Predominant Instrument Recognition in Polyphonic Music Using Convolutional Recurrent Neural Networks

Lekshmi. C.R and Rajeev Rajan

College of Engineering Trivandrum
APJ Abdul Kalam Technological University
clekshmir04@gmail.com,rajeev@cet.ac.in

Abstract. In this paper, predominant instrument recognition in polyphonic music is addressed using convolutional recurrent neural networks (CRNN) through Mel-spectrogram, modgdgram, and its fusion. Modgdgram, a visual representation is obtained by stacking modified group delay functions of consecutive frames successively. Convolutional neural networks (CNN) learn the distinctive local characteristics from the visual representation and recurrent neural networks (RNN) integrate the extracted features over time and classify the instrument to the group where it belongs. The proposed system is systematically evaluated using the IRMAS dataset. A wave generative adversarial network (WaveGAN) architecture is also employed to generate audio files for data augmentation. We experimented with two CRNN architectures, convolutional long short-term memory (C-LSTM) and convolutional gated recurring unit (C-GRU). The fusion experiment C-GRU reports a micro and macro F1 score of 0.69 and 0.60, respectively. These metrics are 7.81% and 9.09% higher than those obtained by the state-of-the-art Han's model. The architectural choice of CRNN with score-level fusion on Mel-spectro/modgd-gram has merit in recognizing the predominant instrument in polyphonic music.

Keywords: predominant, Mel-spectrogram, modgdgram,convolutional gated recurring unit.

1 Introduction

Predominant instrument recognition refers to the problem where the prominent instrument is identified from a mixture of instruments being played together [16]. In polyphonic music, the interference of simultaneously occurring sounds makes instrument recognition harder. Automatic identification of lead instrument is important since the performance of the source separation can be improved significantly by knowing the type of the instrument [16].

Han *et al.* [16] employed Mel-spectrogram-CNN approach for instrument recognition. Pons *et al.* [22] analyzed the architecture of Han *et al.* in order to formulate an efficient design strategy to capture the relevant information about timbre. Detecting the activity of music instruments using a deep neural network (DNN) through a temporal max-pooling aggregation is addressed in [15]. Dongyan *et al.* [31] employed a network with an auxiliary classification scheme to learn the instrument categories

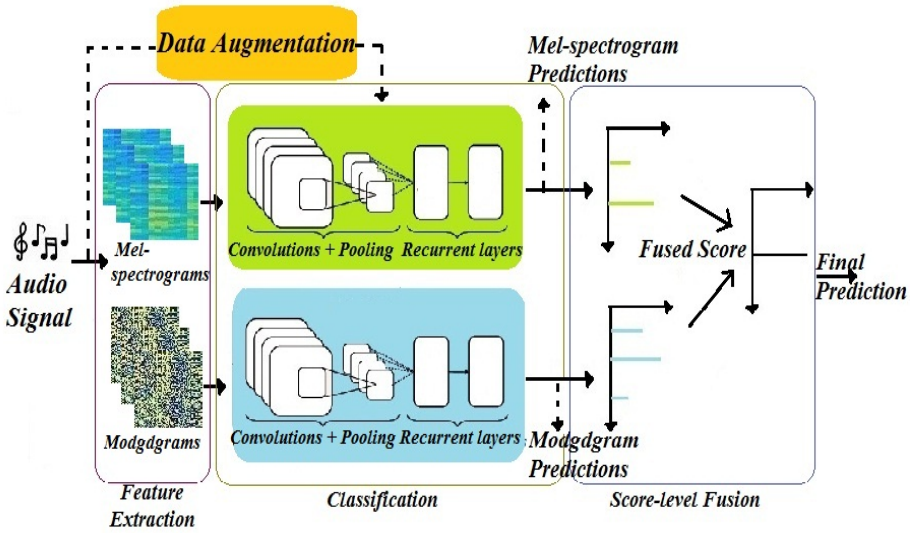

Fig. 1. Block diagram of proposed method of predominant instrument recognition.

through multitasking learning. Gomez *et al.* [14] investigated the role of two source separation algorithms as pre-processing steps to improve the performance in the context of predominant instrument detection tasks. In [18], the Hilbert-Huang transform (HHT) is employed to map one-dimensional audio data into two-dimensional matrix format, followed by CNN to learn the effective features for the task. In [17] an ensemble of VGG-like CNN classifiers is trained on non-augmented, pitch-synchronized, tempo-synchronized, and genre-similar excerpts of IRMAS for the proposed task.

