Musical Instrument Classification from Mel-Frequency Cepstral Coefficients Using a Convolutional Neural Network

1. Introduction

There are countless use cases for sound classification, making it one of the leading applications of audio deep learning. As the name suggests, sound classification uses learning to classify sounds based on their unique features and to predict the category of a sound. Although sound classification is often overlooked by trendier fields such as NLP or computer vision, sound classification has many groundbreaking applications transforming our daily lives. One such application is the identification of instruments from music clips, which is needed for music source separation and can be applied to karaoke, mixing music, instrument-wise equalization of a music recording or other sound engineering tasks.[1] Another area where instrument identification is critical is in the study of music. Take for example an orchestral piece, identifying the correct instruments in the piece can help one understand how the usage and combination of those instruments developed overtime. It can also be essential for the synthesis of a new piece based on a particular music style or time period of music. However, the task of instrument identification remains a challenging areas in music information retrieval as there are minor differences between some instruments.

Automatic identification algorithms can be model-driven or data-driven. Nevertheless, most researchers prefer data-driven machine learning techniques for instrument identification since the algorithm only needs to classify the instrument after extracting identifiable features. Mel-frequency cepstral coefficients, MFCCs, are the most typically extracted features that are fed as input into a neural network in music information retrieval models. Therefore, by employing machine learning, it is possible to implement a classifier for instrument recognition.

In our work, we extract features from audio samples by computing the MFCCs which are then used to train a Convolutional Neural Network to classify the features to the corresponding instrument labels.

2. Related work

The identification of musical instruments from music excerpts has a notable role in various classification tasks in audio deep learning. One prevailing example is music genre classification, since music genre can be distinguished by the instruments present in the piece. For instance, the string banjo and the fiddle are almost exclusively used in the bluegrass/country music genre and the saxophone and cello are most often seen in jazz music.

Because music instrument identification is needed to build content-based recommendation systems and in music genre identification, there has been a lot of significant prior work. Tzanetakis and Cook (2002) [3] pioneered using machine learning methods for music genre classification: Gaussian classifiers, Gaussian mixture models and k-nearest neighbors. They also discussed how features can be extracted from audio: from Mel-frequency cepstral coefficients (MFCC), spectral contract and spectral centroids. Some more recent popular approaches use SVMs, Hidden Markov Models and Neural Networks for the classification task. The input type and architecture employed in Blaszke, Koszewski, and Zaporowski (2019) [9] study provided a good model for our work. Their research showed that converting audio signals into MFCCs and imputing the features into a CNN performed exceptionally well, with an 92.24% accuracy. Table I summarizes the recent literature on instrument classification and highlights the inputs, architecture, and results of the related works from 1999-2021.

Table I: Summary of Related Works to Musical Instrument Classification

Title	Year	Task	Input Type	Architecture	Findings
Deep Convolutional and Recurrent Networks for Polyphonic Instrument Classification from Monophonic Raw Audio Waveforms.	2021	Predominant Instrument Recognition	Raw Audio	RNN, CNN & CRNN	LRAP (label ranking average precision)—0.747, F1 micro—0.608, F1 macro—0.543
Augmentation methods on Monophonic Audio for Instrument Classification in Polyphonic Music.	2021	Instrument Identification	Constant-Q Transform	CNN	LRAP—0.805 F1 micro—0.647 F1 macro—0.546
Research on Music Classification Technology Based on Deep Learning, Security and Communication Networks	2021	Genre Detection	MIDI music	RNN	Accuracy—89.91% F1 macro—0.9
Music Instrument Recognition using Machine Learning Algorithms	2020	Single Instrument Classification	MFCC	ANN & CNN	ANN accuracy—72.08% CNN accuracy—92.24%
Real and Virtual Instruments in Machine Learning—Training and Comparison of Classification Results.	2019	Single Instrument Classification	MFCC	CNN	Precision—0.99 Recall—1.0 F1 score—0.99

