Analyzing quality clarinet sound using deep learning. A preliminary study.

Francisco Chávez de la O
Department DISIT
Centro Universitario de Mérida
University of Extremadura
Santa Teresa de Jornet, 38
Mérida, 06800, Spain

Francisco Fernádez de Vega
Department CCT
Centro Universitario de Mérida
University of Extremadura
Santa Teresa de Jornet, 38
Mérida, 06800, Spain

Francisco J. Rodríguez Díaz
Department DISIT
Centro Universitario de Mérida
University of Extremadura
Santa Teresa de Jornet, 38
Mérida, 06800, Spain

***Abstract*—When a music student begins training, one of the main problems encountered is the proper understanding of specific terms that teachers introduce as a way of analyzing the type of sound produced by the student. The goal of a music teacher is that their students improve the quality of the sound they are emitting, but not in all cases students understand and know how to apply the concepts that the teacher wants the instrument to be able to emit the expected sound. Any tool that allows students to distinguish between sound quality would be helpful. The work presented here is a preliminary study of the quality of the sound emitted by a clarinet using deep learning techniques. It presents a first approximation of what will become a software tool that will allow music students to see that the sound quality they are emitting is correct and in real time. With this type of tools music students will be able to understand and associate the concepts explained by the teacher in a simpler way, and will even serve as a guide to improve their learning when the teacher is not present.**

I. Introduction

Playing a musical instrument is a complex task and requires a number of years of practicing under the guidance of an expert. If a young student in playing an instrument, he has to practice several hours per week with a professional that will take the role of an expert and to teach how the student has to learn to play the instrument. Nevertheless, the language employed by the teacher makes it difficult for the student to properly understand the way of improving. The teacher will use term such as: *color, texture, sound center, focused sound, dark or bright sound, etc*. Although the terms can be connected to physical components of the sound, instrument, and physical actions performed by students, such as overtones of the main frequency, or embouchure, to name but a few, the fact is that the connection is not clearly described by teachers. Students spend a lot of time in relating the terms used by the teacher with the techniques they should use to finally make these techniques improve the final quality of the instrument's sound.

The influence of the physical elements that the student should use and the final sound produced in the clarinet, have been studied by some authors, such as [1], [2], and also how to measure the quality of an instrument, [3]. These works present the relationship between physical elements and the final sound, but, as far as we know, there is no a tool that can tell the student if the sound quality that is generating his instrument in acceptable, good or excellent, that is, provide software which is able to measure the sound quality of a clarinet . This software will allow the student to better understand and apply the concepts explained by the teacher and we believe that it will allow the student to reach a higher level faster.

In this line, we presented a paper where a first approach for qualitatively analyzing some components of the sound was studied, such as the concept of *sound centered* [4]. The technique employed will allow to provide real-time feedback on the quality of the sound, so that the student will have larger capability for reacting and improving, and hopefully reduce the number of years required for playing the clarinet. Although this research have been applied to clarinet, a similar approach could be applied to other wind instruments.

The work presented in [4] uses the *sound spectrogram* and extracts a set of characteristics that allow us to detect if the student is playing a centered sound. This analysis is carried out through the use of a Fuzzy Rule-Based System (FRBS), which thanks to an optimization process, is able to detect the characteristic of the sound being sought. We can say that the technique presented uses the concept of *sound visualization* [5].

Nowadays, one of the most used image analysis techniques and that the best results are giving is the use of convolutional neural networks (CNN) using deep learning [6], [7], [8]. The benefits of the use of deep learning in complex vision problems are amply demonstrated, therefore, in this work we present a preliminary work where we use vision techniques to analyze the sound quality by means of CNNs. The work presented analyzes the sound of a clarinet extracting sections of 1 second. These sections are shown by their spectrogram that is later analyzed by a CNN to determine whether the sound quality is good or not. The results presented in this preliminary work allow us to affirm that thanks to the use of CNNs we can detect, in real time, the sound quality that is emitted by a clarinet. This technique can be implemented in a specific software, that it will help the young students in their learning, allowing them significantly reduce learning time.

The rest of the paper presents the related works in II and the methodology used in III. A complete description of the CNN used is presented in IV. Finally, the results and a discussion

978-1-5386-2726-6/17/$31.00 ©2017 IEEE

of the conclusion are presented in V and VI respectively.