The modified group delay feature (MODGDF) is proposed for pitched musical instrument recognition in an isolated environment in [9]. While the commonly applied mel frequency cepstral coefficients (MFCC) feature is capable of modeling the resonances introduced by the filter of the instrument body, it neglects the spectral characteristics of the vibrating source, which also, play its role in human perception of musical sounds and genre classification [12]. Incorporating phase information is an effective attempt to preserve this neglected component. Some preliminary works on predominant instrument recognition in polyphonic music using group delay functions are discussed in [2, 1]. In [28] a multi-head attention mechanism is employed along with modified group delay functions for proposed task.

In the proposed task, CRNN architecture with score level fusion of Mel-spectrogram and modgdgram is used for recognizing predominant instruments in polyphonic music. Similar approaches combining CNNs and RNNs have been presented recently in many music processing applications [6], [5], [20]. The idea of including modified group delay functions and GAN-based data augmentation strategy are the main contributions of the proposed scheme.

Section 2 explains the system description. Feature extraction is described in Section 3, followed by the model architectures in Section 4. The performance evaluation is

Table 1. Model summary of CNN and CRNN architectures.(* represents the multiplication factor, d_i,f_i,h_i,j_i represents the number of filters used in the networks. (d_i=8,16,24,32,64,128,256,512), (f_i=32,64,128,256), (h_i=8,16,32,64,128,256), (j_i=32,64,128).

*	**Mel-spectrogram-CNN**	**Modgdgram-CNN**	*	**Mel-spectrogram-CRNN**	**Modgdgram-CRNN**
x4	2 X Conv2D (3x3) , d_i	Conv2D (3x3), f_i	x3	2 X Conv2D (3x3) , h_i	Conv2D (3x3), j_i
	Leaky ReLU ($\alpha = 0.33$)	ReLU		Leaky ReLU ($\alpha = 0.33$)	ReLU
	3x3 Max-pooling,stride (3,3)			Batch Normalization	
	Dropout (0.25)			2x2 Max-pooling,stride(2,2)	
	Global Max-pooling			Flatten (1024)	
	Dense (1024)	Dense(512)		2 X Bidirectional LSTM / GRU (32 units)	
	Dropout (0.5)			Flatten (1024)	
	Dense (11), Softmax Activation			Dense (512)	
				Batch Normalization, Dropout (0.5)	
				Dense (11), Softmax Activation	

described in Section 5. The results are analyzed in Section 6. The paper is concluded in Section 7.

2 System Description

The proposed scheme is shown in Fig. 1. In the proposed model, CRNN is used to learn the distinctive characteristics from Mel-spectro/modgd-gram to identify the leading instrument in a polyphonic context. We evaluate the proposed method on the IRMAS dataset and compare its performance to CNN and two variants of RNN-long short-term memory (LSTM) and gated recurring unit (GRU). The performance is also compared with a DNN framework. As a part of data augmentation, additional training files are generated using WaveGAN. During the testing phase, the probability value at the output nodes of the trained model is treated as the score corresponding to the input test file. The input audio file is classified to the node which gives the maximum score during testing. In the fusion framework, the individual scores of Mel-spectro/modgd-gram experiments are fused at the score-level to make a decision. The fusion score S_f, is obtained by,

$$S_f = \beta S_{spectro} + (1 - \beta) S_{modgd} \tag{1}$$

where $S_{spectro}$, S_{modgd}, β are the Mel-spectrogram score, modgdgram score and weighting constant, respectively. The value of β has been empirically chosen to be 0.5. Each phase is explained in detail in the following sections.

3 Feature Extraction

Mel-spectrogram and modgdgram are the inputs used in the proposed scheme. Mel-spectrogram approximates how the human auditory system works and can be seen as the spectrogram smoothed, with high precision in the low frequencies and low precision in the high frequencies [21]. It is computed with a frame size of 50 ms and a hop size of 10 ms with 128 bins for the given task.