Musical Instrument Identification with Supervised Learning.	2019	Single Instrument Classification	MFCC & Warped Linear Prediction Coding	Logistic Regression and SVM (Support Vector Machine)	Accuracy—100%
Instrument Activity Detection in Polyphonic Music using Deep Neural Networks.	2018	Instrument Identification	MFCC	CNN and CRNN	AUC ROC—0.81
Automatic Music Genre Classification Based on Musical Instrument Track Separation	2018	Genre detection	Feature Vector	SVM	Accuracy—72%
Convolutional recurrent neural networks for music classification	2017	Audio Tagging	MFCC	CRNN	ROC AUC (receiver operator characteristic)—0.65-0.98
Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music	2017	Predominant Instrument Recognition	MFCC	CNN	F1 score macro—0.503 F1 score micro—0.602
Timbre analysis of music audio signals with convolutional neural networks	2017	Predominant Instrument Recognition	MFCC	CNN	F1 score micro—0.503 F1 score macro—0.432
Musical Instrument Recognition Using Machine Learning Technique.	2017	Single Instrument Classification	Feature Vector	K-Nearest Neighbors	A system that can listen to the musical instrument tone and recognize it (no metrics shown)
Raw waveform based audio classification using sample level CNN architectures.	2017	Instrument Identification	Raw Audio	CNN	AUC ROC—0.91 Accuracy—86% F1 score—0.45%

Improved Music Genre Classification with Convolutional Neural Networks	2016	Music Genre Classification	STFT Spectrogram	CNN (nnet1 & nnet2)	Accuracy—nnet1—84.8% Accuracy—nnet2—87.4%
Automatic Instrument Recognition in Polyphonic Music Using Convolutional Neural Network	2015	Instrument Identification	Raw Audio, MFCC, CQT	CNN	Accuracy—82.74%
Improving Instrument recognition in polyphonic music through system integration	2014	Instrument Identification	CQT	Missing Feature Approach with Automated Music Transcription	F1—0.52
Musical Instrument Recognition in Polyphonic Audio Using Missing Feature Approach	2013	Instrument Recognition in Polyphonic Audio	A variety of acoustic features	Local Spectral features and Missing-Feature Techniques, Mask Probability Estimation	Accuracy—67.54%
Comparison of Sound Segregation Techniques for Predominant Instrument Recognition in Musical Audio Signals.	2012	Predominant Instrument Recognition	Raw Audio	SVM	F1 score micro—0.503 F1 score macro—0.432
Musical Instrument Recognition in Polyphonic Audio Using Source-Filter Model for Sound Separation	2009	Instrument Recognition in Polyphonic Audio	MFCC	Non-Negative Matrix Factorization & GMM	F1 score—0.62
Musical Instrument Recognition by Pairwise Classification Strategies	2006	Single Instrument Classification	MFCC & FV	Gaussian Mixture Model & SVM	Accuracy—93%

Musical Instrument Classification and Duet Analysis Employing Music Information Retrieval Techniques	2004	Single Instrument Classification	Combined MPEG-7 & Wavelet-Based FVs	ANN	Accuracy—72.24%
Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs	2003	Single Instrument Classification	MFCC	Independent Component Analysis, ML and Hidden Markov model	Accuracy: 62–85%
Musical Instrument Identification Based on F0 Dependent Multivariate Normal Distribution.	2003	Single Instrument Classification	FV	Discriminant Function (Bayes Decision Rule)	Recognition Rate—79.73%
Musical genre classification of audio signals	2002	Genre Detection	MFCC & FV	Subtree Pruning-Regrafting	Accuracy—61%
Representing Musical Instrument Sounds for Their Automatic Classification	2001	Single Instrument Classification	FV	ANN	Accuracy—94.5%
Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features	2000	Single Instrument Classification	FV	K-NN	Accuracy—80%
A Study of Musical Instrument Classification Using Gaussian Mixture Models and Support Vector Machines	1999	Single Instrument Classification	MFCC	GMM & SVM	Error Rate—17%