II. Related Works

In this work we present a novel system of analysis of the sound quality through deep learning, based on the analysis of spectrograms, but other authors have worked in this line of research with different techniques. In the literature we can find works focused on the classification of musical genre or more focused on the analysis of sound quality.

We can find a wide range of works related to the classification of musical genre. These works are based on different characteristics of the sound so that it can be classified [9]. Such is the case of [10] which is based on invariant sound characteristics for classification. The works presented in [11], [12] are based on acoustic characteristics to make the musical classification. Other works such as [13], [14] are based on the rhythm for classification and even use deep learning as [14]. Among the classification works we can even find the work presented in [15] that tries to classify the author of a musical note with a certain instrument. With the same instrument several authors produce the same note and the system tries to classify its author.

On the other hand, currently we can find in the literature works that are beginning to use the CNNs and deep learning to analyze the quality of a piece of music or some specific aspects of it. It is a research line that is beginning to be explored and that can obtain good results thanks to the capacity of this type of neural networks. In [16] the authors present a novel system to improve feature learning for audio data using neural networks. They show that these methods provide significant improvements in training time and the features learn are better than state of the art. Other works such as presented in [17], [18], [19] use deep learning for music classification.

In other previous work presented by our group, we analyzed a characteristic of sound in which teachers focus a lot on the beginning of music studies, which is whether a sound is centered or not [4]. In this work, a FRBS technique was used to detect this characteristic. The work presented here is aimed at this research line, trying to obtain a complete system that analyzes the quality of the sound emitted by a clarinet.

III. Methodology

In this section we present the methodology used to analyze the sound produced by a clarinet that, by means of deep learning vision techniques, the quality of this sound will be detected.

A sound can be presented by a continuous wave in the time-space, but the information included in this wave is difficult to analyze with the vision techniques used by the CNNs. Therefore, we must transfer the sound information from the time-space to the spectral-space, so that we can apply the necessary techniques for its analysis. The training of the CNNs used requires a large number of images to obtain the best results, so we must obtain the largest number of images of the spectrum of the sound that we are going to analyze. In the problem presented it has been decided to section each of the sounds used in parts of 1 second and obtain the spectrogram information of that section. The Figure 1 shows the methodology used.

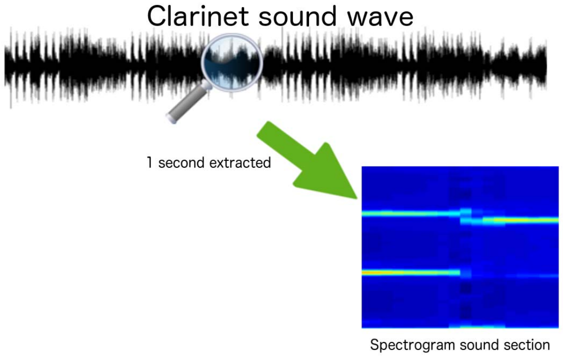


Fig. 1. Methodology used

The spectrogram is a visual representation of the variations of the frequency on the vertical axis, and of the intensity through the levels of colors of the sound that is being represented along the time, represented in the horizontal axis. The spectrogram can reveal features, such as high frequencies or amplitude modulations, that can not be appreciated even though they are within the frequency limits of human ears, hence it can be very useful to detect the sound quality that we are analyzing. To obtain the spectrogram, it is necessary to apply a Fourier transform to the signal. This process can be expensive, but there are libraries that facilitate the extraction of the spectrogram as well as a variety of types useful for the analysis of sound quality. In the work presented in this paper, we have used a Matlab library called *MIRtoolbox*, which allows us to obtain a set of different spectrograms [20], [20].

MIRtoolbox is a library for Matlab that allows us to extract musical features from audio files. This library has a set of functions that work with basic sound operations, feature extractors such as: rhythm, timbre, pitch and tonality; on the other hand it works with high leve features extractions, such as: structured and form, statistics, predictions, similatiry and retrieval and exportation. And finally, MIRtoolbox has a complete set of functions that allows us to work with the sound spectrum.

