Automatic music "listening" for automatic music performance: a grandpiano dynamics classifier

Diego Di Carlo and Antonio Rodà

Sound and Music Computing Group,
Dep. of Information Engineering,
University of Padova, Italy
roda@dei.unipd.it

Abstract. Current computational models for expressive music performance do not consider the role of the acoustic feedback generated by the musical instrument, which instead is usually a very important aspect of the musician-instrument interaction. This paper is a first step toward the design of autonomous systems for playing music able to adapt their behavior depending on the acoustic characteristics of the sounds generated by the instrument. A supervised machine learning approach was followed to implement an automatic classifier of piano dynamics. Firstly, many features were extracted involving current Music Information Retrieval (MIR) algorithms and the psychoacoustic concept of Loudness expressed by the Zwicker's model. Secondly, the most relevant components were selected via the Sequential Feature Selection (SFS) method and, thirdly, some typical algorithms were used to classify the notes into a hierarchy of music dynamics, from very soft to very loud. Results show an accuracy of 97%.

Keywords: Loudness classification, loudness model, feature extraction, machine learning, automatic music performance

1 Introduction

In recent years, several more or less autonomous systems and robotic devices have been developed in order to generate sounds, simulating the musician's main actions (see. [1] for a review). However, *playing music* is a very complex human activity that involves perceptual, cognitive, psychological, and aesthetic processes, besides mechanical motion and sound production. The keyword of human interaction and perception is "feedback". Human music performance, as every kind of artistic performance, is regulated by a sense of self control: musicians usually adapt the quality of their gestures (e.g. the touch on the piano keys) during the performance in relation to the acoustic response of the instrument or the environment in which they are playing. Besides pitches and durations, that are often fixed on the score, the task of these systems is to give *humanity* to automatically played music by varying times, dynamics and timbre of musical events, as musicians do. Despite many differences, all of these systems use an open-chain approach, in the sense that the control of the performance parameters is not

Workshop Proceedings of IAS-13
13th Intl. Conf. on Intelligent Autonomous Systems
Padova (Italy) July 15-19, 2014
ISBN 978-88-95872-06-3 pp. 529-536

regulated by any kind of real-time acoustic feedback, unlike what happens in a human performance.

Modern instruments, such as the Disklavier grandpiano, provide new stimuli in this research: bypassing the sound synthesis step, i.e. the problem associated to the realistic tones, only details regarding the execution of the notes could be spotted. A study of the University of Padova has developed a tool to apply different *music intentions* [2] to a music performance, namely CaRo 2.0 [3]. This system allows a user to manage audio expressive content of a piece, i.e. once the score is fixed, the system plays the song with a chosen intention: hard, soft, sad, bright, etc. However, as these studies proceed, it seems increasingly necessary to endow the automatic system with a feedback loop, to have a real time control of its performance (see Fig. 1).

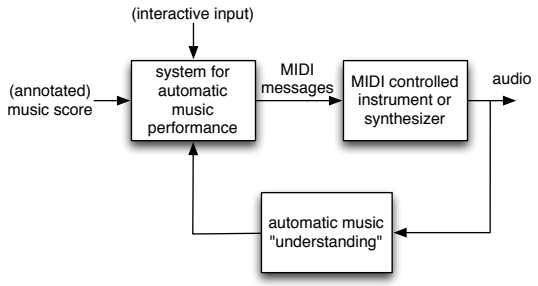

Fig. 1: *Block diagram of an automatic music performance system with a feedback loop for a real time control.*

This paper is a first step to improving the CaRo 2.0 system with a controller on music dynamic (related to the sound loudness) based on a real-time acoustic feedback. Dynamic was chosen among many different music parameters for its importance in music score indications. A dataset of 200 piano tones, played with different dynamics (namely *pp, p, mf, f, ff*), has been recorded and supervised machine learning techniques have been applied in order to classify the acoustic recordings on the bases of their dynamic. The supervised machine learning approach has already been successfully applied to other musical applications, such as the automatic classification of musical genre [4] or affects and emotions recognition of a music piece [5]. This approach is composed by the following steps: feature extraction and selection were performed, then automatic classification algorithms provide an evaluation of the proposed features.