3. Architecture

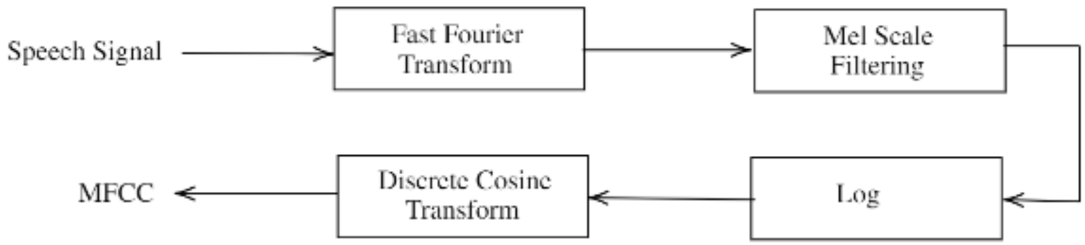
3.1 Data Collection

For the dataset we decided to use IRMAS (https://www.upf.edu/web/mtg/irmas). This dataset contains thousands of audio files representing musical and non-musical instruments (guitar, various sounds…). For this project we restricted our training on 5 instruments: Piano, Voice, Electric Guitar, Trumpet and Clarinet. After filtering and selecting the relevant labels and preprocessing the audio files, we were left with 2700 musical instrument audio samples, at a sampling rate of 22050 Hz, each of them being 3 seconds.

3.2 Feature Extraction-MFCC

In addition to the loudness we took into account the pitch of the raw audio data.
To be able to extract it we used the Mel Frequency Cepstral Coefficients (MFCC).
The MFCC feature variables contain the number of audio data and the heatmap image with the size of 275 times 13 produced using the mfcc() function. As stated previously, MFCCs have been used in a variety of audio mining activities as they demonstrate superior performance compared with other features. MFCCs are the short-term spectral features that are most typically used in the area of music information retrieval as they have been shown to efficiently identify the composition of musical audio signals as well as modeling the subjective frequency and pitch content of those audio signals. This is reasonable as MFCCs are able to capture the phonetically important elements of an audio signal and concisely describe the overall spectral envelope. In our study, we use MFCC features to classify musical instrument signals. We choose MFCCs as our preferred feature based on their superior performance in recent relevant work in instrument/music classification tasks.

As a widely used feature in genre classification systems, MFCC is typically believed to encode timbral information, since it represents short-duration musical textures. According to Wikipedia, “The MFCC is a reputation based on a linear cosine transform of a log power spectrum on a nonlinear Mel-scale of frequency.” [32] Mel is a number which links to a pitch, similar to how a frequency is described by a pitch. The basic flow of calculating the MFCCs is outlined below. [32]

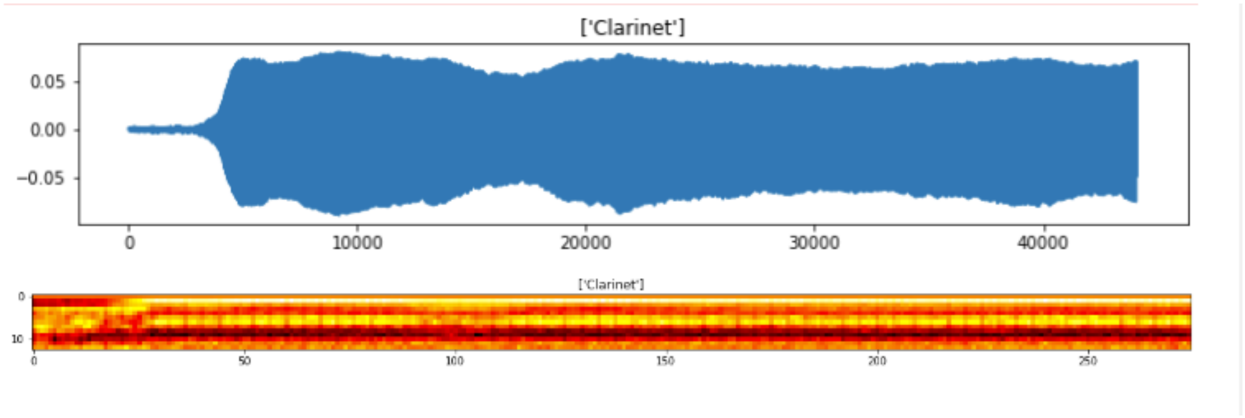


The mathematical formula for frequency-to-mel transform is given as:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right).$$

Therefore, MFCCs are obtained by transforming Hz scale to Mel scale.

