Noname manuscript No.
(will be inserted by the editor)

Distinguishing pianos: The difference in similarity

Himadri Mukherjee[⊕,⋆] · **Matteo Marciano**[⊘,⋆] · **Ankita Dhar**[⊖] · **Sk. Md. Obaidullah** [⊗] · **Kaushik Roy**[⊖]

Received: 18-09-2020 / Accepted:

Abstract Notating a music piece is not a trivial task. It requires training and experience. This is challenging for new and inexperienced musicians. Automated transcription systems can be very useful in such cases. A piece generally consists of multiple instruments and it is essential to separate them prior to transcription. The challenge aggravates even more when multiples modulations or varieties of the same instrument are present. This is common in the nowadays and hence, it is essential to distinguish different varieties of the same instrument prior to transcription. In this paper, a line spectral frequency-based approach is presented for this task. Experiments were performed with clips of different lengths from 6 and a highest accuracy of 97.06% was obtained.

Keywords Inter instrument variety · Music signal processing · Lead instrument identification · Line spectral frequency

1 Introduction

Music information retrieval has made significant progress over the years [20], [9]. The roots have propagated to disparate avenues like automatic music transcription [22], instrument identification [17], instrument family identification [11], groove identification [16], raga identification [1], etc. to name a few. There has also been attempts to identify genre of music as well [5]. A music piece is implicitly polyphonic. It consists of a lead melody and the background music. The background music is consists of chords, percussion, and a baseline in the simplest form. Notating music

[⊕]Department of Computer Science
New York University Abu Dhabi, UAE
Email: himadri.mukherjee@nyu.edu
[⊘]Department of Arts and Humanities, Music Program
New York University Abu Dhabi, UAE
Email: matteo.marciano@nyu.edu
[⊖]Department of Computer Science
West Bengal State University, India
Email: {ankita.ankie, kaushik.mrg}@gmail.com
[⊗]Department of Computer Science and Engineering
Aliah University, India
Email: sk.obaidullah@gmail.com
⋆ Authors contributed equally

piece is not a trivial task and requires experience. This is an extremely important aspect for music practitioners and is one of the primal hurdles for studying music pieces or performing them. Automatically notating a monophonic audio is fairly easy with respect to identifying the played notes at the bare minimum. The problem arises for polyphonic pieces. It is important to identify the sections played by a particular instrument and at times separate it from others in overlapping section. Moreover, different varieties of an instrument are often used in a single piece like different pianos or guitars. This is an extremely challenging task for properly notating pieces in an automatic manner. Most of the available works concentrate on music instrument/ family identification.

Eggink and Brown [6] presented a missing feature-based approach for music instrument identification. They used the concept of missing feature in a gaussian mixture model and tested it on 2-tone chords. Sushen et al. [14] used MFCC along with timbral features for music instrument identification. Experiments were performed for 4 instrument families totalling to 8 instruments including piccolo, saxophone, shenai, flute, sarod, santoor, piano, and guitar. Foomany and Umapathy [8] used wavelet-based timescale features for instrument distinction. They reported an accuracy of 85% for distinguishing 13 instruments. Bhalke et al. [3] used MFCC-based features from isolated notes. Experiments were performed with 6 different instruments which were classified with a dynamic time warping-based approach. Ashwini and Krishna [21] used SVM along with feature selection for distinguishing Indian instruments. Experiments were performed with 7 different instruments and accuracies of 97.1% and 93% were reported without and with feature selection respectively. Giannoulis et al. [13] fused shift invariant latent component analysis and polyphonic instrument recognition with the concept of missing features for improving the instrument distinction performance. Essid et al. [7] attempted to distinguish instruments from polyphonic music using taxonomy. Experiments were performed on the Jazz genre due to its tendency of combining multitudinous instruments and an average accuracy of 53% was reported. Degawa et al. [4] used exemplar-based sparse representation for music instrument recognition. They performed automatic transcription as well for both single and multi instrument scenario. Ghosh et al. [12] transformed signals to a spatial 2D representation for distinguishing instruments. This was followed by classification with decision tree. They reported an accuracies of 84.02% for 9 different instruments. Shreevarhsa et al. [19] presented a system with CNN for distinguishing 8 different musical instruments. Their study also involved MFCC, and zero-crossing-based features. Mousavi and Prasath [15] used disparate features like MFCC, roll-off, spectral centroid, zero crossing rate, and entropy enregy for distinguishing Persian music instruments. Experiments were performed with 7 different instruments namely Tar, Ney, Santoor, Kamancheh, Ud, Sitar, and Tonbak. They used fuzzy entropy measure coupled with multilayer neural networks and reported an accuracy of 82.5%. Ajaykumar and Rajan [2] combined gaussian mixture model with deep neural network for identifying dominant instrument in polyphonic music. An accuracy of 93.20% was reported using this approach.

