Article in Press IJIMAI journal

Motivic Pattern Classification of Music Audio Signals Combining Residual and LSTM Networks

Aitor Arronte Alvarez[1]*, Francisco Gómez[2]

[1] University of Hawaii at Manoa, Honolulu (USA)
[2] Technical University of Madrid, Madrid (Spain)

Received 13 August 2020 | Accepted 14 January 2021 | Published 21 January 2021

Abstract

Motivic pattern classification from music audio recordings is a challenging task. More so in the case of a cappella flamenco *cantes*, characterized by complex melodic variations, pitch instability, timbre changes, extreme vibrato oscillations, microtonal ornamentations, and noisy conditions of the recordings. Convolutional Neural Networks (CNN) have proven to be very effective algorithms in image classification. Recent work in large-scale audio classification has shown that CNN architectures, originally developed for image problems, can be applied successfully to audio event recognition and classification with little or no modifications to the networks. In this paper, CNN architectures are tested in a more nuanced problem: flamenco *cantes* intra-style classification using small motivic patterns. A new architecture is proposed that uses the advantages of residual CNN as feature extractors, and a bidirectional LSTM layer to exploit the sequential nature of musical audio data. We present a full end-to-end pipeline for audio music classification that includes a sequential pattern mining technique and a contour simplification method to extract relevant motifs from audio recordings. Mel-spectrograms of the extracted motifs are then used as the input for the different architectures tested. We investigate the usefulness of motivic patterns for the automatic classification of music recordings and the effect of the length of the audio and corpus size on the overall classification accuracy. Results show a relative accuracy improvement of up to 20.4% when CNN architectures are trained using acoustic representations from motivic patterns.

Keywords

Motivic Patterns, Convolutional Neural Networks, Data Augmentation, Audio Signal Processing, Music Information Retrieval.

DOI: 10.9781/ijimai.2021.01.003

I. Introduction

The automatic extraction, discovery, and classification of motivic patterns from music audio recordings is a task that has gathered the attention of the Artificial Intelligence community in general, and the Music Information Retrieval (MIR) community in particular [1], [2], [3]. Repeated melodic patterns are important in the analysis and understanding of music. More recently, research has shown that repeated small musical patterns that are transformed up to a certain extent, play an important role in establishing music similarity in orally transmitted songs [4].

The computational study of orally transmitted vocal music repertoires present different types of problems associated with the audio signal obtained from such recordings. The high degree of variability in the audio signal has to do with the environmental conditions of the recordings, the improvisatory nature of the singing styles, and the rapid fluctuation of wide vibrato ranges. In flamenco music, these difficulties are even more acute, since intervals are often smaller than the half-tone. A cappella flamenco *cantes* exhibit characteristic melodic features such as conjunct degrees in the melodic movement, high degree of ornamentation, extreme pitch oscillations, microtonal variation, and constant timbre changes. These features make the automatic extraction of motivic patterns from audio recordings an especially challenging task.

The computational study of flamenco music has concentrated on the melodic characterization of *cantes* [5], [6], melodic pattern extraction [2], and the modelling of melodic variation [7]. Pattern extraction methods in *flamenco* research have used humans to extract relevant segments and melodic motifs [2], [5]. To our knowledge, exclusively data-driven approaches for the automatic intra-style classification of music audio signals have not yet been developed in previous research.

Different approaches in the MIR research literature have considered the use of Convolutional Neural Networks (CNN) for music tagging, genre prediction, and music classification. CNN have been used for mood and genre prediction using mel-spectrograms as the input representation [8]; the classes used in this study include genres (classical and pop), and moods (soft, ambient) among other label descriptors. Image classification CNN architectures were used for music classification based on general music style tags [9]. Other transfer learning approaches on MIR tasks include multi-label classification and prediction [10], and general-purpose music classification [11]. In the audio signal processing research in general, CNN architectures were used on large-scale audio event classification [12], showing that image architectures can be reused for audio processing task with some adjustments in the architectures' filter size.

Other applications of deep neural networks to music analysis and its computational understanding include low-level tasks such

* Corresponding author.

E-mail address: arronte@hawaii.edu

Please cite this article in press as:
A. Arronte Alvarez, F. Gómez. Motivic Pattern Classification of Music Audio Signals Combining Residual and LSTM Networks, International Journal of Interactive Multimedia and Artificial Intelligence, (2021), http://dx.doi.org/10.9781/ijimai.2021.01.003

as beat tracking [13], onset detection [14], tempo estimation [15], and chord recognition [16]. These low-level tasks attempt to learn representations of acoustic phenomena directly from the audio signal. Higher-level tasks learn representations that can map acoustic features into more abstract musical concepts such as music style classification [17], and singer identification [18] amongst others. MIR applications of high-level music tasks strongly depend on pre-existing knowledge and domain adaptation. In the approach presented in this article, no hand-crafting or domain adaptation is needed, since motivic patterns are extracted directly from the audio signal without prior knowledge.

