Saul Ivan Rivas Vega (2020). Musical Note and Instrument Family Identification using the Constant-Q Transform, -, V(N), pp. xx-xx, DOI: https://doi.org/xx.xxxx/xxxx.xx

OVERVIEW ARTICLE

# Musical Note and Instrument Family Identification using the Constant-Q Transform

Saul Ivan Rivas Vega*

**Abstract**

In this paper a Naive Bayes Classifier is trained using the Constant-Q Transform coefficients as features. The use of a modern dataset and just the frequency domain features is intended for testing the robustness of these features that can then be of use for applications which perform Musical Instrument Identification and Automatic Music Transcription. The results shows a significantly high accuracy in the train/test split of the dataset and in a further testing with monophonic recordings from two music production applications.

**Keywords:** Pitch Detection; Timbre Estimation; Music Information Retrieval; Constant-Q transform; Naive Bayes

## 1. Introduction

Tasks like Music Instrument Identification and Automatic Music Transcription have been widely studied within Music Information Retrieval (MIR) tasks (Kitahara et al., 2005; Salamon and Gómez, 2012; Gururani et al., 2019), refer to (Benetos et al., 2013, 2019) for a more extensive overview. Being able to tackle these tasks can have a great impact on how we teach, experience and play music. Open Source implementations of such works can be found in GitHub repositories made by the authors, developers[1] and in full development libraries like ESSENTIA Bogdanov et al. (2013) or Librosa McFee et al. (2015).

Musical notes are named pitches, like A4, C3 and G#2. In Schnupp et al. (2011) pitch is defined as a subjective percept of sounds not a physical property of them. That is, pitch is how we perceive a certain sound and it could not be measurable by the sound only. Regardless, we know that periodicity in sound plays a heavy role in our perception. We measure the periodicity in a sound signal by counting how many repetitions of the period are in 1 second (Hertz). This quantification of periodicity is called "fundamental frequency" (f0). Pitch is often used interchangeably with f0, and although f0 can used to measure pitch it does not necessarily correspond to the same. Sounds we perceive with a high pitch normally have a high f0 and vice versa. Thus the complexity of pitch as subjective is often neglected outside psycho-acoustical studies.

As for musical instrument identification or recognition we can view them as sound sources and to differentiate them we now have to talk about timbre. Timbre like pitch is a subjective percept of sound which also is neglected as subjective by assuming certain properties of sound as its physical correlates. Timbre is specially distinct of pitch as for how well is defined using physical properties. In Siedenburg et al. (2019) a review of many definitions is made and mainly discussing the one provided by the American National Standards Institute (ANSI): "That attribute of auditory sensation which enables a listener to judge that two nonidentical sounds, similarly presented and having the same loudness and pitch, are dissimilar." In their discussion they mention of timbre correlating with multidimensional set of sound properties varying in scale which is yet to be defined at it fullest if possible due to the context dependent results many studies have found.

The physical properties associated with timbre often used are the temporal envelope of the sound, spectral and cepstral analysis, which is the envelope of the frequency spectrum of the signal, to mention a few. From the most used properties we find the frequency analysis and as stated in Brown (1991): "The frequencies that have been chosen to make up the scale of Western music are geometrically spaced. Thus the discrete Fourier transform (DFT), although extremely efficient in the fast Fourier transform implementation, yields components which do not map efficiently to musical frequencies." The constant-Q transform is different to the DFT in which there is no constant difference between frequencies but a constant ratio of difference. The frequencies in the music scale are spaced by multiples of its fundamental frequency thus a matching constant-Q transform is equivalent to a 1/24-oct filter bank.

*Universidad Autónoma de México, IIMAS, Mexico City

## 2. Related Work

An overview as the one in Gerhard (2003) or Drugman et al. (2018) has the most common evaluation and classification of pitch estimation and audio analysis methods in general. The features extracted can be of the frequency domain, time domain or calculating special features like the Mel-frequency cepstral coefficients (MFCCs) in Rao and Ghosh (2017) or using the Constant-Q Transform in Argenti et al. (2011). For timbre estimation the computational research has addressed it as part of a better defined task, being it of the musical instrument identification. The overviews Herrera-Boyer et al. (2006); Hall et al. (2012) also consider features in the time and frequency domains, but differing of the previously discussed in the addressing of the multidimensionality of timbre, and specifying how the task can be performed in isolated sounds and in pitched instruments (refer to Fuhrmann (2012) for a review considering a broader experimental settings apart of just pitched instruments). A work making a review of MIR toolboxes is Müller and Zalkow (2019) having a comprehensive and bread overview of the capabilities of such packages in a educational context. The presented approach differs in to test the robustness of the Constant-Q transform as a timbral and pitch physical correlate and descriptor in a more traditional machine learning fashion.

