Analysis of the Music Stimuli Effect on Head Micromotion

M. Boukoutsou, D. F. Kavelidis, D. Natsidou, N. Papageorgiou, I. Roboli and L. J. Hatjileontiadis
Aristotle University of Thessaloniki

***Abstract*—This paper is a comprehensive study on classification of motion capture data based on features extracted from wavelet analysis (using Discrete Wavelet Transform), Higher-Order Spectral Analysis (HOSA) and Cepstral Analysis. More specifically, the analysis presented is a continuation of the open researchon the influence of auditory stimuli on human micromotion, following the MICRO project of the RITMO Centre of the University of Oslo. The idea of the classification is to confirm that this micromotion is affected by the music genre, using new approaches on the specific problem. The emphasis on this paper is given more to the time-series representing the motion of thehead and not the direct correlation with the auditory stimuli, but rather the labels of each of the music genre. From the analysis mentioned above, features were extracted and a classifier was used; HOS-Cepstrum Classifier. The results show that the best accuracy (holdout- 25%) the classifier could achieve is equal to 64.81%.**

***Index Terms*—Discrete Wavelet Transformation, Higher-Order Spectral Analysis, Cepstrum, Motion Capture Data, Classification**

I. Introduction

Music is strongly connected with the natural tendency of humans to respond with motion to the rhythm they listen to [1]. Either it is finger tapping or full-body dancing, it is undeniable that there is something special happening when human brain is exposed to a rhythmic pattern. These spontaneous and voluntary reactions of humans to music are well known motions in a larger scale. However, movements of the head and body also take place in a micro scale, which can not been easily detected with naked eye. The studies that have analysed the field of involuntary movement when listening to music are limited and the research about micromotions especially is restricted. For this reason, the present paper is an attempt to analyse the micromotions based on the open science research project called MICRO created in 2012 [2], by A. Jensenious.

A. Experiment Setup / Subjects, Task, Stimuli

The data for this project were obtained by an experiment which took place in the University of Oslo twice (2012 and 2015). In the present paper the database from the latter experiment is used, where 108 participants were asked to stand still for 6 minutes. During this time period, different auditory stimuli were used, beginning with 60 seconds of silence and followed by alternated 60-seconds segments of music and silence. Specifically, the different music genre were Electronic Dance Music (EDM), Meditation Music, and Salsa. To further reduce the possibility of voluntary motion during the experiment, its setup was in a form of a competition were the winner was defined by the minimum motion during the 6 minutes. Moreover, the participants were divided into groups of 3-12 people at a time and they were exposed in auditory stimuli in different orders, playing from speakers. The idea behind this was to detect possible correspondences due to the different order of the stimuli i.e. do people seem to move in the same manner if they listen to Meditation Music first and then EDM as if they would for the reverse order? However, in this paper, this characteristic is not used, as our classification is purely based on the 4 different classes of stimuli (including silence). Therefore, the data are split in segments of length of the 1/6 of the original data, getting label signals for each of the 4 classes (Silence, Meditation, Salsa, EDM).

B. Micromotion Data

A marker was placed in each of the participant's head to record the position of the head using a Qualisys infrared motion capture system (at 100 Hz). For the handling of the motion capture data, the MoCap Toolbox for MATLAB [3] was used. For every room respectively, there is a corresponding *tsv* file. Each file consists of the 3 time-series of every participant in the room describing the x,y and z axis respectively. Every signal has a length of 36000 samples, before it is split in 6000-samples-long labeled segments. Also, in every file, there are markers on non-moving objects in order to obtain the noise of the measurements. As for the prepossessing of the data, all time series were centered to a zero mean.

C. Previous Research

Most of the research in the field for the specific problem was done by the RITMO researchers themselves. As described in [4], in most of the studies the main results of the proposed analysis was derived from calculations of the *Quantity of Motion*:

$$QoM = \frac{1}{T}\sum_{n=2}^{N} ||p(n) - p(n-1)|| \tag{1}$$

where p is a position vector to describe either the XY plane or the three-dimensional (XYZ) space, N is the number of samples and T is the total duration of the recording. Thus, it is a metric describing a cumulative speed-like magnitude. The experiment showed that there is a constant movement at the scale of millimeters, even when there is no music, in all recordings [5]. Furthermore, it was deducted that the quantity of motion had a linear distribution over time which means that