This paper investigates the usefulness of motivic patterns for the automatic classification of different styles of flamenco music by using different CNN architectures originally conceived for image classification tasks. This research also extends the computational study of motivic patterns in flamenco music by presenting a pipeline for motivic contour extraction from audio recordings based on an approximation scheme. Then classification task is performed from the patterns obtained using the log mel-spectrograms extracted from the recordings' raw audio signals by using different CNN. The contributions of this research are the following: 1) We propose a motivic extraction pipeline as a preprocess step, which improves the classification accuracy of all the architectures tested. 2) It is shown that CNN architectures from very different domains can achieve competitive results with state-of-the-art algorithms while simplifying the learning process and making it computationally more efficient, mostly because of the pipeline introduced in this article. 3) A neural architecture is presented that is able to use some of the advantages of image classification CNN models, particularly as audio feature extractors, while at the same time adding recurrent layers with bidirectional LSTMs that are able to process musically relevant sequential data, adding more explanatory power to the results. 4) We make code and data of the experiments publicly available1.

The different sections of this article are organized as follows: Section II presents the corpus of flamenco recordings (COFLA) and describes its contents and music characteristics. Section III sets forth the motivic pattern extraction method and audio features used as the input of the different architectures. Section IV describes the CNN models used as baselines in this research and the hybrid recurrent model introduced in this article. Section V presents the experiments and data used to test the different CNN architectures. Section VI outlines the results of the experiments and discusses the main findings, improvements, and shortcomings. Section VII concludes by listing the main contributions of this research and possible future lines of work.

II. Corpus of Flamenco Recordings

Flamenco is an orally transmitted musical tradition from Andalusia, a region in the south of Spain. Its rich history and musical characteristics are derived from the region's cultural exchanges amongst various populations over centuries, most notably Andalusian-Romani, Jews, and Arabs. Some of the key characteristics of flamenco music such as pitch instability, the use of intervals smaller than the half-tone, the amount of variation from phrase to phrase and from singer to singer, are derived from its improvisatory nature. Even though improvisation plays a very important role in the conception of flamenco music, it is a highly structured and elaborated musical tradition [19].

Flamenco music centers around the singing voice usually accompanied by guitar, hand-clapping, and other percussion instruments like the *cajón.* Melodies are characterized by a combination of short and long notes with syllabic ornamentations (melismas), that are placed in specific locations in a phrase [20]. Flamenco singers learn melodies belonging to different styles and acquire singing techniques by oral transmission.

The main focus in the computational study of flamenco music is the development of algorithms that target the analysis of the singing voice [19]. Flamenco music, like most orally transmitted musical cultures, lacks music transcriptions of the repertoire. For that reason, corpora of audio recordings, with their corresponding meta-data, are the main source of research data. In this article corpus COFLA is used [20]. The corpus consists of more than 1,800 music recordings taken from flamenco anthologies. This corpus follows the research corpora principles formulated by Serra [21]. The main characteristics of the corpus, as summarized by its authors [20], are:

- Exhaustiveness: the corpus is composed of all anthologies published on CD during the 20th century, and are considered references for music critics and musicologists.
- Representation: each anthology represents a wide variety of styles and their variants.
- Sound quality: the audio quality varies greatly amongst recordings, but all recordings comply with a minimum standard.
- Commercial availability: all recordings are available to the general public, which facilitates the acquisition and allows for the establishment of ground truth data.

In this research, the following styles and substyles are used from the corpus COFLA: *tonás* (*deblas, martinetes, and saetas*), and *fandangos.* Stylistically, the *tonás* is an important group of a capella *cantes* sung in free rhythm, where singers choose their own reference pitch and perform variations on a given melody. A *toná* normally is composed of four verses of eight syllables each. Tempo is not strictly kept during a single piece and ornamentation is heavily used by singers. In the *tonás* style, *deblas* are characterized by melismatic ornamentations with more abrupt changes than the rest of the compositions in the *tonás* style. *Martinetes,* also a *toná* variant, differ slightly in its melodic model from the *debla* and, even though it is mostly sung without accompaniment, it uses a hammer and anvil as percussion instruments. *Saetas,* another *toná* variant, have a religious content in its lyrics and is stylistically closer to the *debla* in its usage of long and sustained notes combined with melismatic ornamentations. The style of *fandango* is more differentiated from the variants in the *tonás.* A *fandango* is a musical style associated with a dance and is rhythmically more complex than the *tonás.*

We select a sample from the corpus COFLA consisting of 13 *deblas,* 12 *saetas,* and 50 *martinetes.* The *martinete* subsample contains a wider variety of singing styles, and to some researchers it can be decomposed into 2 subtypes [22]. The current sample presents different stylistic challenges and difficulties for the automatic classification of motifs based on their substyle. First of all, 3 of the classes belong to the same genre (*deblas, martinetes, and saetas*), which means that these substyles share more musical traits with each other than with the *fandango.* This will add another level of complexity to the computational analysis, considering that previous studies have dealt only with the classification of different genres of music. In this paper the analysis is restricted to a very specific genre of music, namely flamenco, but also it is restricted to unaccompanied vocal music of different subgenres of flamenco.

