Proceedings of the 3rd International Conference on Music & Emotion (ICME3), Jyväskylä, Finland, 11th - 15th June 2013. Geoff Luck & Olivier Brabant (Eds.)

INDUCING RULES OF ENSEMBLE MUSIC PERFORMANCE: A MACHINE LEARNING APPROACH

Marco Marchini, Rafael Ramirez, Panos Papiotis, Esteban Maestre

Music Technology Group, Universitat Pompeu Fabra, Spain
marco.marchini@upf.edu

Abstract

Previous research in expressive music performance has described how solo musicians intuitively shape each note in relation to local/global score contexts. However, expression in ensemble performances, where each individual voice is played simultaneously with other voices, has been little explored. We present an exploratory study in which the performance of a string quartet is recorded and analysed by a computer. We use contact microphones to acquire four audio signals from which a set of audio descriptors is extracted individually for each musician. Moreover, we use motion capture to extract bowing descriptors (bow velocity/force) from each of the four performers. The gathered multimodal data is used to align the performance to the score. Then, from the aligned data streams, we obtain a note-by-note description of the performance by extracting note performance parameters. We apply machine-learning algorithms to induce human-readable rules emerging from the data. The dataset consists of three performances of Beethoven's quartet n° 4 in C minor by a group of professional musicians: a "normal", a "mechanical" and an "over-emphasized" execution. We run our analysis on the three conditions separately as well as jointly, deriving rules specific to each condition and rules of general domain. Apart from encoding knowledge of expressive performance, the results shed light on how musicians' roles in ensemble performance.

Keywords: ensemble music performance, machine learning, expressive music performance

1. Introduction

In western music tradition, the notation provides the height and the duration of each note in a fairly explicit way. However, intensity and tone quality are represented only approximately. This leaves to the performer enough freedom in deciding how to interpret the music's content. Deviations in timing are also introduced, which render the performance more human and expressive. Musicians always introduce such deviations, even when playing mechanically (Palmer 1997).

The phenomena have been studied in the past from a computational approach leading to models of expressive performance capable of emulating human expression. Previous research on expressive performance used machine learning techniques to build models from real data of piano performances (Widmer & Goebl 2004). Other instruments have been considered in few other works (L Mantaras et al. and Ramirez et al.) by also considering additional expressive transformations that are absent in piano technique (e.g. vibrato and glissando).

Despite the abundance of applications of machine learning to expressive performance, most works are focused on solo performance and do not address the problem of performing in an ensemble. In classical music ensembles, each performer interprets their own part as dictated by the score. Since it is the sum of the parts that makes the whole, relations among individual parts implicitly define the role and the task of each member within the group.

The problem of playing in ensemble has been studied by narrowing down music performance to very specific tasks. Some studies focus on synchronization on the task of tapping together (Repp 2005) or on very specific musical skills (Moore et al. 2010). In (Goebl & Palmer 2009) the synchronization among musicians is studied also taking into account the role and the auditory feedback.

The only work addressing expressive performance in ensemble is devoted to string quartets (Sundberg et. al 1989) but from an analysis-by-synthesis approach. This means that the rules that Sunberg defined were tested directly by creating synthetic performances and were not evaluated on real recordings.

In this work we use machine learning to derive rules of expressive performance from recording of string quartet performances. We aim to understand how relations among parts affect the performance of each musician. For this reason we extract a set of score contextual descriptors including information about the relationship of each note with other parts in the score. We use such score descriptors to predict several note level performance parameters. We compare the predictive power of the machine-learning algorithm in two cases: when relationships among parts are considered or ignored. We then discuss the results and present some rules derived by the system in the cases where relationship among parts proved useful for the prediction.

The rest of the paper is structured as follows. Section 2 describes the recorded material and the acquisition of the data. Section 3 introduces the descriptors extracted from the score and the parameters extracted from the performance. We then explain our method in Section 4, present the results in Section 5 and discuss them in Section 6. Finally, in Section 7, we conclude providing directions for future work.

2. Data acquisition

We recorded a professional string quartet executing Beethoven's quartet n° 4 in C minor (opus 19 n°4, allegro-prestissimo movement).

