Towards E-Motion Based Music Retrieval

A study of Affective Gesture Recognition

Denis Amelynck, Maarten Grachten, Leon Van Noorden and Marc Leman

Abstract—The widespread availability of digitised music collections and mobile music players have enabled us to listen to music during many of our daily activities, such as exercise, commuting, relaxation, and many people enjoy that opportunity. A practical problem that comes along with the wish to listen to music is that of music retrieval, the selection of desired music from a music collection. In this paper we propose a new approach to facilitate music retrieval. Modern smart phones are commonly used as music players, and are already equipped with inertial sensors that are suitable for obtaining motion information. In the proposed approach, emotion is derived automatically from arm gestures, and is used to query a music collection. We set-up predictive models for valence and arousal from empirical data, gathered in an experimental setup where inertial data recorded from arm movements is coupled to musical emotion. Part of the experiment is a preliminary study confirming that human subjects are generally capable of recognising affect from arm gestures. Model validation in the main study confirmed the predictive capabilities of the models.

Index Terms—Affect detection, Expressive gestures, Music retrieval, Human computer interfaces

This is a close to final preprint version of the paper that appeared as: D. Amelynck, M. Grachten, L. Van Noorden, M. Leman (2012). Towards E-Motion Based Music Retrieval: A study of Affective Gesture Recognition. IEEE Transactions on Affective Computing. Vol. 3 (2), pp. 250-259.

1 Introduction

The widespread availability of digitised music collections and mobile music players has enabled us to listen to music during many of our daily activities, such as exercise, commuting, relaxation, and many people enjoy that opportunity. A practical problem people face when they want to listen to music is the selection of desired music from a music collection. The bibliographic, text-based interface to music-collections that is prevalent in mobile music players as we know them today, is not an optimal solution to this problem for two major reasons. Firstly, bibliographic indices to music, such as artist and album names, are only useful when the user is familiar with the music she is looking for. Secondly, a text-based visual interface is often impractical to handle on the small screens of mobile devices. It requires full attention and fine motor control of the user, which can be cumbersome and even dangerous in everyday life situations.

The basic tenet of affective computing, as stated by Calvo and D'Mello [1], is that automatically recognising and responding to a user's affective states during interactions with a computer can enhance the quality of the interaction, thereby making a computer interface more usable, enjoyable, and effective. This may be particularly true in the context of interfaces for music players, since music and affect are strongly related. Not only is it natural for people to describe music in affective terms; studies have also suggested that the most common goal of musical experiences and in particular for music listening is to influence emotions: People use music to change emotions, to enjoy or comfort themselves, and to relieve stress [2].

Of the various forms an affection-based interface to music players might take, motion-driven approaches seem especially promising. One reason for this is that there is ample evidence that corporal gestures are a very effective way of communicating affect among people (see section 2). Another, more pragmatic reason is that many smart phones that people use as music players nowadays, are equipped with inertial sensors that make it possible to capture movements of the user.

In the envisioned interface, users can search through music collections by the affective character of the music, where the character of the desired music is expressed through corporal gestures. In this way we aim to implement the conceptual framework of *embodied music cognition* and *mediation technology* [3], and reduce the gap between the fundamentally corporeal aspects of music and the disembodied, bibliographical way of interacting with music collections that is common practice today.

The work presented in this paper is intended as the foundation for a such a motion based affective user interface for music retrieval. We present a linear regression model that predicts the affective character of music, based on the arm movements of people expressing that character. The model is derived from empirical data that is gathered from an experiment, as described in section 3. A motion based interface can employ this model to interpret arm movements of the user in terms of affective character, so that the movement can be matched to the affective character of music. For this, it is also necessary to have a music collection that is annotated in affective terms. Automated affective description of music is beyond the scope of this paper, but this is an active field of research in its own right (see e.g. [4], [5], [6]).

It is commonly acknowledged that the notion of *affect*, and subsumed notions such as *emotion*, and *mood*, are notoriously intricate. The study of affect in combination with music is by no means less intricate and controversial. First of all, some studies question whether the emotions music evokes should be considered as basic emotions [7], whereas others consider this view mislead [2]. There is also disagreement on the question whether music is more adequately described as inducing *mood* (a relatively vague and long-lasting form of affect), or *emotion* (more instantaneous and focused forms of affect) [2]. Furthermore, perception of music has been shown to influence neuroaffective processes [8]. Other studies show however, that the strength of emotions induced by music is relatively low compared to emotions induced by personal memories [9].

In the light of these controversies, it is useful to clarify our use of notions related to affect in this paper, and the corresponding assumptions we make. To begin with, we focus on the affective character of the music, as perceived by the listener. We will also refer to these as the *emotional* character of the music, because of the instanteous and concrete nature of the music. More specifically, we adhere to the *valence-arousal* model [10] to represent the emotional character of music, in line with other studies concerning emotion in music[4]. Although extended versions of this model have been proposed that include a third dimension representing *tension*, there is evidence that the three-dimensional model does not account better for perceived emotions in music [11].

Although we focus on emotions *perceived* by listeners, there is a possible confounding effect of emotions that are *induced* by the music. We consider this risk unproblematic however, for several reasons. Firstly, the setup of the experiment, in particular the short duration and rapid succession of musical fragments (see section 3) do not foster a deep emotional involvement of the subject with the music. Secondly, as far as induced emotions may affect the arm movement of listeners, the we assume that the induced emotion is most likely to correspond to the perceived emotion of the music. In other words, although it is possible that a musical fragment of a particular emotional character produces a non-congruent emotional state in a subject due to a strong memory association, we believe such cases to be too infrequent to be of significance.

The next section contains a brief review related work concerning both human and automated affect detection from human movement. In section 3, we describe the experiments carried out to gather the movement data for affect recognition. In section 4, we present the data-modelling process, and the results of the model validation. A discussion of the results is presented in section 5, followed by conclusions and directions for future work, in section 6.