III. Audio Features and Contour Extraction Method

From the sample of songs described in Section II, we extract musical motifs following a pipeline based on two main components: a contour simplification method and the BIDE pattern mining algorithm [18]. The purpose of this pipeline is to extract statistically and musically relevant motifs from flamenco audio recordings characterized by high

[1] https://rb.gy/q3ppg0

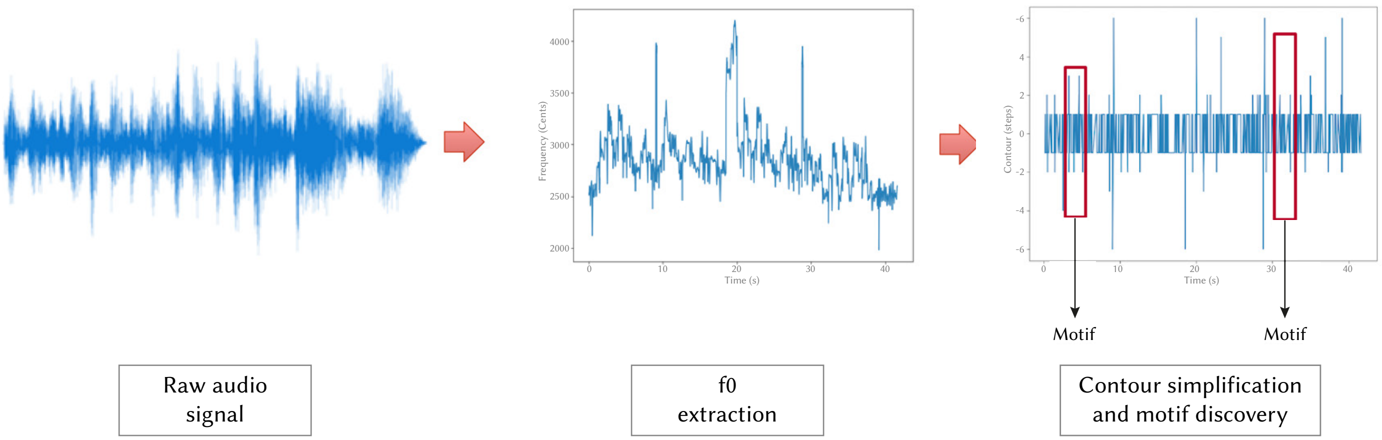

Fig. 1. Motivic pipeline for the extraction of patterns from raw audio signals.

instability of pitch. The pipeline here described attempts to solve the problem of reducing the pitch variability in the audio signal, with an approximation method that uses ranges of pitch distances instead of a fully tempered system such as the one used in western classical music. The steps of this motivic pipeline, as shown in Fig. 1, are the following:

1. Extract the fundamental frequency f_0 from the audio signals of the songs
2. Apply a melodic contour simplification function $C(f_0)$ based on the extracted f_0 for each one of the songs, thus obtaining an approximation of f_0
3. Apply the BIDE algorithm on the melodic contours obtaining a dictionary of motifs for the entire collection
4. Generate log mel-spectrograms for each motif in the dictionary

The fundamental frequency is extracted from raw audio signals by using a sinusoid extraction and salience function [24]. A sampling rate of 44.1 KHz and a window step of 256 samples are used. Then, a melodic contour simplification procedure is performed to extract meaningful motivic representations of a cappella flamenco *cantes* from the fundamental frequency. In previous studies [6], contour simplification procedures have been used to obtain consistent representations of flamenco melodic segments by converting complex pitch fluctuations to equal-step segments. Since we are studying motivic patterns in complex flamenco vocal pieces, we are interested in exploring the unequal microtonal nature of this type of music. In order to accomplish this goal, our contour simplification process takes into account ranges of cent-based distances instead of set of pitches as presented in previous work [6].

We follow these steps to find a curve approximation to f_0 given a step length of ε=66 cents based on previous approximation approaches [20]:

- Given a set of points *P* in f_0 we say that a line segment *L* is bounded by all points in *P* given a maximum accepted step size of ε.
- The output of this procedure is a contour simplification function of *f0*.

Once this output is computed, a contour *C* is obtained based on the following distance specification in cents:

- If the distance *d* between two points <=66 then, *d*=1
- If the distance *d* between two points >66 or *d* <=132 then, *d*=2
- If the distance *d* between two points >132 <=198 then, *d*=3
- 4 otherwise

The result is a vector of contour points represented in the time domain. The signs + and – are used to specify whether the direction of the contour ascends (+) or descends (-). Sudden jumps in frequency are eliminated due to external noise conditions.

Once the approximation function is created, the BIDE algorithm is used to discover motifs in the contour sequences. Motifs that are repeated at least 3 times in a single song are kept. From the dictionary of motifs, log mel-spectrograms are computed from the 2D time-frequency motivic patches, with hop and window sizes of 25 ms. The input size for all samples is 128x426, zero-padding smaller audio files.

