Emotion recognition and brain activity synchronization across individuals

Ayoub Hajlaoui

Thèse

Emotion recognition and brain activity synchronization across individuals

By

AYOUB HAJLAOUI




Department of Engineering Mathematics
UNIVERSITÉ PIERRE ET MARIE CURIE & TÉLÉCOM PARIS

Thesis director : **Pr. Mohamed Chetouani**, *Université Pierre et Marie Curie*
Thesis co-director : **Pr. Slim Essid**, *Télécom Paris*

Jury :

M. Fabien Lotte, CR (Hdr), *Inria Bordeaux Sud-Ouest*	rapporteur
M. Marco Congedo, CR (Hdr), *Grenoble INP*	rapporteur
M. Bruce Denby, Pr, *Université Pierre et Marie Curie*	examiner
M. Slim Essid, Pr, *Télécom Paris*	examiner
M. Mohamed Chetouani, Pr, *Université Pierre et Marie Curie*	examiner

A dissertation submitted to Université Pierre et Marie Curie & Télécom Paris in accordance with the requirements of the degree of DOCTOR in Mechanical sciences, acoustics, electronics & robotics.

SEPTEMBER 2018

DÉDICACE

À ma mère, qui m'a inculqué le goût des sciences depuis ma plus tendre enfance. Merci, Maman, de m'avoir imposé le Bus Magique pendant que les autres regardaient Action Man.

À mon père, qui m'a appris à ne jamais me complaire dans mes facilités, mais au contraire à aller jusqu'au bout de mes capacités. Je peux souffler, maintenant, Papa ?

À Safa, Chaima, et Amine, pour qui mon affection est sans commune mesure avec les trop rares moments que je leur ai consacrés ces dernières années.

À ma famille, et à tous mes amis qui ont sû me soutenir même pendant les moments les plus difficiles de cette tranche de vie qui s'achève.

À Mohamed et Slim, qui ont sû me secouer dans ces moments d'une thèse où le temps s'allonge et où les espoirs s'enlisent.

À mes amis et collègues du laboratoire, avec qui j'ai eu les discussions les plus intéressantes. Tu vois, Manu, maintenant que j'en suis réellement arrivé à cette dédicace dont nous avons tant parlé, je me dégonfle un peu. Peut-être trop mainstream à ton goût, mais on verra ce que tu écriras, toi.

À mes chers élèves, enfin, à qui j'ai eu la chance d'enseigner en TD ou en TP, et qui m'ont permis de me rendre compte à quel point j'affectionnais l'enseignement des mathématiques.

RÉSUMÉ

En fonction du point de vue, une émotion peut être comprise comme la représentation consciente de ce qu'un individu ressent (perspective psychologique), ou réponse complexe et moins consciente du corps à un stimulus émotionnel donné (perspective neuro-psychologique). Dans cette thèse, nous suivons la position décrite dans [1], à savoir "une position médiane tentant de définir comment les changements physiologiques ont lieu quand nos ressentis changent ".

Au sein des recherches sur les émotions humaines, l'informatique affective vise à permettre à des " systèmes intelligents de reconnaître, ressentir, déduire et interpréter " de telles émotions [2]. Une part importante des recherches en informatique affective se focalise sur la prédiction, à partir de données physiologiques, d'émotions produites chez un sujet par le biais de stimuli spécifiques. Usuellement, les sujets annotent l'émotion ressentie selon le plan valence/arousal [3].

C'est dans ce cadre que s'inscrit la mise en place de systèmes de reconnaissance automatique d'émotions, en parallèle avec la constitution de bases de données émotionnelles.

La reconnaissance automatique d'émotions s'effectue généralement de la manière suivante :

— Des émotions sont produites chez le participant par le biais de stimuli spécifiques. Dans le cadre de cette thèse, nous nous intéressons aux stimuli audiovisuels (vidéos). En parallèle de l'enregistrement de signaux physiologiques, le participant annote l'émotion ressentie. L'annotation peut se faire pendant ou après la stimulation.
— Ensuite, une représentation de caractéristiques (features) est choisie. Selon cette représentation, des features sont extraites de l'EEG du participant. Bien entendu, le choix de la représentation est crucial aussi bien pour la performance de la classification que pour l'explication physique des features.
— En utilisant les features extraites à l'étape précdente et les annotations du participant, un classifieur d'émotions est alors appris sur un set d'entraînemen et évalué sur un set de test, selon une métrique d'évaluation donnée.

Les travaux sur la reconnaissance automatique d'émotion sont basés principale-

ment sur des modalités comme la parole, l'expression faciale ou le regard [4–6]. Ces modalités sont principalement limitées par leur altérabilité, qu'elle soit volontaire ou non [7], limitation dont ne souffrent pas des signaux physiologiques comme les électro-encéphalogrammes (EEG). Ces derniers permettent de capturer des informations non observable de manière externe. C'est pourquoi l'EEG, dont il a été démontré qu'elle contient des indices précieux pour la classification d'émotion [8], attire l'attention des chercheurs en informatique affective. Et c'est pourquoi nous nous focalisons, dans cette thèse, sur la reconnaissance d'émotion à base d'EEG.

Traditionnellement, la reconnaissance d'émotion via EEG se fait par extraction de caractéristiques dans des bandes de fréquence prédéfinies, connues en neuro-sciences pour leur lien avec l'émotion : bandes alpha, bêta, gamma... Cette approche traditionnelle ne tient pas compte de la forte variabilité inter-sujet des réponses EEG à un même stimulus, en plus de nécessiter des connaissances a priori quant aux bandes de fréquence à considérer.

Une problématique centrale de la reconnaissance d'émotion à base d'EEG est la variabilité des réponses individuelles aux stimuli, que ce soit au niveau émotionnel ou physiologique. En effet, d'un sujet à l'autre :

— le même stimulus peut produire des émotions différentes [9]
— une même émotion annotée peut correspondre à différentes réponses physiologiques d'un sujet à l'autre [10].

La tendance en machine learning consiste en l'apprentissage de représentations adaptées à la tâche de classification. Un cadre robuste d'extraction automatique de features devrait permettre de résoudre le problème de la dépendance de l'EEG aux sujets. Dans cette optique, un dictionnaire commun représentant les données peut être appris à partir du set d'entraînement. Ensuite, les données sont projetées sur ce dictionnaire pour obtenir des features de classification. Par l'apprentissage de dictionnaire, on recherche la " reprśentation appropriée de sets de données par le biais de sous-espaces à dimensions réduites. " [11].

Dans ce contexte, nous utilisons la Nonnegative Matrix Factorization (NMF) [12] qui permet, à partir de la matrice de densité spectrale de puissance, d'extraire un dictionnaire d'atomes fréquentiels et une matrice d'activation de ces atomes. L'activation des atomes est ensuite utilisée pour entraîner des classifieurs de valence/arousal.

Bien que l'utilisation de la NMF mène globalement à une amélioration des résultats (en comparaison avec une baseline de features traditionnelles) sur les bases de données HCI MAHNOB [13] et EMOEEG [14], cette amélioration n'est pas encore satisfaisante. En intra-sujet, les résultats de classification varient encore beaucoup d'un sujet à l'autre. En inter-sujet, les améliorations observées dépendent la baseline et de la dimension (valence/arousal).

D'où l'idée, en inter-sujet, de rendre l'apprentissage de représentation sensible aux variations entre sujets. Une variante de la NMF, la Group NMF [15, 16], permet une telle considération. Il s'agit de faire en sorte que certains atomes du dictionnaire appris présentent une certaine similarité s'ils sont extraits à partir des données du même sujet. Mais en comparaison avec la NMF simple, une telle configuration de GNMF n'améliore par les résultats de classification. Par ailleurs, on se rend compte que les résultats dépendent beaucoup du niveau d'émotion annotée. Ainsi, la classification est moins performante lorsque l'arousal est faible.

Le constat de cette dépendance vis-à-vis de la nature de l'émotion a motivé notre étude de l'effet de cette dernière sur la corrélation des signaux EEG entre sujets qui ont regardé le même stimulus. Pour quantifier cette corrélation, nous avons utilisé l'Inter-Subject Correlation (ISC) [17–19], en proposant différents schémas de calcul. En étudiant les variations du score d'ISC en fonction du niveau de valence et d'arousal annotés, nous avons constaté une augmentation significative du score d'ISC lorsque l'arousal augmente, et une diminution de ce score lorsque la valence augmente. Cela permet de fournir une explication quant à la dépendance observée des performances de classification vis-à-vis de l'émotion.

Forts de cette nouvelle information, nous avons alors décidé de redéfinir notre manière d'utiliser la GNMF. Au lieu de définir les groupes par sujet ou session, nous les définissons désormais par le niveau de valence et d'arousal annotés. L'apprentissage de features se fait alors de manière multi-tâche, l'information relative aussi bien à la valence et l'arousal servant à l'apprentissage de features pour classifier les deux dimensions. Cependant, dans les fonctions objectif à minimiser, les paramètres relatifs aux similarités de la GNMF varient pour la classification de chacune desdites dimensions. Cette nouvelle GNMF (GNMF-val/aro) offre de bien meilleurs résultats que la précédente. L'apprentissage de features par niveau d'émotion semble donc plus porter ses fruits que celui par sujet.

Cette utilisation de l'ISC est indirecte : en effet, la variation de l'ISC en fonction de la valence et de l'arousal nous a donné l'idée de définir nos groupes en fonction de ces dernières. Dès lors, pourquoi ne pas définir directement les groupes de la GNMF par le score d'ISC, au lieu de passer par l'intermédiaire valence/arousal ? C'est ce que nous avons fait sur la base de données HCI (les sujets de EMOEEG n'ayant pas tous vu les mêmes vidéos).

Dans un premier temps, nous avons pris en compte l'ISC de manière légère, dans l'étape d'apprentissage du classifieur, en pondérant les observations par le score d'ISC. Cette première initiative n'a pas donné des résultats sensiblement différents de GNMF-val/aro.

Ensuite, nous avons décidé de prendre l'ISC en compte plus en amont, lors de l'apprentissage de features. Au lieu de discriminer les features en fonction des niveaux de

valence/arousal, nous les discriminons uniquement en fonction du niveau d'ISC discrétisé (bas/haut). Ce nouveau schéma de GNMF (GNMF-ISC), où les groupes sont définis en fonction du niveau d'ISC, donne des scores encore plus hauts que GNMF-val/aro.

Ces résultats placent donc l'ISC au coeur de la problématique de la reconnaissance de l'émotion via EEG. Des travaux futurs porteront non plus sur l'utilisation de l'ISC discrétisé pour définir des groupes de GNMF, mais sur l'incorporation directe du score d'ISC continu dans la fonction objectif de la NMF.

Au-delà de l'effet de l'ISC à proprement parler, la reconnaissance d'émotion via EEG reste fortement tributaire de la taille des bases de données utilisées, que ce soit en nombre de sujets ou en nombre de stimuli présentés à chaque sujet. C'est ce nombre-là qui, décuplé, pourrait permettre une plus grande efficacité de la GNMF, en s'assurant que les données soient assez nombreuses pour que l'extraction de features se fasse avec précision.

Une autre question peu approfondie au cours de cette thèse concerne les différences entre techniques d'annotation de l'émotion. Bien que nous nous soyons focalisés sur les dimensions classiques valence/arousal, l'utilisation de descripteurs plus qualitatifs (emotional words) pourrait modifier la manière de concevoir la GNMF.

Remerciements

Je tiens à remercier chaleureusement mes encadrants de thèse, Mohamed Chetouani et Slim Essid, qui m'ont fait profiter de leurs connaissances, de leurs bons conseils, et ont surtout sû m'apporter la motivation nécessaire tout au long de cette thèse.

Je remercie aussi tous ceux qui, à l'ISIR et à Télécom Paris, m'ont transmis des informations cruciales, ont été de bon conseil, et avec qui j'ai eu les discussions les plus riches. Notamment Anne-Lise Jouen, qui a guidé mes touts-premiers pas expérimentaux en EEG, et Anne-Claire Conneau, dont le travail remarquable sur la base de données EMOEEG a été très utile pour cette thèse. Je remercie aussi André-Marie Pez et Fanny Roussel pour leur contribution technique.

Enfin, je remercie Marco Congedo et Fabien Lotte d'avoir accepté d'être rapporteurs de cette thèse, et Bruce Denby d'avoir accepté de faire partie du jury.

TABLE OF CONTENTS

LISTE DES TABLEAUX

TABLE DES FIGURES

CHAPITRE 1

INTRODUCTION

Emotion can be defined as a "distinct, integrated, psycho-physiological response system", an "organized highly structured reaction to an event that is relevant to the needs, goals, or survival of the organism" [20]. In particular, an emotion must be distinguished from a mood, which is also a transient episode of affect. The main differences are that :

— the duration of the episode is typically shorter in the case of the emotion
— emotions are response systems activated by specific stimuli. The use of such stimuli to this end is referred to as emotion elicitation.

Depending on the standpoint, an emotion can be understood either as the conscious representation of what an individual feels (psychological perspective), or the complex and less conscious body response to a given emotional stimulus (neuro-psychological perspective). In this thesis, we follow the position described in [1], that is "a middle position by trying to define how physiological changes occur when our feelings change", the assumption being that "when participants recognize their emotions well, the association between physiological data and perception of different feelings will be reliable."

Among the research on human emotions, affective computing is a large field that aims at enabling "intelligent systems to recognize, feel, infer and interpret" these emotions [2]. Therefore, a significant part of the investigations in affective computing research seeks to predict the emotions elicited from a subject using specific stimuli based on the subject's physiological responses to such stimuli. In line with this effort, automatic emotion recognition systems are set up, motivating the constitution of numerous emotional

databases.

Contributions to automatic emotion recognition mainly rely on modalities such as speech, facial expressions, or eye gaze [4–6]. The main limitation of these modalities is their alterability, whether voluntary or not [7]. On the other hand, physiological modalities such as Electroencephalography (EEG) do not suffer from such a drawback. As stated in [21], "EEG signals are directly recorded from human's brain cortex and hence they could be more reliable in reflecting the inner emotional states of the brain", with a remarkable advantage in comparison to other physiological modalities : the information EEG can capture is not necessarily observable externally. Thus, EEG has attracted the attention of researchers in the field of affective computing and it has been shown to hold precious cues for emotion classification [8]. This has motivated the focus on EEG-based emotion classification in this thesis.

To perform EEG-based emotion classification, one has to cope with the variability of individual responses to stimuli, whether it be at the emotion level or at the physiological signal level. Indeed, from one subject to another :

— the same stimulus can elicit different emotions [9] ;
— the same elicited emotion translates into different physiological responses across participants [10].

Many factors of decision can have an impact to address this stability issue. Which stimuli to use ? Should emotion classification be done individually or in an inter-subject fashion ? Which features to extract from the physiological data ? How should such features be normalized ? And, more deeply, how to take into account the variabilities exposed above in the chosen feature representation ? These have been our focuses in this thesis.

1.1 Stimuli choice

Even if they induce harder emotion classification tasks than image stimuli, audiovisual stimuli offer the advantage of eliciting dynamic emotions, which is more consistent with realworld applications. Therefore, this thesis focuses on the use of audiovisual stimuli.

The duration of such stimuli is chosen such that it is both not too long to hamper the subject's concentration, and not too short in order to capture the dynamics of emotion. As

the order of magnitude of emotional reactions length was found to be around 10 seconds [22], a usual order of magnitude for the duration of audiovisual stimuli is 20-30 seconds.

Across the available emotional databases, different types of audiovisual stimuli are used, according to the considered task. Such stimuli are often short movie excerpts [13] or music videos [23], which induces different elicitation. Some databases can also focus on specific types of emotions, such as negative ones [14].

1.2 Emotion annotation

Given a choice of stimuli, emotion has to be accurately translated by the participant. This raises the issue of emotion annotation, that is to say the assessment by the participant of the emotion he/she felt as a result of each stimulus.

To be more accurate, emotion annotation can take the form of a verbal description using specific keywords. However, a scalar representation offers the advantage of both systematizing and simplifying the annotation. To this end, emotions are often represented in a two-dimensional valence-arousal space [3], which respectively describe the pleasure or displeasure felt by a person and her degree of excitement. In Figure 1.1, some specific emotions are placed onto this space.

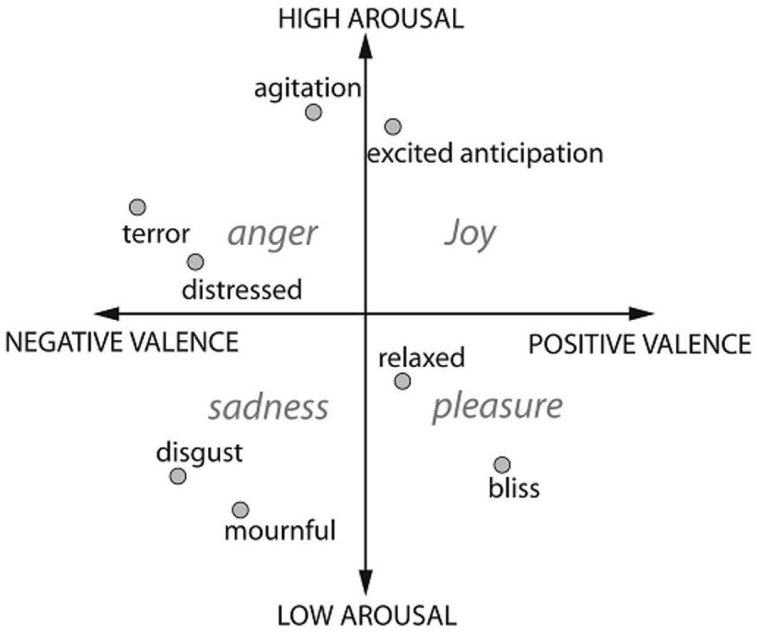

FIGURE 1.1 – The valence-arousal space

In this thesis, emotion annotation corresponds to a double scalar annotation, that is to say valence and arousal information. Valence and arousal annotation can either be continuous or discretized. Even though a discrete representation on the valence and arousal axes "may not reflect the subtlety and complexity of the affective states" [24], such a discretization is a straightforward way of obtaining meaningful labels with a view to elicited emotion classification. Most discretization models decompose each axis into two or three labels, respectively low/high and low/average/high.

The emotional state (valence-arousal) self-assessment by the participants can be made in an online fashion while watching the stimulus, or after the end of its exposure. While the first option can help capture the variations of valence and arousal more accurately in the audio-visually stimulated case, it might hamper the participants' concentration towards the stimuli.

As for the annotation itself, it can either globally describe the stimulus, or be decomposed so as to describe sub-parts of the stimulus, in order to capture the dynamics of emotion.

1.3 Factors of variability for the EEG response

As stated earlier, numerous factors affect the EEG response stability, making EEG-based emotion classification a challenging task.

The same stimulus can elicit different emotions among individuals. For instance, as made clear in [25], age differences can induce valence/arousal rating differences for the same stimulus. Gender differences have also been found to have an effect in the rating of negative emotions [26]. Differences of annotation can be caused either by theses differences in the emotion felt or by inter-subject variability of emotion representation, that is to say the way subjects interpret emotional keywords or valence/arousal axes. This results in high inter-subject variability of the EEG responses to the same stimuli [27].

Along with other works, the distinction made in [26] between negative and positive emotions when it comes to gender differences makes it clear that inter-subject variability depends on the emotion type. This is also emphasized by many classification results such as the ones obtained in [13], which shows differences of classification performance according to the emotion type.

Naturally, this high inter-subject variability of EEG responses results in two setbacks for emotion classification :

— intra-subject emotion classification performance varies a lot from one subject to another [28]

— inter-subject emotion classification tasks are complex because the generalization of features across subjects is difficult. Therefore, compared to intra-subject tasks, inter-subject classification performance is deteriorated.

In order to perform valid intra-subject analysis, we need EEG emotional datasets with enough experimental repetitions for each subject, so that enough subject specific information is available. This raises the issue of the subject's fatigue : if we want both enough repetitions per subject and to avoid any fatigue, multiple sessions for the same subject should be considered. Another related issue is thus raised, that is inter-session variability of the EEG signal. From an inter-subject classification point of view, enough subjects should participate to the experiments so that the problem of inter-subject variability, which remains a challenging issue, could be tackled. More focus on EEG emotional datasets is made in Chapter 2.

1.4 Objective and contributions

This thesis aims at introducing original EEG-based emotion classification methods that take into account factors of variability in EEG responses to audiovisual emotional stimuli. To this end, our contributions to the problem are the following :

— Features that are classically extracted from EEG data to perform emotion classification are the spectral power for each considered electrode in specific frequency bands (theta, slow alpha, alpha, beta, gamma) that are well known for their role in emotional and cognitive processes [29, 30]. Spectral moments of different orders and heuristic spectral shape descriptors have also been used [13, 31]. In the multi-channel case, the spectral power asymmetry between specific pairs of electrodes can be computed in the frequency bands mentioned earlier [32]. Other approaches such as Common Spatial Patterns (CSP) [33–35] rather focus on the spatial aspect of the activity on the skull.

Representations used in previous works have in common the fact that they rely on expert knowledge and a feature engineering effort. The new trend in machine

learning is to learn representations adapted to the subsequent classification stage. Along this line, Nonnegative Matrix Factorization (NMF) [12], which is an an unsupervised feature extraction method, has been mostly used for EEG-based motor imagery classification tasks [36]. **We use NMF to perform intra and inter-subject EEG-based emotion classification, extracting dictionaries of frequency atoms from EEG spectrograms.** The activation information of these atoms are then used as features for emotion classification.

— Noticing the high inter-subject variability of intra-subject classification results and the unsatisfactory inter-subject classification results, we were attracted by Group NMF (GNMF) [15]. Given predefined sub-parts of the data, this method extracts dictionaries separately and constrains specific similarities. **We use GNMF to extract NMF atoms subject-wise, atoms among which some were constrained to be similar across subjects.** No visible improvement is observed compared to NMF.

— Our previous results as well as many results in the literature show different classification performance across levels of valence/arousal. This motivates an analysis of the valence/arousal level effects on the correlation between EEG responses of subjects watching the same stimuli. Thus, **we analyze the effects of valence/arousal on EEG Inter Subject Correlation (ISC) [17–19]. We find significant links between the valence/arousal levels and ISC. A particular care was given to the statistical validity of the observed ISC variation along valence and arousal dimensions, using computationally intensive randomization tests.**

— We adjust our Group NMF model accordingly. Rather than extracting dictionaries of atoms subject-wise as made earlier, **we used Group Nonnegative Matrix Factorization in a multi-task fashion, where both valence and arousal labels are exploited to control valence-related and arousal-related feature learning. Some improvement was observed for emotion classification results. The results are further improved with the explicit use of ISC information in the feature learning stage.**

1.5 Organization of the document

This thesis is organized as follows :

— Chapter 2 presents preexisting EEG emotional databases, as well as commonly

extracted features for EEG-based emotion classification. Classification results obtained using these features on some databases are also exposed.

— In Chapter 3, the NMF and Group NMF approaches are detailed, and we expose our NMF and Group NMF-based emotion classification. In this chapter, GNMF atoms are extracted subject-wise, atoms among which some were constrained to be similar across subjects.

— In Chapter 4, the Inter Subject Correlation framework is exposed. Then, the effect of valence/arousal on Inter Subject Correlation (ISC) is analyzed.

— Finally, following the conclusions of Chapter 4, Chapter 5 presents the adjustment of Group NMF to a valence/arousal-based definition of sub-groups.

CHAPITRE 2

BASELINE EEG EMOTION CLASSIFICATION

In this chapter, we present the procedure classically followed to perform an EEG-based emotion classification task, as well as available databases in the case of audiovisual elicitation, and classification results obtained on such databases. As presented in Figure 2.1, a usual emotion recognition task is carried out as follows :

— Emotion is elicited from a participant by means of specific stimuli. In other words, stimuli are used in order to activate emotional responses in the participant. During emotion elicitation, physiological data concerning the participant - EEG in our case - is recorded. Along with the recording of physiological information, the participant assesses his/her emotional state, either during or after the stimulation.

— Then, one has to choose a feature representation, according to which EEG-based features are extracted from the participant's data. As stated in [21], "the target of emotional EEG feature extraction is to seek a set of optimal features that characterize the emotion information of the raw EEG signals". Naturally, the choice of feature representation is crucial in classification performance, but also in the physical explanation given to the features. In this thesis, the focus is made on this step.

— Using the features extracted in the previous step and the participant's annotations, an emotion classifier is then learned on a training set, and finally evaluated on a test set, according to given evaluation metrics.

In Section 2.1, we present different stimuli used in EEG-based emotion classification tasks, as well as the requirements needed by such tasks. Section 2.2 presents available

EEG-based affective datasets, whereas Section 2.3 reports commonly used features in such tasks. Section 2.4 is a reminder of the usual procedure in classifier training and evaluation metrics. Finally, in Section 2.5, we study the influence of feature and other parameters on classification results.

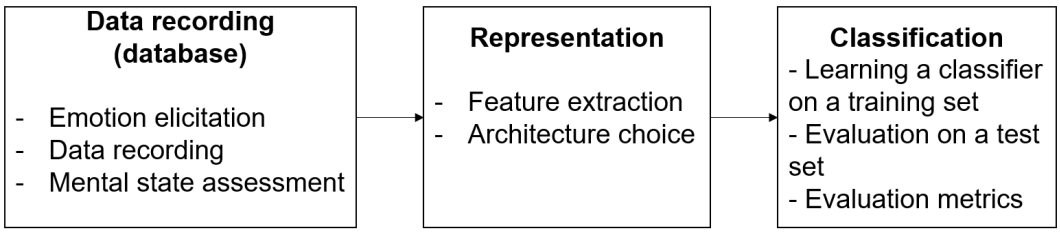

FIGURE 2.1 – Usual steps of an emotion recognition task

2.1 Emotion elicitation and EEG acquisition

For a given subject, we call trial the combination of one elementary emotion elicitation (using one stimulus) and the self annotation (by the subject) of the emotion felt. For instance, as shown in Figure 2.2, the EMOEEG database protocol [14] requires that the participant annotates his/her emotion right after each stimulus. It is also the case for the two other databases used in this thesis : HCI MAHNOB [13] and DEAP [23].

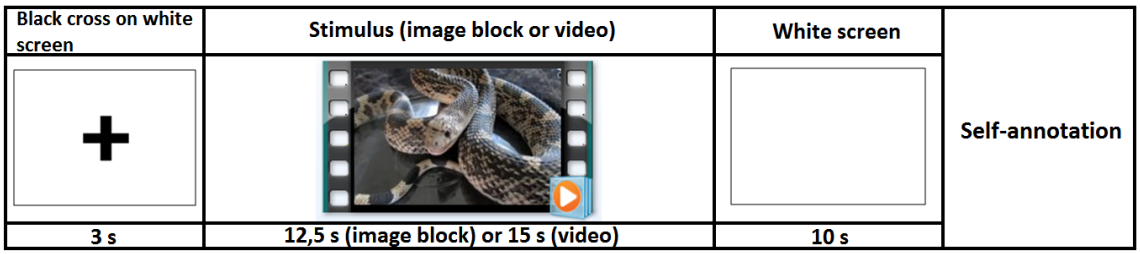

FIGURE 2.2 – Protocol for one trial (EMOEEG)

In the audiovisual stimuli case, an alternative to post-stimulus assessment is to make the participant assess his/her emotional state dynamically, while the stimulus is watched, as it is the case for the Feeltrace [37] and Gtrace [38] annotation methods. Even if such methods enable dynamic annotation, that is to say annotation which takes emotion variation across time, they have two major drawbacks, as stated in [39] :

— as the dynamic annotation has to be made while watching the video, it induces a lack of concentration, that can only be tackled by watching each stimulus twice,

which results in an increase of the experimentation duration an the participant's fatigue.

— watching each stimulus more than once may induce a habituation effect that would influence the participant's annotation

FIGURE 2.3 – Participant during a trial

Emotional stimuli can have different natures, depending on the focus. One can get interested in emotion recognition during music listening [40, 41]. Others have used images or image blocks as stimuli [42–44], using pictures from databases such as the International Affective Picture System (IAPS, [45]). Musical stimuli present the disadvantage that " subjects are prone to misunderstand positive/negative valence as preferred/not preferred " [46]. For instance, a music can be appreciated by the listener even if it makes him/her sad. As for image stimuli, even if they are an efficient way of eliciting emotion, they do not offer dynamic emotional responses. Therefore, in this thesis, the focus is put on audio-visually stimulated emotions, in order to be closer to realworld stimulation.

During each trial, the EEG signal is acquired by means of an EEG headset, as shown in Figure 2.3. An EEG headset is usually composed of 20, 32, or 64 electrodes. The names and positions of each electrode are defined by the 10-20 international system [47]. Figure 2.4 [48] shows the positions of 20 electrodes on the skull.

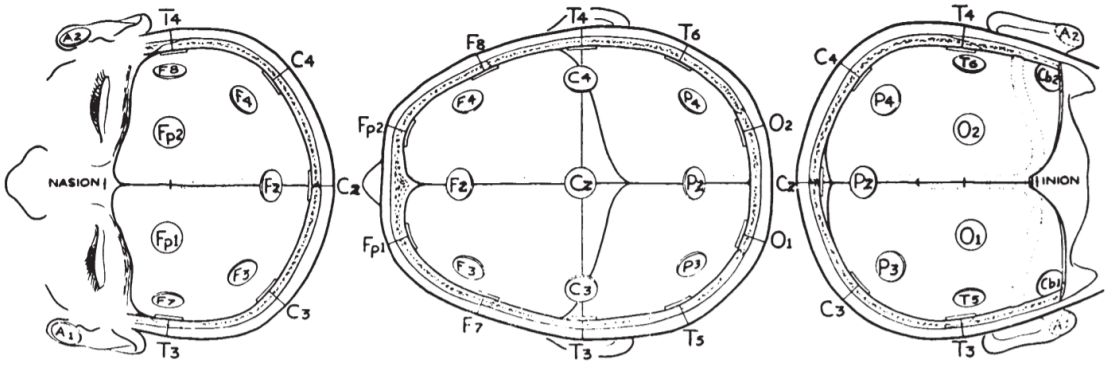

FIGURE 2.4 – Electrodes names and positions following the 10-20 system

2.1.1 Specific requirements

Each trial has to last long enough so that the information extracted from the EEG signals is sufficient. It is all the more important given that within the frequencies of interest, there are relatively low frequency bands such as alpha (8-12 Hz) and theta (4-8 Hz) frequency bands [30, 49]. The duration of a single trial should correspond to enough periods of such considered frequencies. On the other hand, the total duration of an experiment should not be too long so as to avoid participants' loss of concentration over time. Therefore, a usual order of magnitude for the duration of one trial is 15-20 seconds [13, 50].

Then, according to the desired classification task, additional requirements have to be fulfilled :

— if the task is inter-subject classification, enough subjects are needed so that the features generalize well across subjects. To address this challenge of assessing the generalization abilities of EEG-based classification systems across subjects, some existing databases such as HCI MAHNOB and DEAP [13, 23] included a relatively high number of participants (respectively 27 and 32 each).
— if the task is intra-subject classification, enough trials per subject are needed to provide each subject-dependent classifier with enough training data. Other databases such as eNTERFACE'06 and EMOEEG [14, 44] (with respectively 5 and 8 participants) chose to sacrifice the number of subjects for the benefit of this consideration (with respectively 30 and 50-100 trials per participant).

2.2 EEG-based affective datasets

Emotion recognition databases are numerous [51], but they mainly rely on modalities such as speech, facial expressions, or eye gaze. To the best of our knowledge, only a few EEG-based emotion recognition databases are publicly available. Tables 2.1 and 2.2 list those databases. In this thesis, the datasets used are HCI MAHNOB, DEAP, and EMOEEG.

TABLE 2.1 – EEG-based affective datasets

Name	Authors/year	Nature of stimuli
eNTERFACE'06	Savran et al. [44] (2006)	12.5 s image blocks
DEAP	Koelstra et al. [23] (2012)	One-minute music videos
HCI MAHNOB	Soleymani et al. [13] (2012)	Movie and video excerpts
EMOEEG	Conneau et al. [14] (2017)	Movie and video excerpts

EMOEEG, HCI MAHNOB and DEAP are multi-modal datasets where physiological responses to both visual and audiovisual stimuli are recorded, along with videos of the subjects, with a view to developing affective computing systems, especially automatic emotion recognition systems. The experimental setups involve various physiological sensors, among which electroencephalographic, electrocardiographic, electromyographic and electro-oculographic sensors, in addition to skin conductance data.

TABLE 2.2 – EEG-based affective datasets (important figures)

Name	Nb of stimuli per participant	Nb of sessions	EEG channels
eNTERFACE'06	90	5	54
DEAP	40	32	32
HCI MAHNOB	20	27	32
EMOEEG	50	11 sessions (8 subjects)	20

EMOEEG's experiment is performed with 8 participants, 4 from both genders. The stimuli include both sequences of static images from the IAPS dataset, and short video excerpts focusing on negative fear-type emotions. We only use audio-visual trials from this database. The annotation is obtained by participant self assessment, after a calibration phase.

EMOEEG stimuli focus on negative fear-type emotions. This choice is motivated by the development of strategies amenable to the analysis of the impact of violent videos on humans, and possibly treatments for subjects suffering from phobia. Thus, in terms of valence and arousal, there is a bias towards negative emotions in the choice of video

stimuli.
The originality of this database lies in three main aspects :

- an important number of repetitions were performed per subject for the purpose of a reliable intra-subject classification. Indeed, EEG responses are known to be strongly individual-specific
- a calibration phase which allows each participant to become familiar with the emotion annotation axes.
- a novel simplified dynamic annotation strategy used on video stimuli allows to consider the variations over time of felt emotion, and enhance the quality and consistency of the self-assessments.

As for HCI MAHNOB, it contains the recordings of 27 participants. We used 24 of theses sessions for valence classification and 23 for arousal classification. In each session, the participant watches 20 emotional videos. Thus, HCI MAHNOB contains more sessions than EMOEEG but less videos per session.

DEAP contains the recordings of 32 participants, even more than HCI MAHNOB, with more stimuli per participant (40). However, the nature of stimuli - music videos - is quite different from HCI MAHNOB and EMOEEG.

2.3 Commonly used features for EEG-based emotion classification

Features that are used for EEG-based emotion classification can be divided into three categories : time domain features, frequency domain features, and time-frequency domain features. In some reviews like [52] (Kim et al, 2013), such features are divided into only two categories, namely time domain and time-frequency domain features.

2.3.1 Time domain features versus time-frequency domain features

Classic time domain features such as the mean, power, or standard deviation, can be extracted from the EEG signals. More complex features, commonly used in time series analysis, such as first differences, second differences, kurtosis, or Hurst exponent, have also been used. Finally, time domain features were specifically for EEG analysis : for instance, the Hjorth features [53] named activity, mobility and complexity. Table 2.3 lists

previous works where EEG time domain features were used in image and video-elicited emotion classification tasks. The performances obtained using such features are also indicated.

TABLE 2.3 – Time domain features used in EEG-based classification of image and video-elicited emotion (val stands for valence, arsl for arousal, std for standard deviation, skew for skewness)

Authors/year	Electrodes	Features	Classes	Score
Takahashi et al. [54] 2004	3	- Mean, power, std - 1st and 2nd differences (diff.) - Normalized 1st and 2nd diff.	5	0.41 (mean F1-score) Intra-subject
Brown et al. [55] 2011	8	- Max, kurtosis - + freq-domain features	3	82 % (accuracy) Intra-subject
Conneau et al. [31] 2013	54	- Min, max, skew, kurtosis, mean, std, median, mean/max of 1st and 2nd diff. absolute values	2	val 70% (accuracy) Intra-subject Intra-subject Intra-subject
Valenzi et al. [1] 2014	8	- $\delta, \alpha, \beta, \gamma$ PSD (FFT)	4	97.2% (accuracy) Intra-subject
Wang et al. [56] 2014	8	- approximate entropy - Hurst exponent + freq-domain features	2	87.53% (accuracy) Intra-subject
Atkinson et al. [57] 2016	14	- Median, std, kurtosis - Hjorth features (activity, mobility, complexity) + freq-domain features	3	val 66.33% (accuracy) arsl 60.7% Intra-subject (on DEAP)

As for time-frequency domain features, commonly extracted features for EEG-based emotion classification are the Power Spectral Density (PSD) for each considered electrode in specific frequency bands (theta, slow alpha, alpha, beta, gamma) that are well known for their role in emotional and cognitive processes [29, 30]. For instance, "EEG alpha bands reflect attentional processing and beta bands reflect emotional and cognitive processing in the brain", according to Rowland et al. [49] and Klimesch et al. [58]. Spectral moments of different orders and heuristic spectral shape descriptors have also been used [13, 31]. In the multi-channel case, the spectral power asymmetry between specific pairs of electrodes can be computed in the frequency bands mentioned earlier [32]. Other approaches such as Common Spatial Patterns (CSP) [33–35] rather focus on the spatial aspect of the activity on the skull.

TABLE 2.4 – Frequency and time-frequency domain features used in EEG-based classification of image and video-elicited emotion (val stands for valence, arsl for arousal, std for standard deviation, skew for skewness)

Authors and year	Elect.	Features	#classes	Score
Davidson et al. [59] 1992	8	- α PSD	2	Statistical diff.
Murugappan et al. [60] 2007	63/24	- Entropy & energy of 4th level detail coeffs (by DWT)	3	Clustering
Khosrowabadi et al. [61] 2010	8	- Magnitude Squares Coherence Estimate	4	84.5 % (accuracy) Intra-subject
Koelstra et al. [33] 2010	32	- PSD band powers - CSP	2	val 58.8 %, arsl 55.7 % Intra-subject (accuracy)
Murugappan et al. [62] 2010	64	- Energy, power, std - RMS, REE, LREE, ALREE	5	83.3 % (accuracy) Intra-subject
Brown et al. [55] 2011	8	- Peaks of asym. α avg power + time domain feature	3	82 % (accuracy) Intra-subject
Nie et al. [63] 2011	62	- $\delta,\theta,\alpha,\beta,\gamma$ PSD (FFT)	5	83.3 % (accuracy) Intra-subject
Park et al. [64] 2011	32	- α,β,γ PSD (FFT)	5	Statistical diff. Intra-subject
Soleymani et al. [65] 2011 (DEAP)	32	- θ, slow α, α,β,γ PSD - θ, α, β, γ differential asym. (Welch's method)	2	val 0.58 (mean F1) arsl 0.56 Intra-subject
Soleymani et al. [13] 2012 (MAHNOB-HCI)	32	- θ, slow α, α,β,γ PSD - θ, α, β, γ differential asym. (Welch's method)	3	val 0.56 (mean F1) arsl 0.42 Inter-subject
Duan et al. [66] 2013	62	- $\delta,\theta,\alpha,\beta,\gamma$ PSD (FFT) - diff. & rational asymmetries - differential entropy (DE) - DCAU (spatial DE ratios)	2	84.25% (accuracy) Intra-subject
Rozgić et al. [67] 2013	32	- θ, slow α, α,β,γ PSD (FFT) - differential asymmetries (DASM)	2	val 76.9% (accuracy) arsl 69.1% Intra-subject (on DEAP)
Conneau et al. [31] 2013	54	- CSP on θ, α, β, γ (and all frequencies) - Heuristic spectral shape descriptors	2	val 78% (accuracy) Intra-subject
Valenzi et al. [1] 2014	8	- $\delta,\alpha,\beta,\gamma$ PSD (FFT)	4	97.2% (accuracy) Intra-subject
Wang et al. [56] 2014	8	- $\delta,\theta,\alpha,\beta,\gamma$ PSD (FFT) - differential asymmetries - wavelet features + time domain features	2	87.53% (accuracy) Intra-subject
Zheng and Lu [68] 2015	4-12	- $\delta,\theta,\alpha,\beta,\gamma$ PSD (FFT) - diff. & rational asymmetries - DE - DCAU	3	86.7% (accuracy) Intra-subject
Atkinson et al. [57] 2016	14	- θ, slow α, α,β,γ PSD + time domain features	3	val 66.33% (accuracy) arsl 60.7% Intra-subject (on DEAP)

The comparison results obtained by Wang et al. [56] and Conneau et al. [31] (2014) suggest the superiority of power spectrum features (time-frequency domain) over time domain features for EEG-based emotion classification. In addition, in the time-frequency domain, even if wavelet features are often used in EEG analysis, it was shown in [56] that they are inferior to power spectrum features in the case of EEG-based audio-visually stimulated emotion classification.

2.3.2 Exploiting spatial information

Many studies have proven the importance of spatial information in EEG-based emotion classification tasks. Davidson et al. (1982) [69] established a link between frontal EEG asymmetry and valence. Then, Cacioppo (2004) [70] put the emphasis on α band power for this very link. Sammler et al. (2007) [71] have shown that pleasant music is associated with an increase of frontal mid-line theta power. Jenke et al. (2014) [72] underlined the importance of parietal and centro-parietal lobes in EEG-based emotion classification feature engineering. For valence classification, Wang et al. (2014) [56] extracted subject-independent features of interest on right occipital lobe and parietal lobe in α band, parietal lobe and temporal lobe in β band, left frontal lobe and right temporal lobe in γ band. Common Spatial Patterns (CSP) [33–35] takes into account this spatial aspect of the activity on the skull.

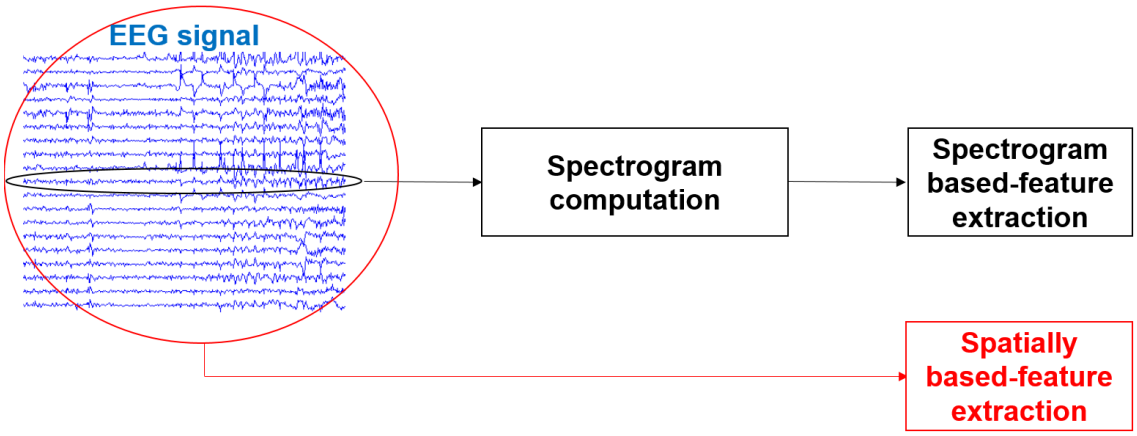

FIGURE 2.5 – Spectrogram-based and spatial distribution-based (spatially based) feature extraction. Spatially based feature extraction is made multi-channel-wise, and can include spectral features.

More recently, novel dynamical graph convolutional neural networks methods have

been used to "learn the intrinsic relationship between different EEG channels " [73], therefore exploiting spatial information to perform more discriminative feature extraction.

However, single-channel based emotion classification opens the way to easier applicability in real-world scenarios with more lightweight devices than full headsets. Therefore, the contributions of this thesis were made in the context of single-channel based emotion classification, and focus more on the spectrogram obtained from one given electrode than on the relationship between electrodes.

2.4 Classifier training and evaluation metrics

Using the extracted features, a classifier is trained on a given subset of the trials (as well as the corresponding annotations) and then tested to classify the remaining trials.

— Let us first study the case of intra-subject classification. If leave-one-out classification is performed, a classifier is trained on all trials but one, and then tested to classify the remaining trial. This procedure is then repeated for each trial, to obtain test labels. These test labels are finally compared to the ground truth, computing a given metric to evaluate classification performance. If k-fold classification is performed (for a given integer k), a classifier is trained on a proportion $\frac{k-1}{k}$ of all the subject's trials, and then tested to classify the remaining $\frac{1}{k}$.

— In the case of inter-subject classification, we use a leave-one-subject-out scheme. For each subject, a classifier is trained on all subjects trials except him/her, and then tested on the remaining subject.

After the classification is performed, let $(C_{i,j}) \quad (1 \leq i,j \leq 2)$ be the confusion matrix, in the case of binary classification. In other words, the scalar $C_{i,j}$ is the number of trials corresponding to a ground truth annotation i, that were classified as j. A commonly used evaluation metric is classification accuracy, which mathematically corresponds to $\frac{C_{1,1}+C_{2,2}}{\sum_{1\leq i,j\leq 2} C_{i,j}}$.

If such metric is appropriate for datasets where labels are balanced, it can give misleading results when there is label imbalance. The macro-averaged F1-score metric, which is defined as follows, is more suited to such a case, and penalizes the classifiers which would perform efficiently on the dominant label, but not on the other one :

$$F_1 = \frac{C_{1,1}}{2C_{1,1}+C_{2,1}+C_{1,2}} + \frac{C_{2,2}}{2C_{2,2}+C_{1,2}+C_{2,1}}$$

2.5 Influence of feature choice and other parameters on classification results

In this section, we present the intra-subject audio-visually elicited emotion binary classification results we obtained on HCI MAHNOB, DEAP and EMOEEG, studying the effects of different parameters on classification performance, and using classical EEG-based features. In the case of EMOEEG, intra-session classification is made. In other words, classification is made separately for each session (even if 3 subjects of this database participated to 2 sessions). Features are normalized by centering and scaling. Tables 2.5 and 2.6 respectively detail the features and classifiers we used.

Unless otherwise specified, the results that are presented correspond to intra-subject (intra-session for EMOEEG) classification tasks, using a leave-one-out scheme. The scores presented are the mean across subjects (resp. sessions) of the subject-wise (resp. session-wise) F1-scores.

TABLE 2.5 – Features we used

Designation	Description
HCI MAHNOB features	- θ, slow α, α, β, γ Power Spectral Density (PSD) - θ, α, β, γ differential asymmetry between electrodes
5 band powers	θ, slow α, α, β, γ PSD
DASM	Differential PSD (5 bands) asymmetry between pairs of electrodes
RASM	Rational PSD (5 bands) asymmetry between pairs of electrodes
DE	Differential Entropy (5 bands)
TDS (Time Domain Statistics)	- power, mean, std, (normalized) 1st and 2nd diff - activity, mobility, complexity

TABLE 2.6 – Classifiers we used

Classifier	Details
Linear Support Vector Machine (SVM)	- Grid search in $2^{[-5:0.5:5]}$ for C parameter
Radial Basis Function (RBF) SVM	- C empirically fixed to 1 - Grid search in $10^{[-2:0.25:1]}$ for gamma parameter

2.5.1 Extending the observation window of the signal

Given the fact that emotion elicitation is not instantaneous, adding a few seconds to the EEG signal after the end of each stimulus could yield more accurate feature computation and better emotion classification results.

TABLE 2.7 – Mean F1-scores obtained by linear SVM on HCI MAHNOB features, without and with extending the observation window (+ 3 seconds, binary intra-session classification task) (valence/arousal)

Database	Nb of sessions used	F1 without adding 3s	F1 with adding 3s
EMOEEG	8	0.57 / 0.55	**0.61** / 0.55
HCI MAHNOB	24	0.58 / 0.57	**0.59** / **0.58**

Table 2.7 shows that adding some seconds to the signal slightly improves classification results, but that such improvement is far from being substantial. As the computation of PSD-based features is averaged over the duration of each stimulus, one can understand why the effect of adding a few seconds to the signal is limited. In addition, it is interesting to observe that the best improvement is obtained for valence classification in the case of EMOEEG. Indeed, as the stimuli of this database are shorter than the ones in HCI MAHNOB, the 3 second-addition has more effect on EMOEEG results.

2.5.2 Impact of feature choice

We then studied the effect of feature choice on classification results. The results obtained using the features we tested are given in Table 2.8.

TABLE 2.8 – Mean F1-scores obtained by linear SVM with different features (val/arsl)

Features/Database	EMOEEG	HCI MAHNOB	DEAP
HCI MAHNOB features	**0.61** / **0.55**	**0.59** / **0.58**	**0.64** / 0.55
5 band powers	0.51 / 0.53	0.58 / 0.54	0.62 / **0.56**
DASM	0.59 / 0.54	0.57 / 0.56	0.63 / 0.54
RASM	0.56 / 0.53	0.56 / **0.58**	0.63 / 0.53
DE	0.51 / 0.53	0.55 / 0.57	0.62 / 0.54
TDS	0.51 / **0.55**	0.57 / 0.57	0.62 / 0.55

These results confirm the superiority of time-frequency domain features on time domain statistics. Globally, the arousal classification task seems more challenging than the valence classification one, in line with previous results exposed in Table 2.4.

Among all the feature sets we used, HCI MAHNOB features, which are a combination of PSD in specific frequency bands and differential asymmetry of such PSD between pairs of electrodes, seem to be the most efficient ones.

2.5.3 Choice of classifier

Using linear SVM or RBF SVM leads to similar results. A more intense RBF SVM tuning effort leads to results that are comparable to linear SVM. That can be explained by the fact there is not enough data for RBF SVM to generalize well.

Therefore, we exclusively train linear SVM classifiers in the remainder of this thesis. This is convenient as linear SVM is a well known reference in classification and classifier choice is not part of our contributions. We are rather interested in feature representation and learning.

2.5.4 Inter-subject classification

The results obtained for intra-subject classification tasks can be improved. Moreover, the fact that F1-scores are computed subject-wise (resp. session-wise) impairs their significance, as each subject (resp. session) corresponds to a limited number of stimuli (20, 30 or 50 depending on the database).

Therefore, even if inter-subject classification is more challenging, it offers two main advantages, in addition to the fact it opens the way to more generalizable systems :

— more data is available to train our classifiers, which are not limited to one subject (resp. one session) anymore

— the significance of F1-scores is increased due to the fact classification is performed on a larger number of trials

Table 2.9 presents the inter-subject classification results obtained in a leave-one-subject-out (resp. leave-one-session-out) fashion. Let us note that in the case of the HCI MAHNOB database, the results are better when emotional classes are determined using emotional keywords rather than valence and arousal levels. However, we consider these valence and arousal levels for the sake of comparison to the other databases.

TABLE 2.9 – F1-scores in the inter-subject classification case (HCI MAHNOB features, linear SVM)

Database	HCI MAHNOB	DEAP	EMOEEG
Valence	0.56	0.55	0.56
Arousal	0.55	0.51	0.51

These slightly better than average scores confirm that inter-subject audio-visually elicited emotion classification tasks are challenging. Arousal is still more difficult to

classify than valence. Naturally, the classification results also depend on the chosen database.

2.5.5 Threshold choice for valence and arousal classes

In the DEAP database, valence and arousal annotation are made on a continuous scale from 1 (the lowest) to 9 (the highest), whereas in HCI MAHNOB and EMOEEG, the annotation is made on a discrete scale where the subject chooses an integer value between 1 and 9.

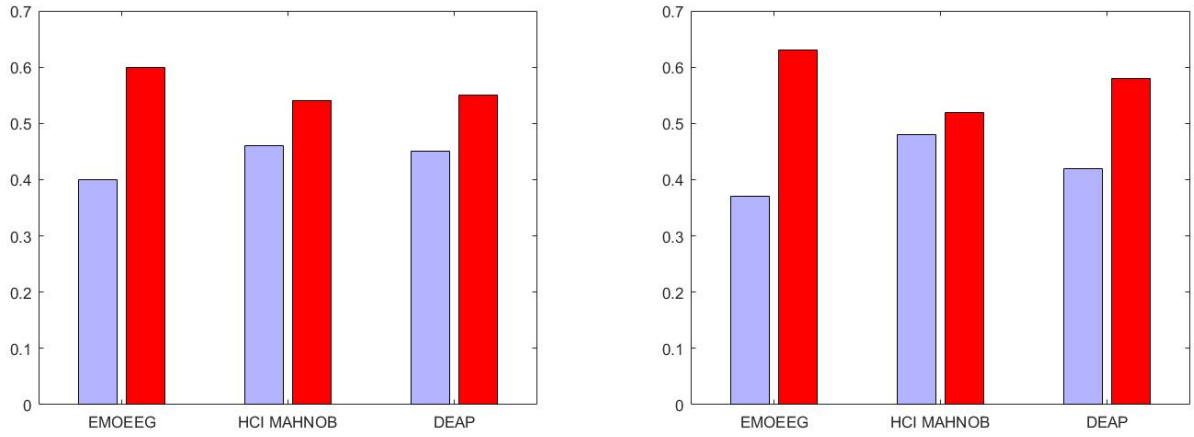

FIGURE 2.6 – Class imbalance in DEAP, HCI MAHNOB and EMOEEG
The left chart presents the proportions of low/high valence annotations (resp. blue/red). The right chart presents the proportions of low/high arousal annotations.

For each database, we have considered two valence and arousal classes, namely low and high valence (resp. arousal). In the case of EMOEEG, where both valence and arousal are biased towards negative values, we have defined $\{1,2,3\}$ as the first class and $\{4,5,6,7,8,9\}$ as the second. As for HCI MAHNOB and DEAP, the low class is composed of values respectively in $\{1,2,3,4\}$ and $\{1,2,3,4,5\}$. Even if this choice was made so as to reduce class imbalance, such imbalance is still present, as shown in Figure 2.5.5.

We can observe that arousal imbalance is more important in the cases of EMOEEG and DEAP, which could explain why the arousal classification task is more difficult on those databases (see Table 2.9).

2.6 Conclusion

In this chapter, we have recalled the usual procedure followed in EEG-based emotion classification, with a focus on audiovisual emotion elicitation scenarios. We have listed available databases as well as baseline features used for such classification. Testing some of these features on HCI MAHNOB, DEAP, and EMOEEG datasets, we have observed the already established superiority of power spectrum-based features. However, the obtained classification results are strongly improvable.

Moreover, they show lower classification scores for arousal, which is consistent with previous results in the literature [13, 23]. As stated in [20], "emotions generally are intense, high-activation states". More specifically, low arousal is more difficult to recognize. Therefore, an alternative feature extraction strategy is required : we choose to follow a feature learning approach.

The obtained intra-subject classification results vary a lot from one subject to another, whereas inter-subject classification results are unsatisfactory. To ease generalization across subjects, the chosen feature representation has to take into account the individuality of each subject from which the EEG signal is extracted.

Chapter 3 seeks a feature representation paradigm that can tackle these issues.

CHAPITRE 3

GROUP NONNEGATIVE MATRIX FACTORIZATION FOR EEG-BASED EMOTION RECOGNITION

The new trend in machine learning is to learn representations adapted to the subsequent classification stage. Along the line of Chapter 2, our approach seeking for more appropriate feature representations differs from most state-of-the-art ones in two ways :

— We focus on emotional states elicited by means of audiovisual stimuli, that is short video excerpts, which is a rather complex task.

— For easier realworld applicability, our case study is based on a single-channel setup. We do not consider spatial scalp information.

Classical power spectrum feature representations rely on neuropsychological prior knowledge concerning which frequency bands of interest to consider, exploiting the results of several studies that have shown the importance of the brain activity in predefined frequency bands, such as the β or γ bands, in emotional and cognitive processes [30, 49]. On the contrary, automatic feature extraction would avoid the need for such priors.

Feature representation includes various approaches such as sparse coding [74] and vector quantization [75]. In this work, we consider the particular dictionary learning technique that is Nonnegative Matrix Factorization (NMF), which has been used successfully in EEG-based motor imagery classification tasks [15, 76]. The method is presented in Section 3.1, whereas Section 3.2 presents the emotion classification results we obtained with this method, both in intra and inter-subject fashions.

In addition, as EEG responses are strongly subject-dependent [9, 10], the specific

frequency bands highlighted by previous research are not equally adapted to every subject. Therefore, in the inter-subject classification framework, the feature extraction method used should take into account the difference of subjects in the feature learning stage and, if possible, focus on subject-independent features. In this regard, the Group NMF principle is presented in Section 3.3, while the results we obtained using this NMF variant are presented and discussed in Section 3.4.

3.1 Nonnegative Matrix Factorization

The recurrent issue of subject dependency can hopefully be tackled by a robust automatic feature extraction framework, as it has been the case in motor imagery EEG-based classification tasks. To this end, a common dictionary that represents the data can be learned from the training set. Then, the data is projected on this dictionary to obtain features for classification. Dictionary learning seeks a "proper representation of data sets by means of reduced dimensionality subspaces" [11].

In this context, Lee and Seung's Nonnegative Matrix Factorization [12] is a well known dictionary learning technique decomposing the data into nonnegative dictionary elements.

3.1.1 General principle

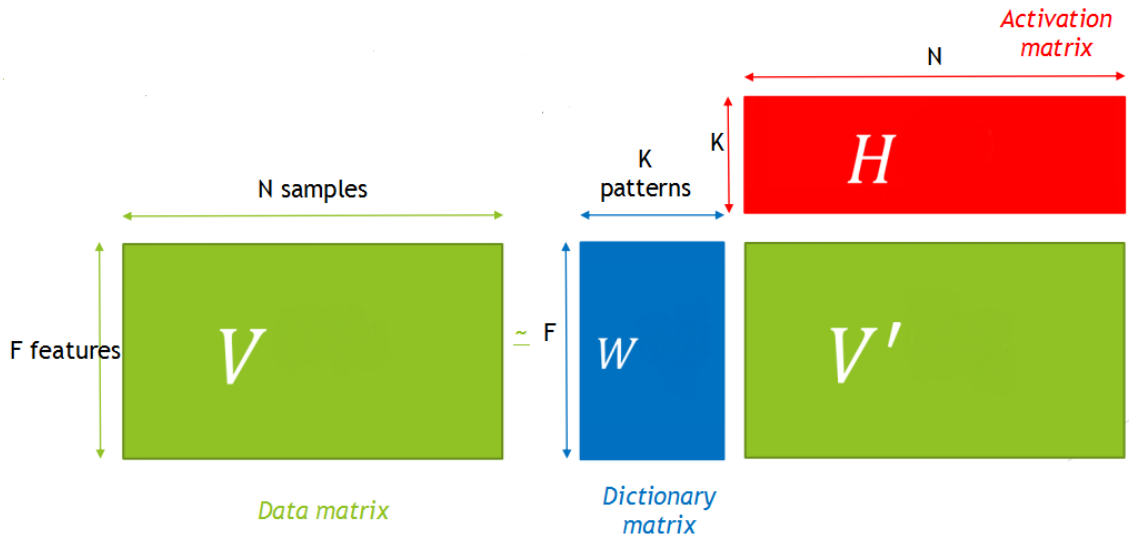

FIGURE 3.1 – Nonnegative Matrix Factorization

Let F be a number of features, N a number of samples, and K a natural number. The idea of NMF is to approximate a given nonnegative matrix $V \in \mathbb{R}_+^{F \times N}$ by a product of non-negative matrices $V' = WH$ with $W \in \mathbb{R}_+^{F \times K}$ and $H \in \mathbb{R}_+^{K \times N}$. Assuming V represents observations (the activity of F features across N time frames), W is a dictionary of K *atoms* (or *patterns*, or *latent variables*) whose activation in time is indicated by the rows of the activation matrix H, that is to say a matrix informing us, at each moment n, how strongly every atom k is activated. This NMF decomposition, represented in Figure 3.1, can be seen as some sort of soft clustering.

Let us note that WH is generally an approximation of V. Indeed, the following inequality holds for matrix ranks, where $\text{rank}(M)$ is the maximal number of independent lines or columns in M :

$$\text{rank}(WH) \leq \min\big(\text{rank}(W), \text{rank}(H)\big) \tag{3.1}$$

As W has K columns and H has K lines, $\text{rank}(W) \leq K$ and $\text{rank}(H) \leq K$. Therefore, if, as it is mostly the case since dimensionality reduction is sought, K is chosen so that $K < \text{rank}(V)$, the following inequality holds :

$$\text{rank}(WH) \leq \min\big(\text{rank}(W), \text{rank}(H)\big) \leq K < \text{rank}(V) \tag{3.2}$$

Thus, in such case, $\text{rank}(WH) < \text{rank}(V)$, which naturally implies that $V \neq WH$.

