Performance to Score Sequence Matching for Automatic Ornament Detection in Jazz Music

Sergio Giraldo and Rafael Ramírez

***Abstract*—Expressive music performance research deals with the analysis and characterization of the performance deviations a musician introduce when performing a musical piece. Most of the previous work in this area focuses on timing and dynamic deviations in classical music. Very few works deal with ornamentation characterization, and most of these works also focus on classical music where ornaments are indicated in the score. However in popular music (e.g. jazz music) ornaments are seldom indicated in the score and it is the performer who introduce them by adding/substituting groups of notes based on melodic, harmonic and rhythmic contexts, as well as on his/her musical background. In this work we present a system for automatically recognizing ornaments in jazz melodies. Based on a set of real performances by a professional jazz guitarist, we apply Dynamic Time Warping to automatically detect ornaments by matching the performed and the score note sequences. For each ornamented note, its musical context description and corresponding ornamentation are stored in a database. We evaluate the alignment of the ornamented performance and score using a ground truth consisting of manual annotations made by jazz musicians.**

***Index Terms*—DTW Alignment, Expressive Music Performance, Ornament Recognition, Jazz Guitar.**

I. INTRODUCTION

Expressive music performance research is usually based in the analysis and characterization of performance deviations from the score that a musician may introduce when playing a piece. Most of the work analyses performance variations such as timing and dynamics deviations, but few research has been done towards ornamentation. Usually, in classical music, ornamentations are indicated in the score and most of the work is targeted to characterize how ornamentation is performed.

Although some authors have reported on systems for automatic ornament detection, based on different musical melodic contexts, very few musical literature can be found about melodic embellishment in jazz. Most of the reported systems utilize classical music theory to decide whether a sequence of notes corresponds to an ornamentation (e.g. Appoggiaturas, trills, mordents, turns, etc.). However this approach does not always apply in jazz music, as melodic embellishment in jazz lays in between archetypical ornamentation and free improvisation. The context in which a musician may use ornaments is usually learnt by copying the playing style of other professional musicians. Furthermore, in popular music, ornaments depend widely on the musician background, taste and current intention, and they are used based on melodic, harmonic and rhythmic context. In the case of jazz music, the performance of a piece usually include the addition of different types of ornaments, such as passing notes, neighbor notes and chord scale notes. Typically, these ornaments may include short musical phrases (also called licks), often used as a preparation for a target note, or to replace long notes.

In this paper we present a system to automatically recognize ornaments in jazz music. By comparing the similarity between a sequence of score notes and its corresponding sequence of performed notes, our aim is to automatically obtain for each performed note (or group of notes) its corresponding *parent* note in the score, as depicted in Figure 1. We use a data set of 27 jazz standards, recorded by a professional musician. We apply Dynamic Time Warping (DTW) to best match the performed note sequence with the score. The alignment reults are validated using manual annotations provided by music experts. Using the identified onrnamentations, a database of ornaments is created, and indexed by the context of the note in which they were used.

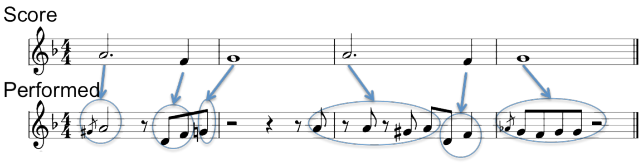

Fig. 1. Ornamentation. Arrows illustrate how circled performed ornaments correspond to its *parent* note in the score[1].

II. RELATED WORK

Most of the work in expressive performance analysis focus on the deviation of onsets, duration and energy, and has been done mainly for piano classical music (for an overview see Goebl et al. (2008) [2]). Expressive music performance in jazz has been investigated by Lopez de

Manuscript received December 29, 2014

S. Giraldo is with the Music Technology Group at Pompeu Fabra University, Roc Boronat 198, 08019, Barcelona, Spain (e-mail: sergio.giraldo@ upf.edu).

