VocalNET: A solution for real-time singing voice vocal register classification

Steven Luo Angel Au* Justin Lam* Samuel Yau

St. Paul's Co-education College AI Project Group (AIPG)

{sp20156401, sp20156021, sp20155021, sp20177461}@spcc.edu.hk

[1] https://youtube.com
[2] VocalSet: https://zenodo.org/record/1193957#.XyOVx577RhE
[3] SingingDatabase: http://isophonics.org/SingingVoiceDataset
[4] Image of Vocal Cords for Head Voice (Gilmore, 2016).
[5] Image of Vocal Cords for Chest Voice (Gilmore, 2016).
[6] container-labels-vocal-tract_orig.jpg (Irene & Harris, n.d.).

Abstract

In recent decades, many researchers have looked into various areas of music using artificial intelligence, including creative music generation, synthesis of singing voice and musical style transfer, etc. Currently, most voice-related classification papers have placed their focus on speech identification rather than singing voice, placing an emphasis on Support Vector Machines (SVMs) and the use of Mel-frequency cepstral coefficients (MFCCs). To explore the area of singing voice, we have created a model to identify and classify various timbres of different vocal registers through quantitative analysis, which could aid amateur singers on their journey of learning about singing and improving their singing techniques.

F

Our model has showed its capacity to be applied in real-time conditions, with an accuracy of 93% in our training and validation dataset. Looking into the future, we look forward to training our model with an extensive dataset and implementing the model in an appropriate medium such as an application.

1 Introduction

Vocal cords are composed of two infoldings of mucous membrane stretched horizontally across the larynx. When humans speak and sing, the vocal cords oscillate and generate sound. The vocal cords can produce different vibratory patterns, each of which produce certain ranges of pitches and sounds. These are commonly known as vocal registers. The vocal registers covered in this study include chest voice, head voice, mixed voice and whistle register, which are the most common ones. Whistle register is the highest phonational register of the human voice, which ranges from C_6 to D_7. Ranging from G_5 to the sixth octave, the head voice is the upper singing register, in which the singer would feel as if the tone is resonating in his/her head. Chest voice is the lower singing register, ranging from the third octave to A_4. In between chest voice and head voice is the mixed voice, which ranges from C_4 to $F^{\#}_5$. These registers will be further explained in this paper.

The best singers maintain a certain vocal register beyond the natural state (Singing Voice Registers, n.d., Vocal Skills). Generally, this is not possible when the notes are beyond the natural boundary of the singer's vocal register. Therefore, singers usually use the next vocal register and make it sound like the same register. They try to keep smooth transitions between the registers, with the ultimate aim of disguising the changes. To the general public, it can be a challenging task to differentiate between different vocal registers, and only professional musicians are capable of identifying the parts of the song where the singer uses another vocal register. Thus, amateur singers may encounter difficulties in maintaining vocal registers at pitches lower or higher than the respective ranges of the registers, requiring extensive training in order to improve. Our model assists the singers by providing automatic classification of vocal registers, which enables them to train their singing voices anywhere, anytime without the presence of a vocal coach and hence increasing the efficiency of their vocal training.

While each vocal register produces sounds in specific vocal timbres, there are a wide range of factors which also contribute to the timbre of a singer. The sound quality and tone color of the performance of a singer can be affected by the singer's mouth shape, pitch, amplitude, gender and other factors. Due to the complexity of these factors, it can be challenging to factor in every vocal characteristic through hard coding. As a result, we have implemented artificial intelligence to take the wide range of factors into consideration.

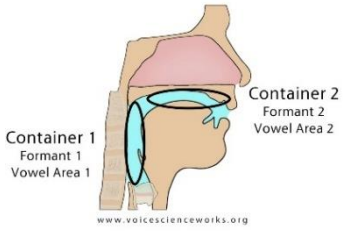
Timbre classification is not an uncommon problem; however, timbre classification relating to voice is significantly less popular. In addition, most voice-related classification papers have placed their focus on speech identification rather than singing voice. Most current methodologies place an emphasis on Support Vector Machines (SVMs) and the use of Mel-frequency cepstral coefficients (MFCCs).