## 3. Model

The model is a Naïve Bayes Classifier trained using the NSynth dataset Engel et al. (2017). The dataset consists in 305,979 monophonic 16kHz audio snippets recorded from acoustic, electronic and synthesized instruments. It is organized by instrument family (Bass, Brass, Flute, Guitar, Keyboard, Mallet, Organ, Reed, String, Synth Lead, Vocal) covering every pitch of a standard MIDI piano (21-108) in five different velocities (25, 50, 75, 100, 127). Not all instruments are capable to produce some pitches and just as a proof of concept it is not necessary to cover all of them. A subset consisting of 1554 audio snippets was used. Where there are 14 acoustic instruments per each of the 3 selected families (string, guitar, brass) with a constant velocity of 127 and the 37 pitches in the range of C1 and C4. Later the Constant-Q transformation is calculated using the implementation in Librosa McFee et al. (2015) based on Schoerkhuber and Anssi (2010) obtaining the coefficients per frame of the audio snippets corresponding to the range from C1 to B7 (84 pitches) as shown in figure 1. Only the coefficients of the first 16 frames are used as features for the Classifier.

$$Y \leftarrow \arg\max_{y_k} P(Y = y_k) \prod_{i}^{d} P(X_i | Y = y_k) \tag{1}$$

In equation 1 the Constant-Q coefficients are the feature vector X with length $d = 84$ assuming a Gaussian Distribution for the values of energy on each one as in

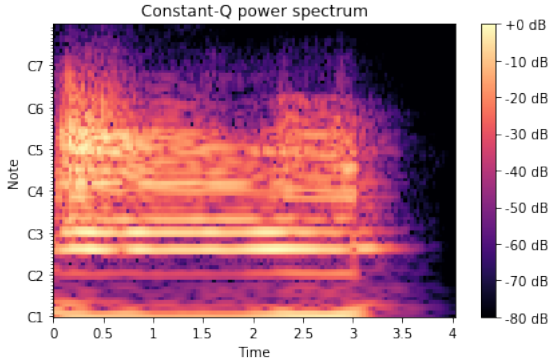


**Figure 1:** Example of the spectrum of the constant-Q transform of a C1 played in a brass acoustic instrument.

the following equation:

$$\mathcal{N}(X_i; \mu; \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(X_i-\mu)^2}{2\sigma^2}} \tag{2}$$

Parameters $\mu$ and $\sigma$ are estimated following:

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ \hat{\sigma}_F^2 &= \frac{1}{n}\sum_{i=1}^{n} (X_i - \hat{\mu})^2 \end{aligned} \tag{3}$$

Where $n$ are the number of samples in the data. The total samples are 24864, 16 frames per pitch, 37 pitches per instrument, 14 instruments per family, with 3 families. We also assume a Categorical Distribution on the pitch and instrument paired label $y_k$, which only consist of dividing the number of samples with such label by the number of total samples. From the total samples 224 correspond for each of the categories, being the 37 pitches for each of the 3 families.

## 4. Evaluation

The model is to be evaluated by running experiments in a K-fold cross validation setting reporting its accuracy overall in the train/test splits of the dataset and a more empirical evaluation of composed monophonic song excerpts.

### 4.1 Experimental Setup

The experiments consist in making 10 repetitions of a stratified K-fold cross validation of 7 partitions to calculate the average of accuracy in the model. In the train/validation split 32 samples will be for validation and 192 for training. Also all the 32 validation samples will be from instrument sources not present in the training split. That is from the 14 instrument sources for each family-pitch paired label 12 will be for training and 2 for validation. The accuracy of the model is reported and a more empirical test with custom recordings as well.

### 4.2 Experimental Results

The model showed an overall accuracy of **93.64%** in the training data and **88.616%** in the test data.

