A Novel Approach to String Instrument Recognition

Anushka Banerjee[1], Alekhya Ghosh[2], Sarbani Palit[3](✉),
and Miguel Angel Ferrer Ballester[4]

[1] Maulana Abul Kalam Azad University of Technology, Kolkata, India
[2] Institute of Radio Physics and Electronics, University of Calcutta, Kolkata, India
[3] Indian Statistical Institute, Kolkata, India
sarbanip@isical.ac.in
[4] Universidad de Las Palmas de Gran Canaria, Las Palmas, Spain

Abstract. In music information retrieval, identifying instruments has always been a challenging aspect for researchers. The proposed approach offers a simple and novel approach with highly accurate results in identifying instruments belonging to the same class, the string family in particular. The method aims to achieve this objective in an efficient manner, without the inclusion of any complex computations. The feature set developed using frequency and wavelet domain analyses has been employed using different prevalent classification algorithms ranging from the primitive k-NN to the recent Random Forest method. The results are extremely encouraging in all the cases. The best results include achieving an accuracy of 89.85% by SVM and 100% accuracy by Random Forest method for four and three instruments respectively. The major contribution of this work is the achievement of a very high level of accuracy of identification from among the same class of instruments, which has not been reported in existing works. Other significant contributions include the construction of only six features which is a major factor in bringing down the data requirements. The ultimate benefit is a substantial reduction of computational complexity as compared to existing approaches.

Keywords: Music information retrieval · Harmonic components
Wavelet coefficients · SVM · Random Forest

1 Introduction

Recognition of musical instruments is quite an easy task for human beings but replicating this process using machines becomes comparatively complex. In the contemporary era of digital music, automatic indexing of music signals is an important requirement. This feature has manifold prospects such as complexity reduction of online music searches and identification of anomaly caused by a particular instrument in a polyphonic music piece.

Automatic recognition of musical instruments primarily requires the extraction of a suitable feature set from the available music signals followed by classification based on them. Over the decades various approaches have been used for

A. Mansouri et al. (Eds.): ICISP 2018, LNCS 10884, pp. 165–175, 2018.
https://doi.org/10.1007/978-3-319-94211-7_19

classification like implementation of Kohonen-self organizing maps using mel-frequency cepstrum coefficients (MFCC) for building timbre spaces [1], using Short term RMS energy envelope and principle component analysis (PCA) with subsequent comparison of an artificial neural network and a nearest neighbor classifier for providing optimum classification ability [2]. Temporal and spectral properties such as rise time, slope of line fitted into RMS-energy curve after attack, crest factor, mean of spectral centroid, average cepstral coefficients during onset etc. were investigated for instrument recognition [3,4]. Mel Frequency Cepstral Coefficients (MFCC) together with linear prediction and delta cepstral coefficients were used yielding a performance of 35% for individual instruments and 77% for instrument families. Gaussian Mixture Models (GMM) and k-Nearest Neighbour (k-NN) model classifier used for instrument classification with line spectral frequencies (LSF) as features [5] produced instrument family recognition of 95% and 90% at the individual instrument level. Instrument recognition by using a source filter model and an augmented Non Negative Matrix (NNM) factorization was also implemented [6]. Uniform discrete cepstrum (UDC) was used to represent timbre of sound sources in a sound mixture [7]. UDC and the corresponding mel-scale variant MUDC significantly outperformed other contemporary representations. A convolutional neural network (CNN) based framework was employed for identification of predominant instrument in polyphonic music [8].

Most of the available works deal with a large set of features, mainly spectral and temporal, such as [9,10] for the classification methods. Moreover, the reported accuracy of identification of instruments within the same family is particularly low. This work aims to design an efficient as well as compact feature set consisting of only six features producing effective classification within the string class of instruments. Hence, the proposed approach not only has the distinction of employing a very small feature set but also, as a consequence of this property, has a low data requirement. Wavelet based features together with the presence and absence of harmonics determined using Fourier analysis, are used to construct the feature set, which is described in Sect. 2. The performance of the proposed algorithm is evaluated over a large dataset of notes and music pieces in Sect. 3. The results establish the approach to be very promising. Section 4 concludes the article.

