Genre Classification of Rap and Gregorian Chant Based on Vocal Qualities
Final Project Report

Corinne Darche
MUMT621: Music Information Acquisition, Preservation, and Retrieval
2 May 2022

Introduction

The human singing voice is one of the oldest musical instruments. It also remains a focal point of music information retrieval (MIR) research due to its complexity. Significant progress has been made in the realm of speech processing, however the same cannot be said about singing voice processing. Since the speaking and singing voice have significant differences, techniques from speech processing cannot be neatly transferred to MIR applications (Lori and Subbaraman 2013).

There are, however, some genres that possess vocal characteristics close to human speech, such as rap and chant. Rap is a "musical form that makes use of rhyme, rhythmic speech, and street vernacular, which is recited or loosely chanted over a musical soundtrack" (Keyes 2004, 1). In contrast, chant music uses a small set of pitches with a unique rhythmic structure, often absent from scores (Stolba 1994). While the two genres are very different in terms of their subject matter and their cultural purpose, they both blur the line between speech and singing. Their focus is less so on the melodic contour in the vocal line, and more on rhythmic structure.

This project's focus was to create a machine learning classifier to identify rap music and chant music based solely on their vocal characteristics. Since these genres are much closer to speech, isolating them from their accompaniment was not anticipated to be a great challenge. In doing so, I aimed to begin to bridge the gap between speech processing and singing processing which has been shown to be a significant challenge in previous vocal classification research (Kim and Whitman 2002; Tsai, Rodgers, and Wang 2004; Regnier and Peeters 2009; Fujihara et al. 2010).

Method

AudioSet was the foundation for this project's data. The dataset, created by a research team at Google, consists of different 10-second clips from YouTube videos across multiple categories and musical genres (Gemmeke et al. 2017). Their data for "rapping" is fairly accurate and high-quality. However, due to the broad definition of "chant", the data was inconsistent and unfocused. Therefore, this data was not used and was replaced by a more specific "Gregorian chant" dataset.

It should be noted that AudioSet's data is only available through .csv format. That is, due to copyright reasons, the audio data is not available for direct download. Fortunately, for her own Master's thesis on audio processing, Aoife McDonagh created a script to download specific audio files from AudioSet from a user-specified class (McDonagh 2019). The final rap dataset used in this project consisted of 1,476 .wav files.

There is no known pre-existing dataset for Gregorian chant audio files. As such, I made my own using an existing Gregorian chant playlist on YouTube (Quang 2020). The initial plan was to use a Python package to automate the downloading process, such as *pytube* or *pafy*. However, YouTube removed its dislike counter on videos in their November 2021 update, and this has not been properly updated in any of the Python packages. So, I extracted all the audio manually, ensuring that the selected videos were proper Gregorian chant as opposed to standard Catholic liturgical music. Once all the main audio files were downloaded, they were segmented into 10-second clips using the Jupyter Notebook script *split.ipynb*. Files under 10 seconds and with minimal vocals were removed to improve the quality of the dataset. The final Gregorian chant data consisted of 1,451 .wav files.

The next step was to extract the vocals. This was done to ensure that the classification only relied on vocal characteristics. There are many complex ways to do this, such as a three-step method involving estimating the fundamental frequency and resynthesizing the melody from there (Fujihara et al. 2010). For the scope of this project, a pre-made function was selected from *librosa*'s documentation. The *librosa* Python package was created for MIR projects, and it was used for this project's audio processing (McFee et al. 2015). In the package's examples, they showcase a vocal extractor which separates the audio into the foreground (i.e., the vocals) and the background through masks. While it does lose certain qualities of the vocal line, the resulting foreground audio still provides the necessary information. In addition, the vocal extractor was able to attenuate the reverb in the Gregorian chant files.

Following vocal separation, nine features were extracted: spectral chroma, spectral centroid, spectral bandwidth, spectral flatness, spectral rolloff, Mel-Frequency Cepstral coefficients (MFCCs), root-mean-square (RMS) values, tempogram, and zero-crossing rate. Spectral qualities are often used for melodic analysis, as timbre can play a significant role in vocal identification studies (Fujihara and Goto 2007). However, as previously mentioned, these two genres are defined by their rhythmic and speech-like qualities. So, the tempograms, the MFCCs, and the zero-crossing rates were also considered and expected to be significant. The means of all these values were calculated for each audio sample and scaled for training.

Two models were trained for this classification process: a Support Vector Machine (SVM) and a Gaussian Mixture Model (GMM). Not only are these models often used in singing voice classification, but they also represent both the supervised learning and unsupervised learning approaches (Loni and Subbaraman 2013). Both models were taken from Python's scikit-learn package. The SVM model used hinge loss and trained with 10-fold cross-validation. The GMM was specified to have two components, representing the two genres, k-means initialization, and tied covariance parameters. The two models, after training and testing in *classifier.ipynb*, were exported and can be tested by users using the Python script *rapChantClassify.py*.

Results

Both models reported very high accuracies (see table 1). The SVM model reported an average of 97.64% accuracy over its cross-validation. The features were also evaluated in this model to identify the most impactful parameters for classification. Of the nine features, the most significant ones (i.e., features with coefficients of an absolute value greater than one) were the spectral centroid, the spectral rolloff, the zero-crossing rate, the MFCC, and the spectral bandwidth.