In the below image, you see the raw waveform and the heatmap of the second data sample. The heatmap image shows the frequency distribution within each time step. The darkest pixel represents lower energy and the lightest one shows higher energy. The classification will use those heatmaps in a similar way to images classification.



The variations in musical instrument patterns can be captured succinctly after a certain form of transform, such as MFCC in our study and can hence be used as the input image for the CNN model. The musical instrument pattern descriptors extracted with CNN are informative for distinguishing musical instruments from audio pieces, allowing for accurate classification.

<u>3.3 CNN</u>

The model type we used to build our CNN was Sequential, which is a relatively straightforward way to build a model in Keras, allowing us to build a model layer by layer. The CNN architecture that we used to identify instruments was comprised of two Conv2D layers, which are convolution layers that will manage our inputs, with 16 and 32 nodes in each layer. We used a kernel size of 3x3, a rectified linear unit (RELU) activation function, a 2D-MaxPooling layer (used for dimensionality reduction and classification), three 0.5 dropout layers, and 3 dense layers (with the RELU activation function and with the last layer having the softmax activation function), and the Adam optimizer.

<u>3.3 Conclusion</u>

After performing test on various classes we concluded that our model works better for some sounds like electric guitar or the voice but was less effective on other sounds like saxophone (not present in our Jupyter file) or trumpet.

References

1. Rafii Z., Liutkus A., Stoter F.R., Mimilakis S.I., FitzGerald D., Pardo B. An overview of lead and accompaniment separation in music. IEEE/ACM Trans Audio Speech Language Process. 2018;26(8):1307-1335.
2. Sejdic, E., Djurovic, I., Jiang, J. Time-frequency feature representation using energy concentation: An overview of recent advances. Digital Signal Processing. 2009, 19(1):153-183.
3. Tzanetakis, G.; Cook, P. Musical genre classification of audio signals. IEEE Trans. Speech Audio Processing 2002, 10, 293–302.
4. Zhang, W., Lei, W., Xu, X., Xing, X. Improved Music Genre Classification with Convolutional Neural Networks. Proc. Interspeech 2016, 3304-3308, doi: 10.21437/Interspeech.2016-1236.
5. Avramidis, K.; Kratimenos, A.; Garoufis, C.; Zlatintsi, A.; Maragos, P. Deep Convolutional and Recurrent Networks for Polyphonic Instrument Classification from Monophonic Raw Audio Waveforms. In Proceedings of the 46th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021), Toronto, Canada, 6–11 June 2021; pp. 3010–3014.
6. Kratimenos, A.; Avramidis, K.; Garoufis, C.; Zlatintsi, A.; Maragos, P. Augmentation methods on monophonic audio for instrument classification in polyphonic music. In Proceedings of the European Signal Processing Conference, Dublin, Ireland, 23−27 August 2021; pp. 156–160.
7. Zhang, F. Research on Music Classification Technology Based on Deep Learning, Security and Communication Networks. Secur. Commun. Netw. 2021, 2021, 7182143.
8. Shreevathsa, P.K.; Harshith, M.; Rao, A. Music Instrument Recognition using Machine Learning Algorithms. In Proceedings of the 2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM), Dubai, United Arab Emirates, 9–11 January 2020; pp. 161–166.
9. Blaszke, M.; Koszewski, D.; Zaporowski, S. Real and Virtual Instruments in Machine Learning—Training and Comparison of Classification Results. In Proceedings of the (SPA) IEEE 2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications, Poznan, Poland, 18−20 September 2019.
10. Das, O. Musical Instrument Identification with Supervised Learning. Comput. Sci. 2019, 1–4.

11. Gururani, S.; Summers, C.; Lerch, A. Instrument Activity Detection in Polyphonic Music using Deep Neural Networks. In Proceedings of the ISMIR, Paris, France, 23–27 September 2018; pp. 569–576.

12. Rosner, A.; Kostek, B. Automatic music genre classification based on musical instrument track separation. J. Intell. Inf. Syst. 2018, 50, 363–384.

13. Choi, K.; Fazekas, G.; Sandler, M.; Cho, K. Convolutional recurrent neural networks for music classification. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2392–2396.

14. Han, Y.; Kim, J.; Lee, K. Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music. IEEE/ACM Trans. Audio Speech Lang. Process. 2017, 25, 208–221.