If we focus on the characteristics extraction based on the spectrum, in the work presented in this paper we have worked with 4 different types of spectrograms. The expression used to obtain the spectrogram in MIRtoolbox is $s = mirspectrum(a)$, but we can used a set of parameters to obtain different spectrograms types, such us:

- *mirspectrum(...,Terhardt) modulates the energy following outer ear model [21].*
- *mirspectrum(..., Bark) redistributes the frequencies along critical band rates (in Bark).*
- *mirspectrum(..., Mask) models masking phenomena in each band: when a certain energy appears at a given*

> *frequency, lower frequencies in the same frequency region may be unheard, following particular equations.*

[22]

With the combination of the parameters offered by MIR-toolbox, we can obtain different spectrograms, such as:

- Spectrogram: The spectrogram for each of the 1 second sections extracted from the sound wave. $sf = mirspectrum(a,'Frame')$
- Maximum Spectrogram: The spectrogram for each of the 1 second sections extracted for the maximum values of the sound wave. $sf = mirspectrum(a,'Frame','Max', maximum)$
- Correlation Matrix: The correlation matrix for each of the 1 second sections extracted from the sound wave.
- Bark spectrogram: The bark spectrogram for each of the 1 second sections extracted from the sound wave. $sb = mirspectrum(sf,'Terhardt','Bark','Mask','dB')$
- Autocorrelation spectrogram: The autocorrelation spectrogram for each of the 1 second sections that consists in looking at local correlation between samples. $as = mirautocor(sf)$
- MEL - Frequency crepstal coefficients spectrogram: The crepstral spectrogram for each of the 1 second sections that MFCC function offers as a description of the spectral shape of the sound. $m = mirmfcc(a,'Frame')$

To obtain the different spectrograms we have used the following algorithm 1. Figure 2 shows a set of images obtained with the algorithm 1. We can see the different spectrograms types obtained, with professional and student quality. The initial idea is to analyze the different types of spectrograms in order to detect with which of them we can obtain the best results and differentiate the sound of a clarinet produced by a professional or student. Once we see which type of spectrogram produces the best classification results, the next step will be to perform a deeper analysis with this spectrogram type. In both cases, for the initial analysis of the different spectrogram types, as well as for the deeper analysis of the most promising spectrogram type, we will use CNNs. As the literature demonstrates, this type of CNNs framed in the deep learning paradigm, will allow us to obtain good results.

In the next section we introduce the CNN used in this work.

IV. CLASSIFICATION USING DEEP CONVOLUTIONAL NEURAL NETWORKS

In the last few years, a new area known as deep learning have emerged in the field of machine learning. This new area encompasses different techniques that are fundamentally characterized by an hierarchical learning process in which high-level structures are automatically build starting from low-level ones across multiple layers, starting from the raw data (i.e. the pixel values of an image).

