Frequency Characteristics,” *Journal of Voice*, vol. 22, no. 5, pp. 533–540, 2008.

[12] P. M. Pestana, S. Vaz-freitas, and M. C. Manso, “Trends in Singing Voice Research: An Innovative Approach,” *Journal of Voice*, in press.

[13] “Voiceprint Download,” 2019. [Online]. Available: https://store.estillvoice.com/collections/clinical-software/products/copy-of-voiceprint-cd-mac-edition-1

[14] N. Miller, “History,” 2014. [Online]. Available: http://www.vocevista.com/history/

[15] “What People Pay for Music Lessons: 2014 National Report,” 2014. [Online]. Available: https://support.takelessons.com/hc/en-us/article_attachments/200377329/What-People-Pay-for-Music-Lessons.pdf

[16] “Thumbtack – Start a Project,” 2018. [Online]. Available: https://www.thumbtack.com/

[17] P. Michaels and S. Wolpin, “Best Cellphone Plans 2018,” 2018. [Online]. Available: https://www.tomsguide.com/us/best-phone-plans,review-2953-2.html

[18] O. Brown, *Discover Your Voice.* Clifton Park, NY: Delmar Cengage Learning, 1996.

[19] J. P. Dworkin, “Laryngitis: Types, Causes, and Treatments,” *Otolaryngologic Clinics of North America*, no. 41, pp. 419–436, 2008.

[20] A. G. Foote, “Contemporary Commercial Music (CCM) Singers: Lifestyle Choices and Acoustic Measures of Voice,” Ph.D. dissertation, University of Buffalo, 2015.

[21] P. Newham, "Jung and Alfred Wolfsohn: Analytical Psychology and the Singing Voice," *Journal of Analytical Psychology*, no. 37, pp. 323–336, 1992.

[22] R. E. Radocy, "Review: Singing and Self: The Psychology of Performing by Sarah Elizabeth Stedman," *Bulletin of the Council for Research in Music Education*, no. 100, 1989.

[23] J. S. Rubin, L. Mathieson, and E. Blake, "Care of the Professional Voice: Posture and Voice," *Journal of Singing – The Official Journal of the National Association of Teachers of Singing*, vol. 60, no. 3, pp. 271–275, 2004.

[24] S. McCoy, "Building the Foundation," *Journal of Singing – The Official Journal of the National Association of Teachers of Singing*, vol. 67, no. 1, pp. 43–46, 2010.

[25] "SINGPRO: Mobile Vocal Training," 2017. [Online]. Available: https://www.singpro.com/

[26] "Vanido: Your Personal Singing Coach," 2017. [Online]. Available: https://vanido.io/

[27] "SING&SEE: See Your Voice - Hear the Difference," 2019. [Online]. Available: https://www.singandsee.com/

[28] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A Comparative Performance Study of Several Pitch Detection Algorithms," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 399–418, 1976.

[29] M. Dziubinski and B. Kostek, "Octave Error Immune and Instantaneous Pitch Detection Algorithm," *Journal of New Music Research*, vol. 34, no. 3, pp. 273–292, 2005.

[30] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed., M. Jordan, J. Kleinberg, and B. Schölkopf, Eds. New York City: Springer Science+Business Media, LLC, 2006.

[31] J. Schmidhuber, "Deep Learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.neunet.2014.09.003

[32] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Singing-Voice Separation From Monaural Recordings Using Deep Recurrent Neural Networks," 2014. [Online]. Available: http://www.ifp.illinois.edu/~huang146/papers/DRNN_ISMIR2014.pdf

[33] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," pp. 1–15, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[34] E. J. Heller, *Why you hear what you hear : an experiential approach to sound, music, and psychoacoustics.* Princeton: Princeton University Press, 2013.

[35] M. S. Mackinlay, *Garcia the Centenarian and His Times.* Edinburgh and London: William Blackwood and Sons, 1908.

[36] A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal*, vol. 3, no. 3, pp. 535–554, 1959.

[37] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2010.