After the quartet had played their first version ("normal") we asked for a "mechanical" and an "exaggerated" execution. The three performances were 15 min. long in total. Within this time, we collected more than 10k individual notes.

Acquired data include four individual audio tracks (one for each musician) coming from piezoelectric contact microphones. Additionally, we acquired bowing motion data via an EMF motion tracking system as carried out in (Maestre 2009). From the bowing motion data we obtain time series of bow velocity (Maestre 2009) and bow force sampled at 240Hz (Marchini et al. 2011). From the audio we extract time series of energy and pitch.

The extracted signals together with the score are given as input to a dynamic programming algorithm (Maestre 2009) to produce a precise note-by-note segmentation of the performance. Note boundaries are inspected manually to correct eventual segmentation errors.

The score was segmented into phrases the help of a professional musicologist leading to an average phrase length of four bars. As in other works of expressive performance modelling (Widmer & Tubodic 2003) we removed the main effect of tempo modulations to study only residual deviations of timing on note duration. The main effect of tempo modulation is obtained by fitting a parabolic tempo curve to the sequence of onsets of all instruments on each phrase. From this we calculate a value of tempo in beat per minutes (BPM) for each note that we later use as a reference for computing deviations on the duration of the performed note.

3. Descriptors

On each note we compute score contextual descriptors and parameters of the performance. The former will serve as features vector for the machine-learning algorithm while the seconds as learning tasks.

3.1. Score Contextual Descriptors

We define two types of score contextual descriptors: horizontal and vertical.

Horizontal note descriptors are computed based solely on a musician's individual part,

Table 1. Feature set with the smallest context length provided to the machine-learning algorithm.

Horizontal Descriptors
• nominalDuration • previousInterval • nextInterval • previousNoteRatio • nextNoteRatio • metricalStrenght • melodicCharge • narmour • nextNarmour
Vertical Descriptors
• harmonicCharge • isHighestMC • otherMusician1_nominalDuration … • otherMusician1_nextNarmour … • otherMusician2_nominalDuration … • otherMusician2_nextNarmour … • otherMusician3_nominalDuration … • otherMusician3_nextNarmour

ignoring the parts of the other musicians. These include both properties of the note itself, and also properties of the neighbouring notes (preceding and subsequent) in the part. Different temporal context windows sizes are considered by adding more or less neighbouring notes to the feature set.

Melodic contour is represented by melodic intervals of one note to the next and by Narmour implication realization class on each group of three notes (Narmour 1990).

Note salience includes the melodic charge, which is defined as the smallest number of steps to get from the tonic to the note in the circle of fifths (a number from zero to six).

Rhythmic information is represented by the metrical strength and rhythmic contour. Metrical strength depends on the position of the note relative to the bar and is encoded by an integer number from 0 to 5 from the strongest to the weakest metrical position. Rhythmic contour is characterized by the ratio between nominal durations of a neighbouring note and the note itself.

Vertical note descriptors include information from the score about the notes played by other musicians concurrently with the note being characterized. Each concurrent note is picked from others' part by selecting the note simultaneous (if any) to the characterized note or the one active at the beat where the characterized note is started. Vertical descriptors are then formed from the horizontal attributes of those picked notes (one for each other musician).

Additionally within vertical descriptors we include harmonic relationships of the note with concurring notes: `isHighestMC` and `harmonicCharge`. The former is a boolean set to "yes" if the note presents the highest value of melodic charge among concurrent note. The latter is the harmonic charge (Friberg 1995) computed on all the notes active within the beat of the characterized note. We compute the harmonic charge on the list of notes by first estimating the chord root note and then computing the average melodic charge of all the notes respect to the root note. To compute the chord root note we use the implementation from the open project music21 (http://mit.edu/music21/).

Table 1 depicts a list of 29 descriptors divided in horizontal and vertical representing the totality of descriptors for the smallest context window size considered. We build larger feature sets by adding in an analogous way descriptors (both horizontal and vertical) referring to additional neighbouring notes by appending to the descriptor name the string: "`previous`", and "`next`" followed by the number of separating notes from the reference.

