Designing a training set for musical instruments identification[⋆]

Daniel Kostrzewa[0000−0003−2781−3709], Blazej Koza, and Pawel Benecki[0000−0003−4674−5393]

Department of Applied Informatics, Silesian University of Technology, Gliwice, Poland {daniel.kostrzewa,pawel.benecki}@polsl.pl

Abstract. This paper presents research on one of the most challenging branches of music information retrieval – musical instruments identification. Millions of songs are available online, so recognizing instruments and tagging them by a human being is nearly impossible. Therefore, it is crucial to develop methods that can automatically assign the instrument to the given sound sample. Unfortunately, the number of well-prepared datasets for training such algorithms is very limited. Here, a series of experiments have been carried out to examine how the mentioned methods' training data should be composed. The tests were focused on assessing the decision confidence, the impact of sound characteristics (different dynamics and articulation), the influence of training data volume, and the impact of data type (real instruments and digitally created sound samples). The outcomes of the tests described in the paper can help make new training datasets and boost research on accurate classifying instruments that are audible in the given recordings.

Keywords: Music information retrieval · Musical instruments identification · Dataset design · Training data · Analog sound · Digital sound

1 Introduction

One of the most challenging parts of music information retrieval is identifying the musical instrument. In order to fulfill this goal, many algorithms for automatic instrument recognition were developed. Different machine learning strategies can be used for automatic instruments classification. They can be entirely computational [4, 5] or perceptual [17, 18]. The task can be carried out with classic classifiers [6, 7], by analyzing fundamental frequency [14], the usage of hidden Markov models [10], and based on deep neural networks [12, 21]. In this research, the simple comparison of the distance of mel-frequency cepstral coefficient (MFCC) vectors was used [11].

The task of automatic musical instruments recognition has a strong practical justification. People often like listening to similar pieces of music. This is dictated

[⋆] This work was supported by the Excellence Initiative - Research University program carried out at the Silesian University of Technology (DK) and research funds for young scientists of Silesian University of Technology, Gliwice, Poland (PB).

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_43

by their musical taste and current mood. As the Internet allows access to millions of tracks, automatic instrument recognition has become a necessity. However, creating high-quality classification methods is not enough, it is crucial to provide appropriate training data to achieve the highest results.

1.1 Related Work

As the possibilities and popularity of deep learning methods have grown in recent years, the need to create a well-prepared dataset for training such algorithms has become urgent. Unfortunately, the task of instrument classification is undertaken by researchers exceptionally rarely due to its difficulty. Therefore, the conducted studies are performed on various, far from being perfect, databases.

There is no single recognized dataset for the task of identifying musical instruments. However, there are several collections worthy of attention. The NSynth Dataset [9], TinySOL [8], and Good-sounds [2, 19] contain 305979, 2913, 8750 single annotated notes, respectively, which is not enough (because of having separate notes) for train meaningful model. These collections provide some limited information about the dynamics, type of play, and quality. Medley DB [3] collects 122 songs with information about instruments, while Medley-solos-DB [1,15] contains 21571 audio clips of 8 instruments with a fixed duration of 2972 milliseconds, that is, 65536 discrete-time samples. However, it does not offer information about the different playing techniques, dynamics, and articulation. Last but not least, the OpenMIC-2018 dataset [13] contains 2000 excerpts (10s each) from 20 instruments with no additional information about dynamics and articulation.

Moreover, according to our best knowledge, there is no research on the impact of different aspects of designing training dataset for musical instruments recognition methods. We believe that broadening training datasets with recordings with different dynamics, articulation, data sources (real instruments and digitally created sound samples) will allow for significant improvement in the quality of musical instruments recognition.