2 The Proposed Approach

2.1 Music Representation: Notes and Pieces

Notes are the basic building blocks of music. The same note played by different instruments sounds different due to the timbre differences of the instruments. The nature of growth and decay of energy with time is different for different instruments. The conventional pattern of Attack-Decay-Sustain-Release (ADSR) envelope is depicted in Fig. 1. On the other hand, solo instrumental music pieces are combination of several notes played by a particular instrument in a rhythmic manner. The notes are often combined in a random fashion so that partial overlapping of the notes occurs. So, in a piece, different sections of two successive notes may overlap resulting in the deviation of growth or decay of energy of the concerned notes from the ideal nature as represented in Fig. 2.

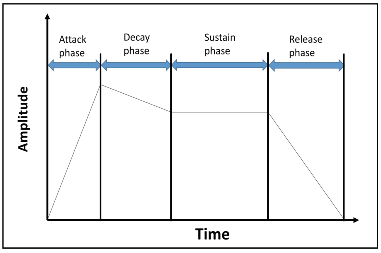


Fig. 1. Time domain ADSR frame model of a single note

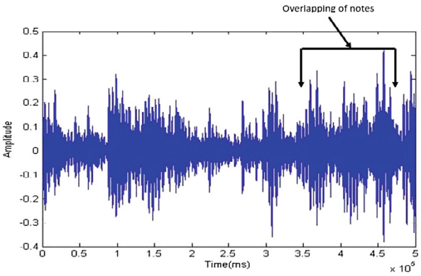


Fig. 2. Time domain representation of a cello instrumental piece showing the overlapping of the notes

It follows that the features that are used for instrument identification will have different values for the two cases. The data considered for this study are:

1. Notes: The dataset [11] comprises of individual music notes for different string instruments namely Cello, Double Bass, Guitar and Violin. For the purpose of analysis, 164 notes of Cello, 154 notes of Double Bass, 210 notes of Guitar and 280 notes of Violin i.e. in total 808 individual music notes have been taken under consideration. Each of these notes is a separate file of nearly 2 s duration in WAV format, with sampling frequency 22,050 Hz.
2. Music pieces: The dataset for music pieces of individual instrument has been obtained from the internet where each music piece is a solo performance of the respective instrument. The duration of each music piece is around 44 s to 60 s. The sampling rate for all the compositions is 44,100 Hz. Prior to the experiment, all these music pieces have been configured using variable window lengths, in order to evaluate and optimize the quality of performance in each case. The window length is chosen to be F_s/k, where F_s is the sampling frequency and $1 \leq k \leq 10$.

2.2 Construction of the Feature Set

The chief mathematical tools which form the backbone of the proposed procedure are the Discrete Fourier Transform (DFT) and the Discrete Wavelet Transform (DWT). Construction of features employing these are now elaborated upon in Sects. 2.2.1 and 2.2.2.

2.2.1 DFT-Based Features

After performing DFT of the entire dataset, the largest DFT coefficient present within a frequency range of 50 Hz. is selected and form a set of similar such coefficients characterizing the data. Next, the harmonic and non-harmonic components are determined and labelled, keeping a tolerance value of 25 Hz. This is done in order to assess the role of the harmonics in characterizing the timbre of the musical note belonging to a particular instrument. Extensive case studies of the notes of various instruments in [11] indicate that the third harmonics as well as the non-harmonic components are well suited to be members of the feature set being developed.

1. Third Harmonic Component: If a frequency present in the frequency spectrum is an integral multiple of the fundamental frequency, where the integer value is 3, it is known as the third harmonic frequency. Presence or absence of this component in the frequency spectrum of a particular instrument has been used as one of the identifying features.
2. Non-Harmonic Component: Those frequency components which are not integral multiples of the fundamental frequency are known as non harmonic components. The presence or absence of these has also been used as a distinguishing feature between the instruments.

These components for the Cello, Double Bass, Guitar and Violin are shown in Figs. 3, 4, 5 and 6 respectively.