3.2. Performance parameters

We apply machine learning techniques to our data set in order to learn models for predicting note-level parameters of the performance. The performer parameters we focus on are: loudness, duration ratio, vibrato amplitude, and bow velocity. Note *loudness* is the maximum RMS value (in dB) within the note boundaries. *Duration ratio* is the ratio between the duration of the performed note and the score duration considering the fitted phrase arc tempo as a reference. *Vibrato amplitude* is extracted from the pitch curve within the note boundaries by

taking only the part with the lowest aperiodicity. A spectral analysis is performed on the selected part, which looks for periodic components in the range 4-8Hz. If no such component is found the vibrato amplitude is set to zero, otherwise it is set to the corresponding amplitude value in pitch cents. *Bow velocity* is computed by taking the interquartile mean on the bow velocity values within the note boundaries (this means that the lowest 25% and higher 25% of the values are discarded to get only the central tendency for the note). The bow velocity is measured in cm/s can be either positive or negative depending on the bow direction.

4. Method

We used machine learning to predict the performer parameters using the introduced score contextual descriptors. We apply the C4.5 decision tree induction algorithm (Quinlan 1992) to obtain a regression tree predicting each of the performance parameters described above. We use the implementation of provided by the Weka machine learning software (Hall et al. 2009).

In order to test all the combinations of features we built a series of different datasets on which we run the algorithm independently. We form the mixed dataset by merging notes from the three expressive intentions into one unique dataset. Performing a regression on the mixed dataset means to find rules that are applicable to the three expressive intentions indistinctively.

For each task and for each musician we have 40 different combinations. Those are obtained by combining in all the possible ways two types of descriptors (horizontal or horizontal+vertical); the four expressive intentions datasets (normal, mechanical, exaggerated or mixed); and different temporal context windows sizes (from one to five neighbouring notes). Furthermore, considering the four learning tasks (loudness, bow velocity, duration ratio and vibrato amplitude) and the four musicians (violin 1, violin 2, viola and cello) we get a total number of 640 datasets.

For each dataset we run the algorithm and compute a value of correlation coefficient using 10-fold cross validation. We use the obtained correlation coefficient to quantify the predictive power of the decision tree algorithm on each musician, each task and each feature set.

5. Results

When training the system on the mixed intentions we obtain a mean correlation coefficient of 69%, 87%, 72%, and 76% on the tasks of loudness, bow velocity, duration ratio, and vibrato amplitude respectively. All the previous values have been computed averaging the correlation coefficients of all the musicians. Table 2 shows the complete set of values obtained on each individual dataset for expressive intentions. In general we see that on the mixed dataset we get a comparable correlation coefficient, when not better than the individual datasets. This means that musicians did not radically change their playing style in the three expressive cases but rather modulated differently the ranges of the deviations (which do not affect correlation).

Table 2. Average correlation coefficient for learning tasks (rows) and expressive intention (columns) pairs.

	Mechanical	Normal	Exaggerated	Mixed
Loudness	65%	68%	72%	69%
Bow Vel.	60%	60%	60%	87%
Duration	52%	56%	45%	72%
Vibrato	48%	77%	74%	76%

We run an ANOVA test on the correlation coefficients finding a significant effect of the following factors: musician, learning task and intention ($p>0.05$). The effect of temporal window size was not significant ($p=0.76$), which means that the smallest considered window of three is sufficient. We thus discard the effect of the temporal window size and focus for a moment on the effect of adding vertical features to the horizontal (feature type).