2 Related Work

A natural way for humans to express affect is by corporal gestures [12], [13]. Communication of affect through gestures (both static and dynamic) is arguably an intrinsic part of social behaviour. This is reflected in numerous studies showing the capability of humans to recognise affect from the corporeal behaviour of others. To a large extent, the quality of movements seem to convey affective information. For example, to recognise emotion from gait, a small number of features describing joint angles and spatial trajectories is sufficient for humans to recognise emotions in animated avatars [14]. Furthermore, Atkinson et. al [15] show that even with very reduced visual representations of the body, such as point-light displays, recognition of emotion by human subjects is still possible (though to different degrees for different emotions). Point-light displays of arm movements of actors expressing affect in everyday movements, like drinking and knocking, also enable observers to recognise the expressed affects [16]. Pollick et al. also found that movement features such as average velocity, peak velocity, acceleration, and jerk were all correlated with the level of activation.

Music-related body movement, such as that of dancers, and the performing musicians, also conveys affect. Brownlow and Dixon [17] state that observers easily can judge happy dances as happier and stronger than sad dances. Again, the observers in their experiment based their judgement solely upon point-light displays of dance, thus excluding recognition of affect by facial expression or other cues. Successful *automatic* recognition of emotions of dance movements has been reported [18]. Vines and Wanderley [19] analysed gestures from professional clarinet performers. They confirm that the visual component (body movement) of the performance carries much of the same structural information as the audio. In some conditions, removing the visual component decreases the judgement of tension (emotion).

These studies strengthen the view that affect can be effectively communicated through human body movement, and therefore, that automatic affect recognition from human motion, even if it is a challenging problem (see [1] for a survey of current research), is not unfeasible.

3 Experimental Set-Up

An experiment was carried out with the goal of building a data set of arm movements expressing the affective character of different pieces of music. The design of the experiment is oriented towards the use case of gesture based music retrieval in mobile devices, in the sense that arm motion is captured using a wireless handheld device equipped with 3D inertial sensors, comparable to the motion-capture technology available in smart phones.

The following setup was designed to link movements to affective descriptions of music: Participants were asked to listen to a musical fragment, and to describe

the emotional character of the music, in terms of valence and arousal. Then, they were asked to listen to the music again, and simultaneously express the emotional character of the music as clearly as possible through the movement of their arm (either the left or right arm, depending on preference). The movement of the arm was observed by three other participants, who had to guess the emotional character of the music being expressed, judging only on the arm movement. One reason for including observing participants in the setup is to encourage the observed participants (called *performers* henceforth) to *communicate* the intended emotion through the movement, rather than making just any movements associated with the music. A second reason is that the degree of agreement between the intended emotion and the emotion recognized by observers serves as an indicator of how clearly the intended emotion is expressed by the movement.

The rest of this section describes the experimental setup in more detail.

Participants

In total 32 persons participated in the experiment. Of these, five groups of four persons participated in the main part of the experiment. The remaining 12 persons participated individually. Their responses were used to validate the model derived from the data obtained in the main experiment, as described below.

Stimuli

The musical material was selected from a pre-existing library of 30 second musical excerpts [20]. In total 24 musical fragments (table 1) were selected divided over four similar sets of six fragments. The sets are separated by a double line in table 1. Similarity of the sets was controlled after the experiment. From the results shown in table 2 it can be verified that the sets were indeed homogeneous and that they spanned the whole valence/arousal range. The arousal and valence scores mentioned in both tables are average appraisal scores collected from the performers.

Each of the 24 musical fragments was rated once in each of the five participant group, resulting in 120 ratings in total, and five ratings per fragment.

Material

For capturing arm movement, a Wii Remote was used. This is a wireless, handheld device commercially available as a gaming interface from Nintendo. It transmits 3D inertial sensor data in realtime via Bluetooth at a sample rate of 100Hz. Musical material was played to the participants from a computer, using wired headphones. Visual recordings of arm movements were were transmitted in real-time, using a digital video camera. Judgments of emotional character were obtained from participants through printed questionaires.

TABLE 1
Musical Fragments and their average Arousal/Valence appraisal scores as given by the Performers

Performer - Title	Arousal	Valence
New Zealand Symphony Orchestra - Many Meetings	1.4	3.6
Midori/Berliner Philharmoniker / Claudio Abbado - Canzonetta. Andante (Concerto for Violin and Orchestra in D major op. 35)	2.6	2.0
Tam Echo Tam - One Step	4.2	4.4
L' Arpeggiata / Christina Pluhar - Ah, vita bella	1.6	1.8
Blur - Song 2	4.8	4.4
DJ Tiësto - Traffic	4.6	3.4
Metallica - St. Anger	4.0	2.8
De Nieuwe Snaar - Achterbank	3.8	5.0
Enya - Orinoco Flow	2.2	3.4
The Cleveland Orchestra/Pierre Boulez - Le Sacre du Printemps	5	2.4
Alberto Gilberto - The girl from Ipanema	1.6	3.8
New Philharmonia Orchestra/Sir John Barbirolli - Adagietto, Sehr langsam Symphony No. 5 in C sharp minor	1.4	1.8
Esa-Pekka Salonen / Philharmonia Orchestra - Car Horn Prelude (Le Grand Macabre)	4.0	1.4
Bob Marley - Corner stone	3.0	5.0
Beyoncé - Naughty Girl	3.8	3.3
Astor Piazzolla - Oblivion	1.2	1.8
Metallica - My World	4.8	1.6
Manu Chao - Mr. Bobby	2.4	4.6
Novastar - Never back down	2.6	3.4
David Hill / Westminster Cathedral Choir - Motectum (Requiem, Officium defunctorum)	1.2	1.2
Usher - Usher	4.6	4.2
Vladimir Ashkenazy - Nocturne in F major op.15 No.1	1.4	2.8
St. Germain - Land of ...	3.8	4.4
Collegium Vocale & La Chapelle Royale/Orchestre des Champs Elysées/Philippe Herreweghe - Dies Irae (Requiem KV 626)	4.4	1.6

Procedure

Within a group of four participants, one set of six musical

TABLE 2
Sets of fragments and their statistical Arousal/Valence dispersion

Set	Arousal : mean ± stdev	Valence : mean ± stdev
1	3.2 ± 1.5	3.3 ± 1.1
2	3.0 ±1.5	3.2 ± 1.1
3	3.2 ± 1.3	3.0 ± 1.6
4	3.0 ± 1.5	2.9 ± 1.3