Group delay features are being employed in numerous speech and music processing applications [24, 26, 23, 25]. The group delay function is defined as the negative derivative of the unwrapped Fourier transform phase with respect to frequency. Modified group delay functions (MODGD), $\tau_m(e^{j\omega})$ are obtained by,

$$\tau_m(e^{j\omega}) = (\frac{\tau_c(e^{j\omega})}{|\tau_c(e^{j\omega})|})(|\tau_c(e^{j\omega})|)^a, \tag{2}$$

where,

$$\tau_c(e^{j\omega}) = \frac{X_R(e^{j\omega})Y_R(e^{j\omega}) + Y_I(e^{j\omega})X_I(e^{j\omega})}{|S(e^{j\omega})|^{2b}}. \tag{3}$$

The subscripts R and I denote the real and imaginary parts, respectively. $X(e^{j\omega})$, $Y(e^{j\omega})$ and $S(e^{j\omega})$ are the Fourier transforms of signal, $x[n]$, n.$x[n]$ ((weighted signal with index), and the cepstrally smoothed version of $X(e^{j\omega})$, respectively. a and b ($0 < a, b \leq 1$) are introduced to control the dynamic range of MODGD [19, 23]. Modgdgram is the visual representation of MODGD with time and frequency in the horizontal and vertical axis, respectively. The amplitude of group delay function at a particular time is represented by the intensity or color in the third dimension. Modgdgrams are computed with a frame size of 50 ms and hop size of 10 ms using a and b values of 0.9 and 0.5 respectively.

4 Model Architectures

CNNs and RNNs are specific instances of the CRNN architecture presented in this section: A CNN is a CRNN with zero recurrent layers, and an RNN is a CRNN with zero convolutional layers. CNN uses a deep architecture similar to [16] with repeated convolution layers followed by max-pooling. The detailed architecture for Mel-spectrogram and modgdgram CNN and CRNN are shown in Table 1.

RNNs are introduced to handle sequence and time-series data and are well suited for various speech and music-related applications [27], [13]. RNN with sophisticated recurrent hidden units like LSTM and GRU is used because such structures are capable of alleviating the vanishing gradient problem. The designed RNN consists of one input layer, two hidden layers which include two LSTM or GRU layers each with 32 nodes, and an output dense layer with eleven nodes for output classes. ReLU activation is used for hidden layers and softmax is used for the output layer.

In order to benefit from both approaches, the two architectures can be combined into a single network with convolutional layers followed by recurrent layers, often referred to as CRNN. The CRNN makes use of the CNN architecture for the task of feature extraction while using LSTM and GRU placed at the end of the architecture to summarise

the temporal information of the extracted features. The main drawback of CNNs is it lacks longer temporal context information. However, RNNs do not easily capture the invariance in the frequency domain, rendering high-level modeling of the data more difficult [5]. In the C-LSTM and C-GRU architectures, batch normalization is employed after convolutional layers to improve the training speed and performance. Two bidirectional LSTM/GRU units are connected after the time-distributed flatten layer. The bidirectional RNN is preferred rather than unidirectional RNN since it considers the future timestamp representations also [8]. The CNN and CRNN networks are trained using Adam optimizer with a learning rate of 0.001.