There has been disparate works towards music instrument identification. However, identifying different varieties of a single instrument has not been explored. This is important for proper transcription of a piece in automatic manner at track level. This is because, a piece often consists of multiple varieties of a single instrument to enhance the listening experience. In this paper, line spectral frequencies have been utilized for this purpose. The proposed technique was tested with disparate varities of piano which is one of the most popular instruments in a composition both in the acoustic and synthesized form.

In the rest of the paper, the dataset is detailed in Section 2 followed by the proposed technique in Section 3.1. The results are discussed in Section 4 and finally we have concluded in Section 5.

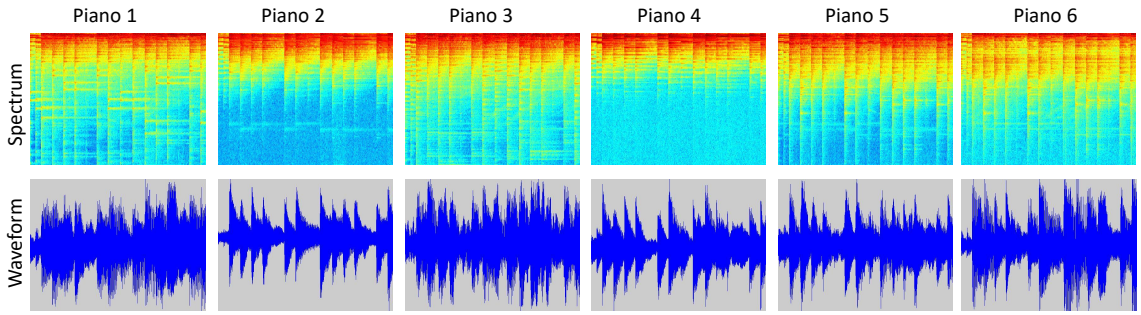
Table 1: Number of clips in each of the engendered datasets

Dataset	# Clips
D_1	1134
D_2	567
D_3	379
D_4	285
D_5	228

2 Dataset description

In this experiment, a dataset of 6 different pianos was used. Each of the pianos were used to play 5 songs. The songs consisted of different dynamics like crescendo and diminuendo and playing styles like staccato, legato, and cantabile. The songs were used to generate 5 datasets having clips of different lengths from 1-5 seconds. The number of clips in each of the datasets is presented in Table 1. There were disparate sections in the pieces which were similar to one another and were in the same key as well. The audios were stored in .wav format at a bitrate of 44100 Hz. All the clips were polyphonic where the diparate layers were played using the same piano which led to resonating effect at times. The percussion sections were ignored while putting together the data. The pianos ranged from acoustic type to electric type There was significant similarity in the tones which is often the case in real World scenarios. The pieces also had distinct notation and time signatures along with the dynamics. This ensured both high intra class difference and interclass similarity. The spectral representation and waveform for a 5 second clip for all the 6 pianos is presented in Figure 1. It is observed that although the same clip was played in terms of notes, tempo, and dynamics but their representations are different. This is primarily due to the timbre variance of the pianos. There was variation in the amount of sustain, loudness, and brightness as well.

Fig. 1: Spectral and waveform-based representation of a 5 second long clip for the different pianos.