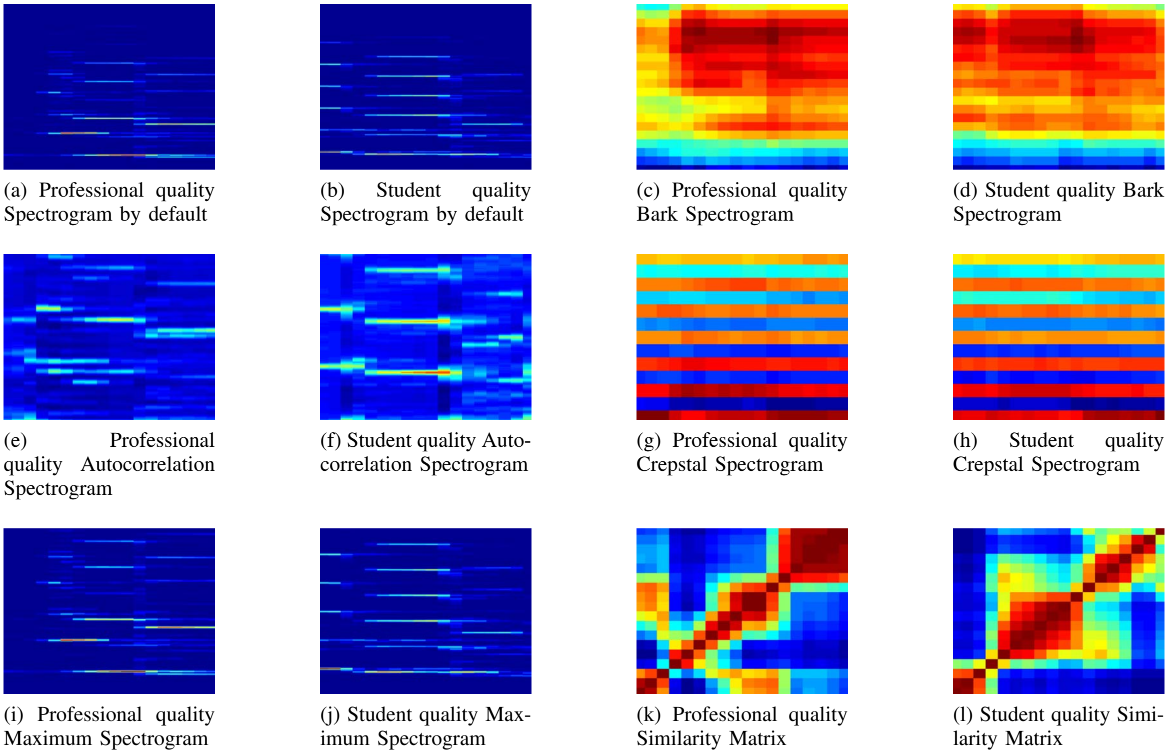
Deep learning appears as an alternative to traditional machine learning methods, which require a carefully selection of hand-designed features from which the classifier can detect patterns by means of one, two at most, non-linear transformations of those features. These classical methods have proven to be quite effective to solve simple or well-delimited problems, but encounter difficulties in dealing with real-world complex problems such as object and speech recognition. By contrast, deep learning techniques have hugely improved the state-of-the-art in such complex tasks.

```
function Extract_Audio_Spectrum_Images( File_name,
 file_part )
    % Load file
    a = miraudio ( File_name, ’Center’, ’Sampling’,
 11025,’ Normal’, ’Trim’ );
    %Spectrogram by default
    sf = mirspectrum ( a, ’Frame’ );
    saveas ( gcf, strcat ( ”, file_part, ”,’spectrum.png’ ) );
    %we take the significant maximums of the
 spectrogram
    maximum= max ( max ( mirgetdata ( mirpeaks ( sf,
 ’Track’, 25 ) ) ) );
    %we obtain the most indicative spectrogram
 according to its maximum
    sf = mirspectrum ( a, ’Frame’, ’Max’, maximum);
    saveas ( gcf, strcat ( file_part, ’_maximum.png’));
    %Similarity Matrix
    sm = mirsimatrix ( sf, ’Distance’, ’cosine’,
 ’Similarity’, ’exponential’)
    saveas( gcf, strcat ( file_part,
 ’_similarity_matrix.png’));
    %Bark spectrogram
    sb = mirspectrum ( sf, ’Terhardt’, ’Bark’, ’Mask’,
 ’dB’);
    saveas ( gcf, strcat ( file_part,
 ’_bark_spectrogram.png’));
    %Autocorrelation spectrogram
    as = mirautocor ( sf );
    saveas ( gcf, strcat ( file_part,
 ’_autocorrelation_spectrogram.png’));
    %MEL - Frequency crepstal coefficients spectrogram
    m = mirmfcc ( a, ’Frame’ );
    saveas ( gcf, strcat ( file_part,
 ’_crepstral_spectrogram.png’));
    close all;
    end
```

Algorithm 1: Matlab algorithm spectrograms extraction

In this work, we focus on a deep learning technique for supervised learning known as deep convolutional neural networks (CNN) [7], [8], which has shown an outstanding performance for visual objects recognition. CNN benefits from the spatial structure of input data, that is, the images. In doing so, CNNs use an architecture based on three key principles [23]:

- *Local receptive fields*: Each neuron of intermediate layers is connected to a small region of the input layer. This kind of layers are called convolutional layers.
- *Shared weights and bias*: All the neurons in a feature map



(a) Professional quality Spectrogram by default

(b) Student quality Spectrogram by default

(c) Professional quality Bark Spectrogram

(d) Student quality Bark Spectrogram

(e) Professional quality Autocorrelation Spectrogram

(f) Student quality Autocorrelation Spectrogram

(g) Professional quality Crepstal Spectrogram

(h) Student quality Crepstal Spectrogram

(i) Professional quality Maximum Spectrogram

(j) Student quality Maximum Spectrogram

(k) Professional quality Similarity Matrix

(l) Student quality Similarity Matrix

Fig. 2. Examples of different spectrograms used for clarinet sound samples with professional quality and student quality.

share these parameters. This way, these neurons become specialized in identifying a particular feature in different regions of the image. At the same time, the number of parameters to be set during the network learning phase is reduced notably.