[38] M. P. Mattson, "Superior pattern processing is the essence of the evolved human brain," *Frontiers in Neuroscience*, vol. 8, no. 265, pp. 1–17, 2014.

[39] T. Dietterich, "Overfitting and undercomputing in machine learning," *ACM Computing Surveys*, vol. 27, no. 3, pp. 326–327, 1995.

[40] W. Gerstner, "Hebbian Learning and Plasticity," in *From Neuron to Cognition via Computational Neuroscience*, M. Arbib and J. Bonaiuto, Eds. Cambridge, MA: The MIT Press, 2016, pp. 199–219. [Online]. Available: https://pdfs.semanticscholar.org/f9fc/99a5c52aa5df1b530dfdeb25dfb6b10bdecf.pdf

[41] W. S. McCulloch and W. Pitts, "A Logical Calculus of the Ideas Immanent in Nervous Activity," *Bulletin of Mathematical Biophysics*, vol. 5, no. 4, pp. 115–133, 1943. [Online]. Available: http://journals2.scholarsportal.info/pdf/00928240/v52i1-2/99{_}alcotiiina.xml

[42] N. Rochester, J. H. Holland, L. H. Haibt, and W. L. Duda, "Tests on a Cell Assembly Theory of the Action of the Brain, Using a Large Digital Computer," *IRE Transactions on Information Theory*, vol. 2, no. 3, pp. 80–93, 1956.

[43] F. Rosenblatt, "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

[44] D. O. Hebb, *The Organization of Behavior: A Neuropsychological Theory.* New York City: John Wiley and Sons, 1949.

[45] P. J. Werbos, *The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting.* J. Wiley & Sons,, 1994. [Online]. Available: http://hdl.handle.net/2027/mdp.39015032742291

[46] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," pp. 1–18, 2012. [Online]. Available: http://arxiv.org/abs/1207.0580

[47] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," 2015.

[48] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How Does Batch Normalization Help Optimization?" in *32nd Conference on Neural Infromation Processing Systems*, Montréal, Canada, 2018. [Online]. Available: http://arxiv.org/abs/1805.11604

[49] "TensorFlow," 2019. [Online]. Available: https://www.tensorflow.org/

[50] "Torch: Scientific Computing for LuaJIT," 2019. [Online]. Available: http://torch.ch/

[51] "Home – Keras Documentation," 2019. [Online]. Available: https://keras.io/

[52] "Google Scholar," 2019. [Online]. Available: https://scholar.google.com/

[53] "HOLLIS," 2019. [Online]. Available: https://hollis.harvard.edu/

[54] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[55] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[56] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *International Conference on Learning Representations*, San Diego, CA, 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[57] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[58] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2017, pp. 131–135.

[59] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[60] A. Graves, A. R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, no. 6. IEEE, 2013, pp. 6645–6649.

[61] H. Lee, Y. Largman, P. Pham, and A. Y. Ng, "Unsupervised Feature Learning for Audio Classification Using Convolutional Deep Belief Networks," in *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, ser. NIPS'09. USA: Curran Associates Inc., 2009, pp. 1096–1104. [Online]. Available: http://dl.acm.org/citation.cfm?id=2984093.2984217

[62] K. J. Piczak, "Environmental Sound Classification with Convolutional Neural Networks," in *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*. Boston: IEEE, 2015.

[63] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of Physiology*, vol. 148, no. 3, pp. 574–591, 1959.

[64] ——, "Receptive Fields and Functional Architecture of Monkey Striate Cortex," *Journal of Physiology*, vol. 195, no. 1, pp. 215–243, 1968.

[65] K. Fukushima, "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position," *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[66] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into Imaging*, vol. 9, no. 4, pp. 611–629, 2018.

[67] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[68] M. García, *Garcia's New Treatise on the Art of Singing*. Boston: Oliver Ditson, 1872.