3 Proposed method

3.1 Feature extraction

Initially, the audio signals were split into shorter segments for proper characterization. This was done to avoid high variations which is generally observed across the entirety of an audio signal. The signals were split into chunks of 256 points with an overlap of 25%. Thereafter, the segments were subjected to hamming window as presented in Equation 1. This was done to avoid jitters post

framing which interfere with frequency-based analysis.

$$H(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \tag{1}$$

where N denotes frame size and $n \in [1, N]$.

This was followed by extraction of Line Spectral Frequency [18]-based features. This is a technique of representing Linear Predictive Coefficients (LPC) and offers which ensures higher interpolation and better quantization. Here, an audio signal is considered as the output of a filter $H(z)$ whose inverse is $A(z)$ (Equation 2).

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3} + ... + a_M z^{-M}, \tag{2}$$

where $a_{1....M}$ are the LPCs up to the order M. $A(z)$ is decomposed into 2 polynomials $A_1(z)$ and $A_2(z)$ (Equations 3 and 4) whose roots are constitute LSF.

$$A_1(z) = A(z) + z^{-(M+1)}A(z^{-1}) \tag{3}$$

$$A_2(z) = A(z) - z^{-(M+1)}A(z^{-1}) \tag{4}$$

Since features were extracted in the frame level, so feature dimensions vary with the length of clips. Though the number of coefficients per frame remain constant (bands) but the variation is introduced by the disparity in number of frames. This is a primal challenge in processing audio signals as in real World audio signals are of disparate lengths. To get an even dimensional feature and also to capture band level information, mean and standard deviation for every band across all the frames was computed and used as features. This produced a feature whose dimension was dependent on the number of extracted dimensions and not on the clip length thereby evening the dimension. In this experiment, 5-25 dimensional features were extracted with a step of 5.

3.2 Classification

A multi layer perceptron-based classifier [10] was used in the present experiment. It is an add-on of feed forward neural network consisting of 3 types of layers - the input layer, hidden layer and output layer. The input layer acquires the input signal to be processed (features). The job of prediction and classification is carried out by the output layer. An appropriate number of hidden layers are allocated in between the input and output layer where the main computation of the architecture lies. Similar to the feed forward network, the data moves in the forward direction from the input to output layer. The neurons are trained with the back propagation learning algorithm. Multi-layer perceptron are formed to estimate any continuous function and can give solution to the problems that are not linearly separable.

The calculations involved at each neuron in the output and hidden layer are as discussed below: The input layer, consists of a set of neurons n_i where $n_i \in \{n_1, n_2, ..., n_k\}$ denoting the input features. Here, k is the feature dimension. Each neuron in the hidden layer converts the values from the previous layer with a weighted linear summation: $w_1 n_1 + w_2 n_2 + \cdots + w_k n_k$, followed by a non-linear activation function. The output layer obtains the values from the last hidden layer and changes them into output values. It basically gives the probability of belongingness of an instance to every class. The number of neurons in the output layer are same as that of the classes. Here, the network was initially trained with a batch size of 100 instances for 500 iterations. The momentum value was set to 0.2 and the learning rate was set to 0.3.

4 Results and analysis

Initially, each of the 5-25 dimensional feature sets from datasets D_1-D_5 were supplied to the multi layer perceptron whose results are tabulated in Table 2. 5 fold cross validation was used to evaluate the system. The best performance was obtained for D_2 with 20 dimensional features. The intercalss confusions is presented in Table 3.

Table 2: Performance of different feature dimensions for disparate datasets.

Feature dimension	Time (seconds)				
	1	2	3	4	5
5	86.1	88.21	90.28	89.53	91.67
10	93.47	95.06	95.47	94.74	95.25
15	96.15	96.5	96.39	96.14	96.27
20	96.65	**96.85**	96.61	96.37	96.2
25	96.71	96.8	96.61	96.55	96.2

Table 3: Interclass confusions for D_2 with 20 dimensional features.

	1	2	3	4	5	6
1	557	1	6	0	0	3
2	2	555	1	5	0	4
3	6	4	553	2	0	2
4	4	4	3	546	2	8
5	2	5	4	7	544	5
6	8	1	5	3	10	540

The best individual accuracy of 98.24% was obtained for piano 1 followed by piano 2 (97.88%). The least accuracy was obtained for piano 6 (95.24%). Pianos 5 and 6 was also among the most confused pair where 15 clips were confused among each other. The sound of these 2 pianos was very close to each other which is a probable cause for such confusions.