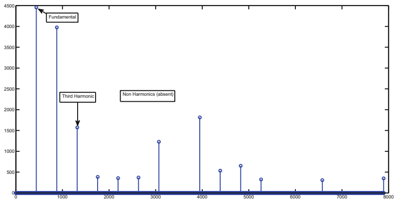


Fig. 3. Cello notes: 3rd harmonic present, non-harmonics absent

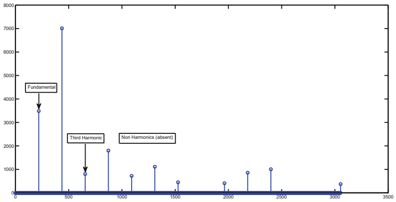


Fig. 4. Double Bass notes: 3rd harmonic present, non-harmonics absent

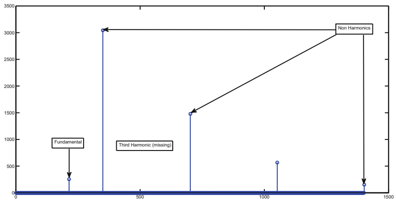


Fig. 5. Guitar notes: 3rd harmonic absent, non-harmonics present

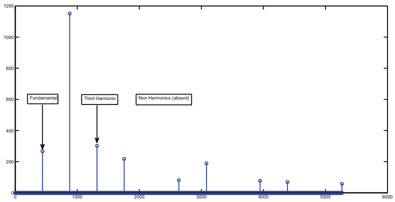


Fig. 6. Violin notes: 3rd harmonic present, non-harmonics absent

2.2.2 DWT-Based Features

After performing a 2-level DWT of the dataset being examined using the Haar wavelet, the mode and standard deviation of the detail coefficients at both Level 1 and 2, are computed. These features exhibit significant differences for different instruments and are hence, included in the feature set.

1. Mode:
   (a) Mode of Level 1: The mode of the detail coefficient vectors of the first level of decomposition for each of the audio signal is calculated. This feature shows significant difference among several instruments. Figure 7 shows the boxplot exhibiting the differences for different instruments.
   (b) Mode of Level 2: Similarly, mode is calculated for the detail coefficients of the second levels also. This feature also shows satisfactory results as depicted in the boxplot of Fig. 8.

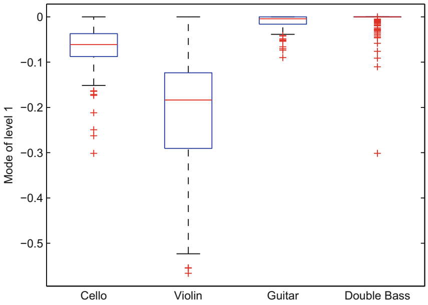


Fig. 7. Boxplot of mode of detail coefficients of Level 1 for solo instrument notes

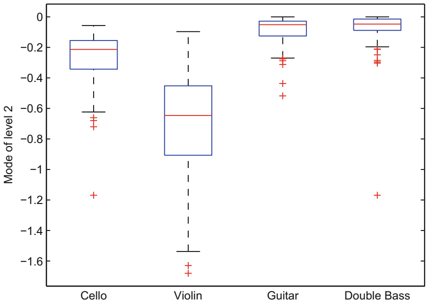


Fig. 8. Boxplot of mode of detail coefficients of Level 2 for solo instrument notes

2. Standard Deviation σ:
   (a) σ of Level 1: The standard deviation of the detail coefficient vector of the first level shows appreciable differences among different instruments as displayed in Fig. 9.
   (b) σ of Level 2: Similarly, standard deviation of the detail coefficient vectors of all the audio signals has been used a feature for the classification. Figure 10 validates the aforementioned statement.

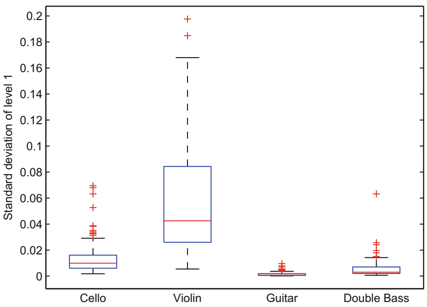


Fig. 9. Boxplot of σ of detail coefficients of Level 1 for solo instrument notes

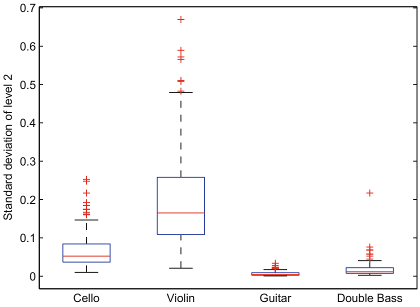


Fig. 10. Boxplot of σ of detail coefficients of Level 2 for solo instrument notes

3 Results

The performance of a simple majority voting scheme using the feature set constructed as explained in Sect. 2.2, was examined as a preliminary exercise. The results are shown in Table 1.