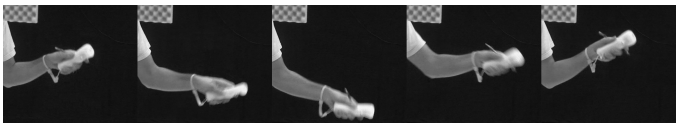

Fig. 1. Video capture of performance as monitored by the observers

fragments was assigned to each participant, such that each fragment was uniquely assigned to a participant within the group. Every participant was asked in turn to listen to each of the fragments assigned to him/her, express this character by arm movements while listening to the fragment again, and judge its emotional character (dealing with one fragment at a time). The arm movements were made while holding the Wii Remote, in front of a camera that was positioned in such a way that only the arm was monitored, as illustrated in figure 1. A small shield was used to prevent the participants' faces from occasionally appearing on the screen.

The instructions for the performer were as follows:

1) Listen to a short musical fragment
2) While listening a second time, express the emotional character of the music as accurately as possible through movements of your arm
3) Rate the emotional character of the music on the provided form

The emotional character of musical fragments was rated in terms of valence and arousal on a 1 to 5 scale. Rather than using the terms valence and arousal directly, the semantics of the two scales was indicated by labeling the extremes of the scales with corresponding adjectives. The adjectives were given in Dutch: *kalm*, *vermoeid* (calm, tired) versus *energetisch*, *gespannen* (energetic,tense) to label the low and high extremes of arousal respectively, and *droevig*, *kwaad* (sad, angry) versus *blij*, *tevreden* (happy, pleased) for valence. It was explained to the participants that a single matching adjective was sufficient to rate a musical fragment correspondingly. For example, it is sufficient for either *kalm* or *vermoeid* to apply, in order to choose that rating.

The three other participants, referred to as *observers*, watched the arm movements of the performer via a monitor in a separate space (figure 1). They did not hear the music fragments the performer heared. The observers were instructed as follows:

1) Monitor the (arm movement of the) performer.
2) Rate the emotions expressed by the arm movement.
3) Describe any cues in the motion that helped you to make your rating (free text)

The remaining 12 subjects participated in the role of performer, as described above. Each subject was assigned again a group of six fragments. This time, no observers were present. The arm movements and the subject's rating of the emotional character of the fragments was recorded as before. The data obtained in this way is used for validation, as described in section 4.

4 RESULTS

The data obtained in the experiment was used to create a regression model for predicting the expressed arousal and valence from arm movements as captured by the 3D inertial sensors of the Wii remote. We aim at a general data model that can easily be ported to other devices, possibly using other sensing technologies. Therefore only predictor variables with a relatively straightforward relationship to movement were considered. This preference of general validity over a fine-tuned model leads us to consider only models with at most five predictor variables.

The Wii Remote measures the acceleration of the device in the direction of three perpendicular axes, relative to the device. Since the way of holding the device was not constrained, similar arm movements may lead to different data, as an effect of the Wii Remote being held in different ways. To compensate for this, the acceleration data for each fragment was projected onto its three principal components by performing a principal component analysis (PCA). In figure 2 we show acceleration data collected from two different subjects performing on the same musical fragment. In this figure an acceleration value of 25 corresponds with 1G (gravity). From this figure it is very difficult to see similarities between the two performances. When the data is translated and rotated to the PCA-axes, the similarity between these two performances becomes apparent (Figure 3). Apart from making data from different subjects easier to compare, an advantage of the PCA transform is that it reveals the intrinsic dimensionality of the movement.

To determine a set of candidate features to compute from the accelerometer data, we made an inventory of the free text responses in which subjects reported useful cues for judging the emotional character of the movements. The cues can be roughly grouped into five complementary aspects of the movement:

1) Roughness: gracefulness, multiple short moves
2) Rhythm: tempo of the music
3) Speed: high speed, low speed, and acceleration
4) Size: large versus small movements
5) Location of the arm: high = happy

Ideally, each of the cue categories should be represented by at least one predictor variable.

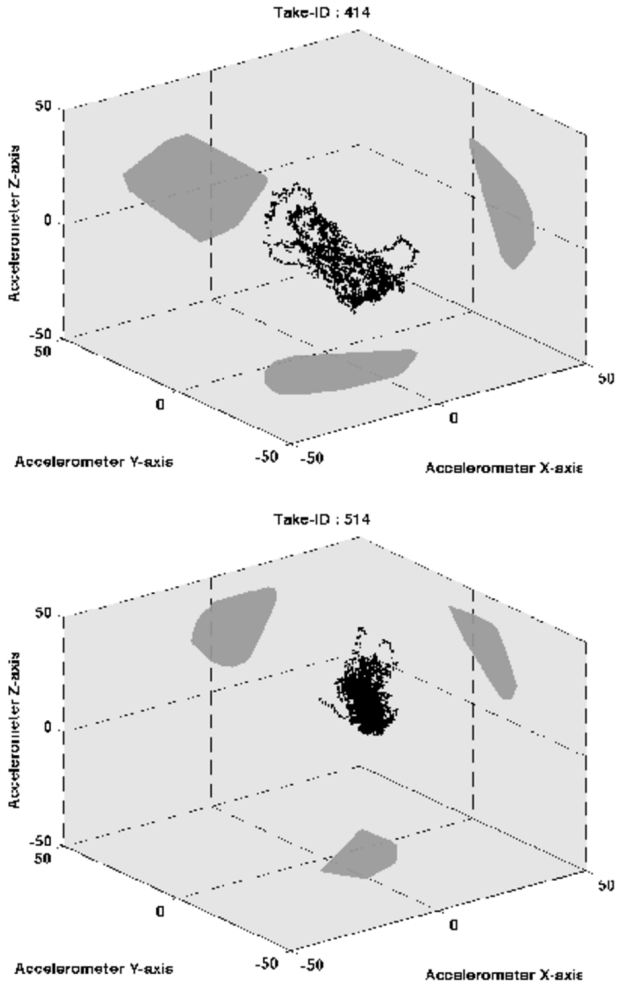

Fig. 2. Raw acceleration data from accelerometer : Performances of two subjects on the same musical fragment (set 1, fragment 4)