R Ramírez, is with the Music Technology Group at Pompeu Fabra University, Roc Boronat 198, 08019, Barcelona, Spain (e-mail: rafael.ramirez@ upf.edu).

[1] Music excerpt of jazz tune "Yesterdays" by J. Kern as performed by jazz guitarrist W. Montgomery.

Mantaras et al. [5], who use case based reasoning to render saxophone performances. Ramirez et al. [8] use evolutionary algorithms to model expressive performances in jazz saxophone, including ornamentation as a performance parameter. In previous work [11], models for ornamentation in jazz guitar context are obtained using machine learning techniques. After manually aligning the score of a music piece with the transcription of the performance of professional musician, ornaments are characterized and stored in a database. This database is later used to generate models to predict note ornamentation and generate ornamented performances from inexpressive scores using machine learning techniques.

Automatic recognition and characterization of ornaments in music has been studied in the past a part of the music expressive analysis. Perez et al. [6] model mordents and triplets in Irish fiddle music with the aid of 3D motion sensors to capture bowing gestures, and time-pitch curves analysis. Trills and appoggiaturas are modeled by Puiggros et al. [7] in bassoon recordings by automatically extracting timing and pitch information from the audio signal, and using machine learning techniques to induce an expressive performance model. Gómez et al. [3] automatically detect ornaments in flamenco music (melismas) categorizing ornaments into six different types, and adapting the Smith-Waterman algorithm [10] for sequence alignment. Casey and Crawford [12] use the MPEG-7 standard audio descriptors to build a Hidden Markov Model classifier to automatically detect a subset of possible ornaments in 18th and 17th century lute music, based on the hypothesis that HMM state transitions occur at higher rates during ornaments than during non-ornamented segments of an audio signal.

III. METHODOLOGY

The general framework of our methodology is depicted in Figure 2. We obtain a MIDI-like machine representation of both, the scores and their respective performance audio files. For each note in the score, we extract a set of melodic descriptors. We use DTW to match the performed and the score note sequences and compare this matching with the one done by human experts for evaluation purposes. Finally we include in a data base each ornamentation annotated with the music context in which it was performed.

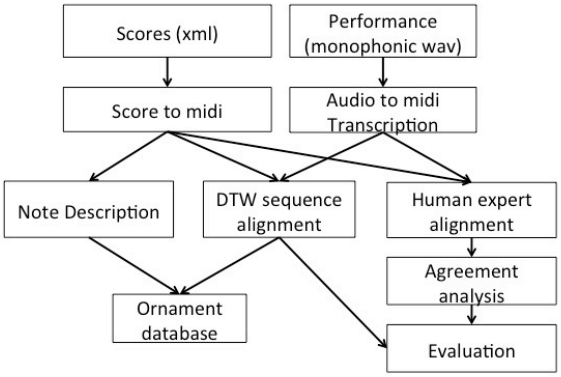

Fig. 2. Methodology framework.

A.Score data.

A set of 27 music scores were obtained from *the real book* [9]. The scores were converted to XML format, which permits to include other extra useful information such as chords. We built our own library of algorithms to parse the XML scores in order to obtain a description of each note in the music pieces.

B.Score note description

Each note in the score was described based on its inherent properties such as pitch, duration and onset, as well as, its musical context in such as previous interval, next interval, chord, key, previous note duration, next note duration, among others. A detailed list of the descriptors used can be found in previous work [11]. Description of the notes gives a framework for automatic score ornamentation that makes possible to apply the same ornament to notes which have similar description. However automatic score ornamentation is out of the scope of this work and will be commented later as future work.

C.Performance recordings and audio to MIDI transcription.