IV. Baseline and Hybrid Architectures

Transfer learning approaches in deep neural networks have shown to be not only computationally more efficient in achieving competitive results, but also show how representations from one task can be transferred to another task. The different CNN architectures developed initially for image classification problems and used in this article's experimental study are, *DenseNet-161,* and *ResNet-50.* A state-of-the-art Convolutional Recurrent Neural Network (CRNN) architecture developed specifically for music classification is also used as a baseline [26]. Filter sizes and strides are kept small, 3x3 and 1x1 respectively. This is mostly because of the small size of the audio input.

A. ResNet-50

Deep residual networks were conceived to address the problem of learning degradation in deep nets. Residual networks are based on the idea of stacking layers and an underlying mapping that is optimized [27]. The model used in this study, *Resnet-50,* is transformed in a similar way as in [12] by removing the stride of 2x2 in the first convolutional layer, and reducing the size of the first convolutional filter from 7x7 to 3x3. In addition to that, and in order to maintain the input tensor size of the mel-spectrogram and to leave the *ResNet-50* architecture intact for baseline purposes, we add an initial convolutional layer with filters of size 3x3 and stride of 1.

B. DenseNet-161

CNN that have shorter connections between layers that are closer to the input and output of a network have shown to be more accurate. This paradigm is followed by the *DenseNet* model [28]. We make the same modifications to the architecture as in *ResNet-50.*

C. CRNN

A model for music audio tagging that has shown state-of-the-art results is the CRNN of Choi et al. [8]. This model utilizes the benefits of CNN as feature extractors and the sequential characteristics of Recurrent Neural Networks (RNN) to summarize time-dependent data as the one obtained from musical pieces.

D. Hybrid Recurrent Architecture

In this work we attempt to use and exploit the advantages of CNN layers as feature extractors and add recurrent components in the last layers to capture sequential characteristics present in music audio data.

Deep learning architectures for audio classification are normally divided into front-end and back-end components [30]. The front-end, is the part of the model that tries to learn a representation based on the input signal. The back-end is in charge of predicting a given output based on the representation obtained in the front-end. Our hybrid model uses shallow residual blocks present in *Resnets* as a front-end, and a recurrent neural model as a back-end. The overall goal of this architecture is to simplify the already high cost of deep learning methods, especially in the front-end, while trying to improve state-of-the-art results by adding domain specific knowledge in the back-end. We try therefore to reduce the number of parameters of very deep networks by adding recurrent layers.

The shallow *residual* network proposed is composed of only 2 residual blocks, which reduces the computational cost and overall training time when compared to denser *Resnet* models normally utilized in the computer vision literature. Small filters of 3x3 with a stride of 1 are used in all convolutional layers to capture local feature-maps, and finer low-level spectral features. As the back-end two-stacked Bidirectional Long Short-Term Memory (BLSTM) layers to capture longer, time-dependent, features [31]. Fig. 2 presents a high-level overview of the architecture described.

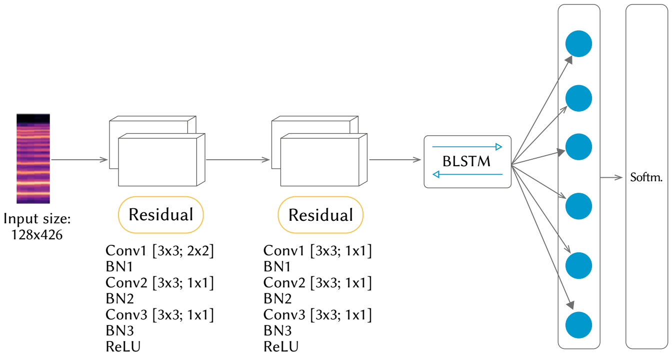

Fig. 2. Neural network architecture overview. Residual blocks contain convolutional layer dimensions (filter size, and stride), and batch size normalization (BN) and ReLU components.

In our back-end, the BLSTM is used to process in both directions (forward and backwards) the embedding obtained from the residual layers. The output of this layer will be a high level, vector representation of the time-dependent features of the motifs. The sequential operation done by the BLTSM can be represented as an input sequence $x = \{x_1, \dots, x_T\}$ that produces an output sequence $y = \{y_1, \dots, y_T\}$ where the input x is a vector of acoustic features at the frame level. A BLTSM is composed of a forward and backward LSTM, where the forward LSTM $\vec{f}$ reads the input sequence as it is ordered, and estimates the forward hidden states $\vec{h}_1, \dots, \vec{h}_T$ from $t = 1$ to T. The backward LSTM $\overleftarrow{f}$ computes the sequence in reverse order obtaining the backward hidden states iterating back from $t = T$ to 1:

$$\vec{h}_t = H(W_{x\vec{h}} x_t + W_{\vec{h}\vec{h}} \vec{h}_{t-1} + b_{\vec{h}}) \tag{1}$$

$$\overleftarrow{h}_t = H(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}) \tag{2}$$

$$y_t = W_{\vec{h}y} \vec{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h}_t + b_y \tag{3}$$

where H is the hidden layer function, W the matrix of weights, and b the bias vector.