Table 1. Confusion matrix (percentage of accuracy) for identification of notes using majority voting.

	Cello	Guitar	Violin	Double Bass	Indecisive
Cello	57.93	0	7.93	18.90	15.24
Guitar	0	71.90	0	10.00	18.10
Violin	11.79	1.79	66.07	0.36	20.00
Double Bass	13.64	9.74	0.65	57.79	18.18

An interesting point to note from Figs. 7, 8, 9 and 10 (obtained from DWT-based features) is that Guitar and Double Bass are quite difficult to distinguish between. However, the confusion matrix of Table 1 shows the misclassification of the two instruments to be a maximum of 10%. This may be attributed to the difference in the timbre characteristics of the two instruments, better reflected in Figs. 4 and 5 (obtained from DFT-based features). A similar situation arises for the Cello, Violin and Double Bass, which have almost identical values of DFT-based features but appreciably different values for the DWT-based features. In both the scenarios, a combination of the features vastly aided the process of classification.

However, as the performance of the aforementioned approach still leaves much to be desired, the use of standard classifiers was now considered. Three different set of experiments were performed for classification, using the k-Nearest

Neighbor method [12], Random Forest [13] and Support Vector Machine (SVM) methods [14]. 50% of the data has been used for training while the rest was employed for testing for both the Random Forest and SVM based methods.

Performance of these classifiers is evaluated in terms of

$$\text{Precision } P = \frac{TP}{TP + FP} \quad ; \qquad \text{Recall } R = \frac{TP}{TP + FN}$$

where, TP is true positive, FP is false positive and FN is false negative. The measures used finally are the micro averaged F1 measure, denoted as $F1_m$ and the macro averaged F1 measure, denoted as $F1_M$. These are defined as

$$F1_m = \frac{2P_m R_m}{P_m + R_m} \quad ; \qquad F1_M = \frac{2P_M R_M}{P_M + R_M}$$

where

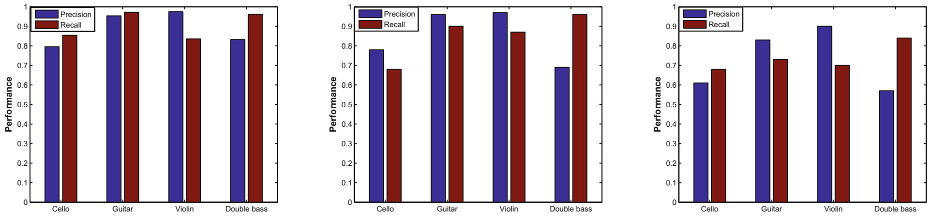
$$\text{micro precision } P_m = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + FP_i} \;\; ; \;\; \text{micro recall } R_m = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + FN_i},$$

$$\text{macro precision } P_M = (1/N) \sum_{i=1}^{N} P_i \quad ; \qquad \text{macro recall } P_R = (1/N) \sum_{i=1}^{N} R_i.$$

N is the total number of instruments while i indicates the ith instrument.

3.1 Identification Based on Notes

Performance of all the methods in terms of precision and recall for all the instruments is presented in Fig. 11 using bar charts. The overall results are clearly the best for the SVM based method, yielding the best precision for Violin and the best recall for Guitar.



(a) SVM (b) Random Forest (c) k-NN

Fig. 11. Performance in terms of precision and recall

Table 2 presents the overall result using SVM, summarized in a confusion matrix. It may be observed that each row corresponds to a different instrument.