1.2 Contribution and Paper Structure

The main goal of the research is to determine the best approach for designing a training set for the task of musical instruments recognition from an audio signal. Several key aspects have been thoroughly examined, i.e., the base confidence of the classification, the impact of the different sound characteristics in the training set, the influence of training data volume, and the impact of data type (analog and digital instruments). Moreover, a simple method for automatic identification of musical instruments was developed. This strategy consists of frequency detection, parameterizing the sound, and classifying it. The analysis is based on samples of five real musical instruments: piano, clarinet, alto saxophone, trumpet, and accordion. All sound samples come from the authors' private recordings. To ensure reliable results, the sound samples have been recorded with different articulation and dynamics to reproduce the instrument's pattern as accurately as possible.

The proposed method created for musical instruments identification is presented in Section 2. Section 3 describes the used dataset, the conducted research, and the obtained results, while the summary and the final conclusions have been included in Section 4.

2 Identification of Musical Instruments

2.1 Extraction of Sound Features

The critical issue in designing a method that enables automatic recognition of the instrument's timbre is the digital representation of sound characteristics, which can be appropriately processed and classified. The way the sound recordings of musical instruments are parameterized has a great influence on the final effectiveness of the algorithms that classify the sound. It is necessary to determine the vector of sound characteristics, whose size, i.e., the number of parameters determining the sound, should be as small as possible. This will allow the representation of strongly correlated parameters to reduce the computational complexity and more accurate ordering of information. This is important because the recordings are easy to be compared with each other.

In digital tone analysis, one of the most frequently used sound parameterization methods is the usage of Mel-frequency cepstral coefficients (MFCCs) [20]. MFCCs are computed in a following steps:

1. Calculate the Fourier transform of the windowed part of the signal.
2. Map the powers of the spectrum obtained to the mel scale using overlapping triangular windows or cosine windows.
3. Perform logarithms of the powers for each mel frequency.
4. Perform a discrete cosine transform of the list of logarithmic mel powers as if it were a signal.
5. The MFCC values are the amplitudes of the resulting spectrum.

In the developed method, the input audio signal's conversion into the MFCC feature vector is done using the Accord.NET library. The resulting vector (i.e., a vector representing an unknown musical instrument) is compared with the vectors obtained from the reference sounds (i.e., all reference vectors of known musical instruments). Distances calculated between the vector of the unknown instrument and reference vectors for each musical instrument are averaged, and this average value is the final distance between the vector representing the unknown instrument and instruments stored in the dataset.

In order to visualize the results, a set of graphs from MFCC vectors was created. For each graph, the horizontal axis determines the frequencies, and the vertical axis indicates the signal power of a particular frequency. The values on the axes do not correspond to the real values of frequencies and decibels. The scaling of the graph is done automatically by the Accord.NET library. Therefore, the graphs are only used to illustrate the sound vector of the instrument (Fig. 1).

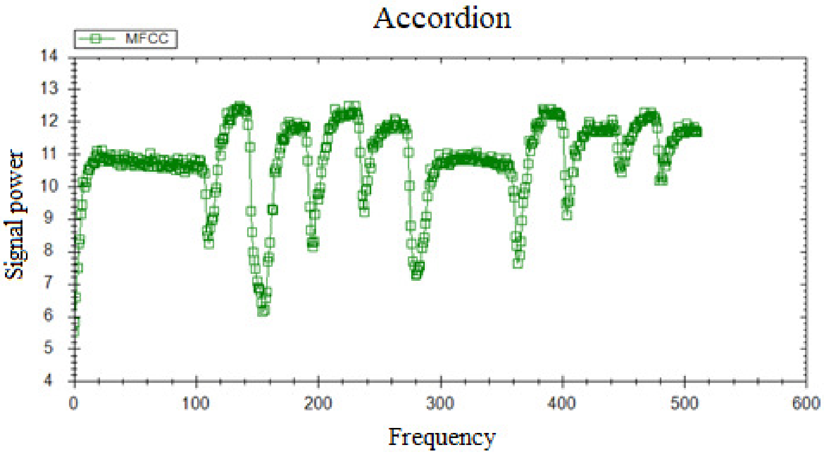

Fig. 1. Sample reference vector for instrument used in experiments.