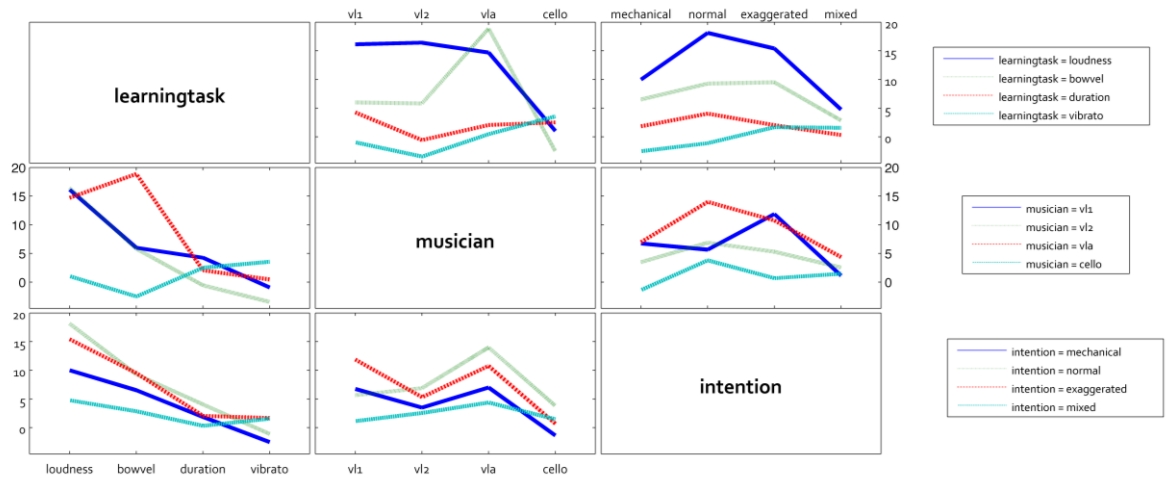

Figure 1 The increase in predictive power of the machine-learning algorithm is shown for the considered learning tasks, musicians and expressive intentions in percentage of improvement (PoI).

We now focus only on the improvement of predicting power when adding vertical descriptors to the feature set. We present the results in terms of *Percentage of Improvement* (PoI) respect to the baseline of only horizontal descriptors. An ANOVA test on PoI shows that this improvement is significantly positive for loudness and bow velocity whereas it is not significant for duration ratio and vibrato amplitude. The ANOVA also proves that the PoI depends not only on the learning task, on the musician and the intention, but also on the interaction between learning task and the musician (all $p<0.05$). This means that the PoI for each learning task follows different directions depending on the musician.

6. Discussion

In Figure 1 we see that we achieve the biggest PoI for loudness followed by bow velocity, duration and vibrato consistently across learning tasks. The second box of the first row of Figure 1 shows that whereas the violins have the highest amount of PoI for loudness, the viola surpasses the others in bow velocity.

The rules derived by the decision tree are sometimes difficult to interpret. We present here some rules derived on datasets where the machine learning algorithm performed significantly better when vertical descriptors where included. Generally the algorithm discovers a set 10-20 rules for each dataset. We report here only the two rules leading to the two more extreme values of the prediction.

We previously observed that the first violinist prediction for loudness seriously improves when adding the vertical features. Rules on the normal intention include the following:

```
IF
        nextNarmour=none > 0.5
        harmonicCharge <= 1.35
THEN
        Loudness_vl1 = 59.8 dB

IF
        vl2_narmour=IP > 0.5
        viola_narmour=VR_,IP,IR,D > 0.5
        cello_nominalDuration <= 1.5
THEN
        Loudness_vl1 = 83.3 dB
```

The first rule means the following: "if the note is the central of three notes that do not define a Narmour group and the harmonic charge is low than play it soft". Regarding the second rule, it is worth noticing how the context of where to play a note loud is defined solely by the properties of other musicians' notes. There is a difference of around 33 dB between the soft and the loud notes of this rule.

In the learning task of bow velocity the system achieves a very good correlation coefficient on the mixed dataset. The following two rules are part of the rules of general domain for the viola.

```
IF
        metricalStrenght_class=3
        vl1_nextNoteRatio <= 1.5
        vl2_interval <= 0.5
        cello_melodicCharge > 0.5
THEN
        BowVel_Viola = -43.4775 cm/s

IF
        nextNoteRatio > 0.75
        nominalDuration <= 0.75
        vl1_nextInterval > -1.5
THEN
        BowVel_Viola = + 49.2557 cm/s
```

Those rules define two contexts: whether to use a clear up-bow or a clear down-bow respectively. Both rules use relationships with other parts.