- *Pooling layer*: This type of layers receive as input each of the feature maps and produce a new simplified feature map. For example, in *max-pooling layers*, each neuron produces an output with the maximum activation value from a small region of feature maps.

In the last few years, multiple deep CNNs models have been designed. For the purpose of the classification system developed in this work, we have chosen Alexnet [24] and GoogleNet [25] networks. Both models were originally build to classify high-resolution images from the IMAGENET dataset[1].

Figure 3 depicts a simplified scheme of Alexnet's architecture, whose network is composed of five convolutional layers and three fully connected. In contrast with convolutional layers, neurons in fully connected layers are linked to all neurons in previous layer. Local-response normalization layers follow the two first convolutional layers. Its aim is to promote competition among nearby groups of neurons by diminishing responses that are uniformly large in the neighborhood and increasing more pronounced responses [24]. Max-pooling layers are placed after the normalization layers and the fifth convolutional layer. Table I summarizes the filter size, the number of filters, and stride for each layer.

TABLE I

ALEXNET'S PARAMETERS FOR CONVOLUTIONAL AND MAX-POOLING LAYERS: FILTER SIZE, NUMBER OF FILTERS AND STRIDE.

Layer	Convolution			Max-Pooling	
	Filter Size	No. of Filters	Stride	Filter Size	Stride
First	11X11	96	4	3x3	2
Second	5x5	256	1	3x3	2
Third	3x3	384	1	-	-
Fourth	3x3	384	1	-	-
Fifth	3x3	256	1	3x3	2

GoogleNet's architecture is composed of 22 layers and places several pieces of the network working in parallel instead of sequentially, as seen in previous architectures. Each of these pieces is called *inception* and performs in parallel 1x1, 3x3, and 5x5 convolutions and max-pooling. At the end, the inception module perform a filter concatenation.

V. Results

This section will be divided into two subsections where we will present the results of the previous study performed with the different types of spectrograms obtained from the sound samples, to analyze which type of spectrogram obtains the best classification results and, with this, to obtain a CNN that allows us to classify the sound of a clarinet emitted by a professional or student.

[1] http://www.image-net.org/

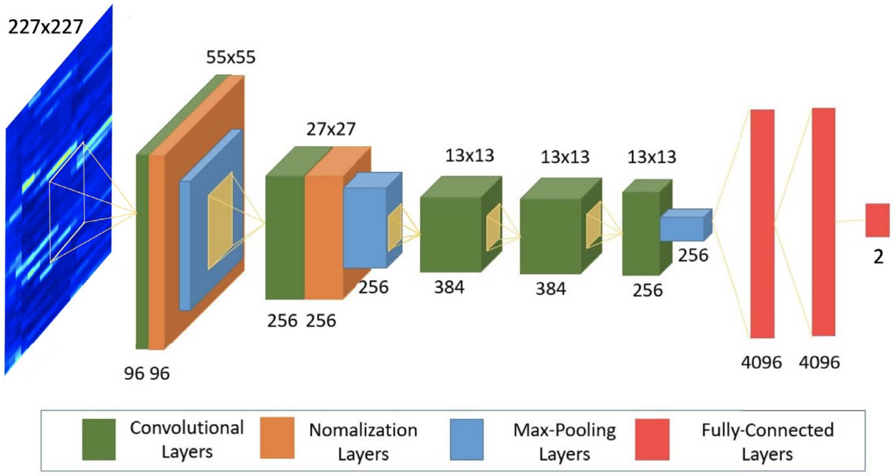


Fig. 3. Alexnet architecture.

A. Preliminary study

For this preliminary study we have selected two musical pieces obtained from a professional and a student, where we can only find the sound of a clarinet. Both pieces, each lasting approximately 3 minutes, have been divided into sections of 1 second. Once the sections have been obtained, the algorithm 1 has been executed and the different types of spectrograms for each section have been obtained, similar to those shown in figure 2.

After obtaining all the different spectrograms of each type of sound, with professional and student quality, we proceed to the adjustment of the CNNs that we have used in this study. Two types of CNNs, AlexNet [26] and GoogleNet [25] have been used. As the results will show, the CNN that obtains the best preliminary results is AlexNet, which is why it is used for the complete study.