Our contribution

- Train classifiers of vocal registers
- Identify traits in audio sample which reflect vocal register
- Identify appropriate data representations which reflect the differences between vocal registers

2 Vocal Registers

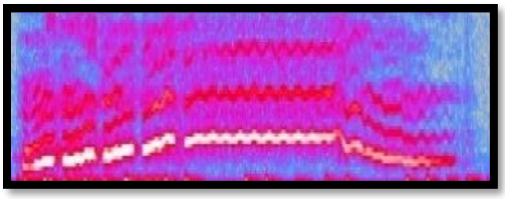
Different vocal registers are results of physical adjustments in the oral cavity (tongue, soft palate and other components), nasal cavity and/or the throat (vocal cords, larynx and pharynx). There are many vocal registers identified, including vocal fry, chest voice, mixed voice, head voice, falsetto register, whistle register. However, there are currently no authoritative vocal register classification systems and vocal registers are often deemed subjective.



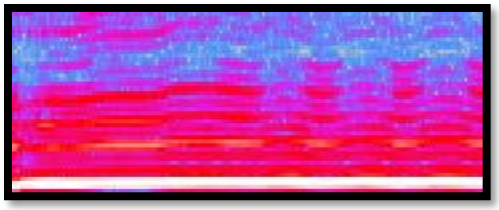
6

In this project, four registers, namely chest voice, mixed voice, head voice and whistle register are chosen to be investigated, since these registers are common and the spectrograms of these registers display distinguishable patterns. The four spectrograms attached below are indicated with six colour bands, each representing varying levels of amplitudes: white, red, magenta, dark blue, light blue and gray. White-coloured bands imply the largest amplitudes while grey-coloured bands imply the smallest amplitudes.

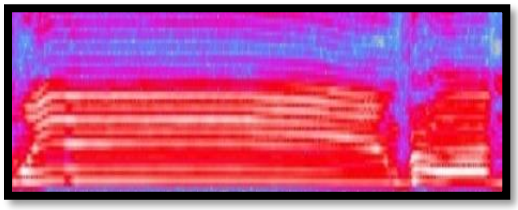
[1] https://youtube.com
[2] VocalSet: https://zenodo.org/record/1193957#.XyOVx577RhE
[3] SingingDatabase: http://isophonics.org/SingingVoiceDataset
[4] Image of Vocal Cords for Head Voice (Gilmore, 2016).
[5] Image of Vocal Cords for Chest Voice (Gilmore, 2016).
[6] container-labels-vocal-tract_orig.jpg (Irene & Harris, n.d.).



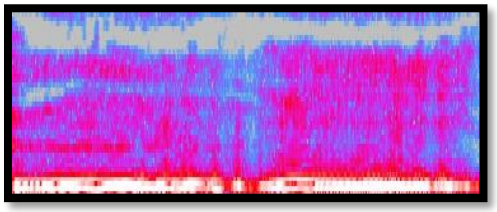
(a) Spectrogram of whistle register



(b) Spectrogram of head voice



(c) Spectrogram of mixed voice



(d) Spectrogram of chest voice

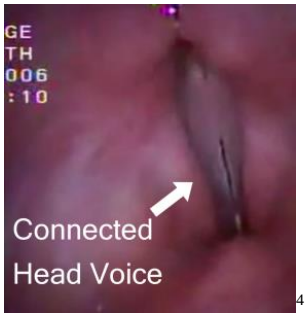
Whistle Register

Whistle register is the highest phonational register of the human voice. The register usually starts at C_6 (any expression of pitch would be under the scientific pitch notation hereinafter) and extends to D_7. The physiology of the whistle register remains the least known amongst the vocal registers as the vocal cords could not be filmed due to the epiglottis closing down over the larynx. There are three theories proposed to explain the register, but no universal consensus was reached. [3] The timbre of notes produced using this register is similar to the sound of a whistle. Some of the singers renowned for their usage of whistle register include Mariah Carey and Minnie Riperton.