Our performance data set consists of 27 jazz standards, recorded by a professional guitarist. The guitarist was instructed to play one note at a time (monophonic) avoiding strumming chord and playing double notes. Pieces were played with no repetitions, i.e. if a song had an AABA form only the A and B section were recorded. Each piece was recorded using the *Aebersold's* commercial accompaniment backing tracks [4]. Monophonic recorded melodies were automatically transcribed into MIDI representation using the methodology for automatic monophonic transcription of previous work [13], in which pitch and onsets are detected from pitch tracking and two heuristic rule filters.

D.Dynamic Time Warping Sequence Matching

For matching performed ornamented notes with score notes we used Dynamic Time Warping (DTW). We depart from the standard implementation of DTW algorithm by defining a similarity cost function between pair of notes belonging to two different sequences, based on the pitch, the duration, the onset and the legato level. The cost function is defined as:

$$\cos t = P_c + D_c + O_c + ILO_c + FLO_c \quad (1)$$

where P_c is the *pitch cost*, D_c is the *duration cost*, O_c is the *onset cost*, ILO_c is the *initial legato onset cost* and FLO_c is the *final legato offset cost*. The purpose of *legato cost* is to prevent the algorithm to split a group of ornament notes into separate parent notes in the score. We assume that a group of notes conforming an ornament is played legato. A key point is to detect which notes belong to the same ornament. Hence, we implemented two functions to track the first and the last note onset of a group of notes based on the legato level (minimum time gap threshold between two consecutive notes). Each of these costs are defined as follows:

- *Pitch Cost* is the distance between the pitches (encoded as MIDI number) of two notes, defined by:

$$P_c = W_p * (pitch_p(i) - pitch_s(j))^2 \quad (2)$$

where W_P is a weight factor, and $pitch_p(i)$ is the pitch of the i^{th} note of the performed sequence and $pitch_s(j)$ is the pitch of the j^{th} note of the score.

- *Duration Cost* is the distance between the duration in seconds of two notes, defined by:

$$D_c = W_d * (dur_p(i) - dur_s(j))^2 \quad (3)$$

where W_d is a weight factor, and $dur_p(i)$ is the duration of the i^{th} note of the performed sequence and $dur_s(j)$ is the duration of the j^{th} note of the score.

- *Onset Cost* is the distance between the onset in seconds of two notes, defined by:

$$O_c = W_o * (onset_p(i) - onset_s(j))^2 \quad (4)$$

where W_o is a weight factor, and $onset_p(i)$ is the onset of the i^{th} note of the performed sequence and $pitch_s(j)$ is the pitch of the j^{th} note of the score.

- *Initial legato onset cost* (ILO_c) is defined as the distance in seconds between the onset of the first note of the ornament note group and its corresponding parent note onset in the score

$$ILO_c = W_{ILO} * (ini\_onset_p(j) - onset_s(i))^2 \quad (5)$$

- *Final legato offset cost* (FLO_c) is defined as the distance in seconds between the onset of the last ornament note group and its corresponding parent note onset in the score.

$$FLO_c = W_{FLO} * (last\_onset_p(j) - onset_s(i))^2 \quad (6)$$

where W_{ILO} and W_{FLO} are weight factors for initial and final onset legato cost respectively, $onset_s(j)$ is the onset of the j^{th} note of the score sequence and $ini\_onset_p(i)$ and $last\_onset_p(i)$ are functions that return the onset of the first/last note of a legated ornament group of notes in which the j^{th} note of the performance occurs.

Finally, we follow the standard procedure of DTW: a similarity matrix $H_{M,N}$ is defined in which M is the length of the performed sequence of notes and N is the length of the sequence of score notes. To calculate similarity, each cell of matrix H is calculated as follow:

$$H_{(i,j)} = \cos t + \min(H_{(i-1,j)}, H_{(i,j-1)}, H_{(i-1,j-1)}) \quad (7)$$

where $cost$ is the one defined in Equation 1, and *min* is a function that returns the minimum value of the preceding cells (up, left, and up-left diagonal). Matrix $H$is indexed by the i^{th} note of the score sequence and the j^{th} note of the performance sequence.