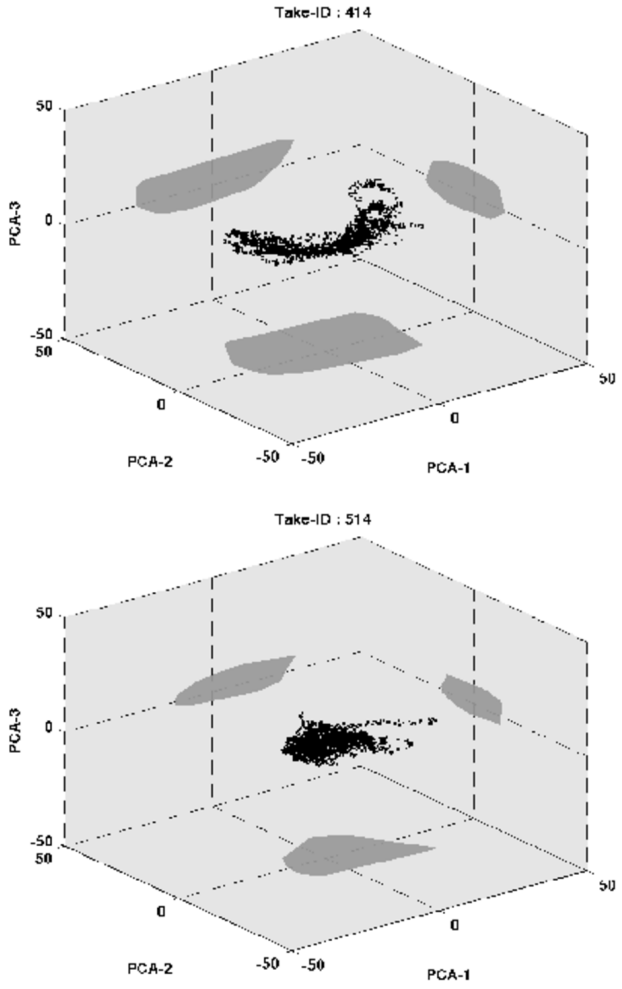

Fig. 3. Acceleration data translated and rotated to PCA-axes : For same performances as shown in Figure 2.

We extract various features that describe various properties of the distribution of acceleration data, in terms of geometry and density (Figure 3). The same was done for jerk (derivative of acceleration) and speed (integral of acceleration). Beside these spatial features we calculated also a number of temporal related features as peak-rates, zero crossings and randomness (runs test). Summarised the following features were extracted :

- Distribution properties for acceleration, jerk and speed along the 3 PCA axes : mean, range, standard deviation, kurtosis, skewness.
- Distribution properties for direction: circular standard deviation, concentration parameter kappa (Von Mises distribution),
- Volume (convhull) of acceleration point cloud.
- Time related variables: speed peak rate, zero crossing, randomness (runs test).

These features can be linked to the cue categories identified before, with the exception of location cues[1].

- Roughness : all jerk related features
- Rhythm : time related variables
- Speed : Speed and acceleration features
- Size : Direction parameters , volume of acceleration point cloud.

To remove any transient effects due to subjects starting or stopping to move, the features are extracted after removing the first and last 5 seconds of each data stream. The remaining stream spanned 20 seconds (corresponding to 2000 samples).

The extracted features were correlated with valence and arousal. There are strong correlations between some features and arousal ($|r| > 0.6$) but in general weaker correlations with valence (all $|r| < 0.4$). Uncorrelated features ($|r| < 0.2$) were discarded from the analysis. Because some features were highly correlated ($|r| > 0.9$). is is necessar to take precautions against multicollinearity.

1. Although the position of the Wii Remote can in principle be estimated by assuming an initial position and tracking acceleration over time, this estimation is unusable in practice, due to cumulative estimation errors.

Using all features in a least squares estimate as a regression model would give two problems [21]. The first problem is prediction accuracy: the least squares estimates often have low bias but large variance. Prediction accuracy can however sometimes be improved by shrinking or setting some coefficients to zero. By doing so we sacrifice a little bit of bias to reduce the variance of the predicted values, and hence may improve the overall prediction accuracy. The second problem is interpretation. With a large number of predictors, we often would like to determine a smaller subset that exhibits the strongest effects.

Because of our restriction to models of maximally five predictor variables, the method of 'best subset selection' was used. By retaining a subset of the predictors and discarding the rest, subset selection produces a model that is interpretable but has possibly lower prediction error than the full model. Best subset selection also gives a hard threshold on how many parameters to keep. Other shrinkage methods (like lasso or ridge regression) may give less variability but the number of parameters is soft-thresholded. The choice for least squares estimates and for best subset selection implied the need of a data validation step to check the generalisation capabilities of the calculated data model.

The regression model was eventually calculated via SPSS using the stepwise method. In this method, each variable is entered in sequence and its value assessed. If adding the variable contributes to the model then it is retained, but all other variables in the model are then retested to see if they are still contributing to the success of the model. If they no longer contribute significantly they are removed. In this way this method ensures that the smallest possible set of predictor variables is included in the model [22].

From here we will make a distinction between data modelling for arousal and data modelling for valence.

4.1 The Regression Model for Arousal

Set-Up of the Arousal Model

The regression model for Arousal was derived from 97 performances out of a total of 120 performances. 23 cases where performer and observers did not agree (Difference > 1) were discarded as mentioned before. In other words, in over 80% of the cases there was an agreement between performer and observer. We use SPSS and the stepwise method to enter the predictor variables, resulting in a model with three predictor variables. The variables (listed in their order of contribution importance) are the following (beta-values in table 3):

SpeedPeakrate: The number of local maxima and local minima for speed (integral of acceleration) divided by the time interval. Speed is calculated as an Euclidean norm.