The final layer of the architecture presented is a fully connected neural network (FCNN) layer with a softmax function to classify the sequences according to the style label. Fig. 3 shows a high-level motivic pipeline overview.

V. Experimental Methodology

We compare all models in the sub-style classification of musical patterns extracted from corpus COFLA, as described in Section III, and use 2D log mel-spectrograms as the input of the networks. The dataset used in this study is composed of 111,076 audio motifs extracted from the 4 sub-collections. We noted in the initial stages of the study that extremely short motifs (<0.5 seconds) do not help in the classification accuracy; for that reason only motifs that are >= 0.5 seconds in duration are kept. This resulted in a corpus of only 10,640 motifs, of which 1,573 were obtained from the *debla* sub-style, 129 from the *fandango,* 5,027 from the *martinete*, and 3, 915 from the *saeta.* We can see how certain sub-styles are richer in motivic patterns than others, and note that the *fandango* sub-collection in particular, is much less varied in longer motivic patterns (>= 0.5 seconds). This unbalanced dataset allows us to test data augmentation techniques in the context of audio musical data.

Unlike previous approaches to music classification and tagging in MIR, the approach presented will only learn a small segment of the entire audio signal. This segmentation based on the extraction of relevant motivic data will greatly benefit the representation learned, and reduce the total training time. From an information-theoretic stand point, it can be argued that reducing the amount of irrelevant information to the task will act as an implicit optimizer for the neural architectures, while at the same time obtaining more explainable results in music terms.

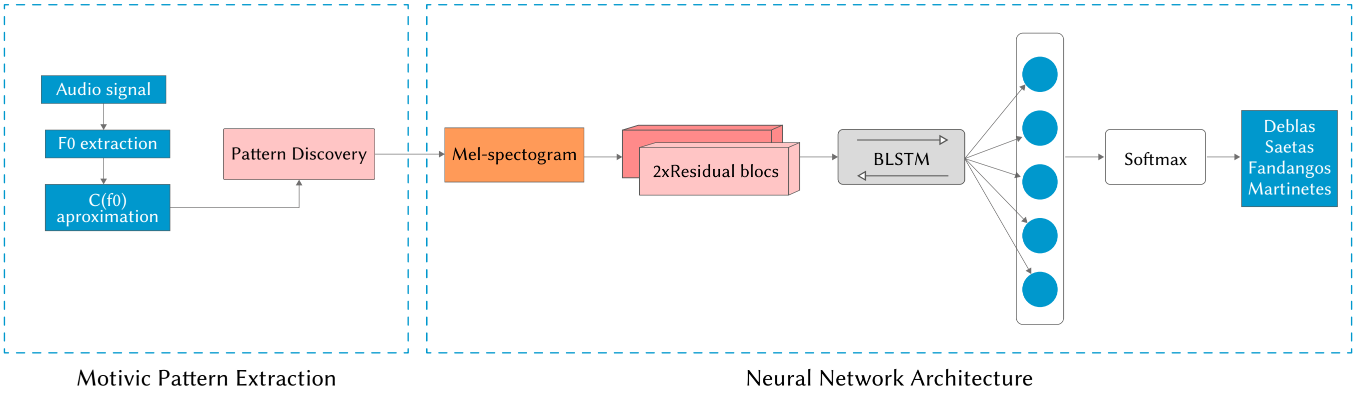

Fig. 3. High-level motivic pipeline overview.

A. Data Augmentation

In this experiment a recent method for data augmentation developed for Automatic Speech Recognition (ASR) called SpecAugment is used [32]. Instead of producing deformations to the raw audio signal like other audio-based data augmentation techniques [33-34], SpecAugment operates directly on the spectrogram by warping it in the time direction, masking frequency channels, and masking blocks of utterances. The method follows a similar rationale as image data augmentation techniques.

We concentrate on the two augmentation policies that seem to be the most effective in ASR tasks [32]: frequency masking, and time masking. Frequency masking works on m consecutive mel frequency channels [m_0, m_0+m], where m is chosen from a uniform distribution from 0 to the frequency mask parameter M. Time masking works in a similar way by applying the masking to t consecutive time steps. We compare the two data augmentation policies with the original unbalanced dataset, and apply the following number of transformations by class:

- For *fandango* style a total of 8 augmentations per spectrogram is performed; 4 of time masking and 4 of frequency masking, resulting in a total subset of 1,032.
- For the rest of the styles we apply one of each augmentations in only 50% of their respective subsets. Resulting in 3,146 *deblas*, 10,054 *martinetes*, and 7,830 *saetas*.

The total dataset after augmentation contains 22,062 spectrograms of motifs. The comparative differences in motivic samples by sub-style are presented in Fig. 4.

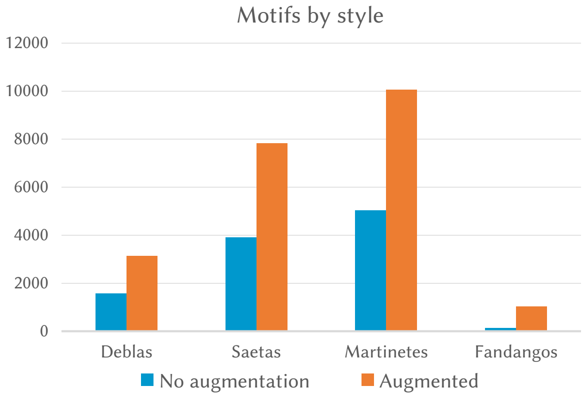

Fig. 4. Motivic patterns by sub-style.