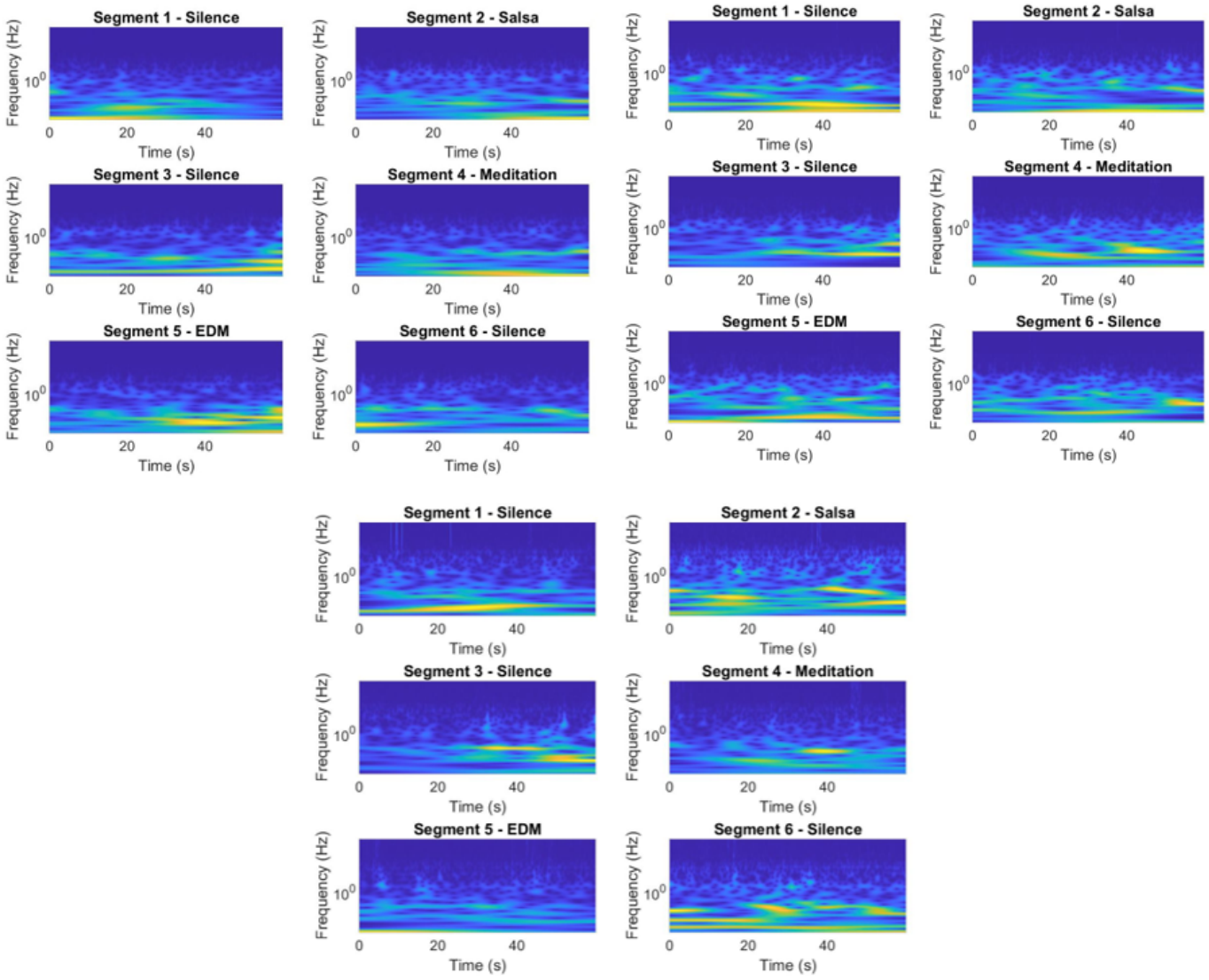

Fig. 1: CWT of X,Y,Z segments for every audio stimuli

participants with a higher tendency to move, preserved it in both silence and auditory stimulus [6].

Additionally, as described in [6], the noise-level was found to be considerably lower than that of the subjects' time-series. Findings in [7] suggest that most of the micromotion was detected in XY plane, with less significance on the Z axis. A periodicity was observed that is probably created from breathing, and a linear regression which indicates tension or fatigue in some recordings. Also, some spikes in the time-series can be explained as postural readjustments [5]. Other conclusion drawn from these researches was that the most significant impact on the micromotion occurred during EDM, and that the most subjects exhibit a very consistent level of micromotion [7].

Another study tried to evaluate the effect of headphones in micromotions. A really important conclusion was that the headphones listening provokes an indisputable increase to the velocity of the head and the body motion as compared to speakers listening. This reaction can be explained given that when the ears are covered, the participants are not capable of listening to the sounds of the environment and so, they are more exposed to the music sounds of the experiment [8]. Last but not least, there was no statistical significance to prove differences among male and female participants, different groups of age, or experience of participants with performing composing or producing music [7].

II. Materials and Methods

A. Use of Continuous Wavelet Transform

Every signal which represents the movement at an axis is divided in 6 segments of silence or music. The Continuous Wavelet Transform (CWT) is applied to every one of them in order to understand if the different types of music offer different pieces of information. Sometimes, it is observed that high frequencies occur via the analysis at the axis x and z and lower at the axis y. Furthermore, it is clear that the three components of movement are not very related. In some cases, even when silence or meditation take place, high frequencies of movement tend to be observed and preserved during the

whole experiment, possibly because of the incapacity of the participant to stand still. To sum up, it is difficult to visually distinguish the kind of music by observing the information provided by the CWT (Figure 1). This is the main reason why machine learning, subsequently, is used in order to end up with more accurate conclusions.