We have also shown how the PoI is not significant in learning the vibrato amplitude. Anyhow, the following two rules are derived by the system for the viola on the normal intention and use vertical descriptors:

```
IF
        harmonicCharge > 2.063
THEN
        vibratoAmp_viola = 2.97 p.cents

IF
        nextNoteRatio <= 0.75
THEN
        vibratoAmp_viola = 12.28 p.cents
```

The latters have to be applied in the order (the first that matches is applied) and thus they mean: "If the following note is consistently shorter, render the note vibrato except when the harmonic charge is high".

7. Future Work

The introduced approach has great potential for understanding roles in ensemble performance and collaboration among musicians.

We found a general improvement in the prediction of expressive deviation when considering the relationships among parts. The amount of improvement depended, however, on the specific performance parameter being predicted. Vertical relationships among parts proved useful for predicting individual musicians' behaviours on loudness and bow velocity.

By considering different expressive intentions we devised specific rules for each case and rules of general domain. A further question still unanswered is how well these rules scale on a larger dataset consisting of more pieces. Also it would be interesting to repeat the same analysis on more experimental conditions such as playing solo vs. playing in ensemble.

In this analysis we focused solely on score information to predict the expressive deviations. We did not consider how the sequence of introduced deviations affects future deviations (as in an autoregressive process). In future work, a more general analysis could also take into account this aspect as a feature and compare how much the introduction of this feature improves the prediction in respect of just score descriptors.

Acknowledgment

This work was partially supported by the EU FP7 FET-Open SIEMPRE Project no. FP7-ICT-2009-C-250026 and by the Catalan Government. We would also like to thank Erika Donald, Alfonso Pérez and Marcelo Wanderley for the support in organizing the experiment, as well as CIRMMT and BRAMS labs at Montreal, Quebec, Canada for hosting them. We thank Rafael Caro for helping us with the musicological analysis.

References

Friberg, A., (1995). *A Quantitative Rule System for Musical Performance*. PhD thesis, Royal Institute of Technology, Stockholm, Sweden.

Goebl, W., & Palmer, C. (2009). Synchronization of Timing and Motion Among Performing Musicians. *Music Perception: An Interdisciplinary Journal*, Vol. 26, No. 5, pp. 427-438

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update; *SIGKDD Explorations*, Volume 11, Issue 1.

Maestre, E. (2009), *Modeling instrumental gestures: an analysis/synthesis framework for violin*

bowing. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain.

Lopez de Mantaras, R. & Arcos, J.L. (2002). AI and music, from composition to expressive performance, *AI Magazine*, 23–3.

Marchini, M., P. Papiotis, A. Pérez, and E. Maestre (2011). A hair ribbon deflection model for low-intrusiveness measurement of bow force in violin performance. In *New Interfaces for Musical Expression Conference*.

Moore, G. P., & Chen, J. (2010). Timing and interactions of skilled musicians. *Biological Cybernetics*, 103, 401–414

Narmour, E. (1990) *The Analysis and Cognition of Basic Melodic Structures: The Implication-Realization Model*. Chicago: University of Chicago Press.

Palmer, C. (1997). Music performance. *Annual Review of Psychology, 48*, 155-138.

Quinlan J. R. (1992). Learning with continuous classes. *Proceedings of the Australian Joint Conference on Artificial Intelligence*. 343--348. World Scientific, Singapore.

R. Ramirez et al., A Tool for Generating and Explaining Expressive Music Performances of Monophonic Jazz Melodies, *International Journal on Artificial Intelligence Tools* **15**(4) (2006), 673–691.

Repp, B. H. (2005), Sensorimotor synchronization: A review of the tapping *literature, Psychonomic Bulletin & Review*, 12 (6), 969-992.

Sundberg, J., Friberg, A., & Frydén, L. (1989). Rules for automated performance of ensemble music, *Contemporary Music Review*, Vol. 3, pp. 89-109.

Widmer G. and Tubodic A. (2003). Playing Mozart by analogy: Learning multi-level timing and dynamics strategies. Journal of New Music Research, 32, pp. 259-268.

Widmer, G., Werner, G. (2004). Computational Models of Expressive Music Performance: The State of the Art. *Journal of New Music Research*, Vol. 33, No. 3, pp. 203–216.