KurtPCA1Speed: Kurtosis of the distribution of speed along the first (main) principal axis. This variable is negatively correlated with arousal. A high value means that intermediate values have become less likely and the central and extreme values have become more likely. In other words low arousal corresponds with long periods of low speed (central values) and other periods of high speed (extreme values). High arousal corresponds with periods of nearly constant speed or where the variation in speed is not huge (intermediate values).

PCA3Std: Standard deviation of the distribution of acceleration along the third principal axis. A small value indicates that the acceleration mainly happens in a plane formed by the two main principal axes.

TABLE 3
Regression analysis for Arousal.

Model	B	SE B	Beta	Sig.
(Constant)	2.096	4.410		0.000
Speedpeakrate	0.560	0.056	0.680	0.000
PCA3Std	0.043	0.014	0.213	0.002
kurtPCA1Speed	-0.585	0.149	-0.212	0.000

R = 0.854 (R^2 = 0.730 adjusted R^2 = 0.721)

The regression analysis did not reveal any outliers. (Criterion used: more than three standard deviations difference). There was however one influential case (group 5 subject 3 fragment 5) that ended up with a high value for DFFit (Difference in Fit). In order to preserve the general character of our model we removed this case and recalculated our regression model. An overview of the recalculated model can be found in table 3.

Because of the correlations between features, the following assumptions were checked:

1) Multicollinearity: VIF (variance inflation factor) average was close to 1 (1.4) and indicated absence of multicollinearity between the 3 predictor variables.
2) Normality for distributed errors: Probability plot for the residuals confirmed normality.

Validation of the Arousal model:

Model validation was done using the data gathered from the 12 individual subjects, who did not participate in the main part of the experiment.

Explanatory capabilities of the model: The variance of the validation data explained by the model : R = 0.754 (R^2 = 0.568). Compared to R = 0.854 (R^2 = 0.730) for the original data, this means a shrinkage with 16%.

The predictive capabilities are presented in Fig. 4. We see that the average prediction from the model deviates most for low arousal values. For other arousal values the prediction is in line with the target value.

Further investigation was done by having a closer look at the residuals. The results of this analysis are in table 4. The large residual value of -4.888 is due to an out of scale prediction of 9.888 for an arousal value of 5. Allowing

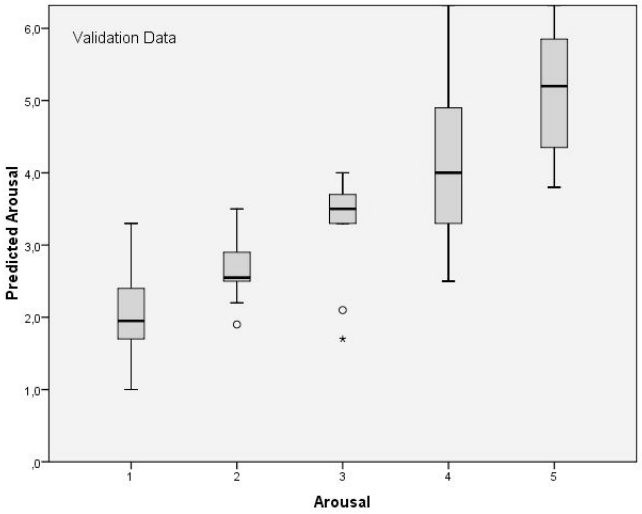

Fig. 4. Validation data. Y-axis predicted Arousal values. X-axis performer Arousal values.

TABLE 4
Residual statistics.

Residuals	Minimum	Maximum	Mean	StdDev
Model	-1.553	1.880	0	0.740
Validation	-4.888	1.506	-0.494	1.083

non-linearity by replacing values outside the boundaries of 1 and 5 with their respective boundary values, reduces the error, and the explanatory value of the model is increased (R^2 = 0.685). This leads to a reduction of only 4.5 % compared to the original data model.

4.2 The Regression Model for Valence

The regression model for valence was derived from 88 out of a total of 120 performances. 32 cases where performer and observers did not agree (Difference > 1) were discarded. We discarded considerable more samples (27 %) than for arousal (19 %).

Just like for the arousal model, we started with the stepwise method to add variables to the model. In a first step we obtained a model with five predictor variables. There were no outliers but however there was one influential case. Group 2 Subject 3 Fragment 5 ended up with a high value for DFFit. In order to preserve the general character of our model we removed that case and recalculated our model. Recalculation led to the removal of two more predictor variables that had no significant contribution. The final result was a model with again three predictor variables. The variables are hereafter listed in their order of contribution importance (beta-value), see table 5:

stdPCA1Jerk: Standard deviation of the derivative of the acceleration (jerk) of the first PCA component and this variable was negatively correlated with valence. If acceleration changes are nearing a random pattern (high standard deviation), this will result into a lower valence.

SpeedPeakrate: See subsection 4.1.

PCA2std: Standard deviation of the second principal component. A small value means that acceleration/movement happens mainly along the axis of the first principal component rather than in a plane. This variable was positively correlated with valence. In other words for low valence the movement is rather one dimensional (1D).

A complete overview of the model can be found in table 5.

TABLE 5
Regression analysis for Valence.

Model	B	SE B	Beta	Sig.
(Constant)	1.083	0.354		0.003
Speedpeakrate	0.751	0.116	1.018	0.000
StdPCA1Jerk	-0.238	0.046	-1.148	0.000
PCA2Std	0.058	0.021	0.461	0.008

R = 0.606 (R^2 = 0.367 adjusted R^2 = 0.344)

Assumptions checked:

1) Multicollinearity: (VIF variance inflation factor). The VIF never exceeded 10, but the average over all variables was well above 1 (4.4). So there might be some moderate bias in the model.
2) Normality for distributed errors: Probability plot indicates that the distribution is slightly skewed left.

The R^2 value of 0.367 for valence is relatively low compared to a value of 0.730 for arousal. A possible reason for this is that the model contains no predictor variable representing location, although observers reported this cue as indicative for valence. Even if location cannot be estimated directly from the accelerometer data, an estimate of position can be made indirectly: Because of the fact that the Wii Remote device is ergonomically designed for one particular way of grasping, in practice subjects held the Wii Remote all in the same position. Additionally, it is reasonable to assume that raising the arm leads to a different angle of the hand than lowering it, due to physiological constraints. By making these extra assumptions, location can be estimated as the rotation of the device along its *pitch* axis, comparable with nose up (pitch>0) or nose down (pitch<0) for a plane. The contribution of the pitch variable to the regression model was slightly below the contribution of the strongest variables. Because the pitch variable did not explain more or additional variance, we did not include it in the model here.

Validation of the Valence model:

Model validation was done again on the validation set.

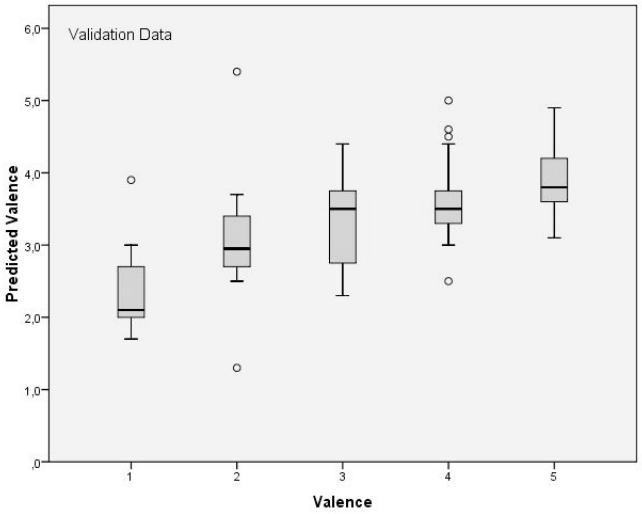

Fig. 5. Validation data. Y-axis predicted valence values . X-axis valence as given by the performer.

TABLE 6
Residual statistics.

Residuals	Minimum	Maximum	Mean	StdDev
Model	-2.031	2.203	0	0.970
Validation	-3.394	1.901	-0.255	1.073

Explanatory capabilities of the model: The variance of the validation data explained by the model : R = 0.532 (R^2 = 0.284). Compared with R = 0.606 (R^2 = 0.367) for the original data, this means a shrinkage with 8.3 %.

The predictive capabilities of the model are presented in Fig. 5. As expected with the lower R^2 values, the predicted (main) values for valence are closer to the mean. Most variation in prediction is found for low valence values.

Further investigation was done by having a closer look at the residuals. The results are in table 6.

Standard Deviation for residuals is 1.073 and that is close to the standard deviation of the model. There is one residual with an excessive value of -3.394. This case is for a low valence value (value = 2).

5 Discussion

The data model for arousal explains 73 % (R^2-statistic) of the variance for the original sample and 68.5 % (R^2-statistic) of the variance for the validation samples. These are high values and endorse the good predicting capabilities of the model. The small shrinkage (4.5 %) from the original sample to the validation data confirms the generalisation capability of this model. The data model for valence resulted in a value for the R^2 statistic of 36.7 % for the original sample and of 28.4 % for the validated data. This is a shrinkage with 8.3 %. The generalisation of the valence model is clearly less than for arousal and its predicting capabilities are also clearly less.

We have tried to remedy a possible cause for this, namely that the accelerometer data do not allow for a good estimate of location. However, an post hoc heuristic to estimate location indirectly did not improve results.

Another explanation for the lower prediction results of valence is that valence related aspects of movement are ambiguous, in the sense that human observers are also less succesful in recognizing valence accurately. This is reflected in the fact that for valence, a larger proportion of the experimental data was discarded due to lack of agreement between intended and observed valence. The higher ambiguity of valence compared to arousal is also on a par with the findings of Pollick et al. who stated that the second dimension of affect, pleasantness, was less correlated with any of the considered movement features [16]. A possible explanation is that sad music is not systematically associated with negative valence [23] [24]. Although sadness is generally considered to be an unpleasant emotion, the classification is not as straightforward in the context of music. Sad music is often considered beautiful, and therefore it may be difficult to perceive sadness in music as unpleasant [11].

The models presented here are based upon motion data from arm gestures as input. To our knowledge, experiments attempting to detect musical affect from movement using inertial sensors are as of yet very scarce. Another Reference is for example made to the study of Yi-Hsuan Yang [25]. In his research a support vector regression model was used based upon timbral texture features (spectral centroid, spectral rolloff, spectral flux and MFCC) and MPEG-7 features They obtained an R^2-statistic of 79.3 % for arousal, and 33.4 % for valence, which is in the same order of magnitude as the results presented here (68.5 % for arousal and 28.4 % for valence).

Apart from the accuracies obtained for predicting arousal and valence, the cue categories identified in observers's responses are likewise similar to those reported in other studies, such as a study on dance movements, where full-body movement was judged [18]. Similar movement cues (irregularity, fluency, speed, amount) were also identified in a study on the visual perception of expressiveness in musicians' body movements [26].

The data regression models for predicting arousal and valence are the key building blocks to form the envisioned application of an affect based music retrieval system. The data models project arm movement data into a point onto the valence/arousal plane used to describe emotion. What is missing for a complete affect based music retrieval system is the annotation of a music library and the construction of playlists. A straight-forward method to construct a playlist is then to select songs that are close to the projected point in the valence/arousal plane. Such a method could possibly be more tolerant to differences in valence than in arousal, to compensate for the lesser quality of the data model

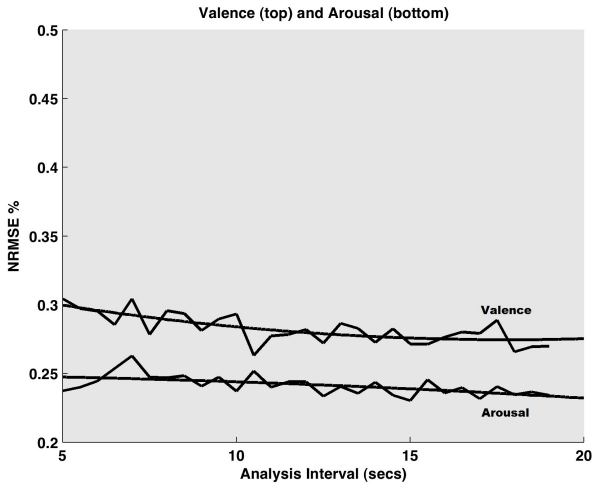

Fig. 6. NRMSE for varying retrieval/analysis intervals. (For convenience of the reader polynomial regression lines were added.)

for valence. This however is an issue of playlist-creation, and what constitutes a "good" playlist also depends on the expectations of the end-user. An end-user study would be required to gain more insight in this domain.