B. Multiresolution Analysis

1) Selection of Wavelet: The Discrete Wavelet Transform (DWT) is used in order to reconstruct the initial signal by isolating the useful information of the approximate and the detailed coefficients that occur. The wavelet selection is based on the maximum correlation measure criterion [9]. In other words, the chosen wavelet maximizes the correlation coefficient between the majority of the signals and the wavelet coefficients:

$$C(X,Y) = \frac{C_{XY}}{\sigma_X \sigma_Y} \tag{2}$$

where C_{XY} is the covariance and σ_X,σ_Y are the standard deviation of signals X and Y. Calculating the correlation coefficient for every signal, in every dimension, a histogram (Figure 2) was developed, showing that the most suitable base wavelet was the discrete Meyer wavelet by 47.04 %.

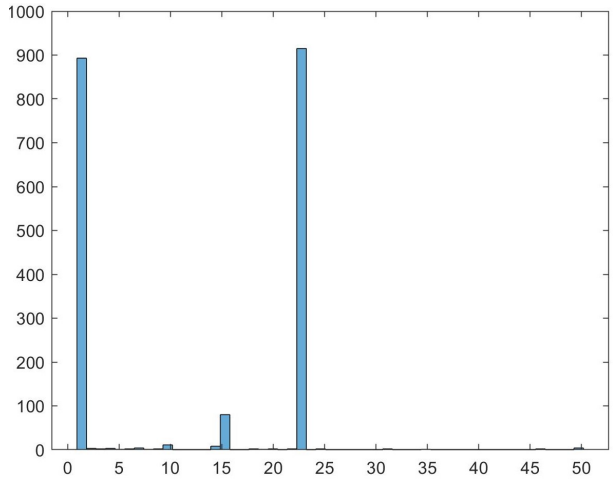

Fig. 2: Histogram showing the prevalence of each Base Wavelet

2) Decomposition and Reconstruction: The goal is to reconstruct the signals to maintain only the micromotion information and then apply high-order statistics and spectral analysis. Therefore, the signal is analysed in levels with DWT, and kurtosis is examined on each level to select the most appropriate one for the reconstruction. The kurtosis is defined by the equation:

$$\gamma_4 = \frac{E(X-m_1)^4}{(E(X-m_1)^2)^2)} - 3 \tag{3}$$

If $x(n)$ is a sample of a random variable following a Gaussian distribution, then kurtosis will approach zero. Since the aim is to maintain the coefficients in the levels where micromotion is illustrated better, and hence they deviate more from Gaussian distribution, this is the reason why the higher kurtosis is, the more valuable information it contains for the following analysis.

Applying a rolling window in the approximate coefficients of every level, the kurtosis of each segment is maintained to create a time-series [10]. Comparing the mean values of these time-series, the most appropriate level of decomposition is chosen according to the maximum mean kurtosis (Figure 3). As a result, a different level is chosen for each signal of the database, and used for the decomposition.

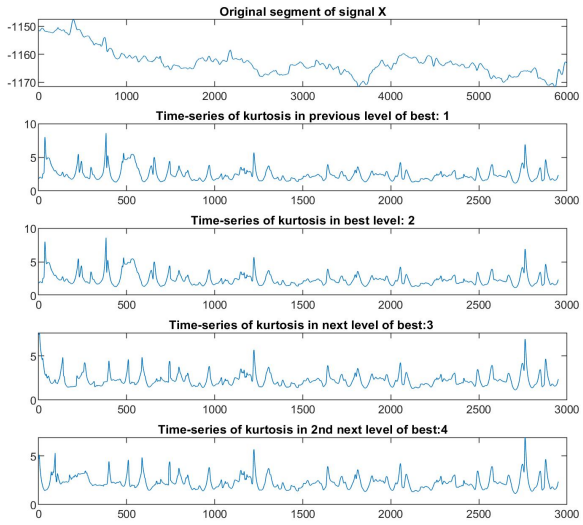

Fig. 3: Time-series of kurtosis in each level

Afterwards, during the decomposition, the kurtosis of the detailed coefficients of each level is compared in order to investigate which ones are useful. The reconstruction is completed using the desired level approximate and all the useful detailed coefficients. In the Figure 4 an original and it's reconstructed signal are illustrated.

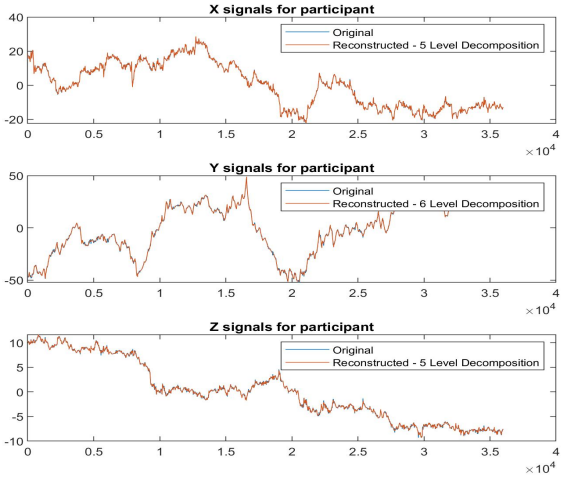

Fig. 4: Original and Reconstructed signal

C. Higher-Order Spectral Analysis (HOSA)

The next step was to examine the existence of Quadratic Phase Coupling (QPC) phenomena. For this purpose, HOSA Toolbox of MATLAB was implemented [11]. More specifically, bicoherence was estimated for the reconstructed signals, which is a normalization of the bispectrum with the power spectrum. Also, it is used to distinguish the non-linearities of the signals. When there are non-linear interactions between frequencies, they are quadratically phase coupled and the power of a third frequency as the sum of the above, is affected. Practically, for the calculation of bicoherence, given a frequency resolution