These results encouraged the evaluation of more general environments. The dataset was sampled by the magenta team in Google but to actually assume these results we must use these exact samples to compose songs. Although possible, many music production software as GarageBand[2] and Ableton Live 10 Lite[3] Live include already sampled instruments in the families we used in this paper. So I extended the evaluation to include 6 compositions using a sampled French Horn for Brass Family, a String Ensemble for String Family and a Nylon Concert Guitar for Guitar Family in GarageBand and a Tenor Trombone for Brass Family, a Cello for String Family and a Nylon Concert Guitar for Guitar Family in Ableton Live 10 Lite using the BBC Symphony Orchestra Discover Plug-in[4].

The score is the same for each composition and can be seen in figure 2. Also the constant-Q Spectral plots for each composition are in figures 3 and 4.

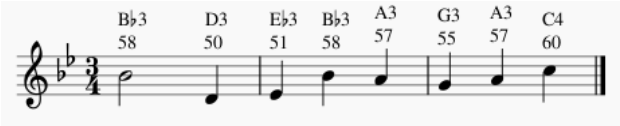


**Figure 2:** Score used for further testing, the note names and MIDI-pitch numbers are shown.

The empirical results point that the system is able to identify the family and note quite precisely having almost all the notes of the score. But more interestingly within the sustain of a note there are some variations on its estimations. In the Guitar recordings it reported some parts as being of a string family instrument and also happened that in the Cello recording it reported some notes as being of a Brass family instrument.

## 5. Discussion

Overall the system did pretty well in the dataset in both train and test splits, but in a general use case, even in still monophonic recordings without noise or reverberation, it behaves quite unstable through time. This can be explained by the lack of features used in training, only the spectral information was used in just a few frames of each recording in the training dataset. Many instruments like the Guitar and Viola share a rather similar sustain, i.e. after the note is played the evolution of the sound until it fades is alike. That temporal features like the temporal envelope considering the attack, decay, sustain and release could improve the stability of the predictions over time. The system although it uses frame-level information, like real-time sound analyzers, it is not suitable for real-time use cases as its predictions take minutes to be computed. It use case could be more of a instrument family recognition and classification of recordings by notes found in it.

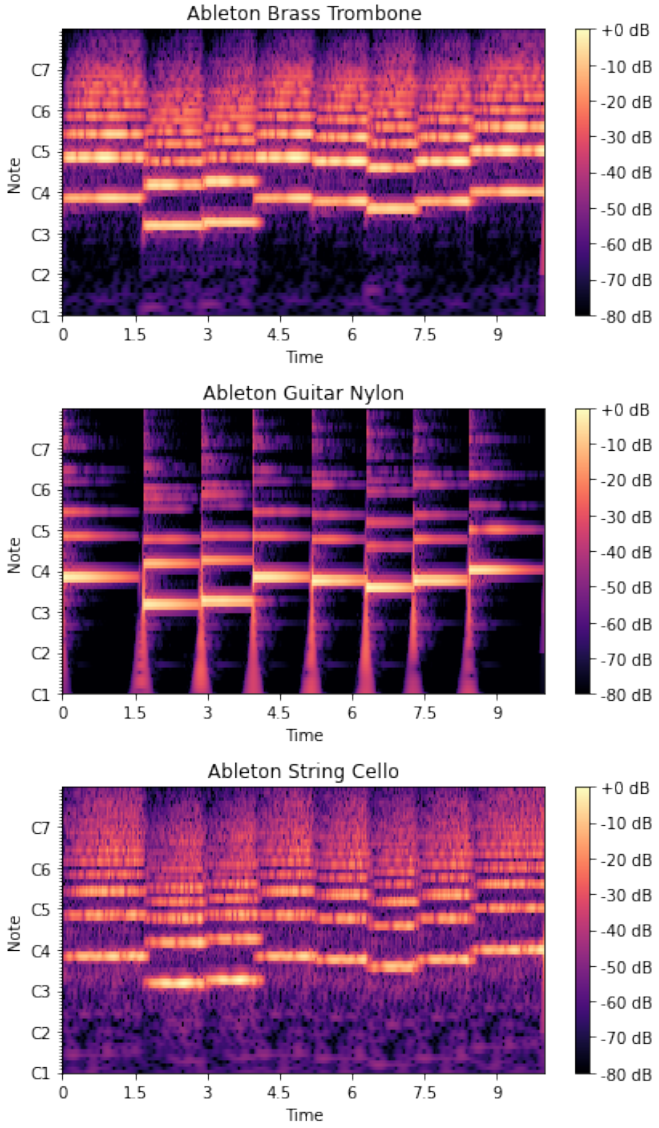


**Figure 3:** Constant-Q plot of the 3 instrument families composed in Ableton Live 10 Lite, from top to bottom: Trombone (Brass), Nylon Guitar (Guitar), Cello (String).