Model	Accuracy
Support Vector Machine (SVM)	97.64%
Gaussian Mixture Model (GMM)	97.61%

TABLE 1*: Rap and Gregorian Chant Classification Accuracy Results*

The GMM produced strong clusters using tied covariance. From comparing the GMM output to the actual labels, it was deduced that the Gregorian chant audio cluster was represented as 1 and the rap audio cluster was represented as 0. Using those identifiers, the accuracy was manually calculated to be 97.61%.

Discussion

It should be noted that there is a difference in the quality and curation of each genre's data. There was effort to have even datasets, both in the number of data samples and in the length of each sample. Since I gathered the Gregorian chant data manually, there was a clearly defined selection process for that data. AudioSet is human annotated, and the rap dataset is estimated to have high quality. However, I did not go through each audio sample to verify this myself. Though the results indicate high accuracy, a stronger rap dataset could be created in the future to ensure the models' validity.

The most valuable features were identified in the Results section using the SVM's coefficients. Despite the higher value given to rhythmic features in both genres, spectral features were still the most useful for classification. However, zero-crossing rate and MFCCs still had a high contribution to the classification. This dynamic could be explained by the stronger melodic contour in Gregorian chant. Rap is undeniably more speech-like than Gregorian chant, as confirmed by the genres' respective definitions. It follows that having a stronger spectral centroid value is an immediate indicator for Gregorian chant in this classification task. This can be further confirmed in future replications of this study.

When building the GMM classifier, the covariance selection had a significant impact on the accuracy results. For example, when using the default full covariance, the model had an accuracy of 57%. The *scikit-learn* documentation explains that tied covariance means that all components have the same general covariance matrix (Pedregosa et al. 2011). That is, both clusters have the same shape with no exact restriction on what that shape could be. Of course, GMM's clustering is much softer than other models, like K-Means. Nevertheless, this indicates that both genres are strongly defined and can be separated based on the provided features.

This project was proposed as a step in bridging the gap between speech processing and singing processing research. There is still much work to be done based on this project's results. Currently, there are under 3,000 data points for these classifiers. To strengthen their accuracy, more data will have to be collected in both genres. In addition, the genres can be compared to speech and more melodic genres, such as opera or pop, to further test and verify the hypotheses surrounding the features. Lastly, while these genres are speech-like, they are not the true middle ground between speech and singing. It would be interesting to incorporate samples from *sprechstimme*, a vocal technique in which the singer recites the text using pitched speech.

Conclusion

This project's objective was to explore the genre classification of speech-like genres with a focus on their respective vocal characteristics. The vocal extraction and classifiers were successful in their respective tasks. While there remains work to be done to improve the datasets and to generalize the models beyond rap and Gregorian chant, this project is a step in reducing the gap in speech processing's advancements and singing processing research.

References

Fujihara, Hiromasa and Masataka Goto. 2007. “A Music Information Retrieval System Based on Singing Voice Timbre.” In *Proceedings of the International Society of Music Information Retrieval*, 467–470.

Fujihara, Hiromasa, Masataka Goto, Tetsuro Kitahara, and Hiroshi G. Okuno. 2010. “A Modeling of Singing Voice Robust to Accompaniment Sounds and Its Application to Singer Identification and Vocal-Timbre-Similarity-Based Music Information Retrieval.” *IEEE Transactions on Audio, Speech, and Language Processing* 18 (3): 638–648.

Gemmeke, Jort F., Daniel P.W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. “AudioSet: An Ontology and Human-Labeled Dataset for Audio Events.” In *2017 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 776–780.

Keyes, Cheryl Lynette. 2004. *Rap Music and Street Consciousness*. Champaign: University of Illinois Press.

Kim, Youngmoo E. and Brian Whitman. 2002. “Singer Identification in Popular Music Recordings Using Voice Coding Features.” In *Proceedings of the 3rd International Conference on Music Information Retrieval* 13, 17–22.

Lena, Jennifer C. 2006. “Social Context and Musical Content of Rap Music, 1979–1995.” *Social Forces* 85 (1): 479–495.

Loni, Deepali Y., and Shaila Subbaraman. 2013. “Extracting Acoustic Features of Singing Voice for Various Applications Related to MIR: A Review.” In *Proceedings of the International Conference on Advances in Signal Processing and Communication*, 66–71.

McDonagh, Aoife. 2019. “AudioSet-Processing.” Source Code. Accessed 18 April 2022. https://github.com/aoifemcdonagh/audioset-processing.

McFee, Brian, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. “librosa: Audio and Music Signal Analysis in Python.” In *Proceedings of the 14th Python in Science Conference*, 18-25.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. “Scitkit-learn: Machine Learning in Python.” *Journal of Machine Learning Research* 12 (85): 2825–2830.

Quang, Vo Duc. 2020. “Gregorian Chants.” *YouTube* playlist. Accessed 18 April 2022. https://www.youtube.com/playlist?list=PL3930DBA3929A5221.

Regnier, Lise, and Geoffroy Peeters. 2009. “Singing Voice Detection in Music Tracks Using Direct Voice Vibrato Detection.” In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1685–1688.

Stolba, K. Marie. 1994. *The Development of Western Music: A History*. 2nd ed. Madison: Brown & Benchmark.

Tsai, Wei-Ho, Dwight Rodgers, and Hsin-Min Wang. 2004. “Blind Clustering of Popular Music Recordings Based on Singer Voice Characteristics.” *Computer Music Journal* 28 (3): 68–78.