Table 2. Confusion matrix (percentage of accuracy) for identification of notes using SVM

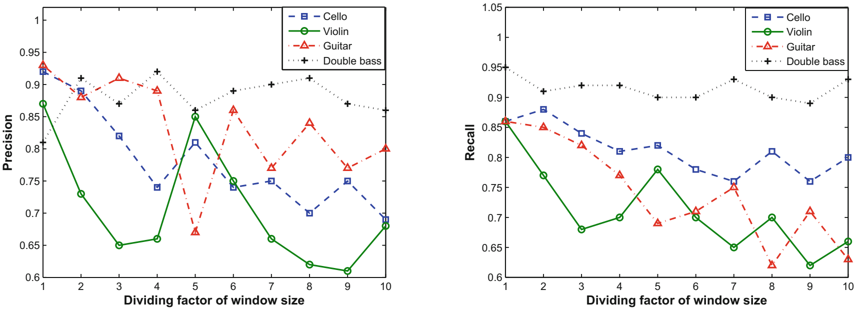
	Cello	Guitar	Violin	Double Bass
Cello	85.37	3.66	3.66	7.32
Guitar	0	97.14	0	2.86
Violin	12.14	0	83.57	4.29
Double Bass	1.30	2.60	0	96.10

Hence, while each row sums to 100%, no such relationship exists between the rows of any particular column.

Similarly, we observe that for all the three classification techniques *viz.*, Random Forest, SVM and k-NN, the best precision is obtained for Violin while Double Bass shows the best recall. The value of recall for Violin and the precision for Guitar is relatively better in Random Forest classification. For all other cases SVM yields better result among the three methods.

3.2 Identification from Solo Instrument Pieces

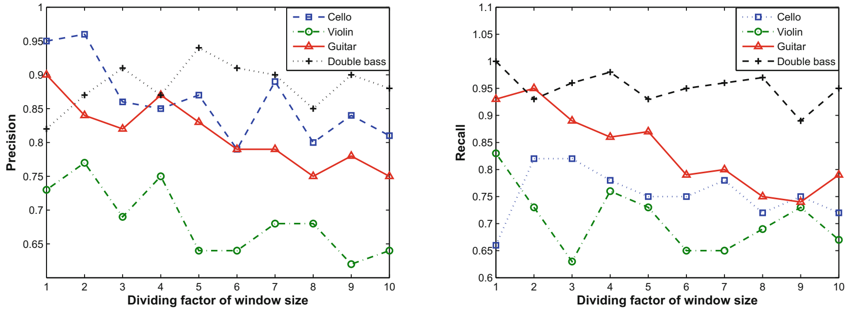
The performance of the Random Forest method for solo instrument pieces is presented in Fig. 12. It may be observed that the precision as well as recall of Cello, Guitar and Violin attain maximum values when window length is maximum (dividing factor is minimum) and gradually decays with decreasing window length. On the other hand, Double Bass maintains a nearly constant high value of precision and recall for different window sizes.



(a) Precision (b) Recall

Fig. 12. Performance of Random Forest method in terms of precision and recall for varying window size (in terms of dividing factor)

Figure 13 shows the performance of the SVM method for solo instrument pieces. It may be concluded that with decreasing window size the recall value



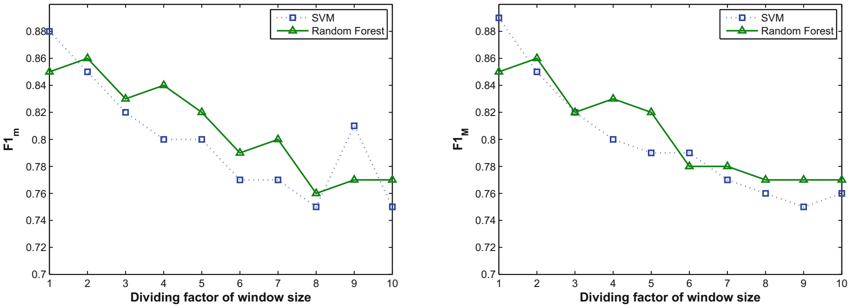
(a) Precision

(b) Recall

Fig. 13. Performance of SVM-based method in terms of precision and recall for varying window size (in terms of dividing factor)

falls for all the instruments except for the Double Bass. This instrument exhibits a comparatively high value of recall amongst all the instruments for all window sizes. The Cello, Violin and Guitar, all show a high precision for larger windows and decreasing precision with decreasing window sizes. Double Bass, however, has a lower precision for large windows which increases with decreasing window size.

From Fig. 14, it can be observed that for higher window sizes, the SVM classifier dominates over Random Forest. But as the window size decreases, the Random Forest method outdoes the SVM method in terms of $F1_m$ and $F1_M$. Considering all the results, a reasonable choice of the window size would be F_S for the SVM and $F_S/2$ for the Random Forest based method. Corresponding results for the k-NN classifier have not been included for the sake of brevity.