[69] C. Callender, "The Vocal Legacy of Oren Brown," *La Scena Musicale*, vol. 10, no. 3, 2004. [Online]. Available: http://www.scena.org/lsm/sm10-3/owen-brown.htm

[70] "Kenneth Bozeman," 2019. [Online]. Available: https://www.lawrence.edu/conservatory/faculty/kenneth_bozeman

[71] C. W. Thorpe, S. J. Cala, J. Chapman, and P. J. Davis, "Patterns of breath support in projection of the singing voice," *Journal of Voice*, vol. 15, no. 1, pp. 86–104, 2001.

[72] M. S. Benninger, "The professional voice," *Journal of Laryngology and Otology*, vol. 125, no. 2, pp. 111–116, 2011.

[73] J. Sundberg, *The Science of the Singing Voice*. DeKalb, IL: Northern Illinois University Press, 1987.

[74] C. Rhodes and T. Crawford, "Breathy or Resonant – a Controlled and Curated Dataset for Phonation Mode Detection in Singing," in *13th International Society for Music Information Retrieval Conference*. Porto, Portugal: International Society for Music Information Retrieval, 2012, pp. 589–594.

[75] H. Hollien, "On Vocal Registers," *Journal of Phonetics*, vol. 2, no. 2, pp. 125–143, 1974.

[76] K. A. Kochis-Jennings, E. M. Finnegan, H. T. Hoffman, and S. Jaiswal, "Laryngeal muscle activity and vocal fold adduction during chest, chestmix, headmix, and head registers in females," *Journal of Voice*, vol. 26, no. 2, pp. 182–193, 2012.

[77] M. Echternach, S. Dippold, J. Sundberg, S. Arndt, M. F. Zander, and B. Richter, "High-speed imaging and electroglottography measurements of the open quotient in untrained male voices' register transitions," *Journal of Voice*, vol. 24, no. 6, pp. 644–650, 2010.

[78] D. G. Miller and H. K. Schutte, "'Mixing' the registers: Glottal source or vocal tract?" *Folia Phoniatrica et Logopaedica*, vol. 57, no. 5-6, pp. 278–291, 2005.

[79] V. E. Negus, O. Jander, and P. Giles, “Falsetto,” 2001. [Online]. Available: http://www.oxfordmusiconline.com/grovemusic/view/10.1093/gmo/9781561592630.001.0001/omo-9781561592630-e-0000009270

[80] N. Y. Li and E. M. Yiu, “Acoustic and perceptual analysis of modal and falsetto registers in females with dysphonia,” *Clinical Linguistics and Phonetics*, vol. 20, no. 6, pp. 463–481, 2006.

[81] I. R. Titze, “Acoustic Interpretation of Resonant Voice,” *Journal of Voice*, vol. 15, no. 4, pp. 519–528, 2001.

[82] G. Bloothooft and R. Plomp, “The sound level of the singer’s formant in professional singing,” *The Journal of the Acoustical Society of America*, vol. 79, no. 2028, pp. 2028–2033, 1986.

[83] J. Sundberg, “Level and center frequency of the singer’s formant,” *Journal of Voice*, vol. 15, no. 2, pp. 176–186, 2001.

[84] S. Komiyama, “A Measurement of Equal-Loudness Level Contours for Tone Burst,” *Acoustical Science and Technology*, vol. 22, no. 1, pp. 35–39, 2001.

[85] K. Gladiné and J. J. Dirckx, “Average middle ear frequency response curves with preservation of curve morphology characteristics,” *Hearing Research*, vol. 363, no. 2018, pp. 39–48, 2018.

[86] W.-H. Su, “An Acoustic Study of the Singer’s Formant: The Comparison Between Western Classical and Traditional Chinese Opera Singing Techniques,” Ph.D. dissertation, Indiana University, 2009.

[87] W. T. Bartholomew, “A Physical Definition of “Good Voice-Quality” in the Male Voice,” *The Journal of the Acoustical Society of America*, vol. 6, no. 25, pp. 224–224, 1934.

[88] A. Kirkpatrick, “Chiaroscuro and the Quest for Optimal Resonance,” *Journal of Singing: The Official Journal of the National Association of Teachers of Singing*, vol. 66, no. 1, pp. 15–21, 2009.