B. Training

All the models use the Adam optimizer [35] and data augmentation dynamically during training. We divide the data in training, and validation subsets, making the 60% and 20% respectively of the entire dataset, leaving the remaining 20% for testing. AUC-ROC, and accuracy scores are used to perform searches over the parameter space. It was found out during training that batch sizes of 20, and a total number of epochs of ~40 performed best in terms of accuracy and computational time. We found no sign of overfitting based on those measures in the validation subset, even in non-augmented sets.

Random initialization versus pretrained weights were also tested during training for all architectures. Results showed that pretrained weights not only perform better overall (>~2% accuracy) than random ones, but also decreased the training time (~10 epochs less to converge). We used pretrained weights from image classification tasks in our experiments.

VI. Results and Discussion

Results in *Table I* show how the motivic pattern dataset has overall better results, with an average accuracy improvement of 13.1% across all models, with a maximum of 14.1% for *Resnet-50*, which indicates a relative improvement of 20.4%. Precision and recall measures also highlight the strength of motivic patterns for all models when compared to non-motivic data, and achieve an 85% precision for the proposed architecture with motivic patterns and no augmentation. These results shed light in the importance of motivic patterns in deep learning for music classification problems. This result can have significant implications in deep learning for MIR tasks, since shorter, more targeted audio data can significantly reduce the already huge computational costs of deep architectures. On the other hand, for multimedia systems in general, and MIR systems in particular, the effective retrieval of relevant audio information from big data can be improved with traditional sequential pattern mining techniques as a pre-step in the computational pipeline.

From a theoretical MIR point of view, our results highlight the importance of musically relevant features in deep learning systems as opposed to merely general audio features. In musically complex systems with melodic variability, microtonal ornamentations and contours, the extraction of relevant patterns can become a challenging task. The proposed contour simplification method takes into account small pitch fluctuations, and extracts small patterns (~0.5 seconds) that highlight particularities of a sub-style within flamenco music. These patterns may reveal vibrato styles, or ornamentation tendencies in singers for a particular style that may be difficult for the human ear to grasp. Further study should concentrate on the exploration of speech features combined with purely musical ones, which may aid the classification and automatic identification not only of styles, but singers as well.

The transferred architectures used in this study show how pretrained image weights can optimize the overall training procedure in music classification tasks and achieve competitive results with less training time. Since we are using mel-spectrograms of an audio signal as the input, the image-like 2-dimensional size of the input seems to be the reason why pretrained weights facilitate the accuracy results in less time when compared with random initialization. The hybrid architecture proposed outperforms the rest in terms of accuracy, AUC, precision, and recall. The performance values for non-motivic datasets with recurrent layers in the architecture indicates that these architectures can indirectly infer the temporal components of the data. Still the motivic dataset outperforms non-motivic ones for all models.

TABLE I. Model Results for the Motivic Patterns and Non-motivic Subsets

Model	*Dataset*	*Accuracy*	*AUC*	*Prec.*	*Rec.*	*F1*
Resnet-50	Motivic	0.832	0.894	0.801	0.769	0.785
Resnet-50	Non-motivic	0.691	0.792	0.631	0.617	0.624
Densenet-161	Motivic	0.817	0.881	0.735	0.731	0.733
Densenet-161	Non-motivic	0.683	0.769	0.61	0.589	0.599
CRNN	Motivic	0.821	0.886	0.78	0.757	0.768
CRNN	Non-motivic	0.796	0.853	0.714	0.711	0.712
ResLSTM	Motivic	**0.911**	**0.91**	**0.848**	**0.813**	**0.83**
ResLSTM	Non-motivic	0.824	0.882	0.816	0.79	0.803

The results in Table II show the data augmentation classification scores for the motivic pattern dataset. An accuracy improvement of 2.4% on the best model when using augmentation, highlights the importance of the data size in deep learning tasks. Since we obtain more than double of the original size from the motivic dataset, the improvements on the classification results seem to be logical.

SpecAugment, however, does not show an improvement as important as the one shown in the original study with speech data [32]. Further research should explore different ranges of masking parameters to determine the quality of the results and its appropriate use with musical vocal data.