Two basic dynamics indications in music are related to the concept of loudness: *piano* (p), meaning soft, and *forte* (f), meaning loud. However, loudness, such as music dynamics, is a relative and subjective measure, often confused with the physical quantities as sound pressure, intensity and power or specific volume levels. Some studies have defined mathematical models of the perceived loudness. Some of these showed good results on simple sounds: one of these is the psychoacoustic Zwicker's Loudess model [6]. The aim of this work is to investigate some other features, besides the amplitude of the audio signal, that can be used to classify the dynamics of Piano sounds. The rest of the paper is organized

as follows: in Section 2 the dataset creation, the feature extraction and selection methods are introduced; resulting features and their classification performance are presented in Section 3 and, finally, Section 4 deals with the conclusions and future directions.

2 Materials and Methods

2.1 *Dataset Construction*

As first part of the investigation, a dataset of 200 items was created: all these items are single notes played with different pitches, durations, key-velocities, and recorded from two different positions. The notes were produced by means of a MIDI file played on a Disklavier grandpiano. In order to collect a good variety of notes, five pitches were chosen: F1 (refered to MIDI code, n=29), C3 (48), G4 (67), D6 (86), and A7 (105). For each tones, four different duration (in milliseconds): 40, 120, 360, 1080; and, finally, the most important, five different MIDI key velocities: 20, 40, 60, 80, 100. To decide the set of values for this last parameter, a professional performer was asked to play some notes with the follow dynamic indications: *pianissimo* (or *pp*, meaning very soft), *piano* (*p*, soft), *mezzo-forte* (*mf*, moderately loud), *forte* (*f*, loud) and *fortissimo* (*ff*, very loud). Every note is followed by 2 seconds of rest, allowing a good sound decay. The MIDI file was compiled with Mtx2Midi software, downloaded on the Disklavier control unit and then directly played by the Disklavier sequencer. Performances were recorded in monophonic digital format at 24bits, using a cardioid AKG C-414B ULS microphone and a TASCAM US 2000 as digital audio converter, in a room with a light natural reverb. To test if the model is invariant to the amplitude of the sounds, two location for the mic were adopted: one inside of the piano, under the lib, at 30 cm from the sound board, the other one outside, at two meters from the instrument and 1.5 meters high. Thanks to the MIDI code, the audio recording was automatically segmented, synchronizing the recording waveform with the temporal description of the MIDI events. To emphasize amplitude invariant features, every item was normalized with respect to the RMS peak.

2.2 *Feature Extraction*

A set of features has been selected on the base of the related litterature. In particular predictors were computed using the Zwicker's Loudness model [6] and the MIRTollbox 1.5 [7].

1. Zwicker's model: Five features were extracted from the Loudness model for time-varying sounds, namely Intantaneous loudness, defined by Zwicker and Fastl [6]. This model of loudness is similar in its principle to the steady sound model. Nevertheless, temporal masking was taken into account and the loudness is calculated as a function of time and not in a global way, as simply sum of the

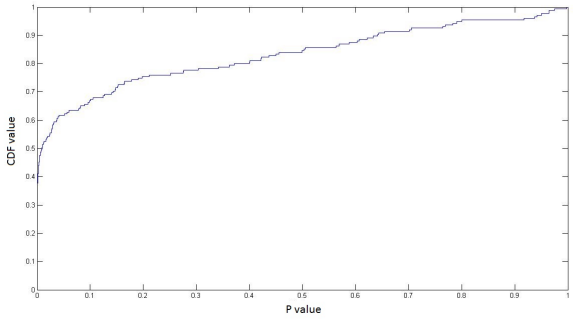

Fig. 2: *Empirical Cumulative Distribution function (CDF) computed comparing p-values of each features on a two random partition of the training set.*

specific loudness over all frequencies. In this paper the LoudnessToolbox 1.2[1] was used as an implementation of Zwicker and Fastl's Instantaneous loudness as a function of time. Both mathematical and statistical descriptors were extracted: peak value, RMS value, signal energy, mean, standard deviation.