[89] B. Delvaux and D. Howard, “A new method to explore the spectral impact of the piriform fossae on the singing voice: Benchmarking using MRI-Based 3D-printed vocal tracts,” *PLoS ONE*, vol. 9, no. 7, 2014.

[90] E. Yanagisawa, J. Estill, S. T. Kmucha, and S. B. Leder, “The Contribution of Aryepiglottic Constriction to with Acoustic Analysis,” *Journal of Voice*, vol. 3, no. 4, pp. 342–350, 1989.

[91] R. Weiss, W. S. Brown, and J. Moris, “Singer’s formant in sopranos: Fact or fiction?” *Journal of Voice*, vol. 15, no. 4, pp. 457–468, 2001.

[92] S.-H. Lee, H.-J. Kwon, H.-J. Choi, N.-H. Lee, S.-J. 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. 92–6, 2008. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2671792{&}tool=pmcentrez{&}rendertype=abstract

[93] J. Sundberg, “Acoustic and psychoacoustic aspects of vocal vibrato,” *StL-QPSR*, vol. 35, pp. 45–68, 1994.

[94] N. Isherwood, “Vocal Vibrato: New Directions,” *Journal of Singing – The Official Journal of the National Association of Teachers of Singing*, vol. 65, no. 3, pp. 271–283, 2009.

[95] Q. Han and R. Zhang, “Acoustic Analyses of the Singing Vibrato in Traditional Peking Opera,” *Journal of Voice*, vol. 31, no. 4, pp. 511.e1–511.e9, 2017.

[96] V. Sublett, “Vibrato or Nonvibrato in Solo and Choral Singing: Is There Room for Both?” *Journal of Singing*, vol. 66, no. 5, pp. 5–6, 2009.

[97] T. Shipp, E. T. Doherty, and S. Haglund, “Physiologic factors in vocal vibrato production,” *The Journal of the Acoustical Society of America*, vol. 4, no. 4, pp. 300–304, 1990.

[98] C. Leydon, J. J. Bauer, and C. R. Larson, “The role of auditory feedback in sustaining vocal vibrato,” *The Journal of the Acoustical Society of America*, vol. 114, no. 3, pp. 1575–1581, 2003.

[99] S. Sapir and K. K. Larson, "Supralaryngeal muscle activity during sustained vibrato in four sopranos: Surface EMG findings," *Journal of Voice*, vol. 7, no. 3, pp. 213–218, 1993.

[100] K. N. Westerman, "The Physiology of Vibrato," *Music Educators Journal*, vol. 24, no. 5, pp. 48–49, 1938.

[101] M. Olson, "Vibrato vs. nonvibrato: The solo singer in the collegiate choral ensemble," *Journal of Singing*, vol. 64, no. 5, pp. 561–564, 2008.

[102] A. von Ramm, "Singing Early Music," *Early Music*, vol. 4, no. 1, pp. 12–15, 1976.

[103] ——, "Style in Early Music Singing," *Early Music*, vol. 8, no. 1, pp. 17–20, 1980. [Online]. Available: http://search.ebscohost.com/login.aspx?direct=true{&}AuthType=ip,shib{&}db=rih{&}AN=1980-01571{&}site=ehost-live

[104] J. Potter, *Vocal Authority: Singing Style and Ideology.* Cambridge, UK: Cambridge University Press, 1998.

[105] D. Katok, "The Versatile Singer: A Guide to Vibrato and Straight Tone," Ph.D. dissertation, The City University of New York, 2016.

[106] H. F. Mitchell and D. T. Kenny, "The impact of 'open throat' technique on vibrato rate, extent and onset in classical singing," *Logopedics Phoniatrics Vocology*, vol. 29, no. 4, pp. 171–182, 2004.