The platform used for testing is a computer DELL Precision T7610, with two high performance Intel ®Xeon ®Processor E5-2600 v2, Super-fast memory: Handle huge data sets with ease with 16GB of 1866MHz ECC memory for blazing-fast performance and professional-grade NVIDIA ®Quadro ®and AMD FirePro$^{\text{TM}}$ graphics. The computer has incorporated a GPU TeslaK20, 1.17 Tflops for double-precision floating-point operations, 3.52 Tflops for single-precision floating-point operations, 5Gb RAM and 2496 CUDA cores. The computer has installed the NVIDIA Deep Learning GPU Training System (DIGITS), version 5.0.0 with Caffe deep learning framework, version 0.15.13.

In table II can see the results obtained for the preliminary study, where we have used the 6 different types of spectrograms, using for each spectrogram type the two CNNs mentioned AlexNet and GoogleNet:

- Spectrogram
- Maximum Spectrogram
- Similarity Matrix
- Bark Spectrogram
- Autocorrelation Spectrogram
- Crepstal Spectrogram

TABLE II
RESULTS OF EXPERIMENTS USING DIFFERENT TYPES OF SPECTROGRAMS IN COMBINATION WITH ALEXNET AND GOOGLENET CNNS

Type of Spectrogram and CNN used	Accuracy (val)	Loss (val)
Spectrum with AlexNet	50.00	0.684164
Spectrum with GoogleNet	46.86	0.729203
Cresptal_Spectrum with AlexNet	62.50	0.686592
Crepstal Spectrum with GoogleNet	53.13	0.688844
Autocorrelation Spectrum with AlexNet	**87.50**	**0.604521**
Autocorrelation Spectrum with GoogleNet	50.00	0.689674
Bark Spectrum with AlexNet	56.25	0,690462
Bark Spectrum with GoogleNet	50.00	0.694991
Similarity Matrix with ALexNet	65.63	2.269320
Similarity Matrix with GoogleNet	50.00	0,702758
Maximum Spectrum with AlexNet	46.86	0,694652
Maximum Spectrum with AlexNet	53.13	0.691520

Each of the examples is executed with CNNs a total of 30 epochs, to determine what type of spectrogram together with which type of CNN obtains the best results.

We can observe in Table II that the results of **Autocorrelation Spectrum with AlexNet** obtain the most promising results (**87.50** in Accuracy and **0.604521** in Loss), so we carried out a second study to determine the number of epochs that would be necessary for a better adjustment of the CNN. (See Figures 2e and 2f)

We can observe in Table III that as the number of epochs is increased the results are not better significantly, in addition the loss of the network in validation increases. Therefore, it is decided that 30 will be the optimal number of epochs for the in-depth study.

TABLE III
RESULTS OF THE EXPERIMENT TO DETERMINE THE NUMBER OF EPOCHS USING AUTOCORRELATION SPECTRUM WITH ALEXNET

Epochs	Accuracy (val)	Loss (val)
30	87.50	0.60452
30	81.25	0.61310
30	78.13	0.62899
300	78.13	2.52284
300	75.00	0.87883
300	71.88	2.91720
1000	71.88	2.49165
1000	68.75	5.53477
1000	71.88	3.92327
2000	81.25	3.51644
2000	78.13	6.86708
2000	75.00	3.42359
3000	84.38	2.78385
3000	78.13	3.51644
3000	78.13	8.01011

B. A complete study using Autocorrelation spectrogram with AlexNet CNN

Once that we have determined the type of spectrogram, CNN and the number of epochs more promising to obtain the best classification results, we proceed to carry out a more complete study that allows us to obtain the efficiency of the CNN used in the problem of quality sound classification emitted by a clarinet.

In this study we performed 30 executions using the autocorrelacion spectrogram and AlexNet CNN with an Adjustment process established for 30 epochs. Table IV shows the result of this experiment and Figure 4 shows us the evolution of 2 different executions.

In Table V we can see a summary of the global results obtained in this experiment.

The results presented in this section allow us to affirm that the preliminary system designed based on a CNN will be able to differentiate the sound emitted by a clarinet of a professional as of a student. After making a total of 30 adjustments of the CNN we obtain a success rate average of 76.56%, with a maximum of 87.50%. Due to the fact that we use only one piece of music for each type of interpreter (professional or student), the results allow us to affirm that in future works this success rate average will grow, thus obtaining a more robust classification system.