2.2 Classification Method

Appropriate data classification is another element necessary for automatic recognition of the timbre of musical instruments. Artificial neural networks [12,16,21], minimal-distance algorithms including the closest neighbor method [4,6], are the most often used for this purpose. Since the main goal of this research is to determine the best approach for designing a training set for the task of musical instruments recognition from an audio signal, only the simple closest neighbor algorithm is employed for the task of classification. The comparison consists of finding the absolute distance between the input sound vector and individual pattern vectors (reference sound vectors). The lowest average distance between the input sound vector and patterns of the given instrument means matching the input sample to this particular instrument. The distance between two n-dimensional vectors is calculated from the Euclidean distance:

$$|u - v| = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + ... + (u_n - v_n)^2} \tag{1}$$

where u and v are vectors belonging to the space R^n and represent two different sounds in the MFCC domain.

3 Experiments

3.1 Dataset and Hardware Setup

It was decided to analyze various sounds with different properties in the conducted research. It is also important to compare effectiveness of the classification method on instruments with similar sounds (belonging to the same group, e.g.,

DOI: 10.1007/978-3-031-08751-6_43

woodwind) and instruments belonging to different groups. Therefore, it was decided to use five instruments, which will allow us to evaluate the method and designing training set procedures by different aspects: alto saxophone, clarinet, trumpet, accordion, and piano. Clarinet and saxophone, belonging to the same woodwind group, allow testing the classification of similarly sounding instruments (which can be difficult to distinguish). The trumpet is another instrument of the wind group, but the way it produces sound is different from the clarinet or saxophone and therefore belongs to the brass family. The next two instruments are the piano and the accordion. The piano is an instrument belonging to the chordophones group, where the sound is produced by hitting a hammer on a string. In the accordion, the sound, as in the wind instruments, is produced by air flowing from bellows through reeds activated using a keyboard. The possibility of playing chords determined the choice of the last two instruments. This will make it possible to test the effectiveness of polyphonic and monophonic sounds. Also, various possible sounds of accordions (they have so-called *registers*, which allow changing the timbre of the sound) were recorded to create a robust instrument reference vector.

The first phase of data collection was the recording of instrument sound samples. This was done using the RODE NT2-A microphone, the Focusrite Scarlett 2i2 analog-to-digital converter, and the free software for multi-track sound recording – Audacity. All samples were recorded in a 44.1 kHz .wav format with a resolution of 16 bits. All instruments' sounds were recorded in one room and on an identical hardware configuration to ensure the highest possible reliability of the experiment. To reflect the sample sound, identical sequences were recorded for each of the five instruments: ranges, passages, melodies, and long sounds. In the case of accordion and piano, apart from single sounds, chords were used. Each recording was made with different dynamics and articulation. A total of 70 recordings were made for each instrument, 60 were used as the training set, and 10 – as the test set.

The dataset consisting of instrumental sound recordings was additionally enriched with recordings from digital sound synthesizers (Logic Pro X application on Mac OS) in order to be able to classify both types of sound sources for each instrument correctly. This allows generating sounds of many instruments with different dynamics and articulation. For each of the five instruments tested, twenty longer samples were recorded, which were divided into smaller fragments for each experiment at the testing phase. As a result, a total of 450 recordings were prepared (real instruments and synthesizers together).

3.2 Assessment of the decision confidence

The first part of the experiments was to determine the overall assessment of the proposed method's confidence. This was necessary to determine whether the algorithms were implemented correctly and whether further research makes sense. It was assumed that the tested sound sample belongs to a given instrument if the distance between the reference vector, i.e., averaged MFCC vector for the given instrument, and the tested instrument vector is the smallest. The result of the analysis process is the distance between reference and sample vectors. The determination of the decision confidence required a relative percentage of confidence based on the distance obtained. For this purpose, all the distances between vectors were collected, and the smallest and the largest were selected. The lowest value was determined as 100% confidence, and the highest value corresponds to 0%. Since simple accuracy shows only classification results, we introduced a more insightful metric. The idea behind creating such a metric shows how instruments can be differentiated from one another.