## 6. Conclusion

In this paper a model for musical note and instrument family is presented. Various limitations were imposed for more of a proof-of-concept experimentation. The results give an optimistic view of how the spectral information given by the Constant-Q transform are robust enough to have good results in both the dataset and a more general use case. Many improvements can be made like considering a larger identification space than only 37 notes, the temporal envelope and a larger set of instrument families. A link to the source code repository for further analysis can be found in the Reproducibility section of this paper.

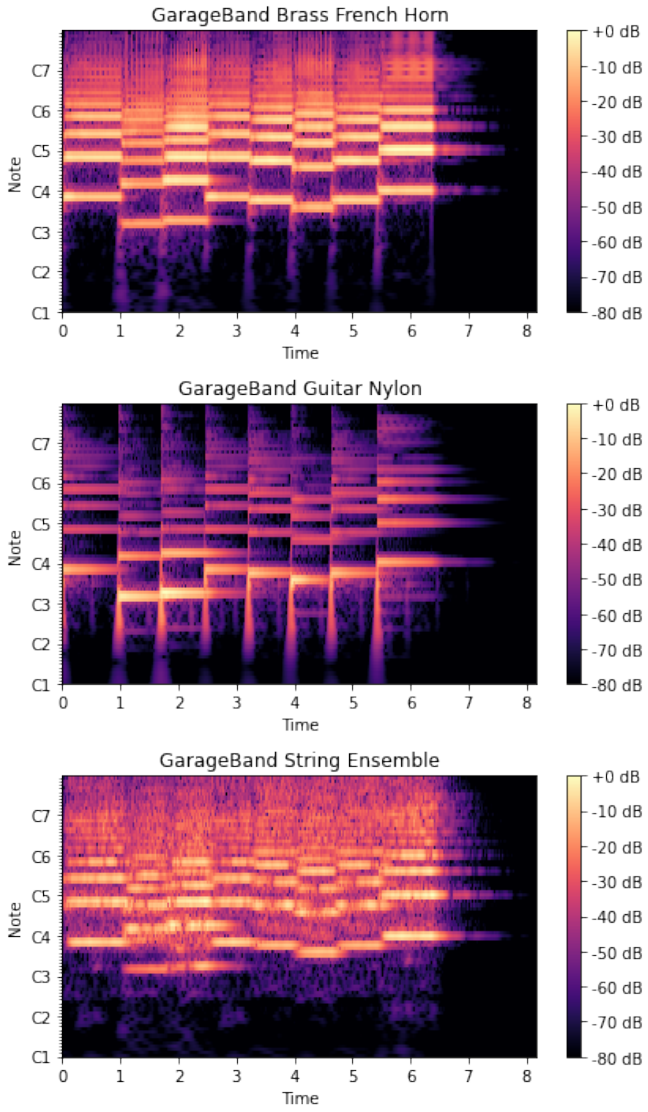


**Figure 4:** Constant-Q plot of the 3 instrument families composed in GarageBand, from top to bottom: French Horn (Brass), Nylon Guitar (Guitar), String Ensemble (String).

## Reproducibility

`https://github.com/Sirivasv/MCC-AA/tree/master/ProyectoFinal`

## Notes

[1] Sevag Hanssian's repository with a compilation of pitch detection methods: `https://github.com/sevagh/pitch-detection`

[2] GarageBand for iOS, `https://www.apple.com/ios/garageband/`

[3] Ableton Live 10 Lite, `https://www.ableton.com/en/products/live-lite/`

[4] BBC Symphony Orchestra Discover, `https://www.spitfireaudio.com/shop/a-z/bbc-symphony-orchestra-discover/`

## Acknowledgements

I gratefully acknowledge the scholarship from CONACyT to pursue my postgraduate studies.

## Competing interests

The author has no competing interests to declare.

## References

Argenti, F., Nesi, P., and Pantaleo, G. (2011). Automatic Transcription of Polyphonic Music Based on the Constant-Q Bispectral Analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1610–1630. Conference Name: IEEE Transactions on Audio, Speech, and Language Processing.

Benetos, E., Dixon, S., Duan, Z., and Ewert, S. (2019). Automatic Music Transcription: An Overview. *IEEE Signal Processing Magazine*, 36(1):20–30. Conference Name: IEEE Signal Processing Magazine.

Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H., and Klapuri, A. (2013). Automatic music transcription: challenges and future directions. *Journal of Intelligent Information Systems*, 41(3):407–434. Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 3 Publisher: Springer US.

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., and Serra, X. (2013). ESSENTIA: an Audio Analysis Library for Music Information Retrieval. In *Proc. International Society for Music Information Retrieval Conference*.

Brown, J. C. (1991). Calculation of a constant Q spectral transform. *The Journal of the Acoustical Society of America*, 89(1):425–434.

Drugman, T., Huybrechts, G., Klimkov, V., and Moinet, A. (2018). Traditional Machine Learning for Pitch Detection. *IEEE Signal Processing Letters*, 25(11):1745–1749. Conference Name: IEEE Signal Processing Letters.

Engel, J., Resnick, C., Roberts, A., Dieleman, S., Eck, D., Simonyan, K., and Norouzi, M. (2017). Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. *arXiv:1704.01279 [cs]*. arXiv: 1704.01279.

Fuhrmann, F. (2012). Automatic musical instrument recognition from polyphonic music audio signals. *Doctoral dissertation*, page 265.

Gerhard, D. (2003). *Pitch Extraction and Fundamental Frequency: History and Current Techniques*. Department of Computer Science, University of Regina. Journal Abbreviation: Department of Computer Science, University of Regina, Regina, Canada Publication Title: Department of Computer Science, University of Regina, Regina, Canada.

Gururani, S., Sharma, M., and Lerch, A. (2019). An attention mechanism for musical instrument recognition. In *Proc. International Society for Music Information Retrieval Conference*.

Hall, G. E., Ezzaidi, H., and Bahoura, M. (2012). Study of Feature Categories for Musical Instrument Recognition. In Hassanien, A. E., Salem, A.-B. M., Ramadan, R., and Kim, T.-h., editors, *Advanced Machine Learning Technologies and Applications*, Communications in Computer and Information Science, pages 152–161, Berlin, Heidelberg. Springer.

Herrera-Boyer, P., Klapuri, A., and Davy, M. (2006). Automatic Classification of Pitched Musical Instrument Sounds. In Klapuri, A. and Davy, M., editors, *Signal Processing Methods for Music Transcription*, pages 163–200. Springer US, Boston, MA.

Kitahara, T., Goto, M., Komatani, K., Ogata, T., and Okuno, H. G. (2005). Instrument identification in polyphonic music: Feature weighting with mixed sounds, pitch-dependent timbre modeling, and use of musical context. In *Proc. International Society for Music Information Retrieval Conference*.

McFee, B., Raffel, C., Liang, D., Daniel, E., McVicar, M., Battenberg, E., and Nieto, O. (2015). librosa: Audio and Music Signal Analysis in Python. In Kathryn Huff and James Bergstra, editors, *Proceedings of the 14th Python in Science Conference*, pages 18 – 24.

Müller, M. and Zalkow, F. (2019). FMP Notebooks: Educational Material for Teaching and Learning Fundamentals of Music Processing. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Delft, The Netherlands.

Rao, M. V. A. and Ghosh, P. K. (2017). Pitch prediction from Mel-frequency cepstral coefficients using sparse spectrum recovery. In *2017 Twenty-third National Conference on Communications (NCC)*, pages 1–6.

Salamon, J. and Gómez, E. (2012). Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20:1759–1770.

Schnupp, J., Nelken, I., and King, A. (2011). *Auditory neuroscience: Making sense of sound*. Auditory neuroscience: Making sense of sound. MIT Press, Cambridge, MA, US. Pages: x, 356.

Schoerkhuber, C. and Anssi, K. (2010). Constant-Q transform toolbox for music processing. *7th Sound and Music Computing Conference*.

Siedenburg, K., Saitis, C., McAdams, S., Popper, A. N., and Fay, R. R., editors (2019). *Timbre: Acoustics, Perception, and Cognition*. Springer Handbook of Auditory Research. Springer International Publishing.

Stol, K.-J., Ali Babar, M., and Avgeriou, P. (2011). The Importance of Architectural Knowledge in Integrating Open Source Software. In Hissam, S. A., Russo, B., de Mendonça Neto, M. G., and Kon, F., editors, *Open Source Systems: Grounding Research*, IFIP Advances in Information and Communication Technology, pages 142–158, Berlin, Heidelberg. Springer.