VI. CONCLUSION

In this paper we present the first preliminary system of quality analysis of the sound emitted by a clarinet through deep learning, using CNNs. We have analyzed two music pieces where we can only find the sound of a clarinet, created by a professional and by a student. The final objective is to equip to the clarinet students, with a tool that allows them to associate the concepts taught by their teachers, which must corresponds to the sound that the clarinet emits, to reach the desired level. Thanks to this tool, the students will be able to associate concepts and sounds, in such a way that it is visualized if its sound resembles that of a professional.

TABLE IV
RESULTS OF COMPLETE EXPERIMENT

Iteration	Accuracy (val)	Loss (val)
1	87.50	0.60452
2	81.25	0.61310
3	78.13	0.62899
4	84.38	0.60814
5	78.13	0.59095
6	78.13	0.60326
7	78.13	0.61402
8	78.13	0.63070
9	59.38	0.61947
10	62.50	0.61580
11	68.75	0.62276
12	84.38	0.58990
13	75.00	0.61808
14	84.38	0.61864
15	75.00	0.61749
16	84.38	0.62690
17	68.75	0.60468
18	59.38	0.63440
19	62.50	0.61437
20	62.50	0.58888
21	81.25	0.59393
22	81.25	0.62397
23	87.50	0.59185
24	84.38	0.59661
25	71.88	0.61314
26	84.38	0.60653
27	81.25	0.60122
28	84.38	0.58967
29	62.50	0.61600
30	87.50	0.57696

TABLE V
SUMMARY RESULTS

	Accuracy (val)	Loss (val)
Average	76.56	0.,609165
Standard Deviation	9.13	0.014502
Maximum	87.50	0.634399
Minimum	59.38	0.576956

To reach our final objective, we present a new system that allows us to analyze the sound quality through its spectrogram. The system transforms the information of the musical wave, expressed in the time domain, to its domain of frequency, by means of its spectrograms. These spectrograms visually represent the sound that will be analyzed by a CNN. We adjusted the CNN called Alexnet to model their weights and adjust them to be able to differentiate between professional quality spectrograms and students' quality spectrograms.

The results presented in this work demonstrate that it is possible to design tools that analyze the sound quality, thus allowing a better and faster learning to music students. The novel model presented in this paper reaches an average of 76.56%, with a maximum quota of 87.50%. This model has been adjusted using only one piece of music by each type of interpreter, professional and student. In future works, we will analyze more than one piece of music for each type of performer, in addition to incorporate new levels of learning, which will allow us to obtain more robust models.

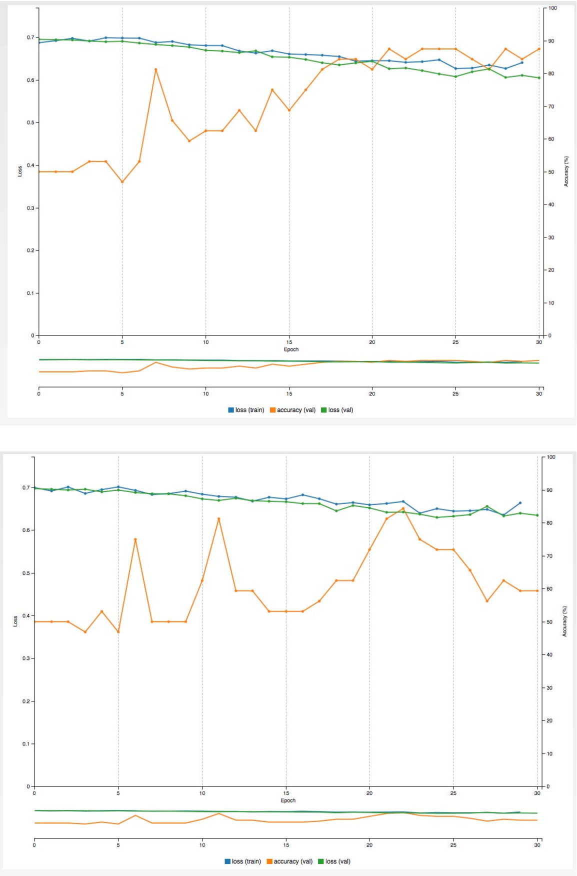


Fig. 4. Evolution of the adjustment CNN process

Acknowledgment

We acknowledge support from Spanish Ministry of Economy and Competitiveness and European Regional Development Fund (FEDER) under project EphemeCH (TIN2014-56494-C4-{1,2,3}-P),and projects IB16035, GR15068 and GR15130 of the Regional Government of Extremadura, Department of Commerce and Economy, conceded by the European Regional Development Fund, ”A way to build Europe”.