2. MIRtoolbox: For audio/music feature extraction and statistical descriptor, MIRTollbox 1.5 was used [7]. Each frame-based feature was rapresented by its statistics, such as mean, standard deviation and slope (i.e. the derivative of the line that would best fit the values of the feature as function of frames), and the peak-based features are rapresented by the statistics of peak position and peak mignitude. It results 176-dimensional features vector each item.

2.3 *Feature Selection*

A hybrid supervised feature selection procedure was employed in three different steps: first, a filter method was used as pre-processing; then a forward SFS method was applied with a 10-fold cross-validation to rank and select some suboptimal sets of features; as a final step, some heuristics were taken to achieve the final features set.

1. Filters and Wrappers: Filter method provides a feature ranking rather than returns an explicit best feature subset. It is used for a first simple, fast and raw reduction of the features set. First, an hold-out partition 40/160 of the items set was made and a t-test was performed on two random partition of training data. The resulting p-values for each feature were compared and Figure 2 shows the empirical Comulative Distribution Function (CDF) of them, giving a general idea of how well-separated the two groups are by each feature. Result is that nearly 50% of the features have p-values close to zero, below the 0.01 threshold, meaning up to 88 features among the original 176 features that have strong discrimination power, and so, they could be used for classification.

The above computation does not even consider interaction between features, in fact an unvariate criterion was applied separetely on each of them; besides,

[1] Genesis S.A.: *"Loudness Toolbox"*, Aix en Provence, France, 2009.

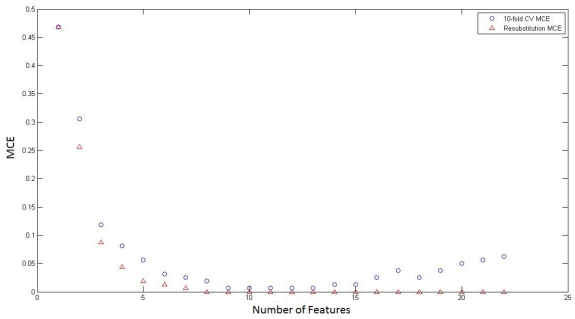

Fig. 3: *Misclassification Error function (MCE) versus the number of the features, resulting from cross-validated Forward Sequential Feature Selection algorithm.*

selected features may contain redundant information, so that not all features are needed. SFS is one of the most used techniques to decide how many feature are needed, selecting a subset of them by sequentially adding one feature at a time until certain stopping conditions are satisfied. In this paper forward SFS in a wrapper fashion was used. The feature selection method performs a sequential search using the Minimum Classification Error (MCE) of the Quadratic Discriminant Analysis (QDA) learning algorithm on each candidate feature subset as the performance indicator for that subset. The training set is used to select the features and to fit the QDA model, while the test set is used to evaluate the performance of the finally selected feature. As evaluation method, a stratified 10-fold cross-validation to the training set was used (see Fig. 3). The SFS algorithm stops when first local minimum of the cross-validation MCE is found, so returns a candidate suboptimal features subset. Figure 3 shows the cross-validation MCE of the first 22 features. The graph reaches the minimum value when 9 features are used, stays flat over a little rang and the curve goes up when more than 15 features are used, which means overfitting occurs there.

2. Heuristics: In this paper many iterations of the SFS method were performed and, in order to select the best feature vector, some heuristics were taken. First, the best set should contain at least one of the most correlated features to the key velocity parameter. It was provided by analysing Pearson's correlation coefficients (PCC, indicated with ρ) and their p-values ($\rho \geq 0.7$ and $p \leq 0.001$ were take as threshold). Second, the good subset should ensure the minimum number of features achiving the minimum value of MCE, so fixed threshold for length and MCE were taken. The last criterion is based on the following hypothesis "Good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other" [8]. This hypothesis is represented by the merit function of a feature subset S consisting of k features [8]:

$$Merit_{S_k} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}, \tag{1}$$

where $\overline{r_{cf}}$ is the average value of all feature-classification correlations, and $\overline{r_{ff}}$ is the average value of all feature-feature correlations. It can be seen as the predictive power of a feature subset divided by the redundancy between its own features. This last criterion provides the suboptimal feature subset, choosing the one with highest merit value. In the next section results are shown.