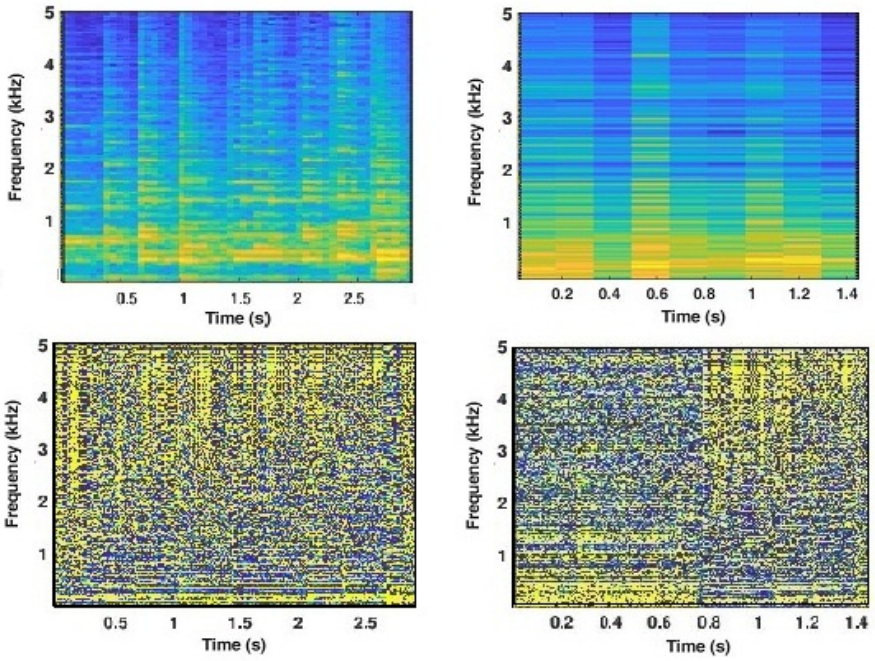

Fig. 2. Visual representation of an audio excerpt with acoustic guitar as leading, Mel-spectrogram of original and WaveGAN-generated (Upper pane left and right). Modgdgram of original and WaveGAN-generated (Lower pane left and right).

A DNN framework on musical texture features (MTF) is also experimented with to examine the performance of deep learning methodology on handcrafted features. MTF includes MFCC-13 dim, spectral centroid, spectral bandwidth, root mean square energy, spectral roll-off, and chroma STFT. The features are computed with a frame size of 40 ms and a hop size of 10 ms using Librosa framework [1]. DNN consists of seven layers, with increasing units from 8 to 512. ReLU has been chosen for hidden layers

[1] https://librosa.org/doc/latest/tutorial.html

and softmax for the output layer. The network is trained using categorical cross-entropy loss function for 500 epochs using Adam optimizer with a learning rate of 0.001.

5 Performance Evaluation

5.1 Dataset

IRMAS dataset [11], comprising eleven classes, is used for the evaluation. The classes include cello (Cel), clarinet (Cla), flute (Flu), acoustic guitar (Gac), electric guitar (Gel), organ (Org), piano (Pia), saxophone (Sax), trumpet (Tru), violin (Vio) and human singing voice (Voice). The training data are single-labeled and consists of 6705 audio files with excerpts of 3 s from more than 2000 distinct recordings. On the other hand, the testing data are multi-labeled and consist of 2874 audio files with lengths between 5 s and 20 s and contain the presence of multiple predominant instruments.

5.2 Data Augmentation using WaveGAN

WaveGAN v2 is used here to generate polyphonic files with the leading instrument required for training. WaveGAN is similar to DCGAN, which is used for Mel-spectrogram generation, in various music processing applications. The transposed convolution operation of DCGAN is modified to widen its receptive field in WaveGAN. For training, the WaveGAN optimizes WGAN-GP using Adam for both generator and discriminator. A constant learning rate of 0.0001 is used with $\beta_1 = 0.5$ and $\beta_2 = 0.9$ [10]. WaveGAN is trained for 2000 epochs on the three sec audio files of each class to generate similar audio files and a total of 6585 audio files with cello (625), clarinet (482), flute (433), acoustic guitar (594), electric guitar (732), organ (657), piano (698), saxophone (597), trumpet (521), violin (526) and voice (720) are generated. The generated files are denoted by $Train_g$ and training files available in the corpus are denoted by $Train_d$. Mel-spectrogram and modgdgram of natural and generated audio files for acoustic guitar are shown in Fig. 2. The experiment details and a few audio files can be accessed at https://sites.google.com/view/audiosamples-2020/home/instrument

The quality of generated files is evaluated using a perception test. It is conducted with ten listeners to assess the quality of generated files for 275 files covering all classes. Listeners are asked to grade the quality by choosing one among the five opinion grades varying from poor to excellent quality (scores, 1 to 5). A mean opinion score of 3.64 is obtained. This value is comparable to the mos score obtained in [10] and [3] using WaveGAN.

5.3 Experimental Set-up

The experiment is progressed in three phases namely Mel-spectrogram-based, modgdgram-based, and score-level fusion-based. Han's sliding window baseline model [16] is implemented for the given experiment with 1 s slice length for performance comparison [2]. We used the same aggregation strategy (S2) as that of Han's model, by summing

[2] https://github.com/Veleslavia/EUSIPCO2017

Table 2. F1 score for the experiments with data augmentation ($Train_d + Train_g$).

SL. No	Class	MTF DNN	Han's Model	Fusion CNN	Fusion LSTM	Fusion GRU	Fusion C-LSTM	Fusion C-GRU
		F1	F1	F1	F1	F1	F1	F1
1	Cel	0.15	0.55	0.55	0.15	0.36	0.42	0.50
2	Cla	0.26	0.18	0.36	0.13	0.36	0.48	0.39
3	Flu	0.27	0.43	0.55	0.32	0.62	0.34	0.31
4	Gac	0.43	0.72	0.63	0.44	0.54	0.51	0.70
5	Gel	0.36	0.69	0.67	0.50	0.49	0.62	0.74
6	Org	0.28	0.45	0.55	0.37	0.49	0.66	0.51
7	Pia	0.36	0.67	0.62	0.50	0.57	0.78	0.78
8	Sax	0.28	0.61	0.58	0.25	0.55	0.47	0.50
9	Tru	0.18	0.44	0.65	0.33	0.62	0.43	0.60
10	Vio	0.22	0.48	0.68	0.38	0.49	0.64	0.69
11	Voice	0.32	0.85	0.73	0.60	0.58	0.85	0.88
	Macro	**0.28**	**0.55**	**0.60**	**0.36**	**0.52**	**0.56**	**0.60**
	Micro	**0.32**	**0.64**	**0.65**	**0.43**	**0.55**	**0.65**	**0.69**

all the softmax predictions followed by normalization and applying a threshold of 0.5. Mel-spectrograms and modgdgrams of input size 128x100x1, corresponding to a window size of 1 s are applied to the corresponding network. The experiments are repeated for CNN, RNN with LSTM and GRU, CRNN with C-LSTM, and C-GRU respectively. Since the number of annotations for each class was not equal, we computed precision, recall, and F1 measures for both the micro and the macro averages. For the micro averages, we calculated the metrics globally, thus giving more weight to the instrument with a higher number of appearances. On the other hand, we calculated the metrics for each label and found their unweighted average for the macro averages.

6 Results and Analysis

Several studies [30, 29] have demonstrated that by consolidating information from multiple sources, better performance can be achieved than uni-modal systems which motivated us to perform the score-level fusion. The standard metrics for various algorithms on the IRMAS corpus are reported in Table 3. Fusion network C-GRU achieved micro and macro F1 measures of 0.69 and 0.60, respectively, which is 7.81% and 9.09% higher than those obtained for the state-of-the-art Han's model. Han employed Mel-spectrogram-CNN for the proposed task. Conventionally, the spectrum-related features used in instrument recognition take into account merely the magnitude information. However, there is often additional information concealed in the phase, which could be beneficial for recognition [9]. The experimental results validate the claim in [9]. Our Fusion-CNN with data augmentation reports a micro and macro F1 score of 0.65 and 0.60 respectively which is 1.56% and 5.26% higher than that obtained for our Mel-

Table 3. Performance comparison on IRMAS dataset

SL.No	**Model**	**F1 Micro**	**F1 Macro**
1	Bosch *et al.* [4]	0.50	0.43
2	Han *et al.* [16]	**0.65**	0.50
3	Pons *et al.* [22]	0.65	0.52
4	Kratimenos *et al.* [17]	0.65	**0.55**
5	**MTF-DNN** ($Train_d + Train_g$)	0.32	0.28
6	**Han Model** ($Train_d + Train_g$)	0.64	0.55
7	**Proposed Mel-spectrogram-CNN** ($Train_d + Train_g$)	0.64	**0.57**
8	**Proposed Modgdgram-CNN** ($Train_d + Train_g$)	0.54	0.53
9	**Proposed Fusion-CNN** ($Train_d + Train_g$)	**0.65**	**0.60**
10	**Proposed Fusion-C-LSTM** ($Train_d + Train_g$)	**0.65**	**0.56**
11	**Proposed Fusion-C-GRU** ($Train_d$)	0.62	0.53
12	**Proposed Mel-spectrogram-C-GRU** ($Train_d + Train_g$)	**0.66**	**0.59**
13	**Proposed Modgdgram-CGRU** ($Train_d + Train_g$)	0.55	0.53
14	**Proposed Fusion-C-GRU** ($Train_d + Train_g$)	**0.69**	**0.60**

spectrogram-CNN with data augmentation. It is evident that modgdgram added complementary information to the spectrogram approach and the importance of the fusion framework for the proposed task. Han's model and the proposed Mel-spectrogram-CNN approach show similar performance with better performance for the proposed architectural choice.

The F1 score of different fusion experiments is tabulated in Table 2. Fusion experiments using RNNs alone do not show improved performance over existing algorithms, however, GRU shows better performance than LSTM. Since we employed the same number of hidden units for both, GRU required less number of trainable parameters and makes faster progress, and reaches the convergence earlier than LSTM. Fusion experiments C-LSTM and CNN show similar performance, but C-GRU outperforms all the models. GRUs train faster and computationally more efficient than LSTM because of fewer trainable parameters. Results of the experiments described in [7] suggest that GRUs perform better than LSTMs on small polyphonic dataset [7]. Our C-LSTM for Mel-spectrogram requires 100224 more trainable parameters compared to C-GRU. It reaches convergence faster without compromising accuracy. The experimental results validate the claim in [7].

Our best model Fusion C-GRU, without data augmentation ($Train_d$) reports micro and macro F1 score of 0.62 and 0.53 respectively. Fusion C-GRU ($Train_d + Train_g$) reports micro and macro F1 scores of 0.69 and 0.60, respectively, with an improvement of 11.29% and 13.21% higher than that obtained by Fusion C-GRU ($Train_d$). This shows the significance of data augmentation in the proposed task.

Our proposed CRNN technique outperformed existing algorithms on the IRMAS dataset for both the micro and the macro F1 measures. The analysis of the experimental frameworks shows the significance of CRNN architecture for the proposed task. Be-

sides, the experiments show the potential of fusion of magnitude and phase information in the proposed task.

7 Conclusion

We presented a CRNN-based predominant instrument recognition system using Mel-spectro/modgd-gram. CRNN is used to capture the instrument-specific characteristics and then do further classification. The proposed method is evaluated on IRMAS dataset. Data augmentation is also performed using WaveGAN. The results show the potential of C-GRU architecture on the score-level fusion of Mel-spectrogram and modgdgram in the proposed task.

References

1. Ajayakumar, R., Rajan, R.: Predominant instrument recognition from polyphonic music using feature fusion. in Proc. of Emerging Trends in Engineering, Science and Technology for Society, Energy and Environment pp. 745–750 (2018)
2. Ajayakumar, R., Rajan, R.: Predominant instrument recognition in polyphonic music using gmm-dnn framework. in Proc.of International Conference on Signal Processing and Communications (SPCOM) pp. 1–5 (2020)
3. Atkar, G., Jayaraju, P.: Speech synthesis using generative adversarial network for improving readability of hindi words to recuperate from dyslexia. Neural Computing and Applications pp. 1–10 (2021)
4. Bosch, J.J., Janer, J., Fuhrmann, F., Herrera, P.: A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals. in Proc. of 13th International Society for Music Information Retrieval Conference (ISMIR) (2012)
5. Cakır, E., Parascandolo, G., Heittola, T., Huttunen, H., Virtanen, T.: Convolutional recurrent neural networks for polyphonic sound event detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing **25**(6), 1291–1303 (2017)
6. Choi, K., Fazekas, G., Sandler, M., Cho, K.: Convolutional recurrent neural networks for music classification. in Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 2392–2396 (2017)
7. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. NIPS 2014 Workshop on Deep Learning, December 2014 (2014)
8. Cui, Z., Ke, R., Pu, Z., Wang, Y.: Deep bidirectional and unidirectional lstm recurrent neural network for network-wide traffic speed prediction. arXiv preprint arXiv:1801.02143 (2018)
9. Diment, A., Rajan, P., Heittola, T., Virtanen, T.: Modified group delay feature for musical instrument recognition. in Proc. of 10th International Symposium on Computer Music Multidisciplinary Reserach, Marseille, France pp. 431–438 (May 2013)
10. Donahue, C., McAuley, J., Puckette, M.: Adversarial audio synthesis. in Proc. of International Conference on Learning Representations pp. 1–16 (2019)
11. Fuhrmann, F., Herrera, P.: Polyphonic instrument recognition for exploring semantic similarities in music. in Proc. of 13th International Conference on Digital Audio Effects DAFx10, Graz, Austria **14**(1), 1–8 (2010)
12. Fuhrmann, F., et al.: Automatic musical instrument recognition from polyphonic music audio signals. Ph.D. thesis, Universitat Pompeu Fabra (2012)

13. Gimeno, P., Viñals, I., Ortega, A., Miguel, A., Lleida, E.: Multiclass audio segmentation based on recurrent neural networks for broadcast domain data. EURASIP Journal on Audio, Speech, and Music Processing **2020**(1), 1–19 (2020)
14. Gómez, J.S., Abeßer, J., Cano, E.: Jazz solo instrument classification with convolutional neural networks, source separation, and transfer learning. in Proc. of International Society for Music Information Retrieval (ISMIR) pp. 577–584 (2018)
15. Gururani, S., Summers, C., Lerch, A.: Instrument activity detection in polyphonic music using deep neural networks. in Proc. of International Society for Music Information Retrieval Conference (ISMIR) pp. 577–584 (2018)
16. 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 (2017)
17. Kratimenos, A., Avramidis, K., Garoufis, C., Zlatintsi, A., Maragos, P.: Augmentation methods on monophonic audio for instrument classification in polyphonic music. in Proc. of 28th European Signal Processing Conference (EUSIPCO) pp. 156–160 (2021)
18. Li, X., Wang, K., Soraghan, J., Ren, J.: Fusion of hilbert-huang transform and deep convolutional neural network for predominant musical instruments recognition. in Proc. of 9th International conference on Artificial Intelligence in Music, Sound, Art and Design (2020)
19. Murthy, H.A., Yegnanarayana, B.: Group delay functions and its application to speech processing. Sadhana **36**(5), 745–782 (2011)
20. Nasrullah, Z., Zhao, Y.: Music artist classification with convolutional recurrent neural networks. in Proc.of International Joint Conference on Neural Networks (IJCNN) pp. 1–8 (2019)
21. O'shaughnessy, D.: Speech communication: human and machine. Universities press pp. 1–5 (1987)
22. Pons, J., Slizovskaia, O., Gong, R., Gómez, E., Serra, X.: Timbre analysis of music audio signals with convolutional neural networks. in Proc. of 25th European Signal Processing Conference (EUSIPCO) pp. 2744–2748 (2017)
23. Rajan, R., Murthy, H.A.: Two-pitch tracking in co-channel speech using modified group delay functions. Speech Communication **89**, 37–46 (2017)
24. Rajan, R., Murthy, H.A.: Group delay based melody monopitch extraction from music. in Proc.of IEEE International Conference on Acoustics, Speech and Signal Processing (ICAASP) pp. 186–190 (2013)
25. Rajan, R., Murthy, H.A.: Melodic pitch extraction from music signals using modified group delay functions. in Proc. National Conference on of the Communications (NCC) pp. 1–5 (February 2013)
26. Rajan, R., Murthy, H.A.: Music genre classification by fusion of modified group delay and melodic features. in Proc.of Twenty-third National Conference on Communications (NCC) pp. 1–6 (2017)
27. Rajesh, S., Nalini, N.: Musical instrument emotion recognition using deep recurrent neural network. Procedia Computer Science **167**, 16–25 (2020)
28. Reghunath, L.C., Rajan, R.: Attention-based predominant instruments recognition in polyphonic music. in Proc. of 18th Sound and Music Computing Conference (SMC) pp. 199–206 (2021)
29. Toh, K., Jiang, X., Yau, W.: Exploiting global and local decisions for multimodal biometrics verification. IEEE Transactions on Signal Processing pp. 3059–3072 (2004)
30. Wang, Y., Tan, T., Jain, A.: Combining face and iris biometrics for identity verification. in Proc. of Fourth International Conference on AVBPA, Guildford, U.K pp. 805–813 (2003)
31. Yu, D., Duan, H., Fang, J., Zeng, B.: Predominant instrument recognition based on deep neural network with auxiliary classification. IEEE/ACM Transactions on Audio, Speech, and Language Processing **28**, 852–861 (2020)