3.1.2 Divergence minimization

In order to choose a proper approximation WH for V, we must choose a "distance" $D(V|WH)$ that the couple (W,H) minimizes. However, the term "distance" can be misleading, since the chosen functions $D(.|.)$ are not necessarily symmetrical. Therefore, they are more generally called divergences or cost functions. D is a double sum of scalar divergences $d(.|.)$ on all the matrices coefficients :

$$D(V|WH) = \sum_{f=1}^{F} \sum_{n=1}^{N} d([V]_{f,n} \,\big|\, [WH]_{f,n}) \tag{3.3}$$

We consider the family of β-divergences introduced in [77] and [78], and extended in [79]. For any $\beta \in \mathbb{R}$, a scalar divergence d_β of this family is defined as follows :

$$d_\beta(x|y) = \begin{cases} \frac{1}{\beta(\beta-1)}\left(x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}\right) & \text{if} \quad \beta \notin \{0,1\} \\ x\log(\frac{x}{y}) - x + y \quad \text{if} \quad \beta = 1 \\ \frac{x}{y} - \log(\frac{x}{y}) - 1 \quad \text{if} \quad \beta = 0 \end{cases} \tag{3.4}$$

Table 3.1 details the expressions of three widespread β-divergences. Following [80], and for the sake of simplification, we abusively call "euclidean distance" the divergence d_2.

TABLE 3.1 – Commonly used β-divergences

β	Name	Expression of $d(x\|y)$
0	Itakura-Saito (IS)	$\frac{x}{y} - \log(\frac{x}{y}) - 1$
1	Kullback-Leibler (KL)	$x\log(\frac{x}{y}) - x + y$
2	Euclidean distance	$\frac{1}{2}(x-y)^2$

Given a matrix $V \in \mathbb{R}_+^{F\times N}$, a number of patterns K, and a β-divergence d_β, the nonnegative matrix factorization problem consists in the following minimization problem :

$$\min_{W\in\mathbb{R}_+^{F\times K}, H\in\mathbb{R}_+^{K\times N}} D(V|WH) = \sum_{f=1}^{F}\sum_{n=1}^{N} d_\beta([V]_{f,n} \,|\, [WH]_{f,n}) \tag{3.5}$$

The optimal W and H minimize a divergence between V and WH, which is the sum of scalar divergences between the coefficients of V and the coefficients of WH.
To determine such matrices W and H, there exists quite efficient multiplicative update rules, introduced by Lee and Seung (1999) as "a good compromise between speed and ease of implementation" [81]. W and H are first randomly initialized, and then updated following multiplicative rules which depend on the chosen divergence. The first rules introduced by Lee and Seung for the euclidean distance are the following :

$$H \longleftarrow H.\frac{W^TV}{W^TWH} \quad \text{and} \quad W \longleftarrow W.\frac{VH^T}{WHH^T} \tag{3.6}$$

where M^T denotes the transpose of matrix M, . represents a term by term multiplication, and $\frac{A}{B}$ denotes the matrix $A.B^{\cdot -1}$ ($M^{\cdot\alpha}$ is a term by term power). These rules were later extended by Févotte and Idier (2011) in the following fashion [80] :

$$H \longleftarrow H.\frac{W^T[(WH)^{\cdot(\beta-2)}.V]}{W^T(WH)^{\cdot(\beta-1)}} \quad \text{and} \quad W \longleftarrow W.\frac{[(WH)^{\cdot(\beta-2)}.V]H^T}{(WH)^{\cdot(\beta-1)}H^T} \tag{3.7}$$

Eventhough they are quite efficient, these algorithms have a noticeable drawback : if the number of atoms K is too big, convergence issues occur more frequently (the optimization

problem is non-convex). But anyway, we have no interest in choosing K too big, as we will see in Subsection 3.1.3. As for the initial choice of W and H, good practice consists in performing the multiplicative update algorithm with many random initializations, and then selecting the result for which the divergence is the lowest.

NMF seeks to exhibit latent variables explaining an observed phenomenon as well as their respective activations over time. Therefore, depending on said phenomenon, one may wish to take other constraints into account in the divergence problem (3.3), corresponding to different requirements on H and K. For instance, sparsity [82] or smoothness [83] conditions can be imposed on the matrix of pattern activations.

One may also want to impose similarity conditions between specific sub-groups of atoms of W, performing the so-called Group NMF [15]. More details on Group NMF are given in Section 3.3.

3.1.3 Specific use to EEG

Motivated by several studies showing the importance of the brain activity in predefined frequency bands, such as the β or γ bands, in emotional and cognitive processes [30, 49], NMF is applied to a time-frequency representation of the EEG data in the EEG-based classification problem. In this case, V is the power spectrogram related to the activity at one particular electrode, as shown in Figure 3.2. The PSD is averaged over each emotion elicitation trial (which corresponds to one video stimulus).

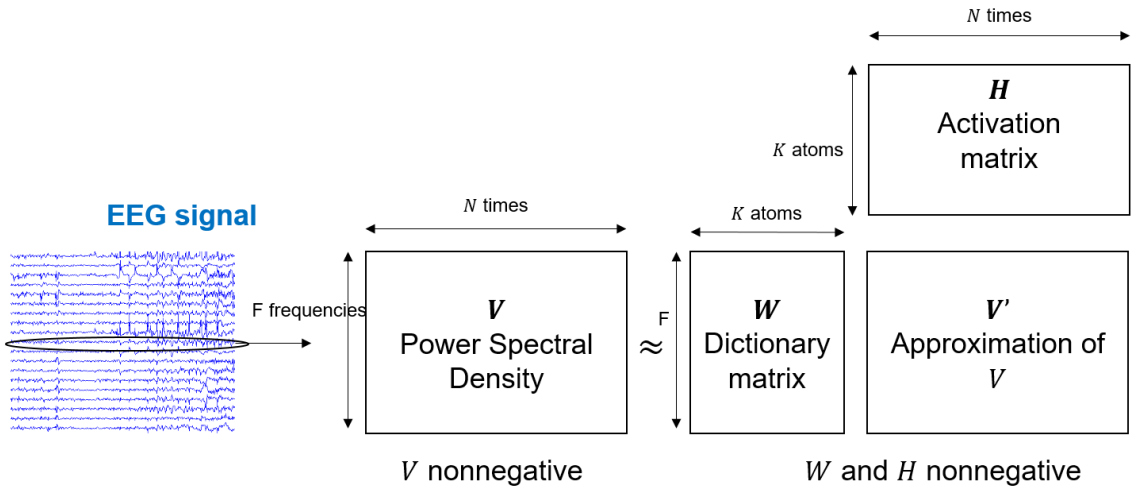

FIGURE 3.2 – Nonnegative Matrix Factorization of a Power Spectral Density Matrix

As for the choice of divergence, we have chosen the Itakura-Saito (IS) one, which has the desirable property of scale invariance [84]. In other words, in the minimization of the divergence between V and WH, no particular advantage is given to high-value coefficients of V at the expense of the low-value ones. This is particularly convenient as the PSD matrix obtained from an EEG channel can present large value differences.

When it comes to the choice of the number of atoms K, it has to offer a good compromise, as :

— a low value for K yields a poor approximation of V

— a high value for K both prevents NMF from performing dimensionality reduction and leads to over-fitting.

As for the number of W and H initializations for the multiplicative update algorithm, we found 10 to be a good compromise between efficiency and computational speed.

Figure 3.3 details the NMF-based feature extraction process. In the intra-session classification scheme, V_{train} corresponds to the PSD matrix of all trials but one, whereas V_{test} corresponds to the PSD matrix of the remaining trial. In the inter-session classification scheme, V_{train} corresponds to the PSD matrix of all sessions but one, whereas V_{test} corresponds to the PSD matrix of the remaining session.

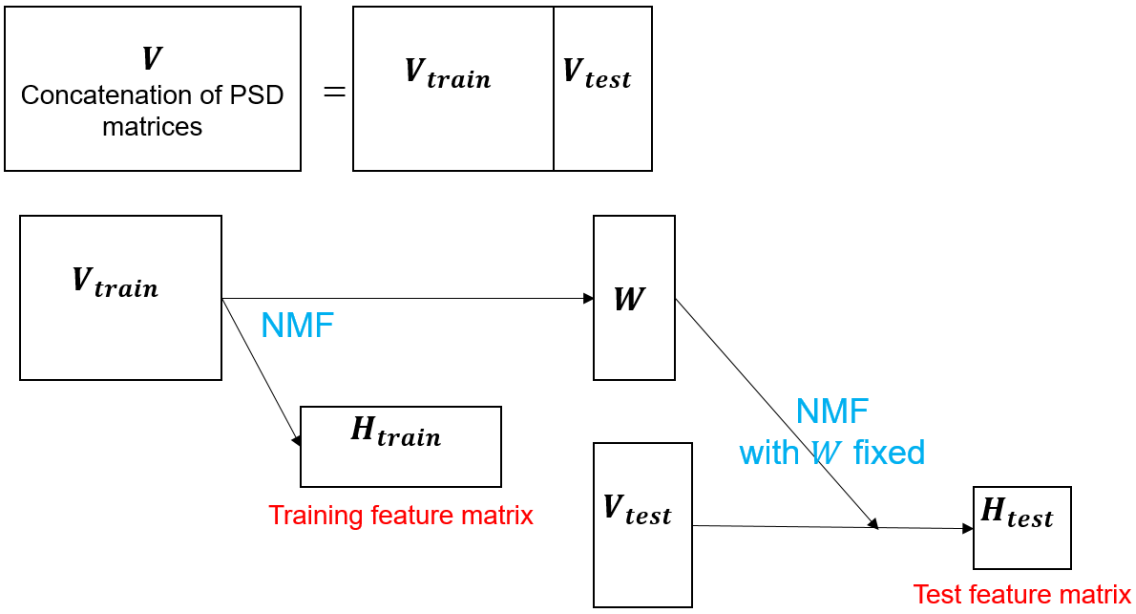

FIGURE 3.3 – Feature extraction with NMF

NMF is first used on a training set V_{train} of the data, to extract both a training activation matrix H_{train} that is used as training feature matrix and a dictionary matrix

W that is then used for the test set V_{test}. Then, NMF with fixed dictionary W is performed on V_{test} to extract the test feature matrix V_{test}.

3.2 Results obtained with NMF and conclusions

In this section, we study the emotion classification performance of NMF on the HCI MAHNOB and EMOEEG databases. We did not use the DEAP database in this part, because of the different nature of stimuli, namely music videos. EMOEEG and HCI MAHNOB are two multi-modal datasets where physiological responses, among which EEG, to audiovisual stimuli were recorded.

We call session the recording of a given subject at a given time of the day. In the case of EMOEEG, most subjects took one session whereas a few took two sessions. As for HCI MAHNOB, each subject took exactly one session, which means intra-session (resp. inter-session) classification is equivalent to intra-subject (resp. inter-subject) classification. Therefore, we talk about intra/inter-session in the case of EMOEEG, and intra/inter-subject in the case of HCI MAHNOB. In the rest of the document, we will mention intra/inter-session in both cases : it will be also understood as intra/inter-session in the HCI MAHNOB case.

TABLE 3.2 – HCI MAHNOB and EMOEEG characteristics

Database	HCI	EMOEEG
Nb of sessions (used for classification)	24	8
Nb of video stimuli per session	20	50
Duration of a video stimulus	$\approx$ 25 s	15 s
Nb of electrodes	32	20

EMOEEG is composed of 11 sessions taken by 8 participants, for a total of 11 sessions. Among these sessions, 8 were kept for emotion classification. There were technical issues in the 3 others and/or annotation abnormalities. For instance, binary classification cannot be performed on a session where only one label was reported. In HCI MAHNOB, the recordings corresponding to 27 sessions (i.e. participants) are available. We used 24 of theses sessions for valence classification and 23 for arousal classification. Table 3.2 summarizes the EMOEEG and HCI MAHNOB database characteristics.

During preliminary tests, we have tested the mid-line electrodes Cz, Pz, and Fz, and selected the central electrode Cz. We have also tested various values of the parameter K,

as shown in Table 3.3. As for the frequencies of interest, we used a 4-45 Hz band-pass filter on the signals, following [23].

TABLE 3.3 – Spectrogram and NMF parameters

Signal frequency	128 Hz
Considered frequencies (band-pass filter)	4 to 45 Hz
Tested electrodes	[Cz Pz Fz]
NMF divergence	Itakura Saito
Number of NMF initializations	10

3.2.1 Intra-session classification

In this scheme, in the case of EMOEEG, as each session is composed of 50 trials and the PSD is averaged over each trial, the matrix V_{train} has $50 - 1 = 49$ columns. In the case of HCI MAHNOB, it has 19 columns. The tested numbers of atoms K are 5,10,15, and 20.

Tables 3.5 and 3.4 present the F1-scores obtained for intra-session emotion classification with NMF, respectively on HCI MAHNOB and EMOEEG. The baseline used corresponds to the band power features named "HCI MAHNOB features" in Table 2.5 (Chapter 2). What the results first show is that NMF does not tackle the inter-subject (resp. inter-session) variability that characterized the baseline results. Even if NMF can turn out to be particularly efficient for some subjects, it performs poorly on others.

TABLE 3.4 – F1-scores for intra-session emotion classification on EMOEEG with NMF

Session	Baseline (val)	Best NMF (val)	K (val)	Baseline (aro)	Best NMF (aro)	K (aro)
1	0.70	**0.88**	10	0.40	**0.48**	20
2	**0.59**	0.58	20	0.53	**0.60**	10
3	**0.49**	**0.49**	15	0.54	**0.58**	20
4	**0.60**	0.47	5	0.52	**0.73**	10
5	**0.56**	0.51	20	0.62	**0.68**	10
6	**0.73**	0.53	20	**0.57**	**0.57**	10
7	0.52	**0.64**	10	0.56	**0.62**	10
8	0.59	0.56	10	0.64	0.60	10
Mean F1	**0.61**	0.58	-	0.55	**0.61**	-

Overall, the performance of NMF is comparable to that of the band power baseline, with slight differences between both databases and dimensions (valence/arousal). Even if there is still way to improve such performance, the fact that NMF, which extracts features coming only from one electrode, can have a performance similar to the extraction of power band features from all electrodes, is encouraging.

TABLE 3.5 – F1-scores for intra-session emotion classification on HCI MAHNOB with NMF

Subject	Baseline (val)	Best NMF (val)	K (val)	Baseline (aro)	Best NMF (aro)	K (aro)
1	**0.75**	**0.75**	15	0.47	**0.57**	15
2	0.39	**0.55**	15	**0.54**	0.49	10
3	**0.75**	0.40	20	**0.64**	0.55	15
4	**0.84**	0.49	10	**0.60**	0.57	15
5	0.49	**0.58**	15	**0.47**	0.40	15
6	**0.73**	0.48	15	**0.76**	0.69	15
7	**0.80**	0.52	15	0.41	**0.49**	15
8	0.63	**0.64**	15	**1**	0.48	15
9	0.50	**0.60**	15	**0.64**	0.60	15
10	0.41	**0.73**	15	**0.87**	0.64	10
11	0.73	**0.90**	20	**0.57**	0.50	15
12	**0.69**	0.65	10	-	-	-
13	0.58	**0.73**	10	**0.58**	0.52	15
14	**0.63**	0.52	15	**0.60**	0.49	10
15	0.23	**0.58**	15	**0.67**	0.63	10
16	**0.65**	0.60	10	0.69	**0.70**	10
17	0.49	**0.67**	20	0.58	**0.64**	15
18	0.54	**0.65**	10	0.44	**0.46**	10
19	**0.74**	0.55	10	**0.40**	0.35	20
20	**0.60**	0.52	15	0.49	**0.83**	10
21	**0.54**	0.49	10	0.40	**0.55**	5
22	**0.52**	0.49	15	0.45	**0.64**	10
23	**0.74**	0.69	10	**0.60**	0.52	10
24	0.23	**0.39**	10	0.40	**0.50**	5
Mean F1	**0.59**	**0.59**	-	**0.58**	0.56	-

However, as the best performing number of atoms K varies from one subject (resp. session) to another, we can anticipate that the inter-subject (resp. session) NMF-based classification task will be difficult.

3.2.2 Inter-session classification

In this scheme, emotion classification is made in a one-session-out fashion. Following preliminary experiments, K is chosen to be equal to 100. This number is higher than in the intra-session classification case, as PSD matrices are bigger, since each of them is composed of the data of all sessions but one.

TABLE 3.6 – F1-scores for inter-session emotion classification with NMF

Database	Baseline (valence)	NMF (valence)	Baseline (arousal)	NMF (arousal)
EMOEEG	0.56	**0.57**	0.51	**0.53**
HCI MAHNOB	0.56	**0.68**	0.55	**0.56**

Table 3.6 shows that inter-session classification results are slightly improved by NMF in the case of arousal, whereas the improvement is more noticeable for valence classification, at least in the case of HCI MAHNOB.

What is quite surprising is the fact NMF performs substantially better in the HCI MAHNOB inter-session valence classification task than in the intra-session classification one. Also, the arousal classification results are not deteriorated from intra to inter-session classification. This can be explained by the following observation : in parallel with the increased difficulty of inter-session classification tasks, more data is available in their case. The NMF extraction that was performed session by session (each session being composed of 20 trials) is now performed on all sessions but one (which equates to 20 × (24 - 1) = 460 trials. NMF-based classification has clearly benefited more from this enlarged dataset than band power-based classification.

Quite naturally, this improvement is not as striking in the EMOEEG inter-session emotion classification task. Indeed, much fewer sessions (8) are used in this case.

To conclude, there is still way to improvement, especially for the EMOEEG database and the arousal dimension. Since NMF seems to benefit from the use of data across different sessions/subjects, we naturally decide to take into account the differences of sessions/subjects in the NMF feature learning stage.

3.3 Group NMF

To take such differences into consideration in the feature learning stage, we exploit the Group NMF (GNMF) model, which allows us to account for similarity between groups of atoms [15].

3.3.1 General method

Again, we wish to approximate a given nonnegative matrix $V \in \mathbb{R}_+^{F \times N}$ by a product of non-negative matrices $V' = WH$ with $W \in \mathbb{R}_+^{F \times K}$ and $H \in \mathbb{R}_+^{K \times N}$, with W a dictionary of K atoms whose activation in time is indicated by the rows of the activation matrix H.

A group of V is a subset of columns of V that were selected according to specific conditions. Given a definition of groups of V, GNMF extracts atoms separately for each group. However, it adds other constraints to the classic NMF constraints. More precisely, it adds to the objective function (to be minimized) some terms controlling similarity

between atoms across groups. In the original formulation proposed by Lee and Choi [15], two constraints are added :

— a constraint of similarity between some atoms across groups. These atoms are called group-independent.

— a constraint of dissimilarity between other atoms across groups. These atoms are called group-dependent.

Let there be L groups and $\{V_1, V_2, ..., V_L\}$ the corresponding partition of V (each V_i is a sub-matrix of V composed of the columns corresponding to group i). Likewise, $W_1, W_2, ...W_L$ are the corresponding sub-dictionaries, and $H_1, H_2, ..., H_L$ the corresponding lines of the activation matrix H. Each sub-dictionary W_i (and each H_i) is decomposed into three parts :

— W_i^C (C for common) is composed of atoms that have to be similar to the other W_j^C ($j \neq i$)

— W_i^I (I for group-independent) is composed of atoms that have to be dissimilar to the other W_j^I

— W_i^R (R for residual) is composed of atoms upon which no specific constraints are added (in addition to classic NMF constraints)

The objective function can then be expressed as follows :

$$\mathscr{F}_{GNMF} = \sum_{l=1}^{L} D_1(V_l|W_l H_l) + \frac{\lambda}{2} \sum_{l=1}^{L} \sum_{j \neq l} D_2(W_i^C|W_j^C) - \frac{\mu}{2} \sum_{l=1}^{L} \sum_{j \neq l} D_2(W_i^I|W_j^I) \tag{3.8}$$

where D_1 and D_2 are two matrix divergences, and λ and μ are positive parameters.

We rather use the following model, proposed by Serizel et al. in [16], and derived from the first. It can tackle two types of dependencies [16], that is to say two kinds of groups at the same time. In this new formulation, two kinds of groups are considered. For instance, applying this model to a speaker identification task, Serizel et al. defined the first kind of group as speakers, and the second kind as speaking sessions.

Let there be L groups of the first kind, and M groups of the second kind. $\{V_{l,m}\}_{l \leq L, m \leq M}$ is the corresponding partition of V. Each $V_{l,m}$ is a sub-matrix of V composed of the columns corresponding to the couple (l, m). Likewise, $\{W_{l,m}\}_{l \leq L, m \leq M}$ are the corresponding sub-dictionaries, and $\{H_{l,m}\}_{l \leq L, m \leq M}$ the corresponding lines of the activation matrix H. Each sub-dictionary $W_{l,m}$ (and each $H_{l,m}$) is decomposed into three parts :

— $W_{l,m}^{C_1}$ is composed of atoms that have to be similar to the other $W_{l,m_2}^{C_1}$ ($m_2 \neq m$)

— $W_{l,m}^{C_2}$ is composed of atoms that have to be similar to the other $W_{l_2,m}^{C_2}$ ($l_2 \neq l$)

— $W_{l,m}^{R}$ (R for residual) is composed of atoms upon which no specific constraints are added (as in the first formulation)

With these notations, the objective function is now expressed as follows :

$$\mathscr{F}_{GNMF} = \sum_{l=1}^{L}\sum_{m=1}^{M} D_1(V_{l,m}|W_{l,m}H_{l,m}) + \frac{\lambda_1}{2}\sum_{l=1}^{L}\sum_{m_1=1}^{M}\sum_{m_2\neq m_1} D_2(W_{l,m_1}^{C_1}|W_{l,m_2}^{C_1}) + \frac{\lambda_2}{2}\sum_{m=1}^{M}\sum_{l_1=1}^{L}\sum_{l_2\neq l_1} D_2(W_{l_1,m}^{C_2}|W_{l_2,m}^{C_2}) \tag{3.9}$$

It is noticeable that in (3.9), there is no specific need to introduce a dissimilarity term as in (3.9). Indeed, the similarity wanted across one kind of groups is balanced by the similarity wanted across the second kind of groups (each being controlled by the parameters λ_1 and λ_2).

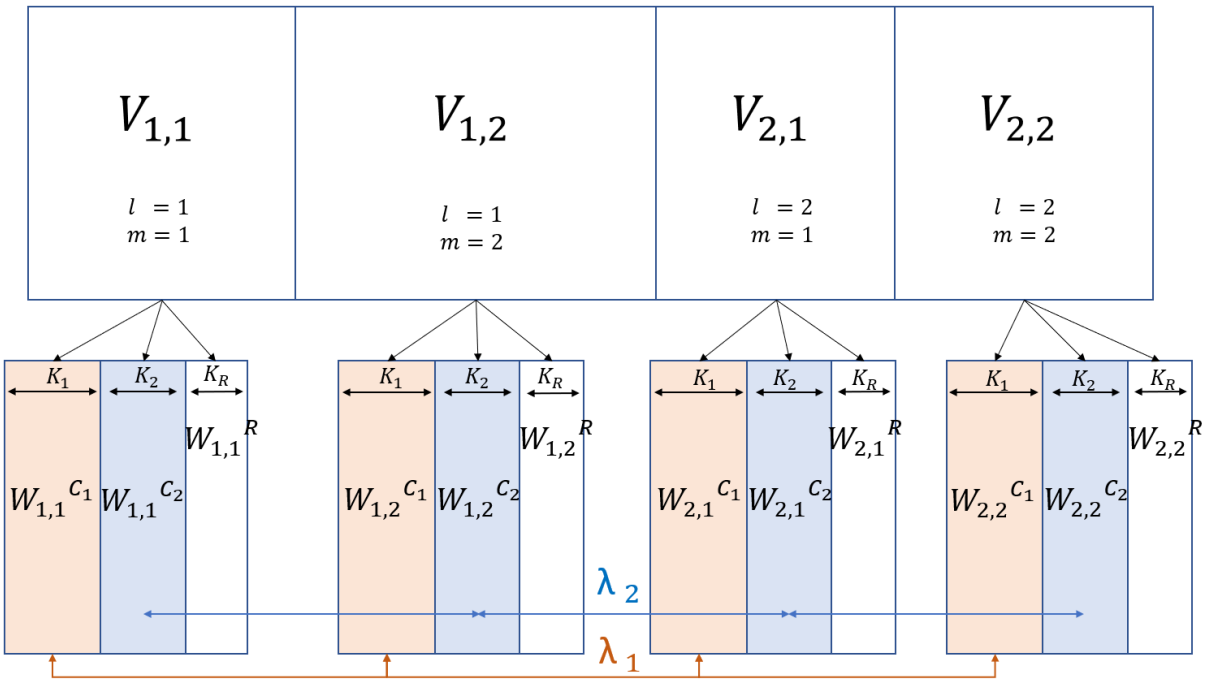

FIGURE 3.4 – Learning a dictionary matrix with GNMF (two kinds of groups, two groups of each kind)

Figure 3.4 shows how sub-dictionaries of W are extracted from each sub-matrix of the matrix V. Parameters λ_1 and λ_2 constrain the colored parts of the same color to be similar. Residual parts $\{W_{l,m}^{R}\}_{l\leq 2,m\leq 2}$ relax the similarity constraints and prevent them from hampering the original NMF approximation.

3.3.2 Specific use to EEG

We use GNMF to perform EEG-based emotion classification tasks in a supervised fashion. In other words, for valence classification, we consider two kinds of groups that are the valence label $v \in \{0,1\}$ (for low and high valence) and the session label $s \in \{1,...24\}$ for HCI MAHNOB (resp. $\{1,...8\}$ for EMOEEG). In the case of arousal classification, the valence label v is replaced by the arousal label $a \in \{0,1\}$.

Let $V_{v,s}$ be the sub-matrix of V_{train} corresponding to valence label v and session s (that is to say, the chunk of the signal corresponding to trials of session s that were given the valence annotation v by the participant). Let $W_{v,s}$ be the sub-dictionary corresponding to valence label v and session s. In such sub-dictionary :

— $W_{v,s}^{\text{val}}$ is composed of K_{val} atoms that must be similar to other W_{v,s_2}^{val} $(s_2 \neq s)$

— $W_{v,s}^{\text{sess}}$ is composed of K_{sess} atoms that must be similar to other $W_{v_2,s}^{\text{sess}}$ $(v_2 \neq v)$

— $W_{v,s}^{\text{res}}$ is composed of K_{res} atoms upon which no additional constraints are added

Then, in an inter-session valence classification scheme on HCI MAHNOB, learning a dictionary matrix W on the 23 first sessions (to use it for feature extraction on the 24^{th}) comes down to minimizing the following objective function :

$$\begin{aligned} \mathscr{F}_{GNMF} &= \sum_{v=0}^{1}\sum_{s=1}^{23} D_1(V_{v,s}|W_{v,s}H_{v,s}) + \frac{\lambda_{\text{val}}}{2}\sum_{v=0}^{1}\sum_{s_1=1}^{23}\sum_{s_2\neq s_1} D_2(W_{v,s_1}^{\text{val}}|W_{v,s_2}^{\text{val}}) \\ &+ \lambda_{\text{sess}}\sum_{s=1}^{23} D_2(W_{0,s}^{\text{sess}}|W_{1,s}^{\text{sess}}) \end{aligned} \tag{3.10}$$

The term λ_{val} controls the similarity between sub-dictionaries corresponding to the same valence label, whereas λ_{sess} controls the similarity between sub-dictionaries corresponding to the same session. We proceed similarly for arousal classification, and for the EMOEEG database.

We keep using the Itakura Saito divergence for the original NMF divergence (D1). Following the framework described in [16], similarities between valence and session-related atoms are expressed in terms of Euclidean distance (D_2).

3.4 Results obtained with GNMF and conclusions

In this section, we consider inter-session emotion classification in a one-session-out fashion. Using different values for the numbers of atoms and the similarity parameters λ, the values of these parameters which yielded the best scores are presented in Table 3.7. The left (resp. right) part of the table corresponds to the valence (resp. arousal)

classification task. K_{total} is the sum of atoms on all extracted sub-dictionaries in the training phase.

TABLE 3.7 – GNMF parameters

Dataset	K_{val}	K_{sess}	K_{res}	K_{total}	λ_{val}	λ_{sess}	K_{aro}	K_{sess}	K_{res}	K_{total}	λ_{aro}	λ_{sess}
EMOEEG	1	1	1	42	0.01	0.01	1	1	1	42	0.01	0.1
HCI	1	1	1	138	10^{-4}	10^{-5}	2	2	2	276	10^{-4}	10^{-5}

TABLE 3.8 – F1-scores for inter-session emotion classification with GNMF

Database	NMF (valence)	GNMF (valence)	NMF (arousal)	GNMF (arousal)
EMOEEG	**0.57**	**0.57**	**0.53**	0.51
HCI MAHNOB	**0.68**	0.66	**0.56**	0.55

As there are 2 valence (resp. arousal) labels and 24 HCI MAHNOB sessions, K_{total} is equal to $2 \times (24-1) \times (K_{\text{val}} + K_{\text{sess}} + K_{\text{res}}) = 46(K_{\text{val}} + K_{\text{sess}} + K_{\text{res}})$ for the HCI MAHNOB valence classification task. Because we use 8 EMOEEG sessions, K_{total} is equal to $2 \times 7 \times (K_{\text{val}} + K_{\text{sess}} + K_{\text{res}}) = 14(K_{\text{val}} + K_{\text{sess}} + K_{\text{res}})$ for the EMOEEG valence classification task. The total numbers of atoms in the arousal classification tasks can be computed similarly.

TABLE 3.9 – F1-scores per class for arousal classification with GNMF

Database	F1-score (low arousal)	F1-score (high arousal)
EMOEEG	0.45	**0.57**
HCI MAHNOB	0.52	**0.59**

Table 3.8 shows the F1-scores obtained using GNMF with the parameters in Table 3.7. The results are globally similar to the ones yielded by NMF, with a degradation in the case of HCI MAHNOB. Obviously, this use of GNMF did not improve emotion classification results. Among the possible reasons why such strategy did not turn out to be efficient, two main ones drew our attention :

— There is a relatively high number of sub-dictionaries (due to the fact each one corresponds to a couple (session,label)). Therefore, each dictionary is learned on a very limited part of the data, which could hamper generalization. Such GNMF could be more suitable with much more experimental repetitions per session.

— Aside from this question of data subdivision, the use of sessions as groups is not necessarily the most judicious segmentation.

As a matter of fact, we have noticed that, for the arousal dimension (which remains the most challenging), classification was less efficient in the case of the low class, as made clear in Table 3.9.

Why is classification more challenging in the case of low arousal? How does it translate if we compare the EEG signals of different subjects watching the same stimuli? We deal with these issues in Chapter 4.

CHAPITRE 4

EEG-based Inter-Subject Correlation Schemes in a Stimuli-Shared Framework : Interplay with Valence and Arousal

In our attempt to improve EEG-based emotion recognition by taking the subject-dependent nature of emotional responses into consideration, we have noticed that the complexity of the task varies according to the emotional level, and therefore according to the stimulus. Therefore it is interesting to study the EEG reactions of different users to the same stimuli, according to the emotional nature of each stimulus.

More than just studying the effects of valence/arousal level on annotation agreement using metrics such as the Cohen's kappa score [85, 86], we want to study this effect in depth, at the EEG level. Hence the idea of addressing the inter-subject variation issue from an interaction perspective, adopting a stimulus-centered study of synchrony between EEG signals, in the same fashion as the robot-centered approach in robotics [87]. In other words, we study the correlations between EEG signals of different subjects who watched the same videos, even if they did not watch them simultaneously.

In addition to being driven by the wish to improve EEG-based emotion classification, two other reasons motivate this approach :

— Shared experiences, such as the exposure to the same audiovisual content, play an important part in the interactions between individuals.

— For complex tasks such as stimulus-based emotion elicitation, single-trial EEG analysis is often a necessity. Therefore, analyzing the signals recorded from dif-

ferent subjects and obtaining insights about their differences and commonalities can make the results more generalizable.

To simultaneously analyze the EEG signals of different subjects, we use the Inter Subject Correlation (ISC) framework, as described in previous studies [17–19]. Dependencies between ISC of EEG recorded during audiovisual stimuli and subject conditions such as age or sex have been established. For instance, decrease in ISC of EEG has been shown as ages of the subjects increase [88]. Others have established links between ISC of functional Magnetic Resonance Imaging (MRI) and emotion, showing that ISC increases for specific regions of the brain when the stimulus elicits high arousal or low valence [89]. Replicating such results with EEG signals would both prove consistency and allow their usability with more lightweight devices.

In line with these previous works, and having acknowledged inter-subject and inter-stimuli variations [9], we propose various schemes to study the effects of valence and arousal variations on ISC of EEG recorded from different subjects watching the same videos : on all the dataset, stimulus-wise, subject-pairwise, or both stimulus-wise and subject-pairwise. Those schemes are detailed in Section 4.2.

In addition to the establishment of a link between ISC of EEG signals and valence/arousal levels which is, to the best of our knowledge, completely novel, our main contributions are :

— the proposal and comparison of various ISC computational schemes
— the assessment of the statistical validity of the observed ISC variation along valence and arousal dimensions, using computationally intensive randomization tests.

Section 4.1 is a reminder of the ISC framework. Section 4.2 presents and discusses different ISC computational schemes, whereas Section 4.3 raises the issue of interpretation of ISC results. Sections 4.4 and 4.5 show the results obtained with different schemes respectively on the HCI MAHNOB [13] and DEAP [23] databases. Finally, section 4.6 emphasizes some limitations of our work and explains observed differences between the databases.

4.1 The ISC principle

To simplify the presentation, we introduce the principle of ISC by directly instantiating it on our use-case : N_{sub} subjects watch N_{vid} video stimuli. All subjects watch the same videos. The videos are not watched simultaneously. During each stimulus, EEG

signals are recorded from the scalp of each subject with a N_{cha}-channel EEG headset. Figure 4.1 illustrates the situation.

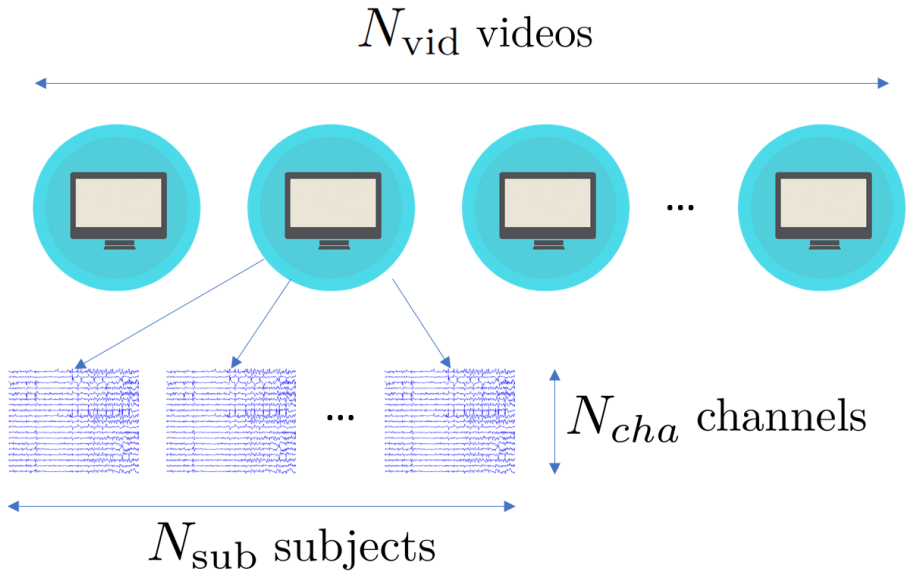

FIGURE 4.1 – Stimulus-centered study of EEG signals

For each video, each subject annotates the emotion felt using the valence and arousal dimensions. The annotation scale can be either discrete or continuous.

4.1.1 ISC score computation

Let $X_{i,v}$ denote the EEG data matrix recorded from subject i while he/she was watching video v. i ranges from 1 to N_{sub} while v ranges from 1 to N_{vid}. $X_{i,v}$ is a $N_{\text{cha}} \times T_v$ matrix, where T_v is the number of EEG signal samples recorded for each channel, which depends on the length of the video v.

Given the matrices R_{ij} of size $N_{\text{cha}} \times N_{\text{cha}}$ which each measure the cross-covariance of all electrodes in subject i with all electrodes in subject j, the pooled within-subject covariance R_w and the pooled between-subject cross-covariance R_b are defined as follows :

$$R_{ij} = \sum_{t=1}^{T} (X_{i,v}(:,t) - \bar{X}_{i,v})(X_{j,v}(:,t) - \bar{X}_{j,v})' ; \tag{4.1}$$

$$R_w = \frac{1}{N_{\text{sub}}} \sum_{i=1}^{N_{\text{sub}}} R_{ii} ; \tag{4.2}$$

$$R_b = \frac{1}{N_{\text{sub}}(N_{\text{sub}} - 1)} \sum_{i=1}^{N_{\text{sub}}} \sum_{j \neq i} R_{ij}. \tag{4.3}$$

where X' denotes the transpose of X and $\bar{X}$ denotes the vector corresponding to the mean over time of X. In Section 4.1.2, a focus is made on a pairwise definition of R_w and R_b, that is to say pooled over each pair of subjects.

Given the matrices R_b and R_w, the eigenvectors e_k of $R_w^{-1}R_b$ are computed and ranked in decreasing order of associated eigenvalue. These eigenvectors are then used to compute the correlation strengths C_k in the following fashion :

$$C_k = \frac{e_k' R_b e_k}{e_k' R_w e_k}. \tag{4.4}$$

C_k accounts for the ratio of the projection strength of e_k on R_b to its projection strength on R_w. Following previous studies that concluded that the choice of the three first components is a good compromise [18, 19, 88], we define the ISC score as $C_1 + C_2 + C_3$.

4.1.2 Averaging R_{ij} to compute ISC eigenvectors

Actually, what is usually done in the EEG-based ISC domain is the averaging of matrices R_{ij} across all stimuli, or across both all stimuli and all pairs of subjects (when ISC are considered pairwise). This only concerns the eigenvectors computation step [88]. For instance, when the averaging is done across all stimuli, the averaged matrices $\mathbf{R_{ij}}$ are computed, for each pair of subjects (i, j), in the following manner :

$$\mathbf{R_{ij}} = \frac{1}{N_{\text{vid}}} \sum_{v=1}^{N_{\text{vid}}} R_{ij} \tag{4.5}$$

Then, following (4.2) and (4.3), $R_{b_{\text{global}}}$ and $R_{w_{\text{global}}}$ are computed from the averaged matrices $\mathbf{R_{ij}}$. Eigenvectors e_k are then computed from $R_{w_{\text{global}}}^{-1} R_{b_{\text{global}}}$.

4.1.3 Shrinkage

As proposed in [90] for Linear Discriminant Analysis-based single-trial ERP classification, $R_{w_{\text{global}}}$ may be shrunk to improve robustness to outliers. Let γ be a regularization parameter between 0 and 1 and $\bar{\lambda}$ the mean eigenvalue of $R_{w_{\text{global}}}$:

$$R_{w_{\text{global}}} \leftarrow (1-\gamma) R_{w_{\text{global}}} + \gamma \bar{\lambda} I \tag{4.6}$$

When estimating a big covariance matrix, large eigenvalues are estimated too large, and small eigenvalues are estimated too small [90]. Shrinkage modifies extreme eigenvalues towards the average eigenvalue. What is convenient is that shrinkage does not change

the eigenvectors of such covariance matrices. In addition to dampening the effect of outliers by this modification, shrinkage allows to compute the inverse of the shrunk $R_{w_{\text{global}}}$ when $R^{-1}_{w_{\text{global}}}$ cannot be computed.

4.2 Different ISC computational schemes

In this chapter, we exploit our shared stimuli framework, to define different ISC computational schemes following theses perspectives :

— whether to compare the EEG signals of the subjects pairwise or globally ;

— how to combine the data on which to compute the eigenvectors of $R_w^{-1}R_b$? : that is whether to consider all the dataset, stimulus-wise, subject-pairwise, or both stimulus-wise and subject-pairwise data batches.

4.2.1 Comparing subject signals globally vs pairwise

Computing ISC eigenvectors using the signal recordings of all N_{sub} subjects globally suits the case when we wish to compare each subject to the group. In this case, ISC scores are computed for each subject i using the following expressions :

$$(C_k)_i = \frac{e_k'(R_b)_i e_k}{e_k'(R_w)_i e_k}; \tag{4.7}$$

$$\text{where } (R_b)_i = \frac{1}{N-1}\sum_{j\neq i}(R_{ij}+R_{ji}); \tag{4.8}$$

$$\text{and } (R_w)_i = \frac{1}{N-1}\sum_{j\neq i}(R_{ii}+R_{jj}). \tag{4.9}$$

In our attempt to establish a link between emotion and ISC scores, we could compare, for each video, each subject to the rest, and look at the effect of elicited emotion on the ISC score of each subject. However, doing so would compel us to consider annotation agreement globally, whereas considering annotation agreement pairwise allows a finer distinction between agreement and non-agreement. In the pairwise setting, we compute the ISC score for each pair of subjects (i,j) in the following fashion :

$$(C_k)_{ij} = \frac{e_k'(R_b)_{i,j} e_k}{e_k'(R_w)_{i,j} e_k}; \tag{4.10}$$

$$\text{where } (R_b)_{ij} = R_{ij}+R_{ji}; \tag{4.11}$$

$$\text{and } (R_w)_{ij} = R_{ii}+R_{jj}. \tag{4.12}$$

We chose to focus on this pairwise setting. In fact, in addition to allowing one to consider agreement in a pairwise fashion, this multiplies the ISC data on which to study valence and arousal effects.

4.2.2 Choosing the data on which to compute the eigenvectors

— Averaging the matrices R_{ij} across all stimuli, and then computing the eigenvectors e_k from $R_{w_{\text{global}}}^{-1} R_{b_{\text{global}}}$, that is using the whole dataset (all subjects, all stimuli), generalizes such eigenvectors and makes them more robust to outliers. All the available information is used to compute the covariance matrices, thus allowing a better precision. In that fashion, we seek to maximize inter-subject correlation on all the dataset. We refer to this scheme as $\boldsymbol{V}_{\text{all}}$. However, as EEG responses are very subject-dependent and session-dependent, computing the eigenvectors e_k on more specific subsets can also be considered.

— Rather than being computed from $R_{w_{\text{global}}}^{-1} R_{b_{\text{global}}}$, the eigenvectors e_k can be computed stimulus-wise, that that is separately for each stimulus, on all pairs of subjects, therefore taking stimulus-dependency into account. The assumption is that we wish to maximize ISC for each stimulus separately. Practically, it consists in not averaging matrices R_{ij} on all stimuli, but rather in processing each stimulus separately.

This scheme, presented in Figure 4.2, is referred to as $\boldsymbol{V}_{\text{stim}}$.

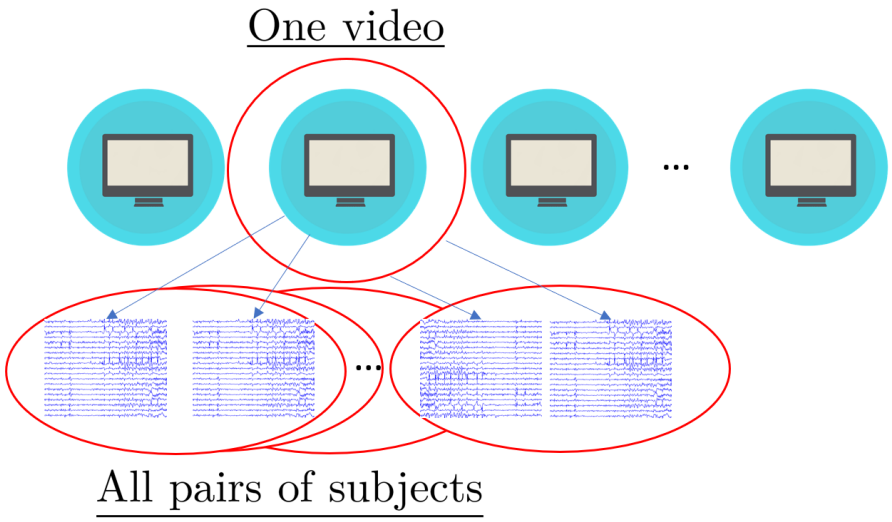

FIGURE 4.2 – Data on which the eigenvectors e_k are computed in the case of $\boldsymbol{V}_{\text{stim}}$

— The eigenvectors e_k can also be computed subject-pairwise, that is separately for each pair of subjects, on all stimuli, as shown in Figure 4.3. Thus, subject-dependency is taken into account. Mathematically, for subjects i and j, this means that the sums in equations (2) and (3) are respectively replaced by $(R_b)_{ij}$ and $(R_w)_{ij}$ (equations (10) and (11)). We refer to this scheme as $\boldsymbol{V}_{\text{pair}}$.

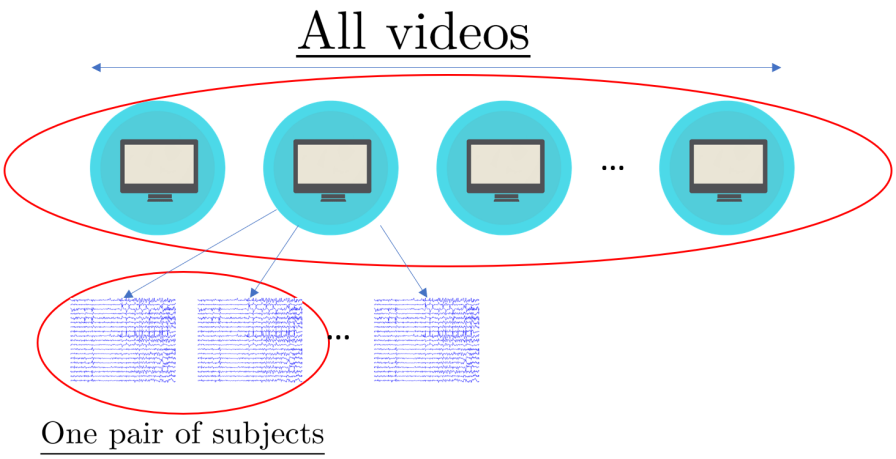

FIGURE 4.3 – Data on which the eigenvectors of $R_{w_{\text{global}}}^{-1} R_{b_{\text{global}}}$ are computed in the case of $\boldsymbol{V}_{\text{pair}}$

— Finally, the eigenvectors e_k can be computed both stimulus-wise and subject-pairwise. This takes both specificities into account, which seems well suited for EEG analysis. However, in this way, covariance matrices are estimated on smaller portions of the dataset, which automatically induces a drop in precision in the estimation of those covariance matrices. We refer to this scheme as $\boldsymbol{V}_{\text{stim/pair}}$.

4.3 Studying the effects of emotion on ISC

There are $N_{\text{pairs}} = \dfrac{N_{\text{sub}}(N_{\text{sub}}-1)}{2}$ pairs of subjects. Regardless of the slicing scheme (Section 4.2), N_{pairs} associated ISC scores are obtained for each video, which makes a total of $N_{\text{pairs}} \times N_{\text{vid}}$ ISC scores. For each pair of subjects, one has to take a decision regarding their agreement on the valence or the arousal annotations, respectively. Indeed, to establish a link between the emotion experienced by two subjects and the ISC score between their EEG signals, we limit the study to the cases where the subjects agree on the annotation of the emotion.

Then, pairs of subjects for which there is agreement should be classified according to the level of valence or arousal that was annotated.

In the HCI MAHNOB database, valence and arousal annotations are discrete values in $\{1,2,...,9\}$. We divide valence and arousal annotations in 3 classes : $\{1,2,3\}$ are considered low, $\{4,5,6\}$ are considered average, and $\{7,8,9\}$ are considered high, following the usual division made in the literature, and more specifically in the paper introducing HCI MAHNOB.

In the DEAP database, valence and arousal annotations are continuous values in $[1;9]$. We again divide valence and arousal annotations in 3 classes : values in $[1;3.5]$ are considered low, values in $]3.5;6.5[$ are considered average, and values in $[6.5;9]$ are considered high.

4.3.1 Assessing pairwise agreement

Assessing the agreement of each pair of subjects is a difficult task that may first seem arbitrary. Previous works often use the Cohen's kappa score as an agreement indicator. However, as this score is suited to multi-annotator cases, its use is less interesting when only computed on a given pair of subjects, which is our case. In addition, we do not wish to assess the agreement of each pair of subjects on all videos, but rather on each video. Therefore, our focus is on the assessment of agreement both subject-pairwise and stimulus-wise. We introduce ad hoc rules for such an assessment, taking into account the non-linearity of agreement [91] :

— For a given stimulus, we assume that two annotations from the same category (low, average, high, as previously defined) are in agreement with each other.
— We consider two annotations from different categories to be in agreement with each other if and only if their difference is lower or equal to 1.

Such rules are chosen both to correspond to the usual categories in the literature (low, average, high) and to allow for some agreement flexibility at the border between two classes.

Figure 4.4 sums up those rules in the form of a decision matrix for the HCI MAHNOB case. For instance, for a given video stimulus, if subject i annotates a valence of 2 and subject j a valence of 4, they are considered in disagreement with each other. On the contrary, if subject i gives an annotation of 7 and subject j an annotation of 9, their annotations are considered to agree with each other.

4.3.2 Assigning a subject pairwise annotation for a given stimulus when there is agreement

When two subjects agree on the annotation of a given stimulus, we want to assign a common label to this video, which is specific to this pair of subjects, in order to establish a link between this label and the ISC score. Previous works use majority decisions to assign a global annotation to each stimulus [92]. However, this is not relevant when only considering two annotators, nor is it justified when the annotations are not binary.

Therefore, for a given stimulus and a given pair of subjects who agree on the annotation of this stimulus, we decide to assign the mean of their two annotations as the pair annotation of this stimulus.

4.3.3 Effects of valence and arousal on ISC

For each category of annotation (low, average, high), the mean ISC of all pairs of subjects who agree on the annotation and whose pairwise mean annotation is in this category is computed, to establish a link between the annotation category and the mean ISC score of this category. To do so, the significance of the difference between the mean

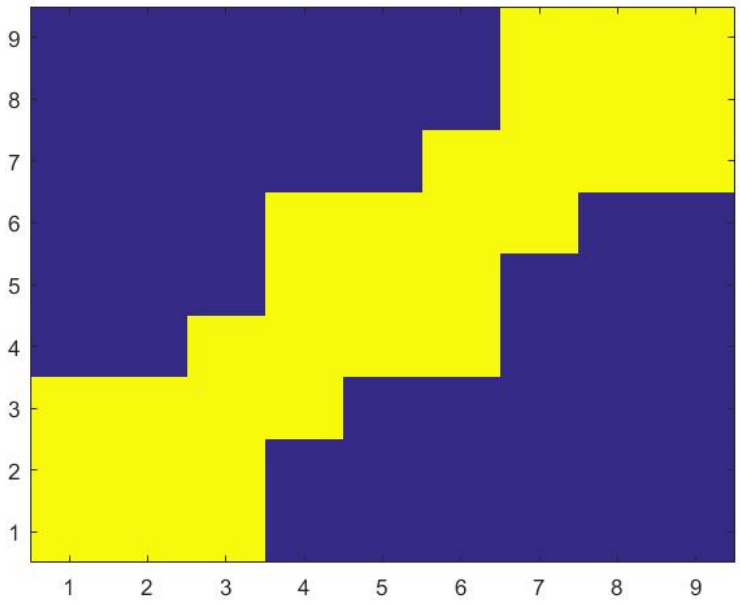

FIGURE 4.4 – Agreement decision matrix (axis values represent annotations from both subjects ; yellow stands for agreement)

ISC scores of different categories has to be assessed. Usually, parametric tests such as t-tests or ANOVA procedures are performed. Even if transformations—such as Fisher's transforms before a t-test—can be applied to make the data better fit the assumptions of the tests, these assumptions are still unwarranted.

Other approaches consist in the comparison of the empirically obtained ISC scores to simulated ISC scores on surrogates of the data. The inconvenient is that for statistical validity to hold, the computation of ISC scores from scratch has to be repeated an important number of times.

Rather, our approach is inspired from the randomization test proposed in [93]. Given the ISC scores separately computed in the 3 valence (or arousal) categories, we shuffle these ISC scores 2^{20} times, reassigning each score randomly to one of the 3 categories (each category's cardinal being kept constant). To assess the significance of the difference between the mean ISC scores obtained for two categories, we look at the number n of the 2^{20} shuffles that gave a higher difference of means than the one experimentally obtained. The significance level of the real ISC difference obtained between the two categories is at most $\frac{n+1}{2^{20}+1}$[94]. This non-parametric test allows us to assess the significance of our results without the need of complex unwarranted hypotheses on ISC score distributions. With this significance test, we are able to assess whether the variations on ISC that we observe as a function of assessed emotion are significant or not.

This procedure is performed to compare ISC scores from different valence or arousal categories, thus trying to assess the dependencies between the valence (resp. arousal) level and the ISC score.

Let us note that significance values not only depend on differences of means, but also on the cardinal of each category, which explains how a slight difference can be more significant than a larger one.

We tested our ISC computational schemes on the HCI MAHNOB [13] and DEAP [23] datasets. The reason why EMOEEG is not part of this study is the fact that, contrary to HCI MAHNOB and DEAP, all the recorded subjects did not watch the same videos, which hampers the stimulus-centered approach.

Even if the nature of the stimuli in DEAP is quite different from those in HCI MAHNOB and EMOEEG, using this dataset will help us back the possible conclusions we can get from the results on HCI MAHNOB, and/or discuss the differences.

4.4 Results on HCI MAHNOB

As stated earlier, HCI MAHNOB is a multi-modal dataset where various physiological signals were recorded from subjects who watched video stimuli. Among these physiological recordings, we are interested in the EEG signals.

Each subject assessed the emotion elicited by each stimulus in terms of valence and arousal. With our notations, $N_{\text{vid}} = 20$ and $N_{\text{sub}} = 24$ (we only took into account the subjects who watched all the videos). This gives a total of 5520 pairwise ISC scores, among which 3685 agreements on valence, and 2968 agreements on arousal. Following 4.3.1, we restrict our computations on pairs of subjects where agreement is obtained.

The focus is made on two specific schemes, that are $\boldsymbol{V}_{\text{all}}$ and $\boldsymbol{V}_{\text{stim/pair}}$. The two remaining schemes are discussed more briefly. Significance results correspond to the upper bounds obtained with the method presented in 4.3.3.

4.4.1 Results with $\boldsymbol{V}_{\text{all}}$

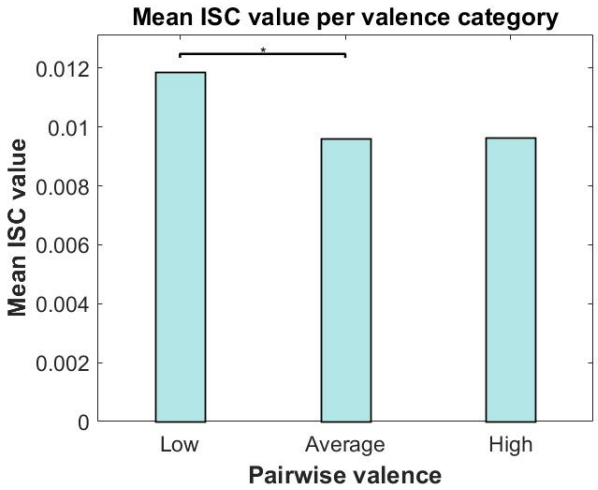

FIGURE 4.5 – Mean ISC score per valence category (low, average, high) for $\boldsymbol{V}_{\text{all}}$ *,**,*** : significance at the respective levels of 5%, 1%, and 0.1% (HCI MAHNOB database)

Figures 4.5 and 4.6 show the means of pairwise ISC scores for each category of annotation (low, average, and high), respectively for valence and arousal, along with information on the significance of the difference between each category. The considered significance levels are 5%, 1% and 0.1%.

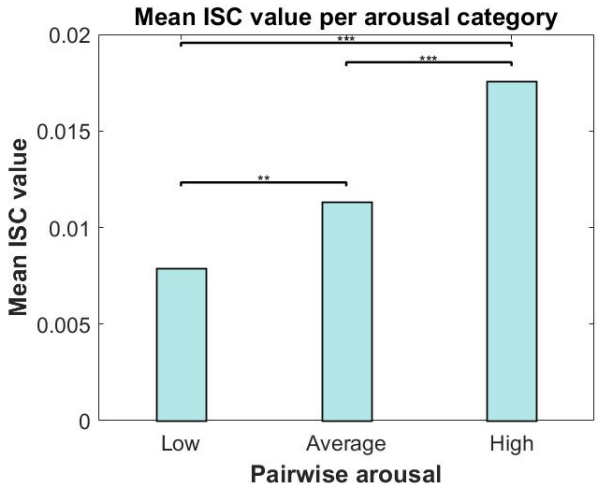

FIGURE 4.6 – Mean ISC score per arousal category ($\boldsymbol{V}_{\text{all}}$, HCI MAHNOB)

As shown in Figure 4.5, ISC scores obtained in this fashion decrease when valence increases. In other words, low valence elicitation induces better Inter Subject Correlation, which echoes the findings of Nummenmaa et al. [89], the latter restricting such variation to specific regions of the brain. However, only the difference between low valence ISC scores and average valence ISC scores is significant at the 5% level.

As for the arousal dimension, Figure 4.6 reveals an increase of ISC scores when arousal increases, which was also expected. In terms of significance, such raise is easier to observe than the decrease of ISC along valence.

4.4.2 Results with $\boldsymbol{V}_{\text{stim/pair}}$

Contrary to $\boldsymbol{V}_{\text{all}}$, this scheme takes into account both subject pair dependency and stimulus dependency. Let us see how the obtained results back the previous ones, despite this dependency change.

Figure 4.7 shows the same tendency as Figure 4.5 in terms of ISC decrease when valence increases. However, differences are better in term of significance. Figure 4.8 also shows the same tendency as Figure 4.5, but the significance level between low arousal ISC and average arousal ISC is decreased.

The monotonicity of ISC as a function of valence and a function of arousal is strengthened as it is observed for both schemes. In addition, one can notice that computing ISC eigenvectors separately for each pair of subjects and each stimulus yields more significant results for valence, whereas it degrades significance for arousal. This could

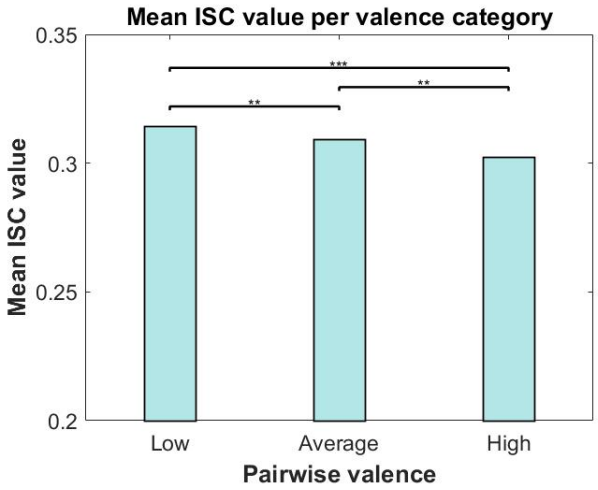

FIGURE 4.7 – Mean ISC score per valence category ($\boldsymbol{V}_{\text{stim/pair}}$)

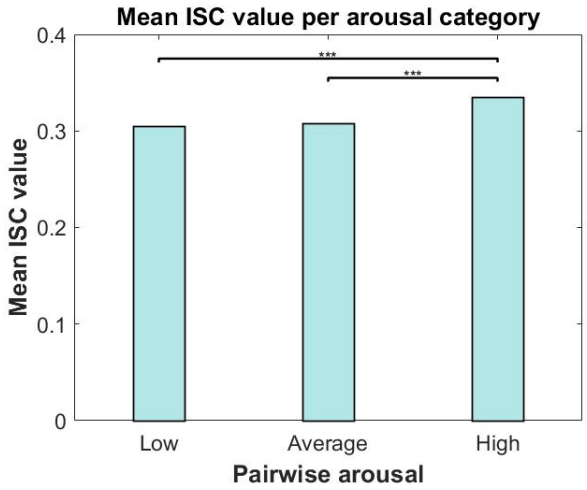

FIGURE 4.8 – Mean ISC score per arousal category ($\boldsymbol{V}_{\text{stim/pair}}$, HCI MAHNOB)

be interpreted by a lesser subject and stimulus dependency of arousal. The following subsection suggests a difference between valence and arousal annotations that could explain the phenomenon.

4.4.3 Linking the ISC level to the annotation agreement

It is worth noticing that among the 5520 HCI MAHNOB data points on which ISC can be computed (276 subject pairs × 20 video stimuli) :

- 3685 correspond to a pairwise valence annotation agreement whereas the remaining 1835 correspond to a pairwise valence annotation disagreement (using the definitions presented in Section 4.3) ;
- 2968 correspond to a pairwise arousal annotation agreement whereas the remaining 2552 correspond to a pairwise arousal annotation disagreement.

At first glance, one could conclude that agreement occurs more easily on valence than on arousal. However, it is more interesting to go in depth with a comparison of ISC levels according to valence (respectively arousal) agreement/disagreement. The results of such a comparison are given in Table 4.1 (ISC scores were computed using the scheme $\boldsymbol{V}_{\text{all}}$, HCI MAHNOB).

TABLE 4.1 – Comparison of mean ISC scores obtained in case of annotation agreement/disagreement

Dimension	Agreement	Disagreement	Significance
Valence	0.0104	0.0106	0.46
Arousal	0.0112	0.0097	0.052

Table 4.1 shows that the mean ISC score is higher on the data subset where agreement on arousal occurs than on the one where there is disagreement on arousal annotation. Such a difference is almost significant at the 5 % level. As for valence annotation, there is almost no ISC difference between agreement and disagreement cases.

This could mean that even if its occurs less frequently, agreement on arousal is more consistent than agreement on valence. Further, it could explain why the ISC monotonicity as a function of valence is more significant when ISC eigenvectors are computed separately for each pair of subject and each stimulus, rather than on the whole dataset.

4.5 Results on DEAP

DEAP is another multi-modal dataset where various physiological signals, among which EEG signals, were recorded from subjects. The main difference with HCI MAHNOB is that the emotions were elicited by the means of music video stimuli. With our notations, $N_{\text{vid}} = 40$ and $N_{\text{sub}} = 32$. This gives a total of 19840 pairwise ISC scores, among which 11126 agreements on valence, and 9184 agreements on arousal.

4.5.1 Results with V_{all}

Figure 4.9 shows that contrary to HCI MAHNOB, mean ISC scores increase when valence increases, even if the significance is only at the level of 5%. Reasons why such a difference is observed are discussed in 4.6.3.

As for the arousal dimension, Figure 4.6 reveals a variation similar to the one obtained for HCI MAHNOB, that is to say an increase of ISC scores when arousal increases, only with a less satisfying significance.

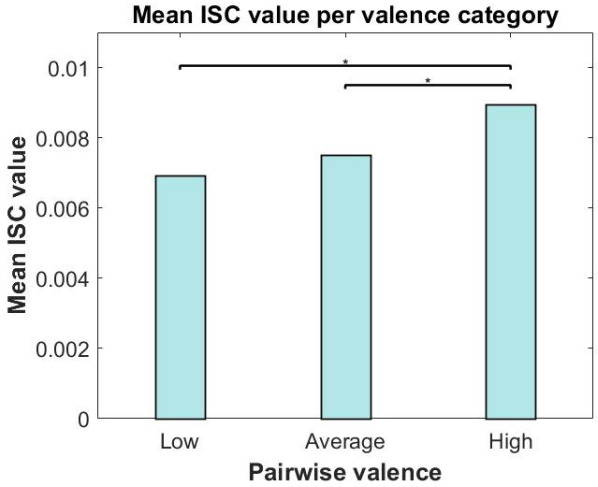

FIGURE 4.9 – Mean ISC score per valence category (V_{all}, DEAP)

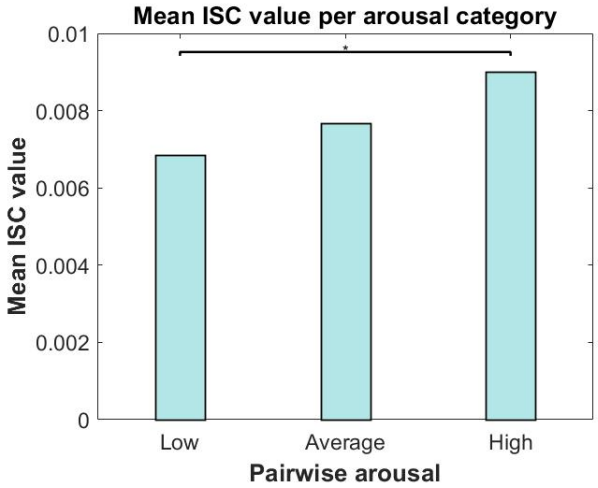

FIGURE 4.10 – Mean ISC score per arousal category (V_{all})

4.5.2 Results with $V_{\text{stim/pair}}$

When ISC eigenvectors are computed subject-pairwise and stimulus-wise, a different pattern of variations is observed for both valence (Figure 4.11) and arousal (Figure 4.12). Indeed, there is a significant ISC decrease for extreme values of valence or arousal. The mean ISC obtained for average valence (resp. arousal) is higher.

However, we can notice something quite consistent with the results concerning HCI MAHNOB, that is to say a significant decrease in ISC between low and high valence, and a significant increase in ISC between low and high arousal.

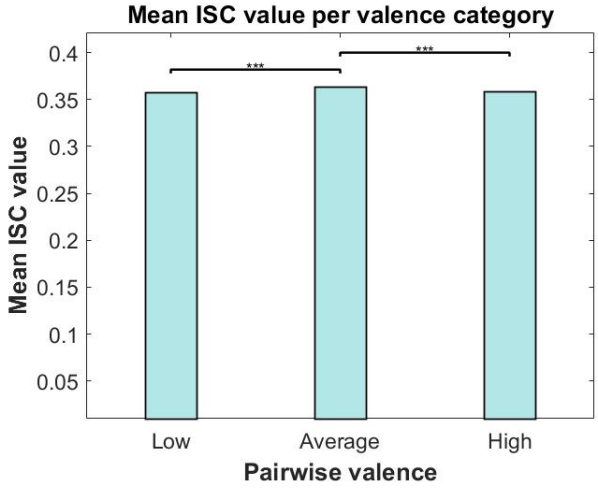

FIGURE 4.11 – Mean ISC score per valence category ($V_{\text{stim/pair}}$, DEAP)

4.6 Further discussion

4.6.1 Agreement is arbitrarily defined

The assessment of subject-pairwise agreement introduced in Section 4.3 follows arbitrary rules, even though they were carefully chosen for consistency. Performing a calibration phase before presenting the stimuli to each participant could help homogenizing the meaning of annotation values among subjects, and therefore mitigate this arbitrary aspect.

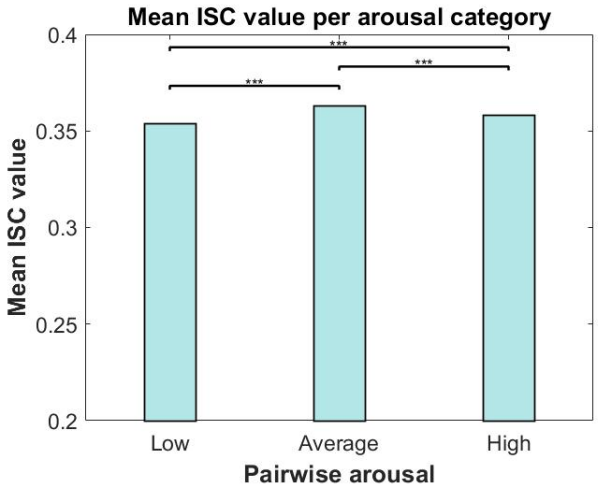

FIGURE 4.12 – Mean ISC score per arousal category ($\boldsymbol{V}_{\text{stim/pair}}$, DEAP)

4.6.2 ISC score variation from one scheme to another

Comparing ISC score levels obtained from the different schemes, one can clearly notice that the more specific the slicing scheme (Section 4.2), the higher the ISC scores. This is quite natural as the correlation is maximized on smaller, more specific subsets of the data.

4.6.3 Differences of ISC score variations along valence between HCI MAHNOB and DEAP

In the case of HCI MAHNOB, the ISC score clearly decreases along the valence dimension (Figures 4.5 and 4.7). However, results are more mitigated in the case of DEAP (Figures 4.9 and 4.11). This can be explained by both the different nature of the stimuli used and the annotation procedure. Annotation is continuous in DEAP, whereas it is discrete in HCI MAHNOB.

But some more striking comparison between HCI MAHNOB and DEAP annotation results could explain this difference better. Table 4.2 shows that the mean absolute valence annotation difference is significantly higher for DEAP than for HCI MAHNOB. Significance is computed using the method described in 4.3.3. One could wonder if the difference observed is simply due to the annotation nature, which is discrete in the case of HCI MAHNOB and continuous for DEAP. However, the same comparison for arousal yields a smaller difference between the two databases, even if the difference is still

significant. Therefore, Table 4.2 shows a difference between the databases that could explain why the ISC score clearly decreases along the valence dimension in the case of HCI MAHNOB, whereas it is more mitigated in the case of DEAP.

TABLE 4.2 – Mean absolute value of pairwise valence annotation difference

Dimension	Valence	Arousal
HCI MAHNOB	1.49	2.02
DEAP	1.69	2.10
HCI/DEAP difference significance	$< 10^{-5}$	2.5×10^{-4}

After that comparison made on the whole databases, it is interesting to compare the same quantities between HCI MAHNOB and DEAP with a restriction to the agreement cases, using the definitions of agreement exposed in 4.3.1. This is relevant as the ISC scores we presented were computed on agreeing pairs of subjects. Such a comparison is made in Table 4.3. Again, this shows that overall, the agreement level is significantly better in the case of HCI MAHNOB than DEAP, with a more significant difference for the valence dimension. This would support the hypothesis that the different valence agreement levels between the two databases explain the difference between ISC variations along valence.

TABLE 4.3 – Mean absolute of pairwise valence annotation difference among cases of agreement

Dimension	Valence	Arousal
HCI MAHNOB	0.77	0.84
DEAP	0.80	0.86
HCI/DEAP difference significance	0.0057	0.05

4.6.4 Effects of shrinkage

As exposed in 4.1.3, $R_{w_{\text{global}}}$ may be shrunk to improve robustness to outliers, by the means of a regularization parameter γ between 0 and 1. This regularization parameter has a limited effect on significance but practically none on the variation itself.

4.7 Conclusions

We have presented and described various schemes to study the effects of valence and arousal on EEG Inter Subject Correlation between participants who watched the

same audiovisual stimuli. We have introduced a definition of agreement so as to limit our study on agreeing subject pairs. Finally, we have presented the obtained results for two schemes on the HCI MAHNOB and DEAP affective datasets [13, 23].

Our results show a consistent increase in ISC scores when arousal increases. Along the valence dimension, a consistent decrease in ISC was obtained in the case of HCI MAHNOB, whereas this conclusion is more mitigated for DEAP. The different nature of the stimuli used in the DEAP dataset (music videos) can explain such drawbacks, as well as the difference between discrete/continuous annotations and, more importantly, the finer agreement level in HCI MAHNOB.

Both the decrease in ISC scores when valence increases and the increase in ISC scores when arousal increases are consistent with previous results on functional MRI in the literature [89].

A great deal of attention was devoted to the significance of such variations, using computationally intensive randomization tests. Of particular note is the fact these results are backed by the different schemes. Even if each scheme focuses on a different dependency (stimuli-wise, subject pairwise...), there is a clear trend when it comes to the variation of ISC score as a function of valence or arousal.

The conclusions of our ISC study help us understand the reasons why the emotion classification results obtained in Chapter 3 were emotion-dependent. Even at the EEG level, we can observe significant variations of inter-subject correlation according to the level of valence/arousal. This gives us a new perspective when it comes to how GNMF should be performed, as we will see in Chapter 5.

CHAPITRE 5

TOWARDS AN ISC-ORIENTED GROUP NONNEGATIVE MATRIX FACTORIZATION FOR EEG-BASED EMOTION RECOGNITION

The conclusions of Chapter 4 are that EEG inter-subject correlation strongly depends on the levels of arousal and valence. Having acknowledged this link, we seek to adapt the GNMF model described in Chapter 3 for improved EEG-based emotion classification. Instead of using GNMF with sessions (resp. subjects) as groups, we choose to focus on emotion labels.

Multi-task feature learning has been used in a subject-to-subject transfer fashion, where priors for feature dictionaries are shared across different subjects. Kang et al. [95] used multi-task feature learning in such a way, improving binary classification accuracy obtained from CSP filters on a motor imagery task. They obtained an average accuracy of 0.54 across all subjects, whereas the average accuracy reached almost 0.57 in the multi-task feature learning case.

Though the classification of valence and arousal levels can be performed independently one from another, it has been shown that multi-task learning, that is, in our case, learning to classify valence and arousal labels jointly, can improve emotion classification performance [96, 97]. The interdependence between valence and arousal [98] can explain such an improvement. We explicitly take it into account in the feature extraction stage, rather than waiting for the classifier training stage. The novelty of our work mainly lies in the exploitation of valence labels to control arousal-related feature learning (and vice

versa) using Group Nonnegative Matrix Factorization (GNMF), motivated by previous works on valence/arousal interdependencies [99].

5.1 Multi-task GNMF-based feature learning

Following the notations of Section 3.3, let $V_{v,a}$ be the sub-matrix of V_{train} corresponding to valence label v and arousal label a (that is to say, the chunk of the data corresponding to the valence annotation v and the arousal annotation a). Let $W_{v,a}$ be the sub-dictionary corresponding to valence label v and arousal label a. In such sub-dictionary :

— $W_{v,a}^{\text{val}}$ is composed of K_{val} atoms that must be similar to the other W_{v,a_2}^{val} ($v_2 \neq v$)
— $W_{v,a}^{\text{aro}}$ is composed of K_{sess} atoms that must be similar to other $W_{v_2,s}^{\text{sess}}$ ($v_2 \neq v$)
— $W_{v,a}^{\text{res}}$ is composed of K_{res} atoms upon which no additional constraints are added

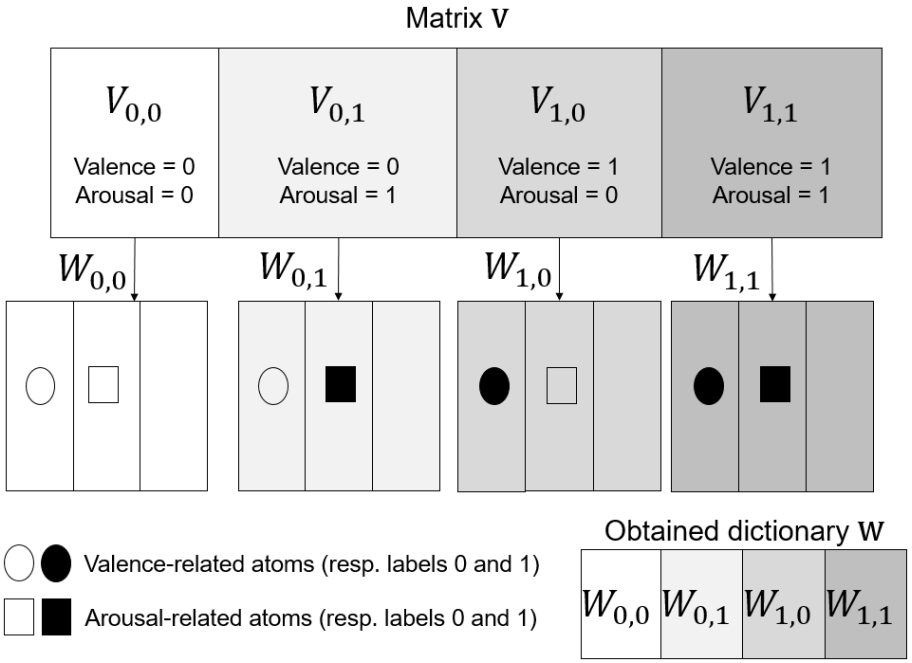

FIGURE 5.1 – Learning a dictionary matrix $\mathbf{W}$ with GNMF (valence/arousal groups)

Then, learning a dictionary matrix W on all sessions but one (to use it for feature extraction on the last) comes down to minimizing the following objective function, where

D_1 and D_2 are respectively the Itakura Saito and euclidean divergences :

$$\begin{aligned} \mathscr{F}_{GNMF} &= \sum_{v=0}^{1}\sum_{a=0}^{1} D_1(V_{v,a}|W_{v,a}H_{v,a}) + \lambda_{\text{val}} \sum_{v=0}^{1} D_2(W_{v,0}^{\text{val}}|W_{v,1}^{\text{val}}) \\ &+ \lambda_{\text{aro}} \sum_{a=0}^{1} D_2(W_{0,a}^{\text{aro}}|W_{1,a}^{\text{aro}}) \end{aligned} \tag{5.1}$$

The term λ_{valence} controls the similarity between sub-dictionaries corresponding to the same valence labels, whereas λ_{arousal} controls the similarity between sub-dictionaries corresponding to the same arousal labels. In Figure 5.1, valence-dependent atoms are constrained to show some similarity between $W_{0,0}$ and $W_{0,1}$ on the one hand, and between $W_{1,0}$ and $W_{1,1}$ on the other hand (same valence, different arousals). Likewise, another constraint lies between arousal-dependent atoms. $\lambda_{\text{valence}} > \lambda_{\text{arousal}}$ for the valence classification task, and vice versa.

In what follows, we call val/aro-GNMF this new version of GNMF, whereas the GNMF used in Chapter 3 is called session-GNMF.

5.2 Results obtained with valence/arousal-based GNMF

In this section, we keep considering inter-session emotion classification in a one-session-out fashion. Using different values for the numbers of atoms and the similarity parameters λ, the values of these parameters which yielded the best scores are presented in Table 5.1. The left (resp. right) part of the table corresponds to the valence (resp. arousal) classification task. K_{total} is the sum of atoms on all extracted sub-dictionaries in the training phase. As there are 2 valence and 2 arousal labels, K_{total} is equal to $4(K_{\text{val}} + K_{\text{sess}} + K_{\text{res}})$.

TABLE 5.1 – val/aro-GNMF parameters

Dataset	K_{val}	K_{aro}	K_{res}	K_{total}	λ_{val}	λ_{aro}	K_{aro}	K_{val}	K_{res}	K_{total}	λ_{aro}	λ_{val}
EMOEEG	15	5	5	100	0.5	0.05	12	6	6	96	0.5	0.05
HCI	15	5	5	100	10^{-4}	10^{-5}	1	1	3	20	0.01	10^{-3}

While the feature learning stage was multi-tasked with GNMF, single-task classifiers were used, that is classifiers for valence and arousal were learned separately. Indeed, we could have used the same parameters λ_{val} and λ_{aro} for both valence and arousal classification tasks, which would have implied learning valence and arousal classifiers

simultaneously. However, quite naturally, efficient valence classification requires λ_{val} to be significantly higher than λ_{aro}, and vice versa, as shown in Table 5.1.