$$\Delta f = f_s \frac{N}{M} \tag{4}$$

where f_s: sampling frequency, N: number of segments into which the signal is divided, and M: elements of signal, the signal must be divided into N segments of length

$$T = M/N/f_s. \tag{5}$$

Throughout the QPC analysis, the Hanning window was utilized with an overlap of 50%. The length of the FFT that was performed in order to calculate the bispectrum, was the next larger power of 2 than T [12].

As it is illustrated at the Figure 5 QPC phenomena take place, given that there are values of bicoherence at the significant area that differ from zero.

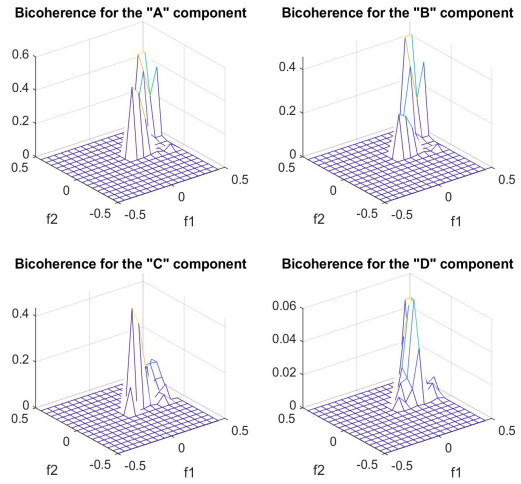

Fig. 5: Bicoherence in Significant Region for Different Audio Stimuli

D. Cepstral Analysis

The term *cepstrum* is derived from reversing the order of the letters in the *spectrum*. Originally, the (power) cepstrum is defined as the spectrum of the logarithmic spectrum [13]. This relationship describing that is:

$$C_p(\tau) = |\mathcal{F}\{logF_{xx}(f)\}|^2 \tag{6}$$

where $F_{xx}(f)$ is the power spectrum of the time signal $f_x(t)$.

However, the most common definition used today is:

$$C(\tau) = \mathcal{F}^{-1}\{log\mathcal{F}(s(n))\} \tag{7}$$

which is simply a scaled version of the power cepstrum, $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote forward and inverse Fourier transforms respectively and log is the complex logarithm.

The real cepstrum is derived by discarding the phase information contained in the imaginary part of the complex logarithm (setting phase to zero), otherwise the complex cepstrum is obtained. Thus:

$$C_r(\tau) = \mathcal{F}^{-1}\{log|\mathcal{F}(s(n))|\}. \tag{8}$$

Most of the applications regarding the use of cepstrum are in the field of speech analysis and earthquakes' signals analysis since it is a very effective tool for periodicity detection. Speech analysis is heavily affected by the phase information, thus usually the complex cepstrum is used. However, in this case, the interest is only for the possible periodicity in the head movement, so the real cepstrum is selected as the domain of the analysis.

In order to detect these periodicities, it is assumed that the signal segments are described by:

$$s(n) = h(n) * v(n) \tag{9}$$

where s is the main signal, h is the impulse response describing system of micromotion of the head, v is an excitement signal and * is the convolution operator. The Fourier transform of this relationship is:

$$S(f) = H(f)V(f) \tag{10}$$

where S, H and V are the spectrum of s, h and v respectively. The main idea behind cepstral analysis and the so-called Homomorphic Deconvolution lies behind the property of the log operator to map multiplication to addition. By substituting (9),(10) to (8):

$$C_r(\tau) = \mathcal{F}^{-1}\{log|H(f)|\} + \mathcal{F}^{-1}\{log|V(f)|\}. \tag{11}$$

Hence, it is clear that with proper *liftering* (filtering in the cepstrum domain), impulse response can be separated from the excitation signal.

The Automated Procedure for Cepstrum Editing (ACEP) used in this problem as proposed in [13] is described below:

1) Real Cepstrum Calculation: The original C_r is obtained and stored.

2) Long-Pass Lifter: Given that the modal information is mostly localized at a low *quefrency* [14] (from "frequency"), long-pass liftering is applied to the real cepstrum.

Specifically, the filter is:

$$l_{LP}(n) = \begin{cases} 0, & n = 0 : N_c \\ 1, & n = N_c + 1 : N/2 \end{cases}$$

where N is the signal segment length and N_c is the cut-off quefrency index. In this case, the cutoff quefrency was selected heuristically based on the quality of the results ($N_c = 200$).