References

[1] B. Gazengel, T. Guimezanes, J.-P. Dalmont, J. B. Doc, S. Fagart, and Y. Léveillé, “Experimental investigation of the influence of the mechanical characteristics of the lip on the vibrations of the single reed,” in *Proceedings of the International Symposium on Musical Acoustics, Barcelona, Spain*, 2007.

[2] F. Pinard, B. Laine, and H. Vach, “Musical quality assessment of clarinet reeds using optical holography,” *The Journal of the Acoustical Society of America*, vol. 113, no. 3, pp. 1736–1742, 2003.

[3] R. Pratt and J. Bowsher, “The objective assessment of trombone quality,” *Journal of Sound and Vibration*, vol. 65, no. 4, pp. 521–547, 1979.

[4] F. Chávez and F. Fernández de Vega, “Assessing quality of sound emission in beginning clarinetists using optimization processes,” in *16th International Society for Music Information Retrieval Conference - ISMIR15, LATE-BREAKING DEMO, Málaga, Spain*, 2015.

[5] Y.-H. Kim and J.-W. Choi, *Sound visualization and manipulation*. John Wiley & Sons, 2013.

[6] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[7] K., K. Kavukcuoglu, M. Ranzato, and Y. LeCun, “What is the best multi-stage architecture for object recognition?” in *2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 2146–2153.

[8] Y. LeCun, F. J. Huang, and L. Bottou, “Learning methods for generic object recognition with invariance to pose and lighting,” in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2, 2004, pp. 97–104.

[9] Y.-F. Huang, S.-M. Lin, H.-Y. Wu, and Y.-S. Li, “Music genre classification based on local feature selection using a self-adaptive harmony search algorithm,” *Data and Knowledge Engineering*, vol. 92, pp. 60 – 76, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169023X14000640

[10] C. Zhang, G. Evangelopoulos, S. Voinea, L. Rosasco, and T. Poggio, “A deep representation for invariance and music classification,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 6984–6988.

[11] J. H. Su, T. P. Hong, and Y. T. Chen, “Fast music retrieval with advanced acoustic features,” in *2017 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW)*, June 2017, pp. 357–358.

[12] L. Nanni, Y. M. Costa, A. Lumini, M. Y. Kim, and S. R. Baek, “Combining visual and acoustic features for music genre classification,” *Expert Systems with Applications*, vol. 45, pp. 108 – 117, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417415006326

[13] A. Lykartsis and A. Lerch, “Beat histogram features for rhythm-based musical genre classification using multiple novelty functions,” in *Proceedings of the 16th ISMIR Conference,(JANUARY)*, 2015, pp. 434–440.

[14] A. Pikrakis, “Audio latin music genre classification: a mirex 2013 submission based on a deep learning approach to rhythm modelling,” 2013.

[15] C. E. Kinzer, W. C. McDermott, and S. A. Cheyne, “Experimental investigation of individual musicians on tonal quality for saxophones and clarinets,” *The Journal of the Acoustical Society of America*, vol. 137, no. 4, pp. 2339–2339, 2015.

[16] S. Sigtia and S. Dixon, “Improved music feature learning with deep neural networks,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 6959–6963.

[17] P. Hamel and D. Eck, “Learning features from music audio with deep belief networks.” in *ISMIR*. Utrecht, The Netherlands, 2010, pp. 339–344.

[18] L. Deng, D. Yu *et al.*, “Deep learning: methods and applications,” *Foundations and Trends® in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[19] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, “Unsupervised feature learning for audio classification using convolutional deep belief networks,” in *Advances in neural information processing systems*, 2009, pp. 1096–1104.

[20] O. Lartillot, P. Toiviainen, and T. Eerola, “A matlab toolbox for music information retrieval,” *Data analysis, machine learning and applications*, pp. 261–268, 2008.

[21] E. Terhardt, “Calculating virtual pitch,” *Hearing research*, vol. 1, no. 2, pp. 155–182, 1979.

[22] O. Lartillot, “Mirtoolbox users manual,” *Finnish Centre of Excelence in Interdisciplinary Music Research*, 2011.

[23] M. A. Nielsen, *Neural Networks and Deep Learning*. Determination Press, 2015.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems 25*, P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., 2012, pp. 1106–1114.

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf