# EXPRESSIVENESS DETECTION OF MUSIC PERFORMANCES IN THE KINEMATICS ENERGY SPACE

*Luca Mion and Giovanni De Poli*
University of Padua
Department of Information Engineering
Center of Computational Sonology

**ABSTRACT**

An experiment for the automatic detection of expressiveness in music performances using a perceptive based auditory models is presented. We recognize the intentions with reference to the Kinematics Energy expressive space. Audio features have been firstly extracted using a perception-based analysis, then we have made several analyses on timing and spectral features over overlapping sliding windows, estimating average and variance for each one of the features. Using a naive Bayesian classifier we investigated which features are most relevant for expression detection. This experiment also yielded interesting contributions for tuning the Kinematics Energy space with new features.

## 1. INTRODUCTION

In verbal and non-verbal communication, two channels have been distinguished: one transmits explicit messages, which is represented by the text or the musical score; the other transmits implicit messages about the expressive intentions of speaker or performer. Both research and technology have invested enormous efforts in understanding the first, explicit channel, but the second is not understood just as well. Understanding the expressive intentions is one of the key tasks associated with the second, implicit channel. Most music performances would involve *expressive intention* from the performers side regarding what the music should express to the listeners. Various studies on performance led to suggesting models that could render synthesized music less monotonous and mechanical. A small number of attempts have been made to go beyond this, in order to identify an expressive model that could render different expressive intentions of a human performer [5].

Recently, studies started being developed in order to recognize the expressive and emotional intention in music performances. In particular, Dannenberg [4] proposed a style classifier for interactive performance systems, employing a machine learning approach. The features he used to classify are simple parameters that can be extracted from trumpet performances played by one performer and recorded as MIDI data. The classified styles consist of a range of performance intentions: frantic, lyrical, pointillistic, syncopated, high, low, quote and blues.

Friberg [6] developed a system that combines a low-level cue extraction algorithm with a listener model to predict what emotion the performer is trying to convey in his or her performance. One or several types of "listener panels" can be stored as models which are used to simulate judgments of new performances based on results from previous listening experiments. From audio input data the following parameters are computed for each tone: inter-onset duration, relative articulation, peak sound level, attack velocity, and spectral ratio. The spectral ratio is simply defined as the difference in sound level below and above 1000 Hz. The acoustic cues are obtained by computing running averages and standard deviations of the parameters. An estimation of the strength of each intended emotion (happy, angry, sad) is obtained from a regression equation taking the standardized cue values as input variables.

Mion [11] employed Bayesian Networks for the recognition of expressive content in musical improvisations. From MIDI piano improvisations, the extracted features are: note number, intensity, articulation, inter-onset duration, features' pattern. The following expressive intentions described by sensorial adjectives are recognized: slanted, heavy, hopping, vacuous, bold, hollow, fluid, tender. The intentions are derived from the Laban's basic effort theory of expressive movement.

There are various ways of expression categorization. In this work, we want to classify the expressive content of audio musical performances using a machine learning approach. We refer to the expressive intention categorization induced by the Kinematics Energy space, that proved to be relevant for sensorial intentional adjectives description [2]. Thus, the expressive intentions are distinguished into four main categories: High Energy ($HE$), Low Energy ($LE$), High Kinematics ($HK$), Low Kinematics ($LK$).

## 2. THE KINEMATICS ENERGY SPACE

In order to understand the sensory expressive intentions of a human performer, measurements of perceptive nature can be used. Performances played according to different expressive intentions are evaluated in listening experiments. Then, a low dimensional structure is derived by multivariate analysis of response data. In [2], using sensorial adjectives to describe expressive intentions, the anal-

ysis of results lead to two quite distinct expressive factors. From a musical point of view, the first factor sets rapid Tempo against slow. The second factor is mainly correlated to Energy-related parameters as Intensity. The listeners' successful identification of the player's intention is demonstrated by the fact that each performance is placed near semantically related adjectives. The positions of performances induce the associations "light vs. heavy" and "soft vs. hard" with the axes. An interpretation can be applied to this space. The first factor, associated to "light vs. heavy" is related to Kinematics; the second factor is associated to "soft vs. hard" and related to Energy. These results were confirmed in other experiments and lead us to conclude that performances played according to sensorial adjectives can be well represented in the *Kinematics Energy space*. Moreover, the space proved to be effective in interactive control of expressivity in synthetic music performance [1, 3].

## 3. EXPRESSIVE INTENTIONS CLASSIFICATION

The cues we decided to extract were found to be important for discriminating different emotions in previous listening experiments [7], and they have been used to classify the content in musical performances [4, 6]. A large variety of methods for features detection appears in the literature, especially for the onset-offset detection. In this experiment we based our cues detection on perception-based analyses. Moreover, we decided not to take into account the score in the selection of the cues to obtain more generality and robustness. Thus, in a certain sense, the pieces are classified as they would be improvisations of any kind of musical audio. The audio features are considered in terms of running average and variance within overlapping windows. Derivatives are taken into account to detect the onset and offset intervals.

The main motivation for using perceptual-based music analysis is that much of the musical audio productions have no score representation. Also, even in musical pieces played with score it has just a small role in the musical communication itself, and most of the real musical meaning is described by audio features as perceived by the human ear. We used for the perception-based analysis developed by IPEM, University of Ghent [10]. Their motivation for developing this toolbox focuses on a fully integrated approach to physical world, perception, cognition and processing of expressive communication. We believe that such integrated approach is very fruitful for significant understanding of musical features.

### 3.1. Features Extraction

In our experiment we extract the following set of features: roughness, cochlear filter-bank centroid, peak sound level, sound level range, inter onset interval, duration, articulation, number of notes, residual spectral ratios. In particular, using this toolbox we obtain the loudness, roughness, and the cochlear filter-bank centroid. The *loudness* ($A$) extractor is based on a low-pass filter on the amplitude in each auditory filter band, and then summed over all bands. The *roughness* ($R$) is the amplitude after a high-pass filter on the filter-bank output amplitude. Roughness is considered to be a sensory process highly related to texture perception. The estimation should be considered an inference, but the module offers more than just an inference. The calculation method of this module is based on Leman's Synchronization Index Model [9], where roughness is defined as the energy provided by the neuronal synchronization to relevant beating frequencies in the auditory channels. This model is based on phase locking to frequencies that are present in the neural patterns. It assumes that neurons somehow extract the energy of the beating frequencies and form internal images on which the inference is based. The concept of synchronization index refers to the amount of neural activation that is synchronized to the timing of the amplitudes of the beating frequencies in the stimulus.

The computation of the cochlear filter-bank *centroid* ($C$) takes into account the non-linear distribution of the cochlear filter-bank: $C = \sum_i (f_i A_i) / \sum A_i$ , where $A_i$ and $f_i$ are respectively the loudness and central frequency of the $i$-th band. The *peak sound level* $PSL = \max_i(A_i)$ and the *sound level range* $SLR = \max_i(A_i) - \min_i(A_i)$ are computed directly from the loudness profile. For the automatic detection of onset and offset instants we analyze the derivative of the loudness envelope of the pieces as in [8], setting the thresholds properly for the detection of the offset instants. Using the extracted onset and offset times, we compute the following parameters for each tone: *duration* ($D$), *inter-onset interval* ($IOI$), *articulation* ($L$) defined as $L = D/IOI$, and the *number of notes* ($N$) within the time window. Previous works demonstrated that articulation and number of notes are important parameters which characterize how expressive intentions are conveyed in music performances [3, 6]. The last set of cues describes the stochastic residual of the audio signal, obtained by removing the deterministic sinusoidal components, is useful for our sound classification. Using the Spectral Modeling Synthesis [12] we extract the residual over different frequency regions as a time-varying filtered white noise component, and we analyzed it by using the auditory toolbox. We characterize such residual by computing opportune spectral ratios: within each region, we compute the ratio of the loudness of the filter-bank output in the region over the global loudness. We experimented two kind of parameters. The first is obtained by dividing the frequency range into two regions [6], placed below and above a frequency close to 1000Hz. Taking into account the frequencies above this, we obtain one feature called $SRa$. The second parameter is given by a separation into three regions: below 534 Hz, from 534 Hz to 1805 Hz, and over 1805 Hz (this division yields three features, called $SRl$, $SRm$ and $SRh$). These bands are characteristic of the sound production mechanism, and they are derived from the actual frequency separations of the the cochlear filter-bank.

| *features* | D | IOI | L | R | N | SLR | PSL | C | SRl | SRm | SRh | SRa |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $F_{3,196}$ | 2.59 | 1.99 | 0.67 | 65.04 | 5.51 | 0.92 | 28.72 | 5.55 | 1.82 | 2.39 | 4.25 | 5.27 |
| $p$ | - | - | - | $***$ | $**$ | - | $***$ | $**$ | - | - | $**$ | $**$ |

**Table 1**. ANOVA test on features results: Codes: "$-$" non significant; $***$ $p < 0,001$; $**$ $p < 0,01$; $*$ $p < 0,05$. The features that will be used in the second part of the experiment are: roughness, peak sound level, centroid, number of notes, spectral ratio above 1000 Hz, and spectral ratio above 1805 Hz.

| *features* | R | N | C | PSL | SRh | SRa |
|---|---|---|---|---|---|---|
| *vector 1* | 0.15 | -0.13 | 0.33 | 0.5 | 0.54 | 0.55 |
| *vector 2* | -0.64 | 0.38 | -0.54 | 0.31 | 0.16 | 0.14 |

**Table 2**. Weights of the features in the most relevant PCA basis vectors. In vector 1 the most relevant features are peak-sound-level and spectral ratios, in vector 2 features are roughness and centroid.

### 3.2. Classification method

Bayesian classifiers have been used in previous works for the analysis of expressive content in musical audio [4, 11]. As classifier, the naive Bayesian classifier assumes that the features are uncorrelated and normally distributed. Given a vector of features, $F$, we would like to know which classification $E$ is most likely. Using the listed assumptions and Bayes' Theorem, it can be shown that the most likely class is the one whose mean feature vector has the least normalized distance to $F$. The normalized distance is the Euclidean distance after scaling each dimension by its standard deviation:

$$\Delta_E = \sqrt{\sum_{i=1}^{n} \left( \frac{F_i - \mu_{E,i}}{\sigma_{E,i}} \right)^2} \qquad (1)$$

where $E$ indexes classes (expressive intention), $i$ indexes audio features, $\mu_{E,i}$ are means, and $\sigma_{E,i}$ are standard deviations.

## 4. THE EXPERIMENT

### 4.1. Categories

In the Kinematics Energy space we can distinguish four main categories situated at the opposite sides of the axis: High Energy ($HE$), Low Energy ($LE$), High Kinematics ($HK$), Low Kinematics ($LK$). We selected the performances whose projection in this space was closer to the categories we want to classify. Thus we have these correspondences: hard $\rightarrow$ High Energy; soft $\rightarrow$ Low Energy; light $\rightarrow$ High Kinematics; heavy $\rightarrow$ Low Kinematics.

### 4.2. Data collection

Professional performers of various instruments were invited to play musical performances inspired by different expressive intentions; we selected the performances played according the following adjectives: light (L), heavy (He), soft (S) and hard (Ha) and as such, each one of the adjectives had its opposite (soft vs. hard) in order to deliberately induce contrasting performances on the part of the musician. Musical performances were recorded in monophonic digital form at 16 bits and 44100 Hz at the CSC, Padua University. The pieces are the followings: Arcangelo Corelli's Sonata in A major for violin; Mozart's K622 Concert for Clarinet played by clarinet, by violin and by voice and an excerpt from Francesco De Gregori's Alice for voice. We had 5 examples each of 4 adjectives, resulting in 20 performances. Some of them were previously used in perceptual experiments and factor analysis to understand how the listeners organized the pieces in their own minds, and thus to learn how many dimensions could actually be determined (see [2]).

### 4.3. Learning the classifier

We measured performance by training the classifier on 90% randomly chosen data, and then classifying the remaining 10%. Audio features are extracted over overlapping windows, each one with 4s of duration and 0.5s overlapping. After sliding over the audio file, mean values and variance for each feature are calculated. The size of each vector of features results equal to 24 (average and variance of 12 features).

### 4.4. Features selection and analysis

In order to distinguish the relevance of the features, ANOVA test have been made over different significance levels. Table 1 summarizes the results. The significant features are in order: roughness, peak sound level, centroid, number of notes, spectral ratio above 1000 Hz, and spectral ratio above 1805 Hz. These features will be used in the second part of the experiment.

In addiction, we sought to reduce the dimensionality of the feature space while simultaneously decorrelating the features. We used the principal component analysis (PCA) that performs a linear transformation of the original input in a new data-set. Each component represents the projection of the original data over a low dimensional

| | *No ANOVA* | *ANOVA* |
|---|---|---|
| *HARD (HE)* | 91.67 | 83.33 |
| *SOFT (LE)* | 60.42 | 72.92 |
| *HEAVY (LK)* | 80 | 70 |
| *LIGHT (HK)* | 15 | 45 |
| *Average* | 61.77 | 67.81 |

**Table 3**. Percentage of correct classifications before and after selecting features with ANOVA test.

orthogonal basis vector. In our case, we found two basis vectors with eigenvalue $>1$ that explain the 75% of the total variance (resp. 51.17% and 25.63%). The other basis vectors explain respectively 14.86%, 5.94%, 2.12% and 0.29% of variance. Table 2 shows the weights of the features in each of the two relevant basis vectors. Notice that in vector 1 the most relevant features are peak-sound-level and spectral ratios, and in vector 2 features are roughness and centroid. Figure 1 shows an example of projection into the 2-dimensional PCA space.

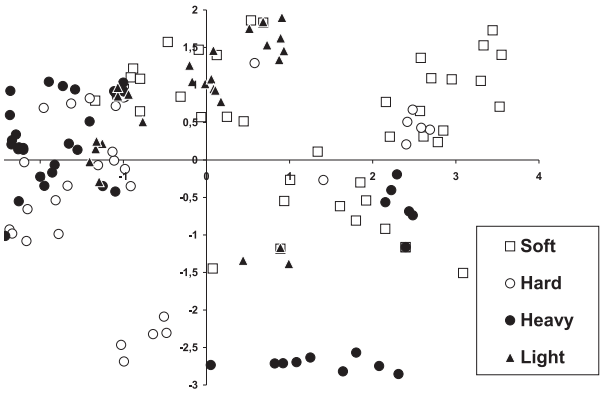


**Figure 1**. Example of projection into the 2-dimensional PCA space.

## 5. RESULTS AND DISCUSSION

We measured performance by randomly dividing the data in ten parts. In turn, we used one part as test set, and the remaining data as training set. We computed the total value of correct answers over the ten tests. We evaluated the classifier performance in two ways: by comparing its behavior when using all the features and only with the 6 selected by the ANOVA test. Table 3 shows the percentage of correct classification over the ten tests. As you can notice *hard* and *heavy* are better recognized without the feature selection, while *soft* and *light* give better results in the second case. Moreover, the average value of percentage of correct classifications increases when using the selected features.

We also tested the classifier after converting the original four-class problem into $n(n-1)/2$ two-class discrimination problems, one for each possible pair of intentions. For each intention pair A-B, the performances of the two

| | *No ANOVA* | *ANOVA* |
|---|---|---|
| $HARD - SOFT$ | 83.33 | 82.14 |
| $HARD - LIGHT$ | 77.63 | 84.21 |
| $HARD - HEAVY$ | 81.58 | 72.37 |
| $SOFT - LIGHT$ | 71.59 | 84.09 |
| $SOFT - HEAVY$ | 78.41 | 87.50 |
| $HEAVY - LIGHT$ | 66.25 | 91.25 |
| *Average* | 76.47 | 83.59 |

**Table 4**. Pair-wise recognition results: percentage of correct classifications before and after selecting features with ANOVA test. The baseline accuracy (the accuracy one would achieve by random guessing) is 50%.

respective intentions of the selected training pieces were used for learning, and the task was to identify the correct intention in a new test piece, where only recordings with intentions A and B were used for testing. Table 4 shows the results after testing the classifier with 2 classes. The average recognition rate over all experiments where a given expressive intention was involved gives a rough measure of the recognizability of that intention. Computing this over all classifiers we get the following ranking. Using all the features (no ANOVA test): *hard* (98.15%), *heavy* (89.17%), *soft* (75.69%), *light* (45.00%).
Using only the features selected with the ANOVA test: *hard* (91.67%) *heavy* (88.33%) *soft* (82.64%) and *light* (73.33%). Notice that *hard* gives impressive values of correct classification, while *light* is still the most difficult intention to be detected. Moreover, percentage of correct classifications increases after applying the feature selection.

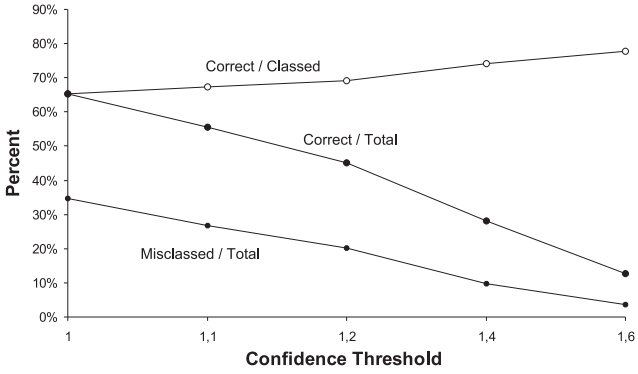


**Figure 2**. The number of misclassified examples decreases faster than total number of classified examples, as the confidence threshold increases. The ratio of correctly classified to all classified examples increases.

We also evaluated the problem of "false positives" (misclassifications which erroneously imply the performer is playing a particular style). An experiment with the classifier suggest that simple confidence measures can dramatically reduce false positives. Recall that this classifier makes decisions based on normalized distances from

means. If the distance to the means of two classes are nearly equal, our confidence in the decision should be reduced. Therefore, we simply reject classifications when the least distance is not less than a given fraction of the next-to-least distance. Figure 2 illustrates the reduction of false positives using this technique. As the confidence threshold increases, the total number of classified examples decreases, but the number of misclassified examples decreases faster, so the ratio of correctly classified to all classified examples increases.

## 6. CONCLUSIONS AND ACKNOWLEDGEMENTS

A classifier able to recognize the expressive intention in musical performances into the Kinematic Energy space has been presented. The more relevant features are roughness, peak sound level, centroid, number of notes, spectral ratio above 1000 Hz, and spectral ratio above 1805 Hz. The PCA analysis showed two rather uncorrelated group of features: peak-sound-level and spectral ratios Vs. roughness and centroid.

This research was supported by the European Network of Excellence “Enactive Interfaces”. We thank G. D’Incà for developing part of the experimental work.

## 7. REFERENCES

[1] Canazza, S., De Poli, G., Drioli, C., Rodà, A., Vidolin, A. (2000). Audio morphing different expressive intentions for Multimedia Systems. IEEE Multimedia, July-September, 7(3), pp. 79-83.

[2] Canazza, S., De Poli, G., Rodà, A., Vidolin, A. (2003). An abstract control space for communication of sensory expressive intentions in music performance, Journal of the New Music Research, 32(3), pp. 281–294.

[3] Canazza, S., De Poli, G., Drioli, C, Rodà A., Vidolin, A. (2004). Modeling and Control of Expressiveness in Music Performance. (invited paper), The Proceedings of the IEEE vol. 92(4), pp. 286–701.

[4] Dannenberg, R., Thom, B., Watson, D. (1997). A Machine Learning Approach to Musical Style Recognition, in Proceedings of the International Computer Music Conference, San Francisco, USA, pp. 344–347.

[5] De Poli, G. (2003). Expressiveness in music performance: analysis and modeling, in Proceedings of the SMAC03 Stockholm Music Acoustics Conference, Stockholm, Sweden, pp. 17–20.

[6] Friberg, A., Schoonderwaldt, E., Juslin, P., Bresin, R. (2002). Automatic Real-Time Extraction of Musical Expression, in Proceedings of the International Computer Music Conference, Göteborg, Sweden, pp. 365–367.

[7] Juslin, P. N. (2001). Communicating emotion in music performance: A review and a theoretical framework. In P. N. Juslin, & J. A. Sloboda (Eds.), Music and emotion: Theory and research (pp. 305–333). New York: Oxford University Press.

[8] Klapuri, A. (1999). Sound onset detection by applying psychoacoustic knowledge. in Proceedings of the IEEE Int. Conf. Acoust., Speech, and Signal Proc. (ICASSP), volume 6, pp. 3089–3092.

[9] Leman, M. (2000). Visualization and calculation of roughness of acoustical musical signals using the synchronization index model (SIM) - in Proceedings of the of the COST G-6 Conference on Digital Audio Effects (DAFX-00), Verona, Italy, pp. 125–130.

[10] Leman, M., Lesaffre, M., Tanghe, K. (2001). An introduction to the IPEM Toolbox for Perception- Based Music Analysis, Mikropolyphonie - The Online Contemporary Music Journal, Volume 7.

[11] Mion, L. (2003). Application of Bayesian Networks to automatic recognition of expressive content of piano improvisations, in Proceedings of the SMAC03 Stockholm Music Acoustics Conference, Stockholm, Sweden, pp. 557–560.

[12] Serra, X. (1997). Musical Sound Modeling with Sinusoids plus Noise. In C. Roads, S. T. Pope, A. Piccialli, & G. D. Poli (Eds.), Musical Signal Processing, pp. 91–122. Lisse: Swets & Zeitlinger.