TABLE II. Results for the Data Augmentation Policies Applied to the Motivic Pattern Dataset

Model	*Augment.*	*Accuracy*	*AUC*	*Prec.*	*Rec.*	*F1*
Resnet-50	Frequency	0.868	0.898	0.811	0.791	0.8
Resnet-50	Time	0.846	0.878	0.767	0.763	0.76
Densenet-161	Frequency	0.852	0.861	0.81	0.804	0.81
Densenet-161	Time	0.831	0.858	0.783	0.766	0.78
CRNN	Frequency	0.87	0.887	0.83	0.796	0.813
CRNN	Time	0.842	0.879	0.782	0.772	0.78
ResLSTM	Frequency	0.935	0.922	0.85	0.839	0.844
ResLSTM	Time	0.921	0.91	0.828	0.821	0.824

VII. Conclusion

Overall the results indicate that the effect of motivic patterns in the classification accuracy of state-of-the-art CNN models is greater than the effect of data augmentation when using SpecAugment. Motivic patterns seem to provide important information in the classification of audio samples by style. Since CNN capture local-level features of a given audio sample, the utilization of motivic patterns seems to highlight higher level melodic features. Recurrent models on the other hand are less sensitive to non-motivic data. We also evaluated the importance of transfer learning in the context of musical audio data. The results of the transferred models are consistent with a recent large-scale audio classification study [12], which also extends the findings to music audio data. We specifically noted the ability of the networks to converge up to a state-of-the-art competitive accuracy with less training when using pretrained weights from image classification tasks. The proposed neural architecture outperforms state-of-the-art CRNN for music classification by taking advantage of the long-term sequence processing that the BLSTM net does. By combining BLSTM with shallow residual blocks, we take advantage of the smaller number of parameters required, and less processing time, when compared with deeper *resnets.*

This study presents an important case of deep learning optimization for audio signal processing, by extracting smaller, more targeted audio samples, discarding irrelevant information from the signal and learning more robust representations. This approach can be particularly interesting for low-resource MIR applications. It can also be easily adapted to sound event recognition and identification, and to speech recognition tasks that have a strong acoustic component such as accent, emotion, and dialect identification.

References

[1] Dannenberg, R. B., and Hu, N. "Pattern discovery techniques for music audio," *Journal of New Music Research*, vol. 32, no.2, pp. 153-163, 2003.

[2] Pikrakis, A., Gómez, F., Oramas, S., Díaz-Báñez, J. M., Mora, J., Escobar-Borrego, F., Gómez, E., and Salamon, J. "Tracking Melodic Patterns in Flamenco Singing by Analyzing Polyphonic Music Recordings," in *International Society for Music Information Retrieval Conference, ISMIR*, Porto, Portugal, 2012, pp. 421-426.

[3] Gulati, S., Serra, J., Ishwar, V., and Serra, X. "Mining melodic patterns in large audio collections of Indian art music," in *2014 Tenth International Conference on Signal-Image Technology and Internet-Based Systems*, Marrakech, Morocco, 2014, pp. 264-271.

[4] Volk, A., Haas, W. B., and Kranenburg, P. "Towards modelling variation in music as foundation for similarity," in *Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music*, Thessaloniki, Greece, 2012, pp. 1085-1094.

[5] Mora, J., Gomez Martin, F., Gómez, E., Escobar-Borrego, F. J., and Díaz-Báñez, J. M. "Characterization and melodic similarity of a cappella flamenco cantes," in *International Society for Music Information Retrieval Conference, ISMIR,* Utrecht, The Netherlands, 2016, pp. 9-13.

[6] Kroher, N., and Díaz-Báñez, J. M. "Audio-based melody categorization: Exploring signal representations and evaluation strategies" *Computer Music Journal*, vol. 41, no. 4, pp. 64-82, 2018.

[7] Kroher, N., and Díaz-Báñez, J. M. "Modelling melodic variation and extracting melodic templates from flamenco singing performances," *Journal of Mathematics and Music*, vol. 13, no. 2, pp. 150-170, 2019.

[8] Choi, K., Fazekas, G., and Sandler, M. "Automatic tagging using deep convolutional neural networks," arXiv preprint arXiv:1606.00298.

[9] Kim, T., Lee, J., and Nam, J. "Sample-level CNN architectures for music auto-tagging using raw waveforms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, 2018, pp. 366-370.

[10] Dieleman, S., and Schrauwen, B. "End-to-end learning for music audio," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Florence, Italy, 2014, pp. 6964-6968.

[11] Choi, K., Fazekas, G., and Sandler, M. "Transfer learning for music classification and regression tasks." arXiv preprint arXiv:1703.09179 2017.

[12] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., and Slaney, M. "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* New Orleans, LA, USA, 2017, pp. 131-135.

[13] Durand, S., Bello J. P., Bertrand D., and Gaël R. "Downbeat tracking with multiple features and deep neural networks," *in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 409-413.

[14] Schlüter, J., and Böck, S. "Improved musical onset detection with convolutional neural networks," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, Florence, Italy, 2014, pp. 6979-6983.

[15] Corbera, F., and Serra, X. "Tempo estimation for music loops and a simple confidence measure," in *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR,* New York, USA, 2016, pp. 269-75.

[16] Korzeniowski, F., and Widmer, G. "A fully convolutional deep auditory model for musical chord recognition," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP),* Salerno, Italy, 2016, pp. 1-6.

[17] Juhan, N., Choi, K., Lee, J., Chou, S., and Yang, Y. "Deep learning for audio-based music classification and tagging: Teaching computers to distinguish rock from bach," *IEEE signal processing magazine*, vol. 36, no. 1, pp. 41-51, 2018.