A *backtrack path* is obtained by finding the lowest cost calculated in the similarity matrix. Starting from the last score/performance note cell, the cell with the minimum cost at positions $H_{(i-1,j)}$, $H_{(i,j-1)}$, and $H_{(i-1,j-1)}$ is stored in a *backtrack path* array. The process iterates until indexes arrive to the first position of the matrix. As explained previously each note in the performance is assigned to a parent note in the score. In Table I we present a *backtrack path output* example of the algorithm for the music excerpt of Figure 1. The correspondence between performed and score parent notes is represented by the column note pairs at the table. As the table shows performed notes 1 and 2 are assigned to note 1 in the score, performed notes 3 and 4 are assigned to note 2 in the score, and so on.

TABLE I: EXAMPLE OF BACKTRACK PATH OUTPUT FOR THE MUSICAL EXCERPT OF FIGURE 1

Score notes	1	1	2	2	3	4	4	4	4	5	5	6	6	6	6	6
Performed notes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

E.Ornament Database

To create the database of ornamentations, we calculate for each pair of notes in the *backtrack output path*, the deviations between the performed note and its corresponding parent note in the score as: *pitch offset* (in semitones), *onset offset* (in beats) and *duration ratio* (fraction of performed duration in beats over score note duration in beats). In Table 2 is shown an example for the music excerpt of Figure 1, where doted lines separate ornaments. We derived a broad characterization of ornaments based on the ornament length: ornaments were labeled as simple, if its length is smaller than three; otherwise they were labeled as complex.

TABLE II: EXAMPLE OF ORNAMET DATABASE FOR THE MUSIC EXCERPT OF FIGURE 1

Score notes	Performed notes	Pitch offset	Onset offset	Duration ratio
1	1	-1	- 1/2	1/6
1	2	0	0	2/3
2	3	-3	- 1/2	1/2
2	4	0	0	1/2
3	5	0	- 1/2	1/8
4	6	0	- 1/2	1/6
4	7	0	1/2	1/6
4	8	-1	1 1/2	1/6
4	9	0	2	1/6

5	10	-3	- 1/2	1/2
5	11	0	0	1
6	12	1	- 1/2	1/8
6	13	0	0	1/8
6	14	-2	1/2	1/8
6	5	0	1	1/8
6	16	0	1 1/2	1/8

IV. RESULTS

Figure 3 presents an example of the resulting similarity matrix obtained for one of the recorded songs, in which the *x-axis* corresponds to the sequence of notes of the score and the *y-axis* corresponds to the sequence of performed notes. The cost of correspondence between all possible pair of notes is depicted in red for the highest cost (less similar) and blue for the lowest cost (most similar). The dots on the graph show the *backtrack path* (or optimal path) found for alignment. Diagonal lines represent notes which were not ornamented, as the correspondence from the performance notes to the parent score notes is one to one. On the contrary vertical lines represent notes that were ornamented, as two or more performed notes correspond to one parent note in the score.

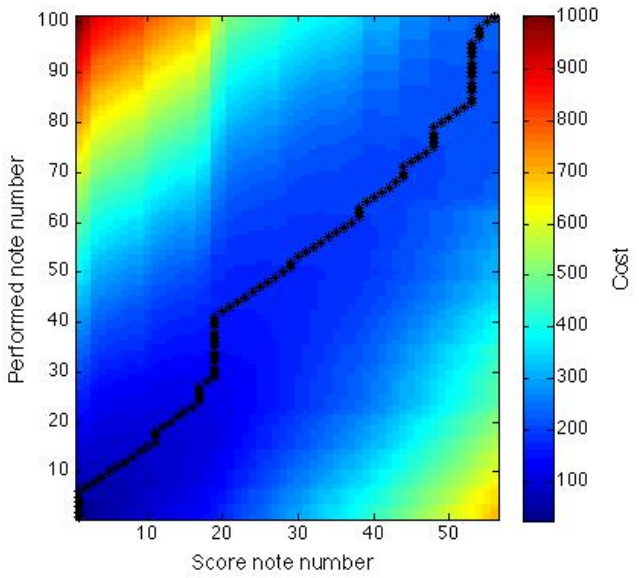

Fig. 3. Score-performance matching using DTW implementation on a song example

A.Human expert alignment.