15. Pons, J.; Slizovskaia, O.; Gong, R.; Gómez, E.; Serra, X. Timbre analysis of music audio signals with convolutional neural networks. In Proceedings of the 25th European Signal Processing Conference (EUSIPCO), Kos, Greece, 28 August−2 September 2017; pp. 2744–2748.

16. Bhojane, S.S.; Labhshetwar, O.G.; Anand, K.; Gulhane, S.R. Musical Instrument Recognition Using Machine Learning Technique. Int. Res. J. Eng. Technol. 2017, 4, 2265–2267.

17. Lee, J.; Kim, T.; Park, J.; Nam, J. Raw waveform based audio classification using sample level CNN architectures. In Proceedings of the Machine Learning for Audio Signal Processing Workshop (ML4Audio), Long Beach, CA, USA, 4−8 December 2017.

18. Li, P.; Qian, J.; Wang, T. Automatic Instrument Recognition in Polyphonic Music Using Convolutional Neural Networks. arXiv Prepr. 2015, arXiv:1511.05520.

19. Giannoulis, D.; Benetos, E.; Klapuri, A.; Plumbley, M.D. Improving Instrument recognition in polyphonic music through system integration. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP), Florence, Italy, 4−9 May 2014.

20. Giannoulis, D.; Klapuri, A. Musical Instrument Recognition in Polyphonic Audio Using Missing Feature Approach. IEEE Trans. Audio Speech Lang. Processing 2013, 21, 1805–1817.

21. Bosch, J.J.; Janer, J.; Fuhrmann, F.; Herrera, P.A. Comparison of Sound Segregation Techniques for Predominant Instrument Recognition in Musical Audio Signals. In Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012), Porto, Portugal, 8–12 October 2012; pp. 559–564.

22. Heittola, T.; Klapuri, A.; Virtanen, T. Musical Instrument Recognition in Polyphonic Audio Using Source-Filter Model for Sound Separation. In Proceedings of the 10th International Society for Music Information Retrieval Conference, Utrecht, The Netherlands, 9–13 August 2009; pp. 327–332.

23. Essid, S.; Richard, G.; David, B. Musical Instrument Recognition by pairwise classification strategies. IEEE Trans. Audio Speech Lang. Processing 2006, 14, 1401–1412.

24. Kostek, B. Musical Instrument Classification and Duet Analysis Employing Music Information Retrieval Techniques. Proc. IEEE 2004, 92, 712–729.

25. Eronen, A. Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs. In Proceedings of the International Symposium on Signal Processing and Its Applications (ISSPA), Paris, France, 1–4 July 2003; pp. 133–136.

26. Kitahara, T.; Goto, M.; Okuno, H. Musical Instrument Identification Based on F0 Dependent Multivariate Normal Distribution. In Proceedings of the 2003 IEEE Int'l Conference on Acoustics, Speech and Signal Processing (ICASSP '03), Honk Kong, China, 6–10 April 2003; pp. 421–424.

27. Kostek, B.; Czyzewski, A. Representing Musical Instrument Sounds for Their Automatic Classification. ˙ J. Audio Eng. Soc. 2001, 49, 768–785.

28. Eronen, A.; Klapuri, A. Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Istanbul, Turkey, 5–9 June 2000; pp. 753–756.

29. Marques, J.; Moreno, P.J. A Study of Musical Instrument Classification Using Gaussian Mixture Models and Support Vector Machines. Camb. Res. Lab. Tech. Rep. Ser. CRL 1999, 4, 143.

30. Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel P. W. Ellis, Xavier Favory, Jordi Pons, Xavier Serra. *General-purpose Tagging of Freesound Audio with AudioSet Labels: Task Description*, *Dataset*, *and Baseline*. In Proceedings of DCASE2018 Workshop, 2018. URL: https://arxiv.org/abs/1807.09902

31. *AudioSet*. (2017). Http://Research.Google.Com/Audioset/Ontology/Index.Html. http://research.google.com/audioset/ontology/index.html

32. https://en.wikipedia.org/wiki/Mel-frequency_cepstrum#:~:text=Mel%2Dfrequency%20cepstral%20coefficients%20(MFCCs,%2Da%2Dspectrum%22).