(a) $F1_m$

(b) $F1_M$

Fig. 14. Comparison of performance of SVM and Random Forest

A comparative study of the performances obtained using the k-NN, Random Forest and SVM classifiers has been presented in Table 3. It may be observed

Table 3. Accuracy (percentage) measure for the three methods.

Instrument	Accuracy (percentage) of 4 instruments			Accuracy (percentage) of 3 instruments		
	k-NN classifier	Random Forest	SVM	k-NN classifier	Random Forest	SVM
Cello	67.68	68.29	**85.36**	92.07	100	95.12
Guitar	72.86	90.48	**97.14**	80.00	100	100
Violin	70.00	**87.14**	83.57	70.00	100	94.29
Double Bass	83.77	**96.10**	**96.10**	–	–	–
Combined	73.57	85.89	89.85	80.69	100	96.33

that the SVM based method delivers better rates of accuracy, precision as well as recall value for almost all the test cases. But it is noteworthy to mention that although accuracy measures might vary, the feature set selected delivers correct identification results for all the three sets of classification methodologies implemented. It may also be noted from Table 3 that the overall performances of all the classification algorithms get better with a smaller set of instruments. The maximum boost in performance is observed in case of the Random Forest based method.

4 Conclusion and Scope for Future Work

Binary representation of harmonic and non-harmonic components combined together with wavelet domain features proved to be efficient in instrument recognition. Although the final performances vary for different classification methods, all of them have shown better classification results than the existing approaches. For four instrument classification for both solo notes and music pieces, SVM proves to be the best method while Random Forest proved to be the best when three instruments were considered.

The variation of the feature values from notes to that of pieces may be attributed to the fact that in the composite piece, the complete attack-decay-sustain-release envelope is not expressed completely. Though there is considerable overlap of envelopes, the sense of timbre for a particular instrument is still conserved and hence the same features yield good classification results. Since preliminary results are greatly encouraging, the incorporation of other time and frequency domain features are being explored for further improvement of performance.

References

1. De Poli, G., Prandoni, P.: Sonological models for timbre characterization. J. New Music Res. **26**(1997), 170–197 (1997)
2. Kaminsky, I., Materka, A.: Automatic source identification of monophonic musical instrument sounds. In: Proceedings of the 1995 IEEE International Conference of Neural Networks, pp. 189–194 (1995)

3. Eronen, A., Klapuri, A.: Musical instrument recognition using cepstral coefficients and temporal features. In: Proceedings of 2000 IEEE International Conference on Acoustics, Speech Signal Processing, vol. 2, pp. II753–II756 (2000)
4. Eronen, A.: Comparison of features for musical instrument recognition. In: 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 19–22 (2001)
5. Krishna, A.G., Sreenivas, T.V.: Music instrument recognition: from isolated notes to solo phrases. In: Proceedings of IEEE International Conference on ASSP, vol. 4, pp. iv-265–iv-268 (2004)
6. Heittola, T., Klapuri, A., Virtanen, T.: Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In: Proceedings of International Society for Music Information Retrieval Conference, pp. 327–332 (2009)
7. Duan, Z., Pardo, B., Daudet, L.: A novel cepstral representation for timbre modeling of sound sources in polyphonic mixtures. In: Proceedings of 2014 IEEE International Conference on ASSP, pp. 7495–7499 (2014)
8. Han, Y., Kim, J., Lee, K.: Deep convolutional neural networks for predominant instrument recognition in polyphonic music. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(1), 208–221 (2017)
9. Foomany, F.H., Umapathy, K.: Classification of music instruments using wavelet-based time-scale features. In: IEEE International Conference on Multimedia & Expo Workshops (2013)
10. Kothe, R.S., Bhalke, D.G.: Musical instrument recognition using wavelet coefficient histograms. In: Proceedings of Emerging Trends in Electronics and Telecommunication Engineering (NCET 2013), pp. 37–41 (2013)
11. Institute for Research and Coordination in Acoustics/Music (IRCAM), Pompidou, France
12. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Trans. Inf. Theor. **13**(1), 21–27 (1967)
13. Breiman, L.: Random forests. Mach. Learn. **45**(1), 532 (2001)
14. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Disc. **2**, 121–167 (1998)