Therefore, the long-pass liftered cepstrum is:

$$C_{LP}(n) = C_r(n)l_{LP}(n). \tag{12}$$

3) Spectral Subtraction: The assumption first is that the noise in the obtained long-pass liftered cepstrum is additive. In order to denoise the signal spectral subtraction is used, considering the other assumptions described in [13]. Spectral Subtraction algorithm will not be analyzed in this paper as it is a very common technique among signal denoising algorithms. The resulting signal is called *Peak-Enhanced Cepstrum*, as it is a cepstrum where peaks are more apparent and sharp, in order to be detected later on.

4) Peak Detection: Taking the peak-enhanced cepstrum C_{res}, a threshold is selected as proposed in [15] as:

$$thres = E[C_{res}(n)] + 3std[C_{res}(n)] \tag{13}$$

where E[.] is the expected value operator and std[.] is the standard deviation. Then, a vector is formed holding the locations of the peaks with a value larger than the threshold:

$$locs = \{\forall n | C_{res}(n) > thres\} \tag{14}$$

that will be used in the following liftering.

5) Comb Liftering: By iterating through the $locs$ vector, a comb lifter $l_c(n)$ is created in order to remove the peaks of the original cepstrum.

$$l_c(n) = \begin{cases} 0, & n = locs(i) - 2 : locs(i) + 2 \\ 1, & else \end{cases}$$

Finally, the desired liftered cepstrum is obtained by liftering the original cepstrum with $l_c(n)$:

$$\hat{C}_h(n) = C_r(n) l_{LP}(n). \tag{15}$$

6) Impulse Response Estimation: The estimated impulse response can be obtained by transforming the liftered cepstrum back to the time-domain:

$$h(n) = \mathcal{F}^{-1}\{e^{\mathcal{F}\{\hat{C}_h(n)\}}\}. \tag{16}$$

It is important to note here that at the first observation of the data in the cepstral domain, the information of peaks are much more apparent in the z axis as shown in the demonstrative example in Fig. 6 and Fig. 7. Also, a demonstration of the whole cepstral analysis with ACEP is shown on Fig. 8.

E. Bicepstrum

The idea behind the analysis of the Bicepstrum domain is similar to the cepstral analysis, meaning that the goal is to obtain the estimated impulse response of the system. The difference now is that this analysis uses parametric models. Let consider an autoreggressive moving average (ARMA) sequence $x(k)$ that has a transfer function $h(k)$ which Z transform is:

$$H(z) = Kz^{-r} I(z^{-1}) O(z) \tag{17}$$

with

$$I(z^{-1}) = \frac{\prod_{i=1}^{L_1}(1 - a_i z^{-1})}{\prod_{i=1}^{L_3}(1 - d_i z^{-1})}, \tag{18}$$

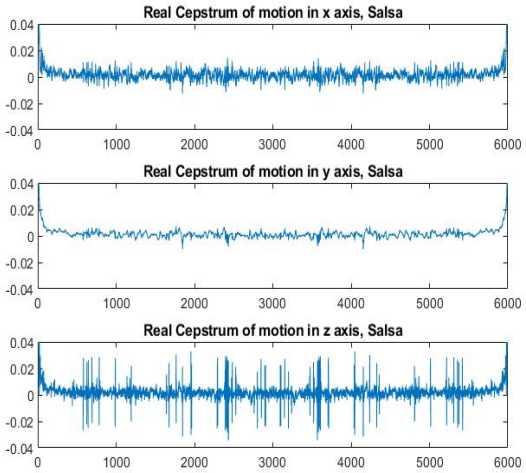

Fig. 6: Real cepstrum in the 3 axis / Salsa

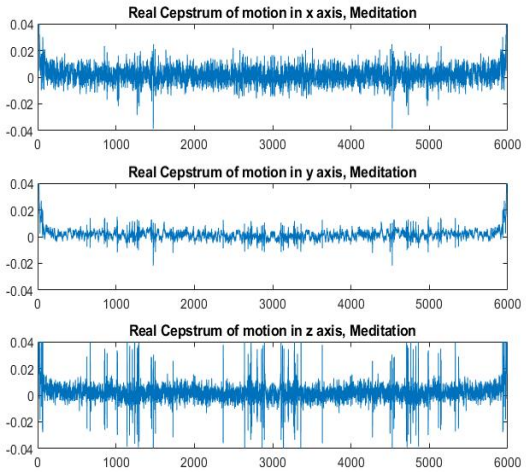

Fig. 7: Real cepstrum in the 3 axis / Meditation

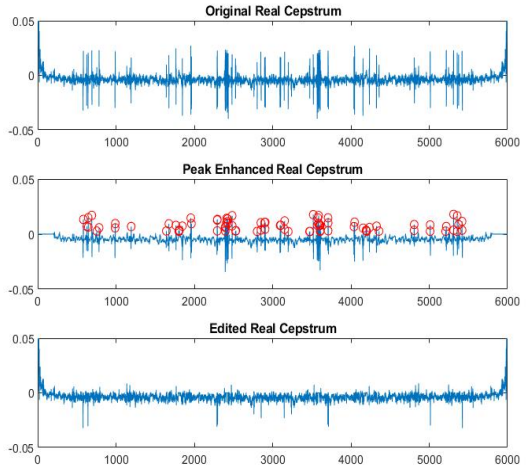

Fig. 8: Automated Cepstrum Editing Procedure (ACEP) demonstration

and

$$O(z^{-1}) = \prod_{i=1}^{L_2}(1 - d_i z^{-1}), \tag{19}$$

where K is a constant, r is a integer, $I(z^{-1})$ is the minimum phase component and $O(z)$ is the maximum phase component. The cepstral parameters, $A^{(m)}$ and $B^{(m)}$ are given by:

$$A^{(m)} = \sum_{i=1}^{L_1} a_i^m - \sum_{i=1}^{L_3} d_i^m \tag{20}$$

and

$$B^{(m)} = \sum_{i=1}^{L_2} b_i^m. \tag{21}$$

As it is presented in [16], the IR for the minimum phase component (causal), $i(k)$ of $H(z)$ is described as:

$$i(k) = -\frac{1}{k}\sum_{m=2}^{k} A^{(m-1)} i(k-m+1),\ k \geq 1 \tag{22}$$

and the maximum phase component (anti-causal) is:

$$o(k) = \frac{1}{k}\sum_{m=k+1}^{0} B^{(1-m)} o(k-m+1),\ k \leq -1 \tag{23}$$

considering $i(0) = o(0) = 1$. Hence, the system IR can be derived from the convolution of the two components:

$$h(k) = i(k) * o(k). \tag{24}$$

The *Bicepstrum* is defined [17] as the 2D Z transform of the log biscpetrum, $C_3^H(z_1, z_2)$ as:

$$b_H(m,n) = Z_2^{-1}[log[C_3^H(z_1, z_2)]]. \tag{25}$$

The reconstruction algorithm used with the FFT-based estimation of $A^{(m)}$ and $B^{(m)}$ is thoroughly described in [18] and will not be further analyzed in this paper since the method was used through the HOSA Toolbox for Matlab [11].

F. Feature Extraction

1) HOS Analysis Features: Taking advantage of the analysis presented previously, about the QPC study and the MRA using the kurtosis data, features can easily been extracted for the classification.

At first, the original data are analysed in segments, that are equal to the 1/6 of the length of the total time series for each movement, measured for the participants of the experiment. To be more specific, 648 segment objects are created, 6 segments for every one of the 108 participants. Each object has the data for the movements on the 3 axes for one minute - the total of the consecutive aural stimuli that the participant responds to. Each segment belongs to a specific class and the 3 time series can be used for the feature extraction - this is the way that the supervised learning for classification is implemented.

Using the results of the kurtosis analysis, each time series is decomposed and the optimal level of reconstruction is estimated. The mean kurtosis, the maximum kurtosis as well as the position of the maximum kurtosis for the optimal level are used as features for the segment. Therefore, 9 feature variables create the feature vector that describes each object. One thing that is worth being mentioned is that the position of the maximum kurtosis is appropriately multiplied by a factor corresponding to the current decomposition level. To be more specific, each maximum position is multiplied by 2^{level}. This is done, so that we take into account the undersampling that takes place on the Multiresolution Analysis.

As for the QPC analysis, each time series is decomposed and reconstructed by the way that has been previously discussed. The bicoherence of every reconstructed signal is estimated and its maximum values are computed. The positions of the maximum values (on the two axes) for the bicoherence of each signal are returned and stored as feature variables. Therefore, a vector with a length equal to 6 (2 variables for the 3 time series of each segment) is used as the object feature vector, as extracted with the QPC method.

2) Cepstral Analysis Features: The idea behind the features of the cepstral analysis is connected with the following classification since the training data set and the test data set (to be used in the classification) are needed. Since the data are split in segments of 1/6 and their corresponding labels, an algorithm that divides the data into 4 groups (according to their label), and computes the impulse response for each signal with both methods (Cepstrum, Bicepstrum) is used. Then, two reference impulse response signals are computed for each class, corresponding to the two methods. These reference signals are computed as the mean signal of all the signals of the class. Then, for every signal segment in the test data, it computes the impulse response from Cepstrum and Bicepstrum and then it does a comparison with the reference signals, providing the following features:

$$i)\ \ dtw(h_{i_{ceps}}, h_{ref_{ceps}}) \tag{26}$$

$$ii)\ \ dtw(h_{i_{bic}}, h_{ref_{bic}}) \tag{27}$$

$$iii)\ \ corr2(h_{i_{ceps}}, h_{ref_{ceps}}) \tag{28}$$

$$iv)\ \ corr2(h_{i_{bic}}, h_{ref_{bic}}) \tag{29}$$

$$v)\ \ HOC(h_{i_{ceps}}) \tag{30}$$

$$vi)\ \ HOC(h_{i_{bic}}) \tag{31}$$

where dtw stands for *Dynamic Time Warping Distance* between two signals, corr2 is the *Correlation Coefficient* of two signals, HOC stands for Higher-Order Crossings [19] with K = 20, $h_{i_{ceps}}$, $h_{i_{bic}}$ denote the impulse responses of the i^{th} segment of the test data obtained by cepstrum and bicepstrum respectively, and $h_{ref_{ceps}}$, $h_{ref_{bic}}$ are the reference signals generated by the algorithm for the feature extraction.