The best performing combination (D_2 with 20-dimensional features) was carried forward for the next phase of experiments. The learning rate was varied from 0.1-0.5 and the best performance was obtained for 0.2 whose interclass confusions is presented in Table 5. In this case, 2 more clips for the 1^{st} piano was misclassified as compared to the default setup. In the case of piano 2, 3, 6 1, 6, and 2 more clips respectively were correctly classified. In the case of piano 4 and 5, there were 2 less correct classifications as compared to the default setup.

Table 4: Performance for different learning rates using 20 dimensional features on D_2.

Learning Rate	0.1	0.2	0.3	0.4	0.5
Accuracy (%)	96.85	96.94	96.85	96.88	96.85

The training iterations was also increased beyond the default of 500 whose results are presented in Table 6. In this case, the learning rate was fixed at the default value. The interclass confusions is presented in Table 7. The performance for piano 1, 4, and 5 was similar to that of the default parameter setup. In the case of piano 6 a slight improvement was observed. However, the best individual accuracy for piano 3 was obtained for a learning rate of 0.2 with 500 training iterations.

Table 5: Interclass confusions for a learning rate of 0.2 on D_2 with 20-dimensional features.

	1	2	3	4	5	6
1	555	3	7	0	0	2
2	0	556	2	5	1	3
3	4	3	559	0	1	0
4	1	4	8	544	3	7
5	1	7	9	3	542	5
6	9	2	6	3	5	542

THe best learning rate (0.2 along with training iteration 1000 iterations) was combined and a lower accuracy of 97% was obtained.

Table 6: Performance of different training iterations for D_2 with 20-dimensional features.

Iterations	500	1000	1500
Accuracy (%)	96.85	97.06	97.00

Table 7: Interclass confusions for 1000 iterations on D_2 with 20-dimensional features.

	1	2	3	4	5	6
1	557	3	5	0	0	2
2	2	555	1	5	0	4
3	6	2	556	2	0	1
4	2	6	3	546	2	8
5	2	4	5	7	544	5
6	8	2	4	2	7	544

Several popular classifiers including Naive Bayes, BayesNet, SVM, RBF, Simple logistic, and Random forest were also applied on the 20-dimensional features computed over D_2 whose results are tabulated in Table 8. The performance of Random Forest was closest to MLP which was followed by Simple Logistic. Thereafter, a sharp decrease in the classification performance was observed for RBF (16.63% less than MLP). The lowest performance was obtained for Naive Bayes wherein an accuracy of only 69.49% was obtained.

Table 8: Performance of different for D_2 with 20-dimensional features.

Classifier	Accuracy (%)
Naive Bayes	69.49
BayesNet	79.54
SVM	75.19
RBF	80.92
Simple Logistic	91.01
Random Forest	95.86
MLP	97.06

5 Conclusion

In this paper, a system has been presented to distinguish different varieties of a single instrument. This is an important aspect in automatic music transcription due to the presence of multiple varieties of a single instrument in a piece. In future, experiments will be performed with a larger dataset consisting of more instruments. Further, more intra class varieties will be introduced. Experiments will be performed in presence of noise to test the system's performance. Further, the audio clips will be parameterized using other handcrafted features. Different feature reductions will be used as well for reducing the feature dimensions so that the system is deployable in resource constrained scenarios. We also plan to explore deep learning-based approaches for characterizing the instruments. The system will also be equipped with real time processing capability and will be deployed in handheld devices.