The regression models were derived from movement on 20 seconds excerpts. For a music retrieval system, requiring 20 seconds of movement to retrieve music is probably unacceptably long. Therefore it is worthwhile to investigate the impact of reducing the query duration. We simulated a shorter query time by stepwise reduction of the analysis interval from 2000 samples (20 seconds) to 500 samples (5 seconds) and investigating the prediction errors at every step. In figure 6 the impact of using shorter retrieval intervals is visualised. The data set used for this investigation is the validation data set. The impact on the prediction errors was measured by the normalised root mean square error (NRMSE):

$$NRMSE = \frac{\sqrt{\frac{\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}{n}}}{(Ymax - Ymin)} \quad (1)$$

With $\hat{Y}_i$ being the value for valence/arousal calculated from the models and Y_i being the real valence/arousal value (appraisal by the performer). Reducing the retrieval time clearly results in less accurate predictions. The loss of precision is however rather small. A usability study should determine the right ratio between retrieval time duration and precision. These findings are valid as well for the arousal model as well as for the valence model.

The data models for valence and arousal were derived from an experiment where subjects could hear the music. This is different from the typical situation of musical playlist creation, where the user does not hear music while making an arm gesture. Instead, he or she must "think" music. Most people intuitively understand what it means to hear a tune in your head. This can be considered as a form of musical (auditory) imagery. Converging evidence indicates that auditory cortical areas can be recruited even in the absence of sound and that this corresponds to phenomenological experience of imagining music [27]. Auditory imagery preserves many structural and temporal properties of auditory stimuli, and generation of auditory imagery appears to involve activation of many brain areas involved in perception of auditory stimuli [28]. We hypothesise that gestures made by subjects to emotionally express the music they hear is triggered by the activation in these brain areas. As a consequence, we expect arm movements made in absence of music but triggered by musical imagery, will be essentially similar to movements that would have been made when the imagined music would have been physically audible. In particular, we assume here that musical emotion can be transmitted through movement independent of the physical presence of the music.

Additional research is needed to gain insight into the role of musical imagery for our application. One important research question is: What is the impact of arousal and valence on musical imagery ? In a study with words, emotional words were consistently better recalled than the neutral words [29]. Does this also apply to music imagery, can we easier imagine music that triggers extreme values for valence and arousal?

6 Conclusions and Future Work

The work presented in this paper is intended as a foundation for a motion based affective user interface for music retrieval. We have derived predictive models for valence and arousal from empirical data, gathered in an experimental setup where inertial data recorded from arm movements is coupled to emotion ratings. This experiment firstly extends previous findings that human subjects are generally capable of recognising affect from arm gestures to the capability of recognising affect from gestures originated by the mood of a musical fragment. Secondly, model validation in the main study confirmed the predictive capabilities of the model regressing musical emotion ratings to arm movement. In line with previous studies we find that arousal is more directly related to arm movement than valence [16]. To our knowledge, attempts to detect affect from movement using inertial sensors are as of yet very scarce[2]. The use of inertial sensors for affect recognition has the crucial advantage that such sensors are readily available in mobile devices nowadays, which makes the use of the developed method in commercial applications a viable option.

Several improvements to the models can be made. A first improvement can come from an individual calibration of the model. Movement on music is an individual expression. Although our general model works, fine-tuning to individual traits of users may increase its accuracy. Studies revealed indeed that for example gender,

2. The exception that proves the rule is [30].

age, musical expertise, active musicianship, broadness of taste and familiarity with music have an influence on the semantic description of music [31].

A second improvement come from the use of other statistical models. In this study, we used linear regression models, but more complex models like support vector regression [32] or reservoir computing [33] may achieve higher prediction accuracies.

A last improvement can result from other and/or more sensing devices. The observers in the experiment indicated that the physical location where the arm movement takes place plays an important role for the determination of the valence. Arm movement performed at higher locations are indicators of joy and consequently of high valence values. Since accelerometer data alone is not sufficient to accurately estimate position, additional sensing techniques (e.g. gyroscopic sensing) will be required.

ACKNOWLEDGMENTS

We thank Kenneth Waegeman and Thomas Faes for their assistance during the experiment, and Micheline Lesaffre for providing the annotated musical excerpts.

REFERENCES

[1] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *Affective Computing, IEEE Transactions on*, vol. 1, no. 1, pp. 18–37, jan. 2010 2010.

[2] P. Juslin and D. Västfjäll, "Emotional responses to music: The need to consider underlying mechanisms," *Behavioral and Brain Sciences*, vol. 31, no. 05, pp. 559–575, 2008.

[3] M. Leman, *Embodied music cognition and mediation technology*. pp. 189–197: The MIT Press, 2008.

[4] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models," in *10th International Society for Music Information Retrieval Conference*. ISMIR, 2009, pp. 621–626.

[5] E. Schmidt and Y. Kim, "Projection of acoustic features to continuous valence-arousal mood labels via regression," in *10th International Society for Music Information Retrieval Conference*. ISMIR, 2009.

[6] E. Schubert, "Modeling perceived emotion with continuous musical features," *Music Perception*, vol. 21, no. 4, pp. 561–585, 2004.

[7] K. R. Scherer, "Why music does not produce basic emotions: A plea for a new approach to measuring emotional effects of music," in *Proceedings of the Stockholm Music Acoustics Conference 2003*, 2003, pp. 25–28.

[8] J. Panksepp and G. Bernatzky, "Emotional sounds and the brain: the neuro-affective foundations of musical appreciation," *Behavioural Processes*, vol. 60, no. 2, pp. 133–155, 2002.