3 Results of the feature selection

3.1 *Feature Selection for Luodness Descriptors*

After features extraction, a matrix of 176-dimensional features vectors per 200 items were optained as a input for the SFS algoritm which was runned with 100 iterations. This should ensure that the final vector will not be biased because of particular partitioning of training and testing, and it result a set of 100 candidate feature subsets. So, the heuristics defined in Section 2.3 were taken in order to select which is the optimal one. First, the PCC analysis has stated that there are three features with correlation coefficients and p-values in respect with the defined threshold; second, the average dimension and the avarage MCE value of the 100 suboptimal feature vectors result respectively 7.09 and 0.087, selecting up to 20 candidate vectors. Finally, the CFS-Merit function was applied and, among all the features, the following were found to be relevant.

1) The mean of the frame-based *Spectral flatness*, or tonality coefficient, which provides a way to quantify how noise-like a sound is, i.e. with a quite flat and smooth spectrum, as opposed to being tone-like [10].
2) The peak of the frame-based *Fluctuation strength*, which indicates the rhythmic periodicities along the different channels, describing on how strong and fast beats are played within the respective frequency bands [11] [12].
3) The RMS values of time-depending *Instantaneous loudness*, which is computed by the auditory model [6].
4) The mean of the frame-based *Root-Mean-Square energy*, the most typical low-level dynamics descriptor which computes the temporal evolution of the energy.
5) The mean of the frame-based *Roll-off*, defined as the frequency under which the 85% of the spectral energy is distributed. It can be see as another measure of spectral shape [4].

3.2 *Classification Result*

To asses these results, ClassificationToolbox 3.1 was used, providing an user-friendly interface for classification with some standard multivariate Supervised Pattern Recognition classifiers [9].

First, a Linear Discriminant Analysis (LDA) e QDA were applied, assuming that the features are Gaussian distributed, with equal prior probabilities, and statistically independent. This last hypothesis is not initially verified in our dataset. However, the SFS approach definited in Section 2.3 allows us to remove

the most correlated features, and therefore satisfy the indipendence among the selected features.

Finally, the k-Nearest Neighbors (k-NN) classifier was tested. Here no assumption is required for probability density functions and features indipendece. When applying k-NN, the optimal value of k must be searched for. A set of k values (e.g. from 1 to 10) was tested with cross-validation and one of the k giving the lowest classification error was choosen as the optimal one. With this dataset, $k = 2$ was obtained. Table 1 shows the MCE resulting by different classifiers. The row labeled RND corresponds to the classification accuracy of randomly classification.

Classifier	MCE
RND	0.79
LDA	0.08
QDA	0.05
k-NN	0.03

Table 1: *Misclassification Error (MCE) for the feature vector defined in Section 3.1, for Random (RDN), Linear (LDA), Quadratic (QDA), k-Nearest Neighbors (k = 2) (k-NN) classifier.*

a	pp	p	mf	f	ff
pp	**97.5**	2.5	0	0	0
p	2.5	**97,5**	0	0	0
mf	0	0	**95**	5	0
f	0	0	0	**97,5**	2.5
ff	0	0	0	2.5	**97,5**

b	pp	p	mf	f	ff
pp	**97.5**	2.5	0	0	0
p	0	**90**	10	0	0
mf	0	5	**72.5**	20	2.5
f	0	0	17.5	**40**	42.5
ff	0	0	5	27.5	**67.5**

Table 2: *Cross-validated confusion matrix for k-NN classifier (k = 2) in percentage. Features defined in Section 3.1 are usend in left table (Table 2a), statistical features of Zwicker's model in right one (Table 2b).*