As there are not clear rules of how embellishments are performed, there is no ground truth for establishing a correspondence between performance and score notes. To overcome this issue, we asked jazz musicians to manually match performed notes with the corresponding parent score notes for each piece. Each of the pieces was annotated by 5 different musicians. Musicians were asked to associate performance notes to score notes by drawing lines in a piano roll representation in a GUI developed for this purpose. Figure 4 shows an example annotation of one user for a piece fragment. The upper note sequence corresponds to the score, and lower note sequence corresponds to the performance. Vertical/diagonal lines were marked by the musician to indicate performance to parent score note correspondence.

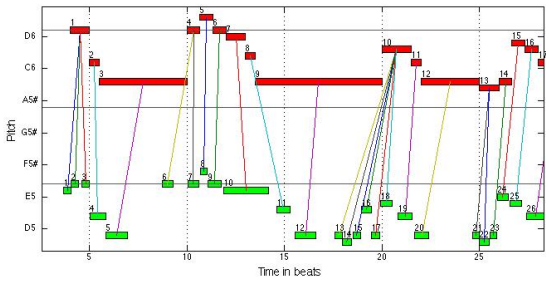

Fig. 4. Human performance to parent note manual annotation example. Score (top), performance (bottom)

B.Agreement analysis.

There are some cases in which the correspondence is ambiguous. Therefore, different musicians may choose to match different performance notes with one score note. In those cases we weighted each couple of linked notes based on how many musicians chose to link that particular pair of notes. Link occurrence count was stored in a matrix defined in a similar way to the matrix $H_{M,N}$ explained in Section III.D. The highest value is given to the pair of notes that were matched by all five musicians, whereas the lowest rating was given to notes for which no musician match the pair. In Figure 5, we present a graph of the link-occurrence-count matrix of one of the pieces in the dataset. In the figure is possible to identify sections in the piece with high agreement and sections with low agreement (ambiguity).

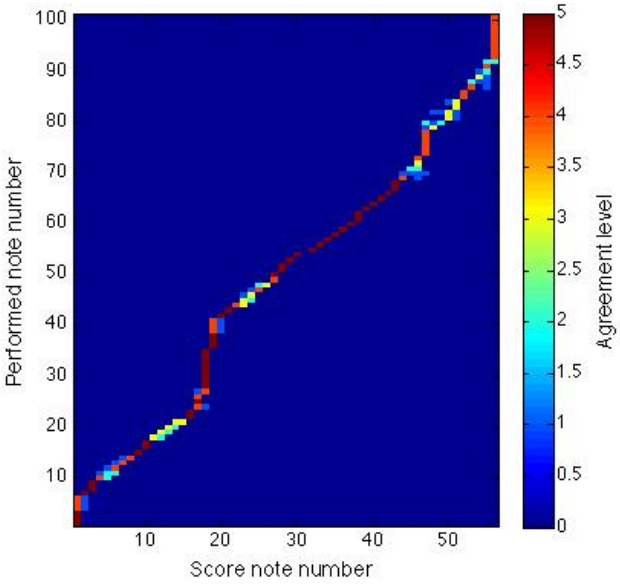

Fig. 5. Agreement level of human annotated score-performance note matching

C.Weight factors optimization

Weight factors were optimized using Genetic Algorithms (GA). Initial values were set in 1 for each weight factor. Maximum and minimum boundaries were set in 0 and 1

respectively. The average accuracy of all songs was defined as de fitness function. The stropping criteria was set to a maximum of 500 iterations, and a relative threshold change in fitness function of 1x10−6. Crossover factor was set to 0.8 and mutation factor was set to 0.02. The weight parameters found after GA optimization are presented in Table III. Notice how legato onset costs (W_{ILO} and W_{IFO}) have a higher weight compare to pitch and duration costs (W_p and W_d).