The general spectrogram of whistle register has distinctive several lines on certain frequencies, with a white line at the frequency of the notes being sung. As the whistle register is used for extremely high notes, the white line is visible at higher frequencies when compared to the position of white lines for other registers. The lines often show zig-zag patterns as singers apply vibratos to their whistle notes.

Head Voice

Head voice is the upper singing register. The register starts at around G_5 and it may extend to the sixth octave depending on the singer. When singing in the head voice, the singer would feel as if the tone is resonating in his/her head. The cricothyroid muscles are mostly used, while the vocal cords would be stretched to be thinner and longer, favouring the production of higher notes. The timbre of notes produced is lighter and thinner. [4]



[4]

The general spectrogram of head voice has a focused, straight white line at the frequency of the notes being sung. At frequencies higher than the main note, high amplitudes could be observed as well, as indicated by the red regions. The overall high amplitudes at high frequencies correspond to the bright timbre of using head voice.

Mixed Voice

Mixed voice is the singing register in between chest voice and head voice. The range of this register depends on the overall vocal range of the singer and may range from C_4 to $F^{\#}_5$. Mixed voice is a mixture between chest voice and head voice. Different ratios of chest voice to head voice would result in different timbres.

The general spectrogram of mixed voice is differentiated by various blurry white lines at different frequencies, surrounded by red regions. Overall medium amplitudes at different frequencies represent the blend of timbres of head voice and chest voice.

Chest Voice

[1] https://youtube.com
[2] VocalSet: https://zenodo.org/record/1193957#.XyOVx577RhE
[3] SingingDatabase: http://isophonics.org/SingingVoiceDataset
[4] Image of Vocal Cords for Head Voice (Gilmore, 2016).
[5] Image of Vocal Cords for Chest Voice (Gilmore, 2016).
[6] container-labels-vocal-tract_orig.jpg (Irene & Harris, n.d.).

Chest voice is the lower singing register. The register starts at around A_4 and it may extend lower to the third octave, depending on the singer. Most women speak in their chest voice. When singing in the chest voice, the singer would feel sensations of vibrations in the chest. [6] The thyroarytenoid muscles are mostly used, while the vocal cords would be stretched to be thicker and shorter. The timbre of notes produced is heavier and fuller. [4]

The general spectrogram of chest voice has scattered white regions at the frequency of notes being sung, with disperse red regions at higher frequencies. The overall high amplitudes at lower frequencies correspond to the full, mellow timbre of using chest voice.

3 Related Works

Timbre classification related

Timbre classification is not an uncommon problem; however, timbre classification relating to voice is significantly less popular. Guven, E., & Ozbayoglu, A. M (2012) studied the use of a Support Vector Machine (SVM) to extract timbre and melody information from spectrograms in their paper "Note and Timbre Classification by Local Features of Spectrogram". In particular, they made use of a frequency-time representation and the formants F1 and F2 to the feature vectors, which they claimed to give results that surpassed previous ones.

H., Sang. (2013) also noted in their paper "Musical Instrument Extraction through Timbre Classification" that a spectral representation of audio seems to be most representative of timbre. They selected a representation with the Mel scale, which is broadly based on how humans perceive sounds. The notable effects of the spectral envelope and changes in the spectral envelope on timbre were also mentioned. Similar to Guven, E., & Ozbayoglu, A. M (2012), they made use of a Support Vector Machine (SVM).

Burred, J., Roebel, A., Rodet, X. (2015) studied an alternative to Mel-frequency cepstral coefficients (MFCC) representations in their paper "An Accurate Timbre Model for Musical Instruments and its Application to Classification.". They claimed that their Envelope Interpolation Method was a superior representation compared to MFCCs.

Singing voice related

Voice-related classification tasks have mainly focused on speech instead of singing.

Raahul, A. et al. (2017) investigated the use of 5 different methods to classify gender based on voice - Linear Discriminant Analysis (LDA), K-Nearest Neighbor (KNN), Classification and Regression Trees (CART), Random Forest (RF), and Support Vector Machine (SVM). They concluded that Support Vector Machines (SVMs) had the highest accuracy out of the 5 different methods.

In "Singing voice detection in music tracks using direct voice vibrato detection", L. Regnier and G. Peeters attempted to extract singing voice from an audio track through identification of key singing voice features, including vibrato and tremolo. They also identified harmonicity and formants as important features in singing voice. They claimed that this method achieved an accuracy similar to the more conventional MFCC approach.

4 Data Collection

Due to the technical difficulties in classifying and labelling vocal registers, no available dataset is provided on the internet. Hence, we have created our own dataset that consists of {'YouTube': 30min; 'opera': 10min; 'volunteer': 5min} minutes of acapella audio samples sung by different female singers. A large portion of the audio samples were collected from various online sources[1], while a small portion of testing samples were recorded by our technically trained choir schoolmates. One thing to note is the quality control of our audio sample collection, as it is only possible to identify vocal registers when the given audio has only one singing voice and without any background accompaniment. The samples were then labelled and cut by Angel Au[1] using Audacity.

YouTube videos

Audios of acapellas (official vocal stems/filtered vocals extracted from a professionally produced song) are converted from YouTube videos. All audios collected consist of vocals of one singer singing at a time only, as audios with background vocals would complicate the task and decrease the training efficiency significantly. The audios are then labelled by our research group members and several volunteers who are amateur singers. Consequently, the labelled audios are split regularly into 0.1 second samples. The decided interval is a compromization between data size and content.

[1] https://youtube.com
[2] VocalSet: https://zenodo.org/record/1193957#.XyOVx577RhE
[3] SingingDatabase: http://isophonics.org/SingingVoiceDataset
[4] Image of Vocal Cords for Head Voice (Gilmore, 2016).
[5] Image of Vocal Cords for Chest Voice (Gilmore, 2016).
[6] container-labels-vocal-tract_orig.jpg (Irene & Harris, n.d.).

Online dataset

VocalSet and SingingDatabase are both public datasets that contain clean monophonic opera singing. We selected only the female audios sets. The addition of opera singing samples provides are dataset with a wider variety of singing voices.

5 Methodology

5.0 Audio Representations

From Section 2, our experimental observations prove that the task of vocal register classification is directly related to timbre. We show that there are general trends that could differentiate the audios of different registers based on patterns shown on STFT graphs. This disproves the stereotypical view of the singing community, which emphasizes the subjective nature of register identification.

Timbre of an audio could be captured by various MIR (musical information retrieval) techniques. [9] provided an extensive investigation into how different audio representations could reflect properties of timbre. In particular, relative amplitudes and frequencies of harmonics and inharmonics influences timbre significantly. Similarly, formants, which traces frequency levels that provide integral pitch identification of an audio sample also affect our perception of timbre. The distance between consecutive formants, or the combination of amplitudes of the formants differentiate brightness, mellowness, and related characteristics of sound.

However, our attempt with using formant and harmonic information in classifying vocal register failed to yield satisfactory results. We processed the audio samples into relative amplitude and frequency information of formants (f0, f1, f2, f3, f4) and harmonic (h1, h2, h3), and fed the information through fully-connected layers. While articles like [9] show that formant frequency and amplitude patterns provide simple but significant influences to the timbre, such as a relatively higher amplitude of f2 results in brighter timbre (general timbre of head voices), our network failed to recognize such relation between formants and register timbre.

An alternative method is to use 2D audio representations as mentioned in [15], such as short-time fourier transform spectrogram (STFT) or mel-spectrograms. These representations are commonly used in complex audio analysis tasks that require deep learning, and promising results are produced. Hence, we utilized STFT for our vocal register task and approach the problem with a set of Convolutional Neural Networks (CNN). CNN's capability to observe local patterns in the 2-dimensional region is especially suited for such timbral identification, as timbral timbral characteristics are best shown through patterns along the frequency dimension. STFTs also provide finer details when compared to pure formants or harmonics, hence increasing the chance of success for our models. The choice of using STFT and CNN will be validated with the excellent results in the following parts, where we evaluate the performance differences in 1D and 2D convolutions, and the real-time capabilities of our model.

5.1 General architecture of models

All models used in the following experimentations follow the same architecture for fairness purposed. The models used a 4-layer convolutional network with a stride of 2 and zero padding. Batch normalization was used after each layer, to improve the speed, performance, and stability of the model. Leaky ReLUs were also used after each layer so as to improve performance. After the 4 convolutional layers, two fully connected Dense layers were used.

We used a categorical cross entropy loss, with the loss function as follows:

$$-\Sigma_{c=1}^{M} y_{i,c} \log(p_{i,c})$$

Soft labelling was also used to improve performance and prevent overfitting. For data sampling, we used a 4:1 train ratio, with the data split based on random sampling.

The models were trained with a Stochastic Gradient Descent learning rate of 5e-4 and a momentum of 0.9, as well as a batch size of 64. All models were trained for 30 epochs.

5.1 1D Vs 2D Convolution

Model	Training Acc	Validation Acc
1D-CNN (time-axis)	95.21%	90.08%
1D-CNN (frequency-axis)	**98.48%**	90.24%
2D-CNN	97.59%	**94.29%**

Table 1: Accuracies of 1D and 2D CNN models

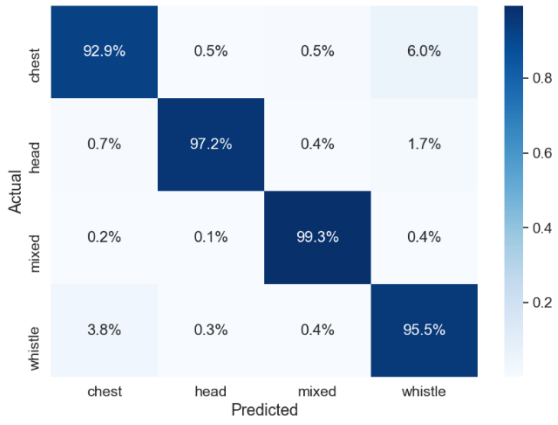


Fig 1: Confusion Matrix of 2D-CNN

1D vs 2D comparison

[1] https://youtube.com
[2] VocalSet: https://zenodo.org/record/1193957#.XyOVx577RhE
[3] SingingDatabase: http://isophonics.org/SingingVoiceDataset
[4] Image of Vocal Cords for Head Voice (Gilmore, 2016).
[5] Image of Vocal Cords for Chest Voice (Gilmore, 2016).
[6] container-labels-vocal-tract_orig.jpg (Irene & Harris, n.d.).

We used both 1D and 2D Convolutional Neural Networks. We first used a 1D CNN that convolved across the frequency axis, which we thought was reflective of timbre characteristics, as timbre is essentially perceived as the pattern of changes in frequency.

Interestingly and much to our surprise, a 1D CNN that convolved across the time axis also achieved a fairly high training accuracy, which suggests that the change of frequency over time is important to timbre as well.

This is also shown through the success of the 2D CNN, which convolves across both the frequency and time domains, indicating that the information about the time domain is important to the classification of timbre/vocal register.

Overall, the 2D CNN performed best out of our three models. We then began to develop the real-time capability of our model, such that it could be used to classify vocal register on a real-time basis.

5.3 Real-time capabilities

The audio intervals we tested (in ms) were{1000, 500, 250, 100, 20}.

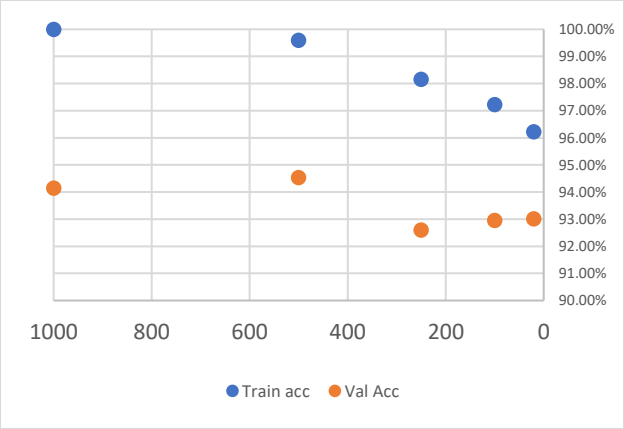
The longer the interval, the more accurate the training and validation was in general. This is to be expected, as a longer interval will give more information about how frequency changes over the time domain. However, all the validation and training accuracies are still quite high, showing that real-time prediction with our model is feasible.

An issue with the short time fourier transform (stft) is that it has a fixed resolution. This means that as the time domain resolution increases, the frequency domain resolution decreases, and vice versa. This means that as the audio interval decreases, there is in fact an increase in the frequency domain resolution.

It is worth noting that the training accuracy drops fairly consistently as the audio interval decreases, but the validation accuracy is more stable and levels off. This indicates that a decrease in audio interval time will not severely affect the performance of the model on unseen data.

Audio interval (ms)	Training Acc	Validation Acc
1000	100.00%	94.14%
500	99.59%	94.53%
250	98.16%	92.60%
100	97.22%	92.95%
20	96.22%	93.01%

Table 2: Accuracies of CNN-models with different audio sample lengths



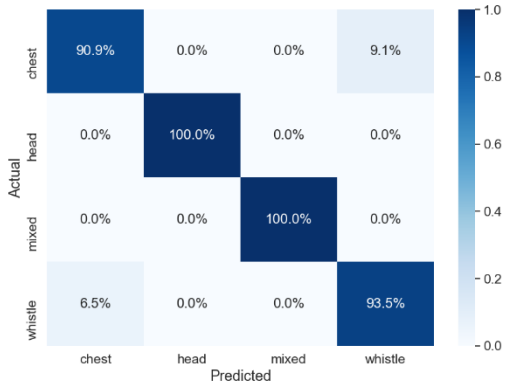
Plot 1: Accuracy progressions with different input interval times (in ms)

The inference time for the GPU is actually slower than the CPU. This is because the GPU only gives benefits to speed during training, due to its capacity for parallel computation. During interference, it is slower, because the clock speed of the GPU is slower.

Regardless, the inference time is generally quite low, which makes real-time inference feasible.

Audio interval (ms)	Audio processing time (ms)	Inference time (ms)	
		CPU[1]	GPU[2]
100	0.41	1.15	2.21
20	0.47	1.08	2.10

Table 2: Inferencing time of 100ms and 20ms model on CPU and GPU



2D-CNN trained with 20ms

[1] Intel Core i7-9700F

[2] Nvidia GTX 1660Ti

[1] https://youtube.com
[2] VocalSet: https://zenodo.org/record/1193957#.XyOVx577RhE
[3] SingingDatabase: http://isophonics.org/SingingVoiceDataset
[4] Image of Vocal Cords for Head Voice (Gilmore, 2016).
[5] Image of Vocal Cords for Chest Voice (Gilmore, 2016).
[6] container-labels-vocal-tract_orig.jpg (Irene & Harris, n.d.).

6 Evaluation

On Opera set

Model	Acc
20ms 2D-CNN	30.54%

Table 3: Accuracies of 20ms 2D-CNN model on our opera set

The accuracy of our model on pop song related data was generally high – often reaching 90%+ accuracy.

However, the accuracy of the model on a dataset derived from opera voices was drastically lower. This may be because of the nature of this dataset – during training, we mainly used data from pop songs, which would naturally be very different from operas. As a result, the model performs much better on pop songs compared to operas.

Although the accuracy of our model on a dataset composed of opera voices is low, this does not indicate that our model is not useful for vocal register classification – in fact, this most likely resulted from our limited amount and scope of data, which can be rectified in the future. By incorporating more types of audio and voices into our training data, we can achieve a more versatile model that can reliably and accurately classify different types of voices, from pop songs to opera.

7 Conclusion and future works

Our model on classifying various timbres of different vocal registers has displayed its capacity in achieving real-time inference, which could be made available to the general public by installing the model into an application, software or website. The model also has high accuracy on the dataset of pop song vocals. However, as labelling the audios according to the vocal registers is time-consuming, a dataset with limited number of samples and variations was used to train the model, resulting in a low accuracy when testing with the opera dataset and dataset contributed by the volunteers.

To improve our model, we aspire to expand our dataset by sampling more audios for our training and validation dataset, as well as collecting more audios from other volunteers. By increasing the variations in the dataset, such as including audios of singing classical pieces and extracted vocals from songs of more genres, the model could be applied in classifying vocal registers in a general setting, instead of capping its possibilities within the realms of pop music. As an extension of this project, we also anticipate creating an application, software or website for installing this model.

[1] https://youtube.com
[2] VocalSet: https://zenodo.org/record/1193957#.XyOVx577RhE
[3] SingingDatabase: http://isophonics.org/SingingVoiceDataset
[4] Image of Vocal Cords for Head Voice (Gilmore, 2016).
[5] Image of Vocal Cords for Chest Voice (Gilmore, 2016).
[6] container-labels-vocal-tract_orig.jpg (Irene & Harris, n.d.).

Reference

1. Vocal Skills. (n.d.). Singing Voice Registers. Retrieved from https://www.vocalskills.co.uk/singing-voice-registers.html
2. Irene, L., & Harris, D. (n.d.). container-labels-vocal-tract_orig.jpg. Retrieved July 31, 2020, from https://www.voicescienceworks.org/vocal-tract.html
3. Garnier, M., Henrich, N., Crevier-Buchman, L., Vincent, C., Smith, J., & Wolfe, J. (2012). Glottal behavior in the high soprano range and the transition to the whistle register. *The Journal of the Acoustical Society of America, 131*(1), 951-962. doi:10.1121/1.3664008
4. The National Center for Voice and Speech. (2017, March 19). Voluntary Register Changes. Retrieved July 31, 2020, from http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/voluntary.html
5. Gilmore, C. (2016, January 22). [Image of Vocal Cords for Head Voice]. Retrieved July 31, 2020, from https://www.powertosing.com/ep-38-singing-in-head-voice-what-is-head-voice/
6. Tarneaud, J. (1933). Study of larynx and of voice by stroboscopy. Clinique (Paris), 28, 337-341.
7. Gilmore, C. (2016, January 15). [Image of Vocal Cords for Chest Voice]. Retrieved July 31, 2020, from https://www.powertosing.com/949-2/
8. Sadolin, C. (2008). *Complete Vocal Technique*. Copenhagen: Shout.
9. Creasey, D. P. (2006). An exploration of sound timbre using perceptual and time-varying frequency spectrum techniques.
10. Guven, E., & Ozbayoglu, A. M. (2012, November 12). Note and Timbre Classification by Local Features of Spectrogram. Retrieved from https://www.sciencedirect.com/science/article/pii/S1877050912006424
11. H., Sang. (2013). Musical Instrument Extraction through Timbre Classification. Retrieved from http://cs229.stanford.edu/proj2013/Park-MusicalInstrumentExtractionThroughTimbreClassfication.pdf
12. Burred, J., Roebel, A., Rodet, X. (2015, June 8). An Accurate Timbre Model for Musical Instruments and its Application to Classification. Retrieved from https://hal.archives-ouvertes.fr/hal-01161413/document
13. Raahul, A. et al. (2017). Voice based gender classification using machine learning. Retrieved from https://iopscience.iop.org/article/10.1088/1757-899X/263/4/042083/pdf
14. L. Regnier and G. Peeters, "Singing voice detection in music tracks using direct voice vibrato detection," 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, 2009, pp. 1685-1688, doi: 10.1109/ICASSP.2009.4959926.
15. Choi, K., Fazekas, G., Cho, K., & Sandler, M. (2017). A tutorial on deep learning for music information retrieval. arXiv preprint arXiv:1709.04396.

[1] https://youtube.com
[2] VocalSet: https://zenodo.org/record/1193957#.XyOVx577RhE
[3] SingingDatabase: http://isophonics.org/SingingVoiceDataset
[4] Image of Vocal Cords for Head Voice (Gilmore, 2016).
[5] Image of Vocal Cords for Chest Voice (Gilmore, 2016).
[6] container-labels-vocal-tract_orig.jpg (Irene & Harris, n.d.).