Confusion Matrices: Table 2a shows more detailed information about the loudness classifier performance in the form of confusion matrix. For clarity, only results of k-NN classifier are represented, other are almost similar in element dispositions and values. The rows correspond to the actual dynamic and the columns to the predicted ones by the classifier. The percentages of correct classification lie in the diagonal. As can be seen from the matrix, misclassification occurs only from adjacent type of dynamics, similar to what human would do. It is interesting to compare these results with those obtained using only the features of the Zwicker's model running the same k-NN classifier. Standard statistical descriptors are taken as feature vector to evaluate the Zwicker's model performance: peak, mean, and standard deviation of the Instantaneous Loudness function. Table 2b shows that in Zwicker's model loud dynamics are mostly misclassified, but it provides a great discrimination between soft and loud, the two basic and most important dynamics: these features alone reach an overall accuracy of 73%.

4 Conclusions

In this article, a features-based method for the automatic classification of the dynamic level of piano tones has been presented and evaluated. A feature selection was performed and the proposed features set was evaluated using statistical pattern recognition, trained with a dataset of single piano notes. Results show that the *Spectral flatness*, the *Fluctuation strength*, the *RMS value* of the Zwicker's *Instantaneous loudness*, the *RMS energy* and the *Roll-off* are good descriptors of the dynamic content. Using these features, a classification rate of 97% has been achieved in a dataset consisting of five different music dynamics.

Future work will focus on a real-time implementation of the proposed classification methods to be integrated in the CaRo 2.0 system. Moreover classification of patterns of notes and chords will be explored.

References

1. Kirke A., Miranda E.: *"An Overview of Computer Systems for Expressive Music Performance"*, Guide to Computing for Expressive Music Performance, Springer, 2013, pp. 1-47.
2. Rodà, A., Canazza, S., De Poli, G., *"Clustering affective qualities of classical music: beyond the valence-arousal plane"*, IEEE Trans. on Affective Computing, (preprint)
3. Rodà, A., Canazza, S.: *"Virtual performance, Actual gesture: A Web 2.0 system for expressive performance of music contents"*, IEEE Conference on Human System Interaction 09, 2009/5, Catania, Italy, pp. 474-480.
4. Tzanetakis, G., Cook, P.: *"Musical genre classification of audio signals"*, IEEE Tr. on Speech and Audio Processing, 2002, 10(5), 293-302
5. Danisman, T., Alpkocak, A.: *"Emotion Classification of Audio Signals Using Ensemble of Support Vector Machines"*, Perception in Multimodal Dialogue Systems, Springer Berlin Heidelberg, Berlin, 2008.
6. Zwicker, E., Fastl, H.: *"Psychoacoustics: Facts and models"*, 2nd Edition, Springer-Verlag, Berlin, 1999.
7. Lartillot, O., Toiviainen, P.: *"MIR in Matlab (II): A Toolbox for Musical Feature Extraction From Audio"*, International Conference on Music Information Retrieval, Vienna, 2007.
8. Hall, M.: *"Correlation-based Feature Selection for Machine Learning"*, PhD Thesis, Departement of Computer Science, Waikato University, New Zeland, 1999.
9. Ballabio, D., Consonni, V.: *"Classification tools in chemistry. Part 1: Linear models. PLS-DA."*, Analytical Methods, 2013.
10. Dubnov, S.: *"Generalization of Spectral Flatness Measure for Non-Gaussian Linear Processes"*, Signal Processing Letters, 11(8), 2004, pp. 698-701.
11. Pampalk, E., Rauber, A., Merkl, D.: *"Content-based organization and visualization of music archives"*, ACM international conference on Multimedia, 2002/12, pp. 570-579.
12. Fastl, H.: *"Psychoacoustic model of the fluctuation strength"*, The Psychoacoustics of Sound-Quality Evaluation, Acta Acustica united with Acustica, 83(5), 1997/10, pp. 754-764(11).