As stated in [20], since the features of the training data are necessary to train the classification model, the proposed

algorithm described above can be used as well for the training data, while doing the comparison with the reference signals without any serious bias issue, to obtain the desired features. This whole procedure is calculating the features for each axis for every segment, in order to have a thorough 3-dimensional analysis, resulting to feature vectors with length $(4+4+4+4+20+20)3 = 108$. However, different tests on which using of only some of the features (or even the axis) have also been performed.

3) Wavelet Scattering Features: As a complementary part of this analysis, one more method for the feature extraction process is tested, necessary for our classification task. The chosen method is the wavelet time scattering method, the idea of which is described and analysed on [21]. In general terms, by consecutively convolving the signal with appropriate wavelet filters and applying a modulus function (signal norm), the coefficients that are called scalogram coefficients are extracted. Then, by convolving each scalogram coefficient with a scaling function, the scattering coefficients are collected, which when handled appropriately can serve as excellent feature extractors - a quite precise classification system can therefore be built on them.

The feature extraction process has many similarities with the general deep convolutional network strategies, that are widely used on the implementations of many deep learning algorithms. The main difference that this method brings to the table is that the scaling function and the wavelet filters are designed a priory - on most of the other deep convolutional networks these factors are learned during the algorithm. This is a big advantage of the wavelet scattering method, as it allows the algorithm to effectively extract features without the need of a large data set.

Initially, the segment structures from the data set are collected. Each segment structure contains the 3 time series describing the movement of one participant for the duration of one minute (before the music genre changes), just like on the previous methods. The total number of segments is 648 - 6 segments for 108 participants. For this algorithm, only the x time series is taken initially for the feature extraction - as it was observed that the result was very satisfactory and further feature extraction was not needed. It was, also, decided that the differences (of order 1) of the time series segments were used, defined as:

$$y(n) = x(n+1) - x(n) \tag{32}$$

where x is the original time series and y is the transformed one. This was done again because it resulted in the improvement of the classification.

For the scattering features, the scattering network was created with an invariance scale (describing the scaling function) equal to 15, 8 filters per octave on the first filter bank and 1 wavelet per octave on the second filter bank. For the algorithms the Wavelet toolbox of MATLAB was used - the wavelet scattering functions are implemented there [22]. Then, by passing the signal through the filters and by the process briefly described previously, the scattering coefficients are extracted. The function used for the extraction critically resamples (downsampling) the coefficients (for computational purposes). Due to this fact, the resolution of the scattering coefficients is equal to 24 in this case, and for each initial data point (which was a time series of length equal to 5999) there are 24 scattering features corresponding in the algorithm.

After getting the final dataset ($24 * 648 = 15552$ scattering feature vectors), a KNN classifier is trained, and validated to use a 5-fold cross validation scheme. The classifier is trained to classify the scattering features, not the initial segment objects, and, as it is already mentioned, there are 24 scattering feature vectors corresponding to one initial segment object. The decision on the classification of the initial segment object is taken by a simple majority vote on the 24 corresponding scattering features - more information about this aspect can be seen on the algorithm results.

III. Results and Discussion

A. HOS - Cepstrum Classifier

Using the features extracted from the HOS and cepstrum analysis, it is now possible to train a classifier so that it can solve the problem. For this aspect, a Linear Support Vector Machine (SVM) model was used, which standardises the predictors. The model is validated using a holdout 25% validation scheme.

The reason why, a holdout validation and not a k-fold cross validation scheme is used, is to create a training data set and a test data set to fit the algorithm for the feature extraction described in section F. 2. The idea is to partition the data with a cross-validation partitioner and then use the two datasets in the algorithm.

In this part, various combinations of the features extracted from the HOS- Cepstrum methods were tested, in order to create the final feature vector:
1) HOSA Features
2) QPC (Bispectrum) Features
3) Cepstrum Features
4) Cepstrum and Bicepstrum Features
5) HOSA and Cepstrum
6) QPC (Bispectrum) and Cepstrum
7) Cepstrum Features of Z axis
8) All features

The results can be seen below:

Method	Validation Accuracy
1	50.00%
2	50.00%
3	64.81%
4	62.34%
5	64.20%
6	64.81%
7	62.96%
8	64.81%

TABLE I: Table to test captions and labels

As it can be observed from the results, the best accuracy that could be achieved was by using the Cepstrum features, and was equal to 64.81%. Worth mentioning is the fact that by using the HOS-QPC features as well, the accuracy was not improved. The classification results are not that great, and it is

evident that the model has a hard time recognising the proper classes.

IV. Future research

The results above show us that the study of the wavelet transforms is a very effective field that can be used for feature extraction on signals and images.

One aspect that we could study in the future is the feature extraction from the CWT of a signal, and not from the DWT and the MRA. We could use image feature extraction algorithms (maybe SIFT and other detectors/descriptors) in order to extract features for our signal using the 2D CWT representation and take an idea on how the CWT can be used for this aspect.

Furthermore, we could use deep learning approaches, like the convolutional neural networks, that we can actually try to train using the CWT data. Of course, a fairly larger dataset should be useful for a better classification and a statistically more accurate analysis.

Also, one serious consideration for future work is to use the audio signals used in the experiment to study direct correlations between the music and the micromotion, as well as locate the events of micromotion in time.

Acknowledgment

The authors would like to thank the RITMO Centre for Interdisciplinary Studies in Rhythm, Time and Motion at the University of Oslo for setting up the experiment and providing access to the Motion Capture Data, as well as the 108 subjects for their voluntary participation in the study.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] B. Burger, M. Thompson, G. Luck, S. Saarikallio, and P. Toiviainen, "Influences of rhythm- and timbre-related musical features on characteristics of music-induced movement," *Frontiers in Psychology*, vol. 4, p. 183, 2013.

[2] Gonzalez, Victor, Zelechowska, Agata, and Jensenius, Alexander Refsum, "MICRO Motion capture data from groups of participants standing still to auditory stimuli (2015)."

[3] B. Burger and P. Toiviainen, "Mocap toolbox - a matlab toolbox for computational analysis of movement data," 08 2013.

[4] A. Jensenius, A. Żelechowska, and V. Gonzalez, "The musical influence on people's micromotion when standing still in groups," 07 2018.

[5] A. Jensenius and K. Bjerkestrand, "Exploring micromovements with motion capture and sonification," vol. 101, 12 2011.

[6] A. Jensenius, K. Nymoen, S. Skogstad, and A. Voldsund, "A study of the noise-level in two infrared marker-based motion capture systems," pp. 258–263, 01 2012.

[7] V. E. Gonzalez-Sanchez, A. Zelechowska, and A. R. Jensenius, "Correspondences between music and involuntary human micromotion during standstill," *Frontiers in Psychology*, vol. 9, p. 1382, 2018.

[8] A. Zelechowska, V. E. Gonzalez-Sanchez, B. Laeng, and A. R. Jensenius, "Headphones or speakers? an exploratory study of their effects on spontaneous body movement to rhythmic music," *Frontiers in Psychology*, vol. 11, p. 698, 2020.

[9] R. Yan and R. Gao, "Base wavelet selection for bearing vibration signal analysis.," *IJWMIP*, vol. 7, pp. 411–426, 07 2009.

[10] C. Saragiotis, L. Hadjileontiadis, A. Savvaidis, C. Papazachos, and S. Panas, "Automatic s-phase arrival determination of seismic signals using nonlinear filtering and higher-order statistics," in *IGARSS 2000. IEEE 2000 International Geoscience and Remote Sensing Symposium. Taking the Pulse of the Planet: The Role of Remote Sensing in Managing the Environment. Proceedings (Cat. No.00CH37120)*, vol. 1, pp. 292–294 vol.1, 2000.

[11] A. Swami, J. M. Mendel, and C. L. Nikias, "Higher-order spectral analysis toolbox," *The Mathworks Inc*, vol. 3, pp. 22–26, 1998.

[12] P. Poloskei, G. Papp, G. Por, L. Horvath, and G. Pokol, "Bicoherence analysis of nonstationary and nonlinear processes," 11 2018.

[13] A. Ompusunggu and T. Bartic, "Automated cepstral editing procedure (acep) for removing discrete components from vibration signals," *International Journal of Condition Monitoring*, vol. 6, pp. 56–61, 09 2016.

[14] J. Tribolet and A. Oppenheim, "Deconvolution of seismic data using homomorphic filtering," p. 11, 08 1977.

[15] C. Peeters, P. Guillaume, and J. Helsen, "A comparison of cepstral editing methods as signal pre-processing techniques for vibration-based bearing fault detection," *Mechanical Systems and Signal Processing*, vol. 91, pp. 354–381, 07 2017.

[16] C. L. Nikias and A. Petropulu, "Higher-order spectra analysis : a nonlinear signal processing framework," 1993.

[17] A. Iturrospe, D. Dornfeld, V. Atxa, and J. Abete, "Bicepstrum based blind identification of the acoustic emission (ae) signal in precision turning," *Mechanical Systems and Signal Processing*, vol. 19, pp. 447–466, 05 2005.

[18] R. Pan and C. L. Nikias, "The complex cepstrum of higher order cumulants and nonminimum phase system identification," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 36, pp. 186–205, 1988.

[19] P. Petrantonakis and L. Hadjileontiadis, "Emotion recognition from eeg using higher order crossings," *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, vol. 14, pp. 186–97, 10 2009.

[20] R. J. Kate, "Using dynamic time warping distances as features for improved time series classification," *Data Mining and Knowledge Discovery*, vol. 30, pp. 283–312, 2015.

[21] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1872–1886, 2013.

[22] M. Misiti, Y. Misiti, G. Oppenheim, and J.-M. Poggi, "Wavelet toolbox," *The MathWorks Inc., Natick, MA*, vol. 15, p. 21, 1996.