In this experiment, all fifty test recordings were used (ten from each instrument), and an average of 60% of the decisions confidence was achieved. The highest confidence was obtained for the clarinet (79%). The piano, accordion, and trumpet were recognized with confidence of 65%, 61%, and 57%, respectively. The saxophone has lowest confidence – 38%.

3.3 Impact of sound characteristics

The next part of the research was to find the optimal method configuration and training samples to build a universal sound pattern for a given instrument. The detection accuracy was tested depending on the constants defining the dimension of a single vector of sound signal features and the number of vectors in a two-dimensional MFCC array. In addition, the influence of the characteristics of the training recordings on the results was analyzed. The sound samples were sorted into groups with different dynamics, articulation, and pitch. The values of constants determining the dimensions of the MFCC matrix were set accordingly: 512 – the dimension of a single vector of sound signal features and 13 – the number of vectors in the MFCC object.

The sound emitted by the instrument depends on the musician's style of playing, habits, or the genre of music performed. For example, the characteristics of a ballad melody are very different from rock or pop music. This is due to various dynamics and articulations in sound. In this experiment, the influence of differentiation of the dataset in terms of sound characteristics on the proposed method's accuracy was observed. The data were divided into three groups: X1 – long sounds with similar dynamics and articulation, X2 – X1 set enriched with samples with different dynamics, and X3 – X1 and X2 sets with added recordings with different articulation. Training sets contained 20 samples each (for each instrument). Importantly, all three groups (i.e., X1, X2, and X3) have 20 samples of each instrument, however, their composition is different. 10 separate recordings (for each instrument) were used for testing. The influence of data differentiation on detection accuracy was observed, and the results of this experiment are presented in Table 1.

After enriching the dataset with samples of different dynamics, the average accuracy increased by 2.2%. The most significant improvement was achieved for the accordion and trumpet. In wind instruments, where a stronger airflow causes higher dynamics, the timbre changes, it is sharper and clearer. In the case of the piano, the change of dynamics does not significantly affect the timbre, and the accuracy of the proposed method has not increased.

DOI: 10.1007/978-3-031-08751-6_43

Table 1. The impact of dataset diversity on the proposed method accuracy.

Instrument	X1	X2	X3
Clarinet	58%	59%	65%
Saxophone	30%	32%	33%
Trumpet	49%	53%	53%
Piano	63%	63%	65%
Accordion	55%	59%	59%
Average	**51%**	**53.2%**	**55%**

The enrichment of the training set with various articulation recordings allowed to increase the average accuracy by 1.8%. The recordings were made in the following techniques: staccato (separate sounds, with shortened values), legato (sounds played smoothly, without any breaks between them), glissando (smooth transition from one sound to another) and, in the case of accordion and piano, arpeggio (chord broken into a sequence of notes). The highest accuracy increase (6%) was achieved for the clarinet, while the trumpet and accordion were recognized with the same accuracy as for the X2 set.

3.4 The influence of training data volume

The variety of instrument sounds is practically endless. The sound of one instrument may vary in many ways and for many reasons. These include the previously described sound characteristics, the musician's playing style, and the materials used to make the instrument. The developed method compares the MFCC signal vector with a previously prepared pattern that is a MFCC vector of individual recordings with different articulation and dynamics of a given instrument. For this reason, the sound patterns of instruments should be as varied as possible, which requires much training data.

In this experiment, the number of training samples on the effectiveness of method accuracy was examined. While maintaining the diversity of sound characteristics, six datasets of sizes 1, 5, 10, 20, 40, and 60 samples were prepared (for each instrument). For the tests, 10 separate recordings (for each instrument) were used. The results obtained in this experiment are presented in Table 2.

Table 2. The influence of the amount of training data on the effectiveness of the proposed method.