References

1. Acharya, S., Devalla, V., Amitesh, O., et al.: Analytical comparison of classification models for raga identification in carnatic classical audio. In: Advances in Speech and Music Technology, pp. 211–222. Springer (2021)
2. Ajayakumar, R., Rajan, R.: Predominant instrument recognition in polyphonic music using gmm-dnn framework. In: 2020 International Conference on Signal Processing and Communications (SPCOM), pp. 1–5. IEEE (2020)
3. Bhalke, D., Rao, C.R., Bormane, D.: Dynamic time warping technique for musical instrument recognition for isolated notes. In: 2011 International Conference on Emerging Trends in Electrical and Computer Technology, pp. 768–771. IEEE (2011)
4. Degawa, I., Sato, K., Ikehara, M.: Multipitch estimation and instrument recognition by exemplar-based sparse representation. In: 2013 Asilomar Conference on Signals, Systems and Computers, pp. 560–564. IEEE (2013)
5. Dhall, A., Murthy, Y.S., Koolagudi, S.G.: Music genre classification with convolutional neural networks and comparison with f, q, and mel spectrogram-based images. In: Advances in Speech and Music Technology, pp. 235–248. Springer (2021)
6. Eggink, J., Brown, G.J.: A missing feature approach to instrument identification in polyphonic music. In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)., vol. 5, pp. V–553. IEEE (2003)
7. Essid, S., Richard, G., David, B.: Instrument recognition in polyphonic music based on automatic taxonomies. IEEE Transactions on Audio, Speech, and Language Processing **14**(1), 68–80 (2005)
8. Foomany, F.H., Umapathy, K.: Classification of music instruments using wavelet-based time-scale features. In: 2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp. 1–4. IEEE (2013)
9. Furner, M., Islam, M.Z., Li, C.T.: Knowledge discovery and visualisation framework using machine learning for music information retrieval from broadcast radio data. Expert Systems with Applications p. 115236 (2021)
10. Gardner, M.W., Dorling, S.: Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric environment **32**(14-15), 2627–2636 (1998)
11. Ghosh, A., Pal, A., Sil, D., Palit, S.: Music instrument identification based on a 2-d representation. In: 2018 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), pp. 509–513. IEEE (2018)
12. Ghosh, A., Pal, A., Sil, D., Palit, S.: Music instrument identification based on a 2-d representation. In: 2018 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), pp. 509–513. IEEE (2018)
13. Giannoulis, D., Benetos, E., Klapuri, A., Plumbley, M.D.: Improving instrument recognition in polyphonic music through system integration. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5222–5226. IEEE (2014)
14. Gulhane, S., Badhe, S., Shirbahadurkar, S.: Cepstral (mfcc) feature and spectral (timbral) features analysis for musical instrument sounds. In: 2018 IEEE global conference on wireless computing and networking (GCWCN), pp. 109–113 (2018)
15. Mousavi, S.M.H., Prasath, V.S., Mousavi, S.M.H.: Persian classical music instrument recognition (pcmir) using a novel persian music database. In: 2019 9th International Conference on Computer and Knowledge Engineering (ICCKE), pp. 122–130. IEEE (2019)
16. Mukherjee, H., Dhar, A., Obaidullah, S., Santosh, K., Phadikar, S., Roy, K., et al.: Segregating bass grooves from audio: A rotation forest-based approach. In: International Conference on Recent Trends in Image Processing and Pattern Recognition, pp. 363–372. Springer (2020)

17. Mukherjee, H., Obaidullah, S.M., Phadikar, S., Roy, K.: Misna-a musical instrument segregation system from noisy audio with lpcc-s features and extreme learning. Multimedia Tools and Applications **77**(21), 27997–28022 (2018)
18. Paliwal, K.K.: On the use of line spectral frequency parameters for speech recognition. Digital signal processing **2**(2), 80–87 (1992)
19. Shreevathsa, P., Harshith, M., Rao, A., et al.: Music instrument recognition using machine learning algorithms. In: 2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM), pp. 161–166. IEEE (2020)
20. Vatolkin, I., Ginsel, P., Rudolph, G.: Advancements in the music information retrieval framework amuse over the last decade. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2383–2389 (2021)
21. Vijaya, K.A., et al.: Feature selection for indian instrument recognition using svm classifier. In: 2020 International Conference on Intelligent Engineering and Management (ICIEM), pp. 277–280. IEEE (2020)
22. Zhang, W., Zhang, Y., She, Y., Shao, J.: Stereo feature enhancement and temporal information extraction network for automatic music transcription. IEEE Signal Processing Letters **28**, 1500–1504 (2021)