TABLE III: OPTIMIZED WEIGHT FACTORS

W_p	W_d	W_o	W_{ILO}	W_{IFO}
0.39	0.03	0.96	0.86	0.64

D.Evaluation

We quantify the overall performance of our approach, based on the *backtrack path output* of the algorithm and the *agreement analysis* of human annotations. Accuracy was calculated by penalizing it when the algorithm output when it diverges form the human agreement. The evaluation criteria was defined as follows:

High Agreement (HA): if the algorithm matches a pair of notes with the highest agreement of the annotated dataset (i.e. all experts agree with the algorithm output), then the penalization is zero.

Medium Agreement (MA): if the automatically matched pair of notes has medium agreement (i.e. some of the experts agree and some do not) the penalization is proportional to the agreement between a certain expert among the rest of the experts. This means that the algorithm makes a mistake which is similar to the one a human expert would make.

Low Agreement (LA): if the predicted matching pair of notes was not annotated by any of the experts then the penalty is 1.

The overall accuracy is then calculated as follows:

$$accuracy = 1 - \frac{HA + MA + LA}{Total\_performed\_notes} \quad (7)$$

The mean accuracy obtained for the 27 recordings set is of 80,76%, with a standard deviation of 0.10 (10%) after optimization. We also quantify the agreement among experts. Each expert alignment was compared to the alignment made by the other 4 experts on each song, following the same evaluation criteria. In Table IV we present the evaluation for each expert. From table IV it can be seen that the performance of the system is comparable to the performance of a human expert. In Figure 6 the calculated accuracy of the algorithm for each individual piece is depicted in the first bar of each song. The second bar represent the average agreement among experts.

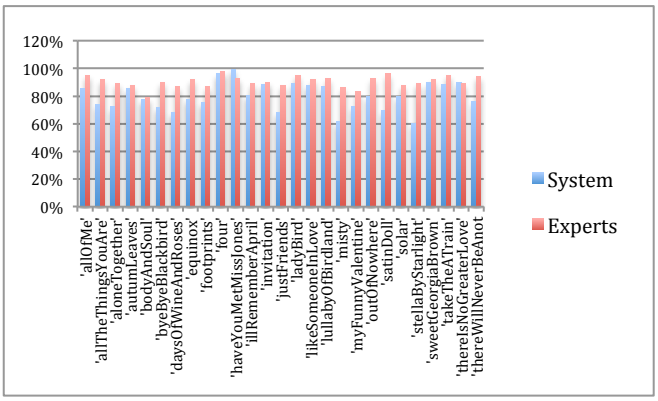

Fig. 6. Accuracy of alignment obtained for each of the 27 jazz pieces.

V. CONCLUSIONS

In this paper we have presented a system to automatically recognize ornamentations in jazz music. We have used a data set of 27 audio recordings of jazz standards performed by a professional guitarist. We have applied Dynamic Time Warping to align the score with the performance of the musician, and match notes of the performance with the corresponding parent notes in the score. Based on the alignment, we have generated a database of embellishments, annotated with the musical context in which they were performed. For evaluation purposes we have analyzed the annotations of jazz musicians to generate an agreement level chart between the performance notes and parent score notes. Based on the experts' annotations, we have estimated the accuracy of the system by creating penalty factors based on how much the output of the algorithm differs from the human experts agreement. Results indicate that the accuracy of our approach is comparable with the accuracy of annotations of music experts.