Instrument	Y1 – 1	Y2 – 5	Y3 – 10	Y4 – 20	Y5 – 40	Y6 – 60
Clarinet	25%	40%	50%	65%	70%	75%
Saxophone	15%	20%	30%	35%	45%	50%
Trumpet	30%	40%	40%	50%	60%	60%
Piano	20%	40%	45%	65%	75%	75%
Accordion	55%	55%	55%	60%	75%	80%
Average	**29%**	**39%**	**44%**	**55%**	**65%**	**68%**

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_43

The obtained results of the proposed method accuracy proved to be strongly dependent on the size of the training dataset. The lowest accuracy was achieved for the Y1 set, where for each instrument, the dataset contained only one sample. The obtained results ranged from 15-55%, and the average recognition was 29%. The lowest 15% result was achieved for the saxophone, which was most often mistaken with a clarinet. This may be since these are instruments from the same family (i.e., woodwind).

The gradual enrichment of the dataset allowed for a considerable improvement in decision-making quality. For the last dataset (Y6), the method recognized instruments with an average accuracy of 68%. The improvement varies from 25% for the accordion up to 55% for the piano. The lowest average performance difference (3%) was observed between the Y5 and Y6 datasets, it can be concluded that the size of the Y6 dataset allows for an impeccable reproduction of the timbre pattern of the instruments and high accuracy sound recognition. A small difference in accuracy may also indicate the upper limit of the amount of data needed to reproduce the timbre pattern.

3.5 The impact of data type

Modern technology allows for sound production through analog musical instruments and the use of software for music production, sound generators, and synthesizers. Training data for timbre recognition cannot be limited to analog instruments only because the pace of digital technology development and the development of electronic music genres can make such a method less and less useful over time.

As a result, it was decided to enrich the set of training data with digital instruments samples and investigate the effectiveness of recognizing such sounds. In the first part of the study, the accuracy of recognizing digital instrument sounds on a set of training data recorded on analog instruments was examined. For this purpose, the Y6 set from the previous experiment with 60 samples size was used. The results obtained during the conducted research are shown in Table 3 (column a), which compares the obtained results with those from the previous experiment (column b).

The average recognition accuracy of digital instruments is 16% lower. This may indicate poor timbre reproduction of instruments by synthesizers. The most significant timbre deviation was observed for the accordion, where the difference was 50%. In the case of saxophone and piano, the accuracy of the method did not change and remained at 50% and 75%, respectively.

Another aim of the experiment was to add samples of digital instruments to the training data and retest the effectiveness of the classification. The dataset was enriched with 10 samples of digital instrument sounds, and the results are presented in Table 3 (column c). The outcomes obtained show that training data from various sources increase the effectiveness of the developed method significantly. After extending the training dataset with samples of digital instruments, an increase of 17% in average accuracy was noted. The final average accuracy

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_43

Table 3. Performance of proposed method for sounds of different types.

	Analog training set		Analog and digital training data
Instrument	a) Digital sounds	b) Analog sounds	c) Digital sounds
Clarinet	65%	75%	75%
Saxophone	50%	50%	65%
Trumpet	40%	60%	60%
Piano	75%	75%	75%
Accordion	30%	80%	70%
Average	**52%**	**68%**	**69%**

(69%) is higher than the average accuracy obtained during the previous experiment, where only analog instruments were used.

This research proved the difference between traditional musical instruments and electronic sounds, which is visible in Fig. 2. Ultimately, it can be stated that the diversity of the dataset of samples of sounds from different sources allows achieving greater accuracy of the developed method.

4 Conclusions

In this paper, we developed a strategy and gave insights how to build a training dataset to identify musical instruments automatically. This strategy can be treated as the pathway to creating an extensive and robust dataset that could be used in a data-centric AI approach. Moreover, the simple classification method was proposed based on comparing the MFCC values. Experiments were carried out for the real instruments belonging to different groups: piano, clarinet, alto saxophone, trumpet, and accordion. We analyzed many aspects of the training data preparation, which were supported by quantitative results. The key conclusion drawn during the research is the impact of the dataset on the effectiveness of the classification method. It was proved that the size of the training dataset directly affects the accuracy of the classification. Significant improvement was also achieved by enriching training data with different sound characteristics (dynamics and articulation) samples. Moreover, due to the rapid growth of the popularity of digital technologies in music, the accuracy in recognizing the sounds of digital instruments was also examined.

The presented method of creating a training set for instrument classification differs from the available benchmark datasets, described in related work, mainly in the diversity of the recordings (different dynamics, articulation, and analog and digital origin of the sound). 450 recordings is a number that is clearly insufficient for modern machine learning methods, especially those using deep neural network architectures. We are aware that this collection needs to be significantly expanded before publication. However, we are confident that such a dataset can be applied to standard classifiers such as random forest, naive Bayes, support vector machine, etc.

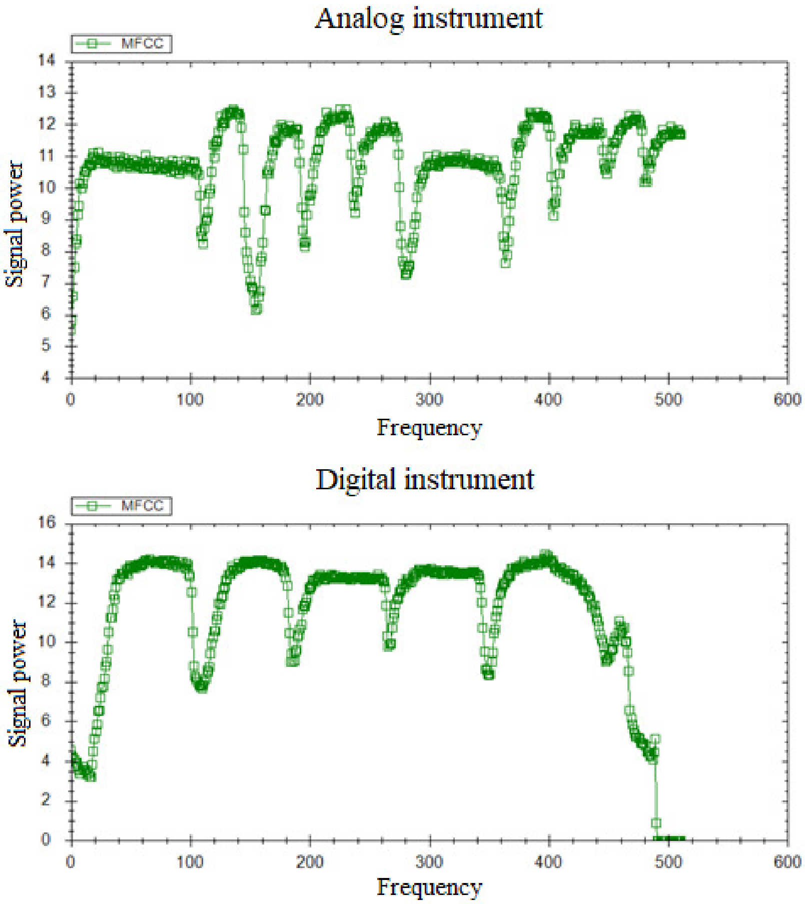

Fig. 2. The comparison of analogue and digital accordion MFCC vectors.

Besides working on a significantly expanding dataset that can be made public, our further research will be focused on unmixing signals from multiple instruments in a single recording. It will make it possible to recognize instruments from authentic musical pieces. However, this is a non-trivial and very complex task, so the number of research in this field is very limited. Moreover, it is necessary to thoroughly examine how sensitive the results will be to choosing different metrics than MFCCs.

Creating proper training dataset is very time-consuming. However, a good preparation of the dataset, consisting of appropriate differentiation of the recordings that will serve as learning data, will achieve very good musical instruments identification results.

References

1. Andén, J., Lostanlen, V., Mallat, S.: Joint time–frequency scattering. IEEE Transactions on Signal Processing **67**(14), 3704–3718 (2019)
2. Bandiera, G., Romani Picas, O., Tokuda, H., Hariya, W., Oishi, K., Serra, X.: Good-sounds.org: A framework to explore goodness in instrumental sounds. In: International Society for Music Information Retrieval Conference. pp. 414–419 (2016)
3. Bittner, R.M., Wilkins, J., Yip, H., Bello, J.P.: Medleydb 2.0: New data and a system for sustainable data collection. ISMIR Late Breaking and Demo Papers (2016)
4. Brown, J.C.: Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. The Journal of the Acoustical Society of America **105**(3), 1933–1941 (1999)
5. Brown, J.C., Houix, O., McAdams, S.: Feature dependence in the automatic identification of musical woodwind instruments. The Journal of the Acoustical Society of America **109**(3), 1064–1072 (2001)
6. Chakraborty, S.S., Parekh, R.: Improved musical instrument classification using cepstral coefficients and neural networks. In: Methodologies and Application Issues of Contemporary Computing Framework, pp. 123–138. Springer (2018)
7. Chandwadkar, D., Sutaone, M.: Role of features and classifiers on accuracy of identification of musical instruments. In: National Conference on Computational Intelligence and Signal Processing. pp. 66–70. IEEE (2012)
8. Emanuele, C., Ghisi, D., Lostanlen, V., Lévy, F., Fineberg, J., Maresz, Y.: TinySOL: an audio dataset of isolated musical notes. https://zenodo.org/record/3685367 (2020)
9. Engel, J., Resnick, C., Roberts, A., Dieleman, S., Eck, D., Simonyan, K., Norouzi, M.: Neural audio synthesis of musical notes with wavenet autoencoders (2017)
10. Eronen, A.: Musical instrument recognition using ica-based transform of features and discriminatively trained hmms. In: International Symposium on Signal Processing and Its Applications. vol. 2, pp. 133–136. IEEE (2003)
11. Gulhane, S.R., Suresh, D.S., Sanjay, S.B.: Identification of musical instruments using mfcc features. In: International Conference On Computational Vision and Bio Inspired Computing. pp. 957–968. Springer (2018)
12. Han, Y., Kim, J., Lee, K.: Deep convolutional neural networks for predominant instrument recognition in polyphonic music. IEEE/ACM Transactions on Audio, Speech, and Language Processing **25**(1), 208–221 (2016)
13. Humphrey, E., Durand, S., McFee, B.: Openmic-2018: An open data-set for multiple instrument recognition. In: ISMIR. pp. 438–444 (2018)
14. Kitahara, T., Goto, M., Okuno, H.G.: Pitch-dependent identification of musical instrument sounds. Applied Intelligence **23**(3), 267–275 (2005)
15. Lostanlen, V., Cella, C.E.: Deep convolutional networks on the pitch spiral for musical instrument recognition. International Society for Music Information Retrieval Conference pp. 1–7 (2016)
16. Loughran, R., Walker, J., O'Neill, M., O'Farrell, M.: The use of mel-frequency cepstral coefficients in musical instrument identification. In: International Computer Music Conference (2008)
17. McAdams, S.: Recognition of sound sources and events. Thinking in Sound: The Cognitive Psychology of Human Audition pp. 146—-198 (1993)
18. McAdams, S.: Musical timbre perception. The psychology of music pp. 35–67 (2013)

19. Romani Picas, O., Parra Rodriguez, H., Dabiri, D., Tokuda, H., Hariya, W., Oishi, K., Serra, X.: A real-time system for measuring sound goodness in instrumental sounds. In: Audio Engineering Society Convention 138 (2015)
20. Sahidullah, M., Saha, G.: Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. Speech communication **54**(4), 543–565 (2012)
21. Solanki, A., Pandey, S.: Music instrument recognition using deep convolutional neural networks. International Journal of Information Technology pp. 1–10 (2019)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_43