[107] S. Melendez, "Google, Mozilla, and the Race to Make Voice Data for Everyone," 2017. [Online]. Available: https://www.fastcompany.com/40449278/google-mozilla-and-the-race-to-make-voice-data-for-everyone

[108] C. Dossman, "Over 1.5 TB's of Labeled Audio Datasets," 2018. [Online]. Available: https://towardsdatascience.com/a-data-lakes-worth-of-audio-datasets-b45b88cd4ad

[109] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[110] J. Wilkins, P. Seetharaman, A. Wahl, and B. Pardo, "VocalSet: a Singing Voice Dataset," in *19th International Society for Music Information Retrieval Conference.* Paris, France: International Society for Music Information Retrieval, 2018, pp. 468–474.

[111] D. A. A. Black, M. Li, and M. Tian, "Automatic identification of Emotional Cues in Chinese Opera Singing," in *13th Int. Conf. on Music Perception and Cognition and the 5th Conference for the Asian-Pacific Society for Cognitive Sciences of Music*, Seoul, South Korea, 2014.

[112] P. Proutskova, C. Rhodes, T. Crawford, G. Wiggins, P. Proutskova, C. Rhodes, T. Crawford, P. Proutskova, C. Rhodes, T. Crawford, and G. Wiggins, "Breathy, Resonant, Pressed – Automatic Detection of Phonation Mode from Audio Recordings of Singing," *Journal of New Music Research*, vol. 42, no. 2, pp. 171–186, 2013.

[113] B. Bozkurt, O. Baysal, and D. Yüret, "A Dataset and Baseline System for Singing Voice Assessment," in *13th International Symposium on Computer Music Multidisciplinary Research*, Matosinhos, Portugal, 2017, pp. 430–438.

[114] K. Choi, G. Fazekas, and M. Sandler, "Explaining Deep Convolutional Neural Networks on Music Classification," 2016. [Online]. Available: http://arxiv.org/abs/1607.02444

[115] L. Wyse, "Audio Spectrogram Representations for Processing with Convolutional Neural Networks," in *Proceedings of the First International Workshop on Deep Learning and Music*, vol. 1, no. 1, Anchorage, AK, 2017, pp. 37–41. [Online]. Available: http://arxiv.org/abs/1706.09559

[116] M. Papakostas and T. Giannakopoulos, "Speech-music discrimination using deep visual feature extractors," *Expert Systems with Applications*, vol. 114, no. May, pp. 334–344, 2018.

[117] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding Data Augmentation for Classification: When to Warp?" *2016 International Conference on Digital Image Computing: Techniques and Applications, DICTA 2016*, 2016.

[118] S. S. Stevens, J. Volkmann, and E. B. Newman, "A Scale for the Measurement of the Psychological Magnitude Pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 185, pp. 185–190, 1937.

[119] J. Pons, R. Gong, and X. Serra, "Score-informed syllable segmentation for a cappella singing voice with convolutional neural networks," 2017. [Online]. Available: http://arxiv.org/abs/1707.03544

[120] D. Scherer, A. Müller, and S. Behnke, "Evaluation of pooling operations in convolutional architectures for object recognition," in *20th International Conference on Artificial Neural Networks*, Thessaloniki, Greece, 2010.

[121] G. E. Hinton and V. Nair, "Rectified Linear Units Improve Restricted Boltzmann Machines," in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

[122] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation Functions: Comparison of trends in Practice and Research for Deep Learning," pp. 1–20, 2018. [Online]. Available: http://arxiv.org/abs/1811.03378

[123] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *International Conference on Learning Representations*, San Diego, CA, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[124] S. McCoy, "Voice Pedagogy : Formantology," *Journal of Singing - The Official Journal of the National Association of Teachers of Singing*, vol. 70, no. 1, pp. 43–48, 2013.

[125] J. E. Sumerau, ""That's What a Man Is Supposed to Do"," *Gender & Society*, vol. 26, no. 3, pp. 461–487, 2012.

[126] S. V. Wel, "Yodeling," 2013. [Online]. Available: http://www.oxfordreference.com/view/10.1093/acref/9780195314281.001.0001/acref-9780195314281-e-9199