[9] V. J. Konečni, A. Brown, and R. A. Wanic, "Comparative effects of musicand recalled life-events on emotional state," *Psychology of Music*, vol. 36, no. 3, pp. 289–308, 2007.

[10] J. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[11] T. Eerola and J. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," *Psychology of Music*, vol. 39, no. 1, p. 18, 2011.

[12] P. Ekman and H. Oster, "Facial expressions of emotion," *Annual review of psychology*, vol. 30, no. 1, pp. 527–554, 1979.

[13] H. Wallbott, "Bodily expression of emotion," *European journal of social psychology*, vol. 28, no. 6, pp. 879–896, 1998.

[14] C. Roether, L. Omlor, A. Christensen, and M. Giese, "Critical features for the perception of emotion from gait," *Journal of Vision*, vol. 9, no. 6, 2009.

[15] A. Atkinson, W. Dittrich, A. Gemmell, and A. Young, "Emotion perception from dynamic and static body expressions in point-light and full-light displays," *Perception*, vol. 33, pp. 717–746, 2004.

[16] F. Pollick, H. Paterson, A. Bruderlin, and A. Sanford, "Perceiving affect from arm movement," *Cognition*, vol. 82, no. 2, pp. B51–B61, 2001.

[17] S. Brownlow, A. Dixon, C. Egbert, and R. Radcliffe, "Perception of movement and dancer characteristics from point-light displays of dance." *The Psychological Record*, vol. 47, no. 3, 1997.

[18] A. Camurri, I. Lagerlöf, and G. Volpe, "Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 213–225, 2003.

[19] B. Vines, M. Wanderley, C. Krumhansl, R. Nuzzo, and D. Levitin, "Performance gestures of musicians: What structural and emotional information do they convey?" *Gesture-based communication in human-computer interaction*, pp. 3887–3887, 2004.

[20] Lesaffre, "Music information retrieval. conceptual framework, annotation and user behaviour." 2005, unpublished PhD.

[21] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, "The elements of statistical learning: data mining, inference and prediction," *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 43–94, 2005.

[22] N. Brace, R. Kemp, and R. Snelgar, *SPSS for Psychologists*. Palgrave, 2003.

[23] E. Bigand, S. Vieillard, F. Madurell, J. Marozeau, and A. Dacquet, "Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts," *Cognition and emotion*, vol. 19, no. 8, pp. 1113–1139, 2005.

[24] G. Kreutz, U. Ott, D. Teichmann, P. Osawa, and D. Vaitl, "Using music to induce emotions: Influences of musical preference and absorption," *Psychology of music*, vol. 36, no. 1, p. 101, 2008.

[25] Y. Yang, Y. Lin, H. Cheng, and H. Chen, "Mr. emo: Music retrieval in the emotion plane," in *Proceeding of the 16th ACM international conference on Multimedia*. ACM, 2008, pp. 1003–1004.

[26] S. Dahl and A. Friberg, "Visual perception of expressiveness in musicians' body movements," *Music Perception*, vol. 24, no. 5, pp. 433–454, 2007.

[27] R. J. Zatorre and A. R. Halpern, "Mental concerts: Musical imagery and auditory cortex," *Neuron*, vol. 47, no. 1, pp. 9 – 12, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/B6WSS-4GJTGTN-4/2/084cd2d103c76453a93338ecf996b2a8

[28] T. Hubbard, "Auditory imagery: empirical findings," *Psychological bulletin*, vol. 136, no. 2, pp. 302–329, 2010.

[29] R. Maddock, A. Garrett, and M. Buonocore, "Posterior cingulate cortex activation by emotional words: fMRI evidence from a valence decision task," *Human Brain Mapping*, vol. 18, no. 1, pp. 30–41, 2003.

[30] J. Wagner, E. André, and F. Jung, "Smart sensor integration: A framework for multimodal emotion recognition in real-time," in *Affective Computing and Intelligent Interaction*, 2009.

[31] M. Lesaffre, L. De Voogdt, M. Leman, B. De Baets, H. De Meyer, and J. Martens, "How potential users of music search and retrieval systems describe the semantic quality of music," *Journal of the American Society for Information Science and Technology*, vol. 59, no. 5, pp. 695–707, 2008.

[32] S. G. V. Vapnik and A. Smola, "Support vector method for function approximation, regression estimation, and signal processing," in *Neural Information Processing Systems*, M. Mozer, M. Jordan, and T. Petsche, Eds. Cambridge, MA.: MIT Press, 1997, vol. 9.

[33] D. Verstraeten, B. Schrauwen, M. D'Haene, and D. Stroobandt, "An experimental unification of reservoir computing methods," *Neural Networks*, vol. 20, no. 3, pp. 391–403, 4 2007.

Denis Amelynck received a master degree in engineering at the University of Ghent. He worked as system and training engineer for several international companies like Alcatel, Honeywell and W.R. Grace. Man Machine Interface is one of his principal interests. Currently as PHD-researcher he makes his expertise available to the Institute for Psycho-acoustics and Electronic Music (IPEM) of the University Ghent.

Maarten Grachten holds a M.Sc. in Artificial Intelligence (University of Groningen, The Netherlands), and a Ph.D. in computer science and digital communication (Pompeu Fabra University, Spain). He is currently a post-doctoral researcher at the Department of Computational Perception of the Johannes Kepler University, Austria. A central topic in his research is the computational analysis of musical expression in sound and motion.

Leon van Noorden holds a PhD in technical sciences of the Technical University Eindhoven. His current research work with IPEM, Ghent University focuses on dynamic and biomechanical aspects of human movement in response to music.

Marc Leman holds a PhD in musicology. He is research professor in systematic musicology at Ghent University and "Laureate of the Methusalem". He is director of IPEM (Institute for Psycho-acoustics and Electronic Music), specialized in methodological and epistemological foundations of musicology. He published books on computational musicology (Springer, 1995) and embodied music cognition research (MIT Press, 2008; Routledge, 2010), and he has been the promoter of a number of national and international projects on the link between music and technology. He has a record of more than 150 journal publications and has been active as key note speaker at several international conferences. In 2007, he was engaged in the design of a roadmap for music and sound computing.