Decoding Affect in Videos Employing the MEG Brain Signal

Mojtaba Khomami Abadi[1,4], Mostafa Kia[1], Ramanathan Subramanian[2], Paolo Avesani[3], Nicu Sebe[1]
[1]University of Trento, Italy
[2]Advanced Digital Sciences Center (ADSC), University of Illinois at Urbana-Champaign, Singapore
[3]Fondazione Bruno Kessler, Trento, Italy
[4]Semantic, Knowledge and Innovation Lab (SKIL), Telecom Italia

Abstract— **This paper presents characterization of affect (valence and arousal) using the Magnetoencephalogram (MEG) brain signal. We attempt single-trial classification of movie and music videos with MEG responses extracted from seven participants. The main findings of this study are that: (i) the MEG signal effectively encodes affective viewer responses, (ii) clip arousal is better predicted than valence employing MEG and (iii) prediction performance is better for movie clips as compared to music videos.**

I. Introduction

Humans perceive emotions from the environment through visual and auditory stimuli- characterized by speech, audio/video music clips, images and movies in the digital world. While many studies have investigated how speech and image signals can effectively elicit emotions in people [15], [18], research on isolating emotional content in music and movie videos began only recently. Past works such as [7], [9] have attempted to identify emotions either by (i) analyzing the content to develop models that link low-level image and audio features to *valence* (emotion type) and *arousal* (emotion intensity) or (ii) analyzing the viewer's facial activity/expressions and correlating these responses with the presented content. While content-based analysis enables discovery of video highlights (typically high-arousal segments), it is inherently not suited for tagging content on the valence-arousal plane. Conversely, while facial expressions can provide some insight regarding emotional video content, they can easily be controlled by the viewer and are therefore, not always reliable. The above shortcomings have prompted researchers to investigate emotional response to affective stimulus considering both peripheral nervous system signals and peripheral physiological signals such as (i) Electroencephalogram (EEG), which measures electrical activity along the scalp, (ii) Electromyogram, measuring electrical activity of skeletal muscles, (iii) heart rate, (iv) galvanic skin response (GSR) measuring skin conductance and (v) skin temperature. These signals have been found to effectively encode emotional responses [12], [10] and are more primitive than facial expressions, which typically denote the conscious manifestation of an emotion.

This work explores a new type of peripheral nervous system signal that records the functional brain activity non-invasively. The purpose of our study is to investigate the use of Magnetoencephalogram (MEG) to decode emotional responses from the brain recording when a subject is exposed to affective videos. MEG is a technology that allows the recording of the magnetic fields produced by the electrical activity of the brain. When a group of neurons is activated, electrical currents along the neurons generate tiny, orthogonally oriented magnetic fields. The sum of these magnetic fields generates a change in magnetic field around the activated part, and constitutes the MEG response.

Furthermore, single trial decoding of MEG responses is attempted. While many EEG studies (*e.g.*, [12], [10], [20]) have successfully decoded affective viewer response to videos, there are no such MEG-based studies. However, the fact that MEG can effectively encode affective responses, similar to EEG, is demonstrated in [17] employing image stimuli. Their results are obtained on analyzing event-related magnetic fields (ERF), where an individual's brain responses are acquired over many trials and averaged. In contrast. This is first study employing single-trial MEG classification for decoding affective viewer responses to videos.

Another aspect investigated in this work is the suitability of different types of video stimuli for emotion elicitation. For a study examining viewers' emotional responses to be successful, the employed stimuli should effectively elicit the emotions targeted by the study. While some works have attempted to identify appropriate video stimuli for studying affect [6], [3], different authors have employed different stimuli for emotion elicitation. [12] presents an affect characterization study using 21 movie clips, while the authors in [10] elicit emotions through music videos.

This work analyzes the suitability of i) movie clips and ii) music videos for measuring affect with MEG signals. We present results of a preliminary study involving seven participants, where the MEG responses of each subject were recorded as they viewed 32 movie clips and 40 music videos over two separate experimental sessions. Every participant also provided valence and arousal ratings for each movie/music clip during the experiment. The MEG responses were then correlated with these ratings to train a classifier for predicting a clip's valence/arousal tag (as 'high' or 'low'). Our experimental results suggest that (i) the MEG signal effectively encodes affective viewer responses, (ii) clip arousal is better predicted than valence employing MEG and (iii) prediction performance is better for movie clips as compared to music videos.

II. Related work

While affective content creators *intend* to convey a certain emotion (or a set of emotions) through a music/movie video, the *actual* emotion induced on viewing the video is influenced by a number of psychological and contextual factors, and is therefore, highly subjective. Therefore, correlating the *observed* emotional response with the *expected* response is a non-trivial problem which is typically simplified in practice employing the following ideas: (1) Most affective studies assume that the entire gamut of human emotions can be represented as a set of points on the valence-arousal plane as demonstrated by Greenwald *et al.* in [5], and (2) To largely ensure that the elicited and expected emotions are consistent, the presentation stimuli are carefully selected based on previous studies, or based on 'ground truth' valance-arousal ratings compiled from a large pool of subjects who evaluate the stimuli prior to the actual experiment.

Emotional states have been found to produce specific types of physiological responses- *e.g.,* excitement is associated with increased heart-beat and respiration rates, and this correlation is exploited in a number of physiology-based affect studies. Heart-rate, skin temperature and conductance level, blood pressure and facial EMG are recorded as subjects view affective imagery in [18]. Their experiments indicate that the responses for anger and fear are uniquely distinctive from the responses to neutral images.

Among physiology-based affective studies with video stimuli, Lisetti and Nasoz [12] employ a two-pronged approach to elicit frustration along with other emotions from 29 participants. They use prototypical movie clips to evoke sadness, anger, amusement, fear and surprise, and induce frustration by asking subjects to solve difficult mathematical questions without pencil and paper. GSR, heart rate, temperature, EMG and heat flow responses are recorded using an armband, and over 80% accuracy is obtained in classifying the aforementioned emotions using extracted features. In the DEAP dataset, Koelstra *et al.* [10] record EEG, GSR, blood volume pressure, respiration rate, skin temperature and Electrooculogram (EOG) patterns as viewers are presented with 40 one-minute music video segments. These responses are correlated with arousal, valence, liking and dominance ratings provided by participants during the experiment. A mean accuracy of over 60% is obtained for single-trial binary classification with EEG and peripheral physiological signals. The MAHNOB-HCI multimodal database compiled by Soleymani *et al.* [20] contains face videos, audio and physiological signals as well as eye-gaze data of 27 participants who watched 20 emotional movie/oline clips in one experiment, and 28 images and 14 short videos in another. Their database facilitates affect computation using single or modalities and determination of the most suitable modalities.

Upon reviewing related literature, one can make the following observations: (1) All these studies, apart from DEAP, derive their conclusions from experiments involving a relatively small number of stimuli. This is because such studies are inherently hard to conduct. One needs to take into account the time required for subject preparation, stimulus viewing and recording user ratings while designing the experiment protocol. Also, the fact that fatigue strongly influences the quality of emotional responses discourages lengthy experiments with many stimuli. (2) While all these approaches have been generally successful in isolating physiological correlates of specific emotions arising from the presented stimuli, no comparison studies have been made to determine which stimulus is ideally suited for affect computation, given the experiment hypotheses and duration. This paper presents one of the first steps in that direction.

III. Experimental Protocol

In this section, we present a brief description of (a) MEG and (b) stimuli selection procedure, before detailing the (c) experimental set-up and protocol, and (d) analysis of self-assessment ratings for the music and movie clips.

A. Magnetoencephalogram

MEG is a recent technology that enables non-invasive recording of brain activity, and is based on SQUIDS (Superconducting Quantum Interference Devices), which enables recording of very low magnetic fields. Magnetic fields produced by the human brain are of the order of pico-Tesla and since sensors are really sensitive to noise, the MEG equipment is located in a shielded room insulated from other electrical/metallic installations. A multiple coils configuration enables measurement of magnetic fields induced by tangential currents, and thus, brain activity in the sulci of the cortex can be recorded.

B. Stimuli selection procedure

As mentioned earlier, many affective studies have been conducted with image stimuli, and there exist standard datasets such as [11] for researchers to conduct experiments and evaluate their findings. However, there exist few affective video datasets, in spite of studies confirming that reliable emotion elicitation is feasible with video stimuli such as movies [8]. An affective music video dataset, comprising 40 music videos, was recently presented in [10]. Our endeavor was to create a large-sized affective movie dataset along those lines owing to the following reasons: (1) The importance of context in emotion perception has been acknowledged by many studies (*e.g.*,[2]). Temporal context can be conveyed effectively by both audio and visual content in movies, whereas context in music videos is predominantly conveyed by the audio, which is supplemented by the visuals; (2) As a result, movies can effectively elicit a larger range of emotions (*e.g.*, including surprise/shock and fear) as compared to music videos.

To this end, we initially compiled a set of 48 Hollywood movie clips, suggested as suitable for affective studies in [6], [3]. These clips were shown to 42 subjects, who self-assessed their emotional state upon viewing each clip, to provide valence and arousal ratings as well as the most appropriate emotion tag (*e.g.*, funny, exciting) for each movie clip. Dividing the valence-arousal plane into 4 quadrants (corresponding to high/low valence and arousal), we finally chose 32 movie clips which obtained the most consistent and representative scores for the experiment, to have a balanced distribution of 8 clips/quadrant. These clips were between 51 sec and 128 sec long ($\mu = 81, \sigma = 21.5$), and were associated with diverse emotional tags such as *funny*, *amusing*, *exciting*, *sad*, *disgusting* and *angering*. To investigate whether MEG-based affect recognition varied with the stimulus type, we also used the 40 one-minute music video highlights suggested in [10] in our experiments.

C. Experimental set-up

1) Materials and set-up: All MEG recordings were performed in a shielded room with controlled illumination. Fig. 1(a) presents an overview of the experimental set-up. Due to the sensitivity of the MEG equipment, all other devices used for data acquisition were placed in an adjacent room, and were controlled by the experimenter. Two PCs were used, one (Intel i7, 8 GB RAM) for stimulus presentation and the other for MEG data recording. The stimulus presentation protocol was developed using MATLAB's Psychtoolbox (http://psychtoolbox.org/) along with some functions adapted from the ASF stimulus presentation framework [19]. Also, synchronization markers were sent from the stimulus presenter PC to the MEG recorder at the beginning and end of each stimulus display. All stimuli were shown at a resolution of 1024×768 pixels and at a screen refresh rate of 60 Hz, and this display was projected onto a screen placed about a meter in front of the subject inside the data acquisition room. All music/movie clips were played to the participant at 20 frames/second, upon normalizing the audio volume to have a maximum power amplitude of 1.

Stereo speakers were also placed in the MEG acquisition room for rendering the audio in the music/movie clips. Also, each participant was provided with a microphone to communicate with the experimenter during the recording or in the case of an emergency. The ***Neuromag*** device, which outputs 306 channels (corresponding to 102 magnetometers and 204 gradiometers) with a sampling frequency of 1 KHz, is used for recording MEG responses.

2) Protocol: 7 university graduate students (4 male, 3 female) participated in the experiments. Data acquisition for each participant was spread over two sessions interspersed by a day- movie clips were presented in one session, while music videos were presented in the other. For four of the subjects, movie clips were shown first, while three viewed music videos before the movie clips. During each session, the music/movie clips were shown in random order, and in such a way that two clips of similar valence and arousal did not follow one another. To avoid fatigue, each acquisition session was split into two halves (with 20 music/16 movie clips shown in each half) and lasted for one hour in total.

At the beginning of each recording session, the participant was first briefed about the experiment, and was asked to remove any metallic objects he/she was wearing before entering the MEG room- this was mandatory as metals would interfere with the magnetic field. Then, a practice trial was conducted so that the subject could acquaint him/herself with the protocol. Each acquisition session involved a series of trials. During each trial, a fixation cross was first shown for 4 seconds to prepare the viewer and to gauge his/her rest-state response. Upon stimulus presentation, the subject conveyed the emotion elicited in him/her by the stimulus to the experimenter through the microphone. Ratings were acquired for (i) arousal ('How intense is your emotional feeling on watching the clip?') on a scale of 0 (calm) to 4 (highly aroused), and (ii) valence ('How pleasant do you feel after watching this clip?') on a scale of -2 (very unpleasant) to 2 (very pleasant). A maximum of 15 seconds was available to the participant to convey each rating. The protocol timeline for each trial is presented in Fig.1(b).

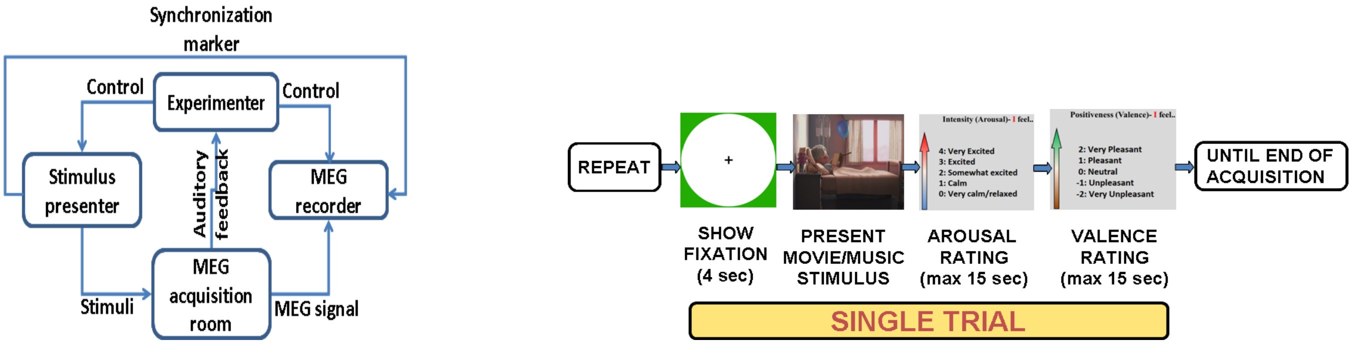
D. Self-assessment ratings: Music vs movie clips

In this section, we compare the valence-arousal ratings provided by participants for music and movie clips. Participant ratings are (i) a conscious reflection of their emotional state upon viewing the stimuli, and therefore, should be correlated with their physiological responses (ii) ultimately used for valence and arousal classification, and the variance in ratings can provide vital cues regarding the best-case classification results and (iii) also indicative of whether the presented music/movie stimuli can effectively evoke an emotional response from viewers.

Fig.2 presents plots of the mean valence-arousal (VA) ratings obtained from 7 participants for the music and movie clips respectively. The blue, cyan, black and red colors are used to respectively denote high arousal, high valence (HAHV), low arousal, high valence (LAHV), low arousal, low valence (LALV) and high arousal, low valence (HALV) stimuli as per the ground-truth ratings. Note that even though the ratings have been compiled from few subjects, we still obtain a C-shape (facing upwards) for both movie and music clips, consistent with previous studies such as [11], [10]. The C-shape is attributed to the fact that it is generally difficult to evoke low-arousal and strong valence responses. This phenomenon is particularly obvious in the case of music clips, where there is considerable overlap between the cyan (LAHV), black (LALV) and red (HAHV) clusters. However, this overlap is not as pronounced for the movie clips.

To further investigate this observation, we performed a Wilcoxon signed-rank test as in [10] to check if high and low arousal stimuli induced a difference in valence ratings. The test showed that a high/low arousal rating significantly influenced valence ratings for music stimuli ($p = 0.005$), while no such influence could be observed for movie clips ($p = 0.791$). Therefore, valence-arousal distinction is clearer for movie clips as compared to music videos.

Noting that significant inter-individual differences could have influenced the observed stimuli distribution, we also performed a second experiment. Assuming that the ground-truth VA ratings were provided by an 'ideal' annotator, we compared the mean agreement between the participant-ground truth ratings using Cohen's Kappa measure. The Cohen's Kappa coefficient measures agreement between two raters who classify N items into C mutually exclusive classes, and is computed as $\kappa = (P(a) - P(e))/(1 - P(e))$, where $P(a)$ denotes relative observed agreement between the raters, while $P(e)$ denotes probability of chance agreement- k increases from 0-1 as the inter-rater agreement increases from random to perfect. For each subject, we thresholded the user ratings based on their mean rating to assign a stimulus to either High/Low valence/arousal. Then,

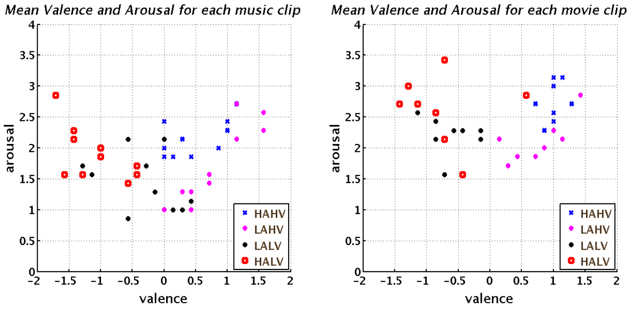


(a) (b)

Fig. 1. (a) Experiment set-up overview and (b) Protocol timeline for a single-trial acquisition.

we computed κ between the determined and the ground-truth labels, with $P(a) = 0.5$. The mean κ over all subjects for the music-valence, music-arousal, movie-valence and movie-arousal were found to be 0.5357, 0.2143, 0.7054 and 0.2321 respectively. This observation demonstrates that inter-rater agreement is higher, especially for valence, between the two subject populations (one used for ground-truth compilation, and the other performing the actual experiment), implying that movie stimuli are more effective in eliciting similar emotions across many subjects as compared to music videos.

To summarize the comparison between music videos and music clips, we observe that the valence-arousal distinction is better perceived for movies, and viewers are able determine clip valence independent of arousal for movie clips. Also, movie clips are rated more consistently across subject populations, indicating that they are more effective for emotion elicitation. We also observe better affect classification with MEG for movies as detailed in section V. The next section details the MEG feature extraction process prior to classification.



(a) (b)

Fig. 2. Mean self-assessment VA ratings for (a) music and (b) movie clips.

IV. MEG Data Analysis

This section describes in detail, the steps involved in (a) preprocessing the MEG data, (b) extracting MEG features and (c) determining the stimulus label for classification. MEG responses for both music and movie stimuli are processed in an identical manner.

A. Preprocessing MEG Data

The data preprocessing consists of four main steps that are handled using the MATLAB Fieldtrip toolbox [14]:

(a) **Trial Segmentation**: Participant responses corresponding to each trial are extracted by segmenting the MEG signal from 1 second prior to stimulus presentation (pre-stimulus) to the end of stimulus presentation. In this way, for each subject, we extract 32 and 40 trials for the movie clips and music videos respectively.

(b) **Baseline correction**: The mean MEG response amplitude in the last 200 ms of pre-stimulus is considered as the baseline for all trials, and this value is subtracted from the trial response. This step corrects stimulus-unrelated MEG signal variations over time.

(c) **Frequency domain filtering**: Low-pass and high-pass filtering with cut-off frequencies of 45 Hz and 1 Hz respectively are performed, as the relevant MEG frequency bands are between 3-45 Hz. Applying the high-pass filter, low frequency noise in the MEG signal generated by moving vehicles is removed. Conversely, the low-pass filter removes some high frequency artifacts generated by muscle activities (between 110Hz- 150 Hz) and electrical noise (50 Hz, 100 Hz and 150 Hz).

(d) **Channel correction**: Dead and bad channels are removed from the MEG data and replaced with interpolated values. Dead channels have zero value over time, while bad channels are outliers with respect to metrics such as signal variance and z-value of signal power over time. To preserve the consistency of MEG data over each trial and subject, removed channels are replaced with the average of neighbor channels.

B. Feature Extraction

Upon segmenting the MEG response for each trial, the most informative content for affect classification needs to be extracted. In MEG studies, the spectral power of certain frequencies is the popularly used feature. There are several methods for computing spectral power of signals like Hanning tapers, multitapers and wavelet. Multi-tapers and wavelet are typically used in order to achieve a better control over the frequency smoothing. In these methods, high frequency smoothing has been found to be principally beneficial when dealing with brain signals above 30 Hz [13], [16]. Therefore, we use the wavelet method to transform our signal to the time-frequency domain. We use a time-step of 1 second for temporal processing of the signal corresponding to each trial and a frequency step of 1 Hz to scan through a frequency range of 1-45 Hz.

Upon applying a wavelet transform on the MEG data, we perform the following steps: (a) In order to better elucidate the MEG response dynamics following stimulus presentation, baseline power corresponding to 1 second pre-stimulus interval is subtracted from the trial power. (b) Since magnetometers are highly prone to environmental noise, we only extract features from the gradiometer channels for better accuracy (*i.e.*, information from 102 of the 306 channels are discarded). (c) Then, we use a standard Fieldtrip function for combining the two planar gradiometers' spectral power for each sensor. This step enables a reduction in the number of spatial MEG features to 102.

Per subject and per movie clip, the output of the above process is a 3-dimensional matrix with the following dimensions: synthetic information of 102 sensors × clip length time points × 45 frequencies. Similarly, for each of the 40 music clips, the output dimensions are 102×60×45. We use this 3-D array to compute 4 different sets of features namely, *i)* full spatial, *ii)* compacted feature, *iii)* 27 DCT coefficients and *iv)* 64 DCT coefficients as follows:

(i) *Full spatial features* are computed by averaging the spectral power over time and four major frequency bands that are: theta (3-7 Hz), alpha (8-13 Hz), beta (14-29 Hz) and gamma (30-45 Hz). As we are preserving the data of all 102 sensors, these features contain full spatial information. Therefore, for each trial, a full spatial feature vector contains 408 (102×4 bands) features. However, full spatial features do not encode any temporal information.

(ii) *Compacted features* are computed by averaging the spectral power over all the channels, time, and each frequency band. Therefore, a compacted feature vector contains only 4 features that represent average spatial and temporal activities over the four frequency bands.

(iii,iv) *DCT features*- we apply a 3D Discrete Cosine Transform (DCT) on time-frequency spectral power for each trial. Then, we use the leading coefficients as our feature vectors. The *27 DCT Coeffs* and *64 DCT Coeffs* feature sets respectively contain 27 and 64 coefficients in their feature vectors. In comparison with the other features, the DCT features incorporate information from the spatial, time and frequency dimensions.

According to Ahmed *et al.*[1], signal information can be approximated effectively with few low-frequency DCT components. DCT is often used in signal, image and speech compression applications due to its strong energy compaction ability. Davis *et al.* [4] showed that the perceptually related aspects of the short-term speech spectrum contributed to superior performance of the mel-frequency cepstrum coefficients in such applications. Inspired by these works, we exploited DCT for compressing information encoded in time-frequency domain over channels. Here, we employ DCT to also retain temporal variations while using only a small number of features. The 3D DCT coefficient matrix, B, is calculated as:

$$B_{pqr} = \alpha_p \alpha_q \alpha_r \sum_{l=0}^{L-1} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A_{lmn} \cos \frac{\pi\,(2l+1)\,p}{2L}$$
$$\cos \frac{\pi\,(2m+1)\,q}{2M} \cos \frac{\pi\,(2n+1)\,r}{2N},$$
$$0 \leq p \leq L-1, 0 \leq q \leq M-1, 0 \leq r \leq N-1$$

where $\alpha_p = \begin{cases} \frac{1}{\sqrt{L}}, & p=0 \\ \sqrt{\frac{2}{L}}, & 0<p \leq L-1 \end{cases}$ $\alpha_q = \begin{cases} \frac{1}{\sqrt{M}}, & q=0 \\ \sqrt{\frac{2}{M}}, & 0<q \leq M-1 \end{cases}$

$$\alpha_r = \begin{cases} \frac{1}{\sqrt{N}}, & q=r \\ \sqrt{\frac{2}{N}}, & 0<r \leq N-1 \end{cases}$$

Here, L, M, N represent the number of spatial, time and frequency steps, respectively. A is the 3 dimensional matrix of power spectral features. Therefore, after computing 3D DCT coefficients, we only use a sub-cube constructed by the first n coefficients from each of 3 feature dimensions. The feature vector for each trial will now contain n^3 entries. For example, in the case of *27 DCT Coeffs*, we assign $n = 3$. The next subsection describes the classification procedure employing the four different feature sets, following which, we discuss the experimental results.

C. Classification procedure

For each of our 7 subjects, we have 32 trials for movie clips and 40 trials for music videos. We have also extracted 4 sets of features for each trial and we want to decode the affective trial responses using these feature representations. To achieve this goal, we solve two binary classification problems- employing the MEG features to differentiate between (i) *low* versus *high* arousal and (ii) *low* versus *high* valence. To this end, we need to associate a label to each of the stimuli for classification.

To assign a label to a particular stimulus, we employed a majority vote-based scheme as follows. We computed the median rating provided by viewers for each movie/music clip. The mean of these medians was then taken as the threshold value separately for movie/music clips. Clips whose median valence/arousal ratings were greater than the threshold were assigned a 'high' (valence/arousal) label, while others were assigned the 'low' label.

The distribution of 'high' and 'low' valence/arousal labels for the music and movie stimuli is presented in Table I. For both music and movie clips, the distribution of the 'high' and 'low' classes is unbalanced for both valence and arousal- the valence distribution for movies is the one that most resembles a balanced distribution. Also, for both music and movie clips, the distribution-bias along arousal is greater than for valence. Given this unbalanced distribution of stimuli, we use F1-scores alongside classification accuracies to report our classification results. Also, similar to [10], we use a naive-Bayes classifier to deal with class imbalance in small training sets. The leave-one-out cross validation scheme is employed in the classification framework. For each participant, we train the model with all-but-one stimulus ratings and the corresponding MEG responses, and use the model to predict the label of the remaining stimulus. The classification results are presented in the next section.

V. Experimental results

Table II shows measured classification accuracy and F1-scores for music videos and movie clips, respectively. To test for significance, the F1-distribution over participants is compared to the 0.5 baseline using an independent one-sample

	Music video clips		Movie video clips	
Class	Arousal	Valence	Arousal	Valence
High	28 (70%)	26 (65%)	12 (37.5%)	14 (44%)
Low	12 (30%)	14 (35%)	20 (62.5%)	18 (56%)

TABLE I

Distribution (numbers and percentage) of samples in each class obtained on stimulus labeling.

Music video clips	Arousal		Valence	
Feature Type	ACC	F1	ACC	F1
Full spatial	0.639	0.381	0.532	0.360
Compact Features	0.657	0.538	0.611	0.564
27 DCT Coefs	0.593	0.476	0.554	0.489
64 DCT Coefs	0.636	0.493	0.632	0.553
Movie video clips	**Arousal**		**Valence**	
Feature Type	**ACC**	**F1**	**ACC**	**F1**
Full spatial	0.567	0.479	0.518	0.480
Compact Features	0.585	0.500	0.531	0.493
27 DCT Coefs	0.656	**0.617***	0.549	0.505
64 DCT Coefs	0.607	**0.553***	0.554	0.503

TABLE II

Average accuracies (ACC) and F1-scores over participants for music video clips and movie video clips. Stars indicate whether the F1-score distribution over subjects is significantly higher than 0.5 according to an independent one-sample t-test (* = p <0.05).

t-test. As evident from the table, none of the classification results obtained for music videos are significant. On the other hand, two significant results (p<0.05) are obtained for arousal classification using DCT features in the case of movie clips. Based on the presented results, we summarize our key observations as follows:

1. The MEG-based affect characterization approach achieves above-chance F1 scores for arousal. Therefore, MEG signals effectively encode affective user responses. The fact that brain signals encode arousal better than valence is also observed in previous EEG studies such as [10]. However, EEG and MEG signals encode complementary aspects of the brain response- more investigation is required to determine whether arousal-related information encoded by MEG is similar to EEG or different.
2. Also, classification results confirm that MEG responses characterize affect better for movie clips than music videos. The fact that evoked emotions are more effective and consistent across subjects for movie clips suggests that they are perhaps, better stimuli to use for affective studies, as compared to music videos.
3. The best F1-measures are obtained with the DCT coefficients. The DCT coefficients encode time-related response patterns in addition to spatial and frequency information. This suggests that time-related information could be a critical factor for encoding affect. Again, investigation with more subjects is required to validate this hypothesis.

VI. Conclusion and Future Work

This paper presents the first work attempting single trial classification of affective video stimuli such as movie clips and music videos. Based on a study conducted with seven subjects, we observe that the MEG signal can effectively encode emotional responses of viewers. Our analysis also suggests that movie clips are more suitable than music videos for eliciting emotions in viewers as (i) valence-arousal ratings are found to be more consistent across subjects and (ii) MEG correlates better with emotional ratings for the movie clips. The obvious limitation of this study is that it involves a small number of subjects and future work involves (i) extending the study including more participants to validate the current findings and (ii) employing a multimodal approach involving peripheral physiological signals such as ECG, EMG and GSR in addition to MEG, to effectively characterize affective viewer responses.

References

[1] N. Ahmed, T. Natarajan, and K. Rao. Discrete cosine transfom. *IEEE Transactions on Computers*, 23:90–93, 1974.

[2] L. Barrett, B. Mesquita, and M. Gendron. Context in emotion perception. *Curr. Dir. Psychological Sc.*, 20(5):286–290, 2011.

[3] E. E. Bartolini. Eliciting emotion with film: Development of a stimulus set. Master's thesis, Wesleyan University, 2001.

[4] S. Davis and P. Mermelstein. Readings in speech recognition. chapter Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, pages 65–74. 1990.

[5] M. Greenwald, E. Cook, and P. Lang. Affective judgement and psychophysiological response: dimensional covariation in the evaluation of pictorial stimuli. *J. Psychophysiology*, 3:51–64, 1989.

[6] J. J. Gross and R. W. Levenson. Emotion elicitation using films. *Cognition & Emotion*, 9(1):87–108, 1995.

[7] A. Hanjalic and L.-Q. Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, 2005.

[8] U. Hasson, R. Malach, and D. Heeger. Reliability of cortical activity during natural stimulation. *Trends in Cogn. Sc.*, 14(1):40 – 48, 2010.

[9] H. Joho, J. Staiano, N. Sebe, and J. M. Jose. Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents. *Multimedia Tools Appl.*, 51(2):505–523, 2011.

[10] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis ;using physiological signals. *T. Affective Computing*, 3(1):18–31, 2012.

[11] P. Lang, M. Bradley, and B. Cuthbert. IAPS: Affective ratings of pictures and instruction manual. Technical report, U. Florida, 2008.

[12] C. L. Lisetti and F. Nasoz. Using noninvasive wearable computers to recognize human emotions from physiological signals. *EURASIP J. Adv. Sig. Proc.*, 2004(11):1672–1687, 2004.

[13] P. Mitra and B. Pesaran. Analysis of dynamic brain imaging data. *Biophysical Journal*, 76(2):691 – 708, 1999.

[14] R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen. Fieldtrip: Open source software for advanced analysis of MEG,EEG, and Invasive Electrophysiological Data. *Computational Intell. Neurosc*, 2011.

[15] P.-Y. Oudeyer. The production and recognition of emotions in speech: features and algorithms. *International Journal of Human Computer Interaction*, 59(1-2):157–183, 2003.

[16] D. Percival and A. Walden. *Spectral Analysis for Physical Applications*. 1993.

[17] P. Peyk, H. T. Schupp, T. Elbert, and M. Junghöfer. Emotion processing in the visual brain: a meg analysis. *Brain Topography*, 20(4):205–215, 2008.

[18] S. R. and P. O.A. Multivariate response patterning of fear and anger. *Cognition and Emotion*, 10(2):173–198, 1996.

[19] J. Schwarzbach. A simple framework (asf) for behavioral and neuroimaging experiments based on the psychophysics toolbox for matlab. *Behavior Research Methods*, pages 1–8, 2011.

[20] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3:42–55, 2012.