As future work we plan to apply machine learning techniques, following the same procedure of previous work in [11], to synthesize an expressive ornamented in the style of a particular guitarist. First, score notes will be classified into ornamented and non-ornamented notes, and for the former we will automatically select suitable (according to similar music contexts) ornaments from the annotated ornamentation database.

VI. ACKNOWLEDGEMENTS

This work has been partly sponsored by the Spanish TIN project TIMUL (TIN2013-48152-C2-2-R)

REFERENCES

[1] H. Crook: How to improvise, Advance Music, UK, 1991.

[2] W. Goebl, S. Dixon, G. De Poli, A. Friberg, R. Bresin, and G. Widmer: "Sense in expressive music performance: Data acquisition, computational studies, and models," Sound to sense-sense to sound: a state of the art in sound and music computing, Logos Verlag, Berlin, pp. 195-242, 2008.

[3] F. Gómez, A. Pikrakis, J. Mora, J. M. Diaz-Baénz, E. Gómez, and F. Escobar: "Automatic detection of ornamentation in flamenco," In Fourth International

Work- shop on Machine Learning and Music MML, 2011.
[4] G. Kennedy and B. Kernfeld: ”Aebersold, Jamey” in B. Kernfeld. The new Grove dictionary of jazz, New York: Grove’s Dictionaries Inc, NY, 2002.
[5] R. L. De Mántaras, J. L. Arcos, X. Serra: “Saxex: A casebased reasoning system for generating expressive musical performances,” Journal of New Music Research, Vol. 27, No. 3, pp. 194–210, 1998.
[6] A. Perez, E. Maestre, S. Kersten, R. Ramírez: “Expressive irish fiddle performance model informed with bowing,” In Proceedings of the international computer music conference., 2008.
[7] M. Puiggòs, E. Gómez, R. Ramírez, X. Serra, R. Bresin: “Automatic characterization of ornamentation from bassoon recordings for expressive synthesis,” In Proceedings of International Conference on Music Perception and Cognition, 2006.
[8] R. Ramírez, A. Hazan, E. Maestre, and X Serra: “A genetic rule-based model of expressive performance for jazz saxophone,” Computer Music Journal, Vol. 32, No. 1, pp. 38–50, 2008.
[9] The Real Book, Hall Leonard, Milwaukee, W, USA, 2004.
[10] T.F. Smith and M.S. Waterman: “Identification of molecular sequences,” J. Molecular Biology, No. 147, p.p. 195-197, 1981.
[11] S. Giraldo. “Modeling Embellishment, Duration and Energy Expressive Transformations in Jazz Guitar”. Masters thesis, Pompeu Fabra University, Barcelona, Spain, 2012.
[12] M. Casey and T. Crawford. "Automatic Location and Measurement of Ornaments in Audio Recordings." Proceedings of the 5th International Symposium of Music Information Retrieval (ISMIR). 2004.
[13] H. Bantulà, S.Giraldo and R. Ramirez. “A Rule-based System to Transcribe Guitar Melodies.” Proceedings of the 7th International Workshop on Machine Learning and Music, Barcelona, Spain, November 28, 2014

Sergio Giraldo received the B.S. degree from los Andes University, Bogotá, Colombia, in 1998, in Mechanical Engineering and the M.S. degree in Sound and Music Computing from the Pompeu Fabra University (UPF) in 2012.

He is currently a PHD student at the Music Technology group from the Pompeu Fabra University, Barcelona, Spain. His current research is in computational models for expressive music performance in jazz music.

Rafael Ramirez received the B.S. degree in mathematics from the National University of Mexico, Mexico City, Mexico, and the M.S. degree in artificial intelligence and Ph.D. degree in computer science, both from the University of Bristol, Bristol, U.K.

He is currently an Associate Professor with the Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain. Prior to joining the Universitat Pompeu Fabra, he was a Lecturer with the Department of Computer Science, National University of Singapore, Singapore. His current research interests include artificial intelligence, music information retrieval, declarative languages, music perception, and cognition.