[18] Murthy, Y., Jeshventh, T. K. R., Zoeb, M., Saumyadip, M., and Shashidhar, G. K. "Singer identification from smaller snippets of audio clips using acoustic features and DNNs," in *2018 Eleventh International Conference on Contemporary Computing (IC3)*, Noida, India, 2018, pp. 1-6.

[19] Gómez, F., Dıaz-Bánez, J. M., Gómez, E., and Mora, J. "Flamenco music and its computational study," in *Mathematical Music Theory: Algebraic, Geometric, Combinatorial, Topological and Applied Approaches to Understanding Musical Phenomena*, World Scientific Publishing, Singapore, ch. 8, pp. 303-315.

[20] Kroher, N., Díaz-Báñez, J. M., Mora, J., and Gómez, E. "Corpus COFLA: a research corpus for the computational study of flamenco music," *Journal on Computing and Cultural Heritage (JOCCH)*, vol. 9, no. 2, pp. 1-21. 2016.

[21] Serra, X. "Creating research corpora for the computational study of music: the case of the Compmusic project," in *Audio engineering society conference: 53rd international conference: Semantic audio*, London, UK, 2014, article number 1-1, [9p.].

[22] Mora, J., Gómez, F., Gómez, E., and Díaz-Báñez, J. M. "Melodic contour and mid-level global features applied to the analysis of flamenco cantes," *Journal of New Music Research*, vol. 45, no. 2, pp. 145-159, 2016.

[23] H. Wang, J., and Han, J. "BIDE: Efficient mining of frequent closed sequences", in *Proceedings of the 20th international conference on data engineering*, Boston, MA, USA, 2004, pp. 79-90.

[24] J. Salamon, E. Gomez, and J. Bonada, "Sinusoid extraction and salience function design for predominant melody estimation," in *International Conference on Digital Audio Effects*, Paris, France, 2011, pp. 73–80.

[25] Díaz-Báñez, J. M., and A. Mesa. "Fitting Rectilinear Polygonal Curves to a Set of Points in the Plane.", in *European Journal of Operational Research* vol. 130, no. 1, pp. 214-222, 2001.

[26] Choi, K., Fazekas, G., Sandler, M., and Cho, K. "Convolutional recurrent neural networks for music classification", in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, US, 2017, pp. 2392-2396.

[27] He, K., Zhang, X., Ren, S., and Sun, J. "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition,* Las Vegas, NV, USA, 2016, pp. 770-778.

[28] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition,* Honolulu, HI, USA, 2017, pp. 4700-4708.

[29] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, USA, 2016, pp. 2818-2826.

[30] Pons Puig, J., Nieto, O., Prockup, M., Schmidt, E. M., Ehmann, A. F., and Serra, X. "End-to-end learning for music audio tagging at scale," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*, Paris, France, 2018, pp. 637-44.

[31] Graves, A., and Schmidhuber, J. "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol.18, no. 5-6, pp. 602-610, 2005.

[32] Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., and Le, Q. V. "Specaugment: A simple data augmentation method for automatic speech recognition". arXiv preprint arXiv:1904.08779. 2019.

[33] Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association,* Dresden, Germany, 2015, pp. 3586-3589.

[34] McFee, B., Humphrey, E. J., and Bello, J. P. "A software framework for musical data augmentation," in *16th International Society for Music Information Retrieval Conference*, Malaga, Spain, 2015, pp. 248-254.

[35] Kingma, D. P., and Ba, J. "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

Aitor Arronte Alvarez

Aitor Arronte Alvarez is a machine learning researcher specializing in Music Information Retrieval, Audio Signal Processing, and Speech Recognition. He works at the University of Hawaii at Manoa at the Center for Language and Technology as a Technology Specialist. Aitor Arronte Alvarez holds a M. Eng. in Decision Systems Engineering and is currently finishing his Ph. D. at the Universidad Politécnica de Madrid.

Francisco Gómez

Francisco Gómez became Full Professor at Technical University of Madrid in 1994. He started doing research on computational geometry, computer graphics and facility location. In 2003 he switched to Music Information Retrieval and Computational Music Theory and has been doing research in this field since then. Francisco Gómez received a Ph.D. in Computer Science from the Technical University of Madrid under the supervision of Godfried Toussaint. His main interests in Music Information Retrieval and Computational Music Theory are music similarity, mathematical measures of rhythm complexity and syncopation, automated analysis of music traditions, especially flamenco music, Afro-Cuban music, Brazilian music and in general African music, teaching mathematics via the arts, and active learning methods in teaching mathematics. He has participated in several research projects funded by several Spanish agencies. Francisco Gómez teaches courses on Computer Graphics, Computational Geometry, Statistics, Algebra, Discrete Mathematics, and Pattern Recognition, among others. He uses innovative teaching methods such as inquiry-based teaching. In particular, he has used a collaborative version of the Moore method, Mazur's method for large audiences, and methods based on writing to teach mathematics.