Tables 5.2 and 5.3 present the F1 scores obtained with the emotion classification based on val/aro-GNMF feature extraction, respectively on HCI MAHNOB and EMOEEG, along with the comparison to the previously obtained scores.

TABLE 5.2 – F1-scores for inter-session emotion classification with GNMF (on EMOEEG)

Dimension	Band power baseline	NMF	Session-GNMF	val/aro-GNMF
Valence	0.56	0.57	0.57	**0.59**
Arousal	0.51	0.53	0.51	**0.55**

TABLE 5.3 – F1-scores for inter-session emotion classification with GNMF (on HCI MAHNOB)

Dimension	Band power baseline	NMF	Session-GNMF	val/aro-GNMF
Valence	0.56	0.68	0.66	**0.69**
Arousal	0.55	0.56	0.53	**0.59**

Overall, val/aro-GNMF performs better than the band power baseline, NMF, and session-GNMF. Therefore, using valence and arousal labels as groups instead of sessions seems more judicious. It is noticeable that the increase of performance from session-GNMF to val/aro-GNMF is more substantial in the case of arousal classification. That can be explained by our findings in Chapter 4, that are the increase of the ISC score along the arousal dimension is more significant than the decrease of this score along the valence dimension.

An interesting comparison point between the two databases is the fact that the band power baseline had comparable performance on EMOEEG and HCI, whereas val/aro-GNMF performed much better for HCI. The reason why EMOEEG did not benefit from val/aro-GNMF as HCI could be that the arousal annotations are less reliable in EMOEEG, as suggested by the weaker baseline arousal classification. Such annotations are used not only for GNMF-based arousal classification, but also for GNMF-based valence classification, which would explain why the valence classification score stagnates in the case of EMOEEG.

5.3 Taking ISC into account explicitly

As seen in the previous section, val/aro-GNMF feature extraction improves emotion classification. The idea of such a scheme came from the observation of sensitivity of ISC

scores to valence and arousal levels. Therefore, one could wonder why the ISC scores are not used directly in the GNMF process. This concerns the HCI MAHNOB database only, on which ISC scores were computed.

A light use of ISC scores would, for instance, consist in weighting each observation in the classification step according to the mean of the corresponding ISC scores : that is, for a given trial of a given subject, the mean of the ISC scores with other subjects on the same stimulus. We have not observed any noticeable effect of this weighting on the classification performance, neither for valence nor for arousal.

However, ISC information can be integrated at an earlier stage. To this effect, we have considered a new GNMF scheme where the ISC scores are taken into consideration in the definition of groups. Namely, we consider two ISC-based labels that are :

— 0 : the mean of the ISC scores where the given trial and subject are involved is lower than the mean of all ISC scores

— 1 : the mean of the ISC scores where the given trial and subject are involved is higher than the mean of all ISC scores

We call this new feature extraction scheme ISC-GNMF. For valence classification, these ISC-based labels replace arousal labels, and vice versa. This means that the parameters K_{aro} (resp. K_{val}) and λ_{aro} (resp. λ_{val}) are replaced by K_{ISC} and λ_{ISC}. The values of these parameters which yielded the best scores are presented in Table 5.4. The left (resp. right) part of the table corresponds to the valence (resp. arousal) classification task.

TABLE 5.4 – ISC-GNMF parameters

Dataset	K_{val}	K_{ISC}	K_{res}	K_{total}	λ_{val}	λ_{ISC}	K_{aro}	K_{ISC}	K_{res}	K_{total}	λ_{aro}	λ_{ISC}
HCI	15	5	5	100	10^{-4}	10^{-5}	0	20	5	100	0	1

What is quite noticeable about Table 5.4 is the fact the best parameters in the valence classification case are the same as the ones with val/aro-GNMF (Table 5.1). As for arousal classification, the best performing combination involves using no arousal-dependent patterns ($K_{\text{aro}} = 0$). That may be an indicator of the fact the arousal information is redundant with the ISC information, which would explain why the ISC information is sufficient.

Such a difference between the two dimensions translates into a better improvement of classification performance for arousal, as shown in Table 5.5, even if the use of ISC in the feature learning stage yields better classification results on both dimensions.

TABLE 5.5 – F1-scores for inter-session emotion classification (HCI MAHNOB)

Dimension	NMF	Session-GNMF	val/aro-GNMF	ISC-GNMF
Valence	0.68	0.66	0.69	**0.71**
Arousal	0.56	0.53	0.59	**0.63**

5.4 Conclusion

The use of GNMF for multi-task feature extraction where atom groups are determined by both valence and arousal labels (val/aro-GNMF) rather than by sessions (session-GNMF) improves classification performance. The tests run on the HCI MAHNOB database further suggest that the introduction of ISC information in the feature extraction step (ISC-GNMF) is beneficial, especially for arousal classification.

Yet, such improvements are still modest. First, one could be tempted to use an all-inclusive version of GNMF, where groups would be defined by session, valence label, arousal label and ISC information altogether. However, the more groups there are, the smaller the data corresponding to each group becomes, thus harming the quality of feature learning. This also explains why we chose binary ISC labels. The more the labels, the smaller the groups.

Consequently, one has to make compromises as for the information to be used in the group slicing. We have also noticed that the ISC-GNMF-based arousal classification task was performed better using only ISC information in the constraint added to the original NMF. This supports the idea that more group information does not necessarily induce better classification performance.

One flaw of our scheme is the quite arbitrary thresholds defining low/high valence, low/high arousal and low/high ISC. A more sophisticated way of combining NMF with ISC, which will be the subject of future work, is to abandon the notion of hard clustering in which GNMF consists, and rather take the continuous ISC score information into account directly in the NMF objective function.

CHAPITRE 6

CONCLUSION

6.1 Conclusion and discussion

EEG-based emotion recognition is a complex task when emotions are elicited by means of audiovisual content. Performing such a task is necessary if we want to be close to real-world stimulation. The complexity is increased in the single EEG channel case, which opens the way to easier lightweight setups, but for which less information is available. Our methods have brought a performance improvement compared to the baselines, which indicates the use of judicious information at the GNMF feature extraction level is promising.

First, we have used the NMF feature extraction method to perform intra and inter-subject EEG-based emotion classification, extracting dictionaries of frequency atoms from EEG spectrograms. The activation information of these atoms are then used as features for emotion classification. Contrary to classic feature representations commonly used in EEG-based emotion recognition, NMF does not rely on expert prior knowledge, and rather seeks to learn a feature representation adapted to the classification stage. Using NMF, we have obtained noticeable classification score improvements, in comparison to the frequency band power baseline features.

However, we noticed the high inter-subject variability of intra-subject classification results and the improvable inter-subject classification results. In an attempt to tackle this problem, we used Group NMF to extract NMF atoms session (resp. subject)-wise. Given predefined sub-parts of the data, this method extracts dictionaries separately

and constrains specific similarities. We used GNMF to extract NMF atoms subject-wise, atoms among which some were constrained to be similar across subjects. Even if the results of such an attempt were still improvable, they showed a discrimination in performance according to the emotional level, especially between low and high arousal, as already established by previous results in the literature.

This motivated an analysis of the valence/arousal level effects on the correlation between EEG responses of subjects watching the same stimuli. Analyzing the effects of valence/arousal on EEG Inter Subject Correlation (ISC), we found significant links between the valence/arousal levels and ISC. A particular interest was given to the statistical validity of the observed ISC variation along valence and arousal dimensions, using computationally intensive randomization tests.

As a consequence of these findings, we modified the Group NMF model we used. Rather than extracting dictionaries of atoms subject-wise as made earlier, we used Group Nonnegative Matrix Factorization in a multi-task fashion, where both valence and arousal labels are exploited to control valence-related and arousal-related feature learning. Some additional improvement was observed for emotion classification results. We also initiated the use of ISC information at the GNMF feature extraction level, which further improved the classification results on the HCI MAHNOB database.

6.2 Outlook

The conclusions of our thesis put ISC at the heart of EEG-based emotion recognition. Pursuing the idea of using ISC scores at the feature extraction level, future work seeks to take the continuous ISC score information into account directly in the NMF objective function, rather than setting arbitrary ISC score thresholds to define the GNMF groups.

Apart from the specific use of ISC information, the exploitation of the GNMF principle at its full potential is limited by the size of the emotional datasets. The more sessions and the more trials per sessions there will be, the more effective GNMF-based methods will turn, as they will include more information in the definition of groups.

In the meantime, the information used at the feature extraction level has to be chosen carefully. Contrary to our initial beliefs about inter-session classification, we have found the separation of dictionaries of atoms onto different levels of valence/arousal and ISC to be more useful than the separation onto sessions. The separation into valence/arousal/ISC classes we performed is binary. While this allows for bigger sub-dictionaries (that are extracted on bigger parts of the data), it limits the precision of the considered information.

Finally, an important issue at stake in EEG-based classification of audio-visually elicited emotion is the importance of the annotation process, than can have a decisive effect on the results. For instance, in addition to valence/arousal annotations, the participants to the HCI MAHNOB database described the emotions elicited by the videos using emotional words. As for the EMOEEG database, it contains dynamical auto-annotations of each video stimulus by each participant, therefore giving an insight of the variation in the felt emotion. Though they have not been handled in great detail in this thesis, where the classic valence/arousal framework was considered, the effects of the annotation strategy on the performance of EEG-based GNMF will be the subject of future work.

Author's publications

— Ayoub Hajlaoui, Mohamed Chetouani and Slim Essid
"Multi-task Feature Learning for EEG-based Emotion Recognition Using Group Nonnegative Matrix Factorization"
Accepted at the European Signal Processing Conference (EUSIPCO), IEEE, 2018

— Ayoub Hajlaoui, Mohamed Chetouani and Slim Essid
"EEG-based Inter-Subject Correlation Schemes in a Stimuli-Shared Framework Interplay with Valence and Arousal"
Submitted to IEEE Transactions on Affective Computing, 2018

— Anne-Claire Conneau, Ayoub Hajlaoui, Mohamed Chetouani and Slim Essid
"EMOEEG : a New Multimodal Dataset for Dynamic EEG-based Emotion Recognition with Audiovisual Elicitation"
Signal Processing Conference (EUSIPCO), 2017 25th European (pp. 738-742). IEEE.

BIBLIOGRAPHIE

[1] Stefano Valenzi, Tanvir Islam, Peter Jurica, and Andrzej Cichocki.
Individual classification of emotions using eeg.
Journal of Biomedical Science and Engineering, 7(08) :604, 2014.

[2] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain.
A review of affective computing : From unimodal analysis to multimodal fusion.
Information Fusion, 37 :98–125, 2017.

[3] A. Mehrabian and J.A. Russell.
An approach to environmental psychology.
the MIT Press, 1974.

[4] D. Keltner, P. Ekman, G.C. Gonzaga, and J. Beer.
Facial expression of emotion.
2003.

[5] B. Schuller, A. Batliner, S. Steidl, and D. Seppi.
Recognising realistic emotions and affect in speech : State of the art and lessons learnt from the first challenge.
Speech Communication, 53(9) :1062–1087, 2011.

[6] E. Bal, E. Harden, D. Lamb, A.V. Van Hecke, J.W. Denver, and S.W. Porges.
Emotion recognition in children with autism spectrum disorders : Relations to eye gaze and autonomic state.
Journal of autism and developmental disorders, 40(3) :358–370, 2010.

[7] G.V. Fiebig and M.W. Kramer.
A framework for the study of emotions in organizational contexts.
Management Communication Quarterly, 11(4) :536–572, 1998.

[8] V. Bajaj and R.B. Pachori.

Human emotion classification from eeg signals using multiwavelet transform.
In *Medical Biometrics, 2014 International Conference on*, pages 125–130. IEEE, 2014.

[9] Mojtaba Khomami Abadi, Jacopo Staiano, Alessandro Cappelletti, Massimo Zancanaro, and Nicu Sebe.
Multimodal engagement classification for affective cinema.
In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 411–416. IEEE, 2013.

[10] Hiroshi Morioka, Atsunori Kanemura, Jun-ichiro Hirayama, Manabu Shikauchi, Takeshi Ogawa, Shigeyuki Ikeda, Motoaki Kawanabe, and Shin Ishii.
Learning a common dictionary for subject-transfer decoding with resting calibration.
NeuroImage, 111 :167–178, 2015.

[11] Ivana Tosic and Pascal Frossard.
Dictionary learning.
IEEE Signal Processing Magazine, 28(2) :27–38, 2011.

[12] D.D. Lee and H.S. Seung.
Learning the parts of objects by non-negative matrix factorization.
Nature, 401(6755) :788–791, 1999.

[13] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic.
A multimodal database for affect recognition and implicit tagging.
IEEE Transactions on Affective Computing, 3(1) :42–55, 2012.

[14] Anne-Claire Conneau, Ayoub Hajlaoui, Mohamed Chetouani, and Slim Essid.
Emoeeg : A new multimodal dataset for dynamic eeg-based emotion recognition with audiovisual elicitation.
In *Signal Processing Conference (EUSIPCO), 2017 25th European*, pages 738–742. IEEE, 2017.

[15] H. Lee and S. Choi.
Group nonnegative matrix factorization for eeg classification.
In *AISTATS*, pages 320–327, 2009.

[16] R. Serizel, S. Essid, et al.

Group nonnegative matrix factorisation with speaker and session variability compensation for speaker identification.
In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5470–5474. IEEE, 2016.

[17] Jacek P Dmochowski, Paul Sajda, Joao Dias, and Lucas C Parra.
Correlated components of ongoing eeg point to emotionally laden attention–a possible marker of engagement ?
Frontiers in human neuroscience, 6 :112, 2012.

[18] Jacek P Dmochowski, Matthew A Bezdek, Brian P Abelson, John S Johnson, Eric H Schumacher, and Lucas C Parra.
Audience preferences are predicted by temporal reliability of neural processing.
Nature communications, 5, 2014.

[19] Jason J Ki, Simon P Kelly, and Lucas C Parra.
Attention strongly modulates reliability of neural responses to naturalistic narrative stimuli.
Journal of Neuroscience, 36(10) :3092–3101, 2016.

[20] David Watson.
Mood and temperament.
Guilford Press, 2000.

[21] Wenming Zheng.
Multichannel eeg-based emotion recognition via group sparse canonical correlation analysis.
IEEE Transactions on Cognitive and Developmental Systems, 9(3) :281–290, 2017.

[22] Franz Konstantin Fuss.
A method for quantifying the emotional intensity and duration of a startle reaction with customized fractal dimensions of eeg signals.
Applied Mathematics, 7(04) :355, 2016.

[23] S. Koelstra, C. Muhl, 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.
IEEE Transactions on Affective Computing, 3(1) :18–31, 2012.

[24] H. Gunes, B. Schuller, M. Pantic, and R. Cowie.
Emotion representation, analysis and synthesis in continuous space : A survey.
In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 827–834. IEEE, 2011.

[25] Dilana Hazer, Xueyao Ma, Stefanie Rukavina, Sascha Gruss, Steffen Walter, and Harald C Traue.
Emotion elicitation using film clips : Effect of age groups on movie choice and emotion rating.
In *International Conference on Human-Computer Interaction*, pages 110–116. Springer, 2015.

[26] Margaret M Bradley, Maurizio Codispoti, Dean Sabatinelli, and Peter J Lang.
Emotion and motivation ii : sex differences in picture processing.
Emotion, 1(3) :300, 2001.

[27] M. Thulasidas, C. Guan, and J. Wu.
Robust classification of eeg signal for brain-computer interface.
IEEE Transactions on Neural Systems and Rehabilitation Engineering, 14(1) :24, 2006.

[28] Irene Winkler, Mark Jäger, Vojkan Mihajlovic, and Tsvetomira Tsoneva.
Frontal eeg asymmetry based classification of emotional valence using common spatial patterns.
World Academy of Science, Engineering and Technology, 45 :373–378, 2010.

[29] R. Yuvaraj, M. Murugappan, N.M. Ibrahim, M.I. Omar, K. Sundaraj, K. Mohamad, R. Palaniappan, and M. Satiyan.
Emotion classification in parkinson's disease by higher-order spectra and power spectrum features using eeg signals : A comparative study.
Journal of integrative neuroscience, 13(01) :89–120, 2014.

[30] Mu Li and Bao-Liang Lu.
Emotion classification based on gamma-band eeg.
In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pages 1223–1226. IEEE, 2009.

[31] Anne-Claire Conneau and Slim Essid.
Assessment of new spectral features for eeg-based emotion recognition.

In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4698–4702. IEEE, 2014.

[32] Y-J. Liu, M. Yu, G. Zhao, J. Song, Y. Ge, and Y. Shi.
Real-time movie-induced discrete emotion recognition from eeg signals.
IEEE Transactions on Affective Computing, 2017.

[33] Sander Koelstra, Ashkan Yazdani, Mohammad Soleymani, Christian Mühl, Jong-Seok Lee, Anton Nijholt, Thierry Pun, Touradj Ebrahimi, and Ioannis Patras.
Single trial classification of eeg and peripheral physiological signals for recognition of emotions induced by music videos.
In *International Conference on Brain Informatics*, pages 89–100. Springer, 2010.

[34] W. Samek, M. Kawanabe, and K.-R. Muller.
Divergence-based framework for common spatial patterns algorithms.
IEEE Reviews in Biomedical Engineering, 7 :50–72, 2014.

[35] A. Dupres, F. Cabestaing, and J. Rouillard.
Sélection par un expert humain des intervalles temps-fréquence dans le signal eeg pour les interfaces cerveau-ordinateur.
In *9e Conférence Handicap (Handicap 2016)*, pages 45–50, 2016.

[36] H. Lee, Y-D. Kim, A. Cichocki, and S. Choi.
Nonnegative tensor factorization for continuous eeg classification.
International journal of neural systems, 17(04) :305–317, 2007.

[37] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou*, Edelle McMahon, Martin Sawey, and Marc Schröder.
'feeltrace' : An instrument for recording perceived emotion in real time.
In *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[38] Roddy Cowie, Martin Sawey, Cian Doherty, Javier Jaimovich, Cavan Fyans, and Paul Stapleton.
Gtrace : General trace program compatible with emotionml.
In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 709–710. IEEE, 2013.

[39] Anne-Claire Conneau.
Reconnaissance automatique de l'émotion à partir de signaux eeg.
Master's thesis, Télécom ParisTech, 6 2016.

[40] Yuan-Pin Lin, Chi-Hong Wang, Tzyy-Ping Jung, Tien-Lin Wu, Shyh-Kang Jeng, Jeng-Ren Duann, and Jyh-Horng Chen.
Eeg-based emotion recognition in music listening.
IEEE Transactions on Biomedical Engineering, 57(7) :1798–1806, 2010.

[41] Nattapong Thammasan, Ken-ichi Fukui, and Masayuki Numao.
Multimodal fusion of eeg and musical features in music-emotion recognition.
In *AAAI*, pages 4991–4992, 2017.

[42] Guillaume Chanel, Julien Kronegg, Didier Grandjean, and Thierry Pun.
Emotion assessment : Arousal evaluation using eeg,Äôs and peripheral physiological signals.
In *International workshop on multimedia content representation, classification and security*, pages 530–537. Springer, 2006.

[43] Kyung Hwan Kim, Seok Won Bang, and Sang Ryong Kim.
Emotion recognition system using short-term monitoring of physiological signals.
Medical and biological engineering and computing, 42(3) :419–427, 2004.

[44] A. Savran, K. Ciftci, G. Chanel, J. Mota, L. Hong Viet, B. Sankur, L. Akarun, A. Caplier, and M. Rombaut.
Emotion detection in the loop from brain signals and facial images.
2006.

[45] Peter J Lang.
International affective picture system (iaps) : Affective ratings of pictures and instruction manual.
Technical report, 2005.

[46] Yi-Hsuan Yang and Homer H Chen.
Machine recognition of music emotion : A review.
ACM Transactions on Intelligent Systems and Technology (TIST), 3(3) :40, 2012.

[47] Richard W Homan, John Herman, and Phillip Purdy.
Cerebral location of international 10–20 system electrode placement.
Electroencephalography and clinical neurophysiology, 66(4) :376–382, 1987.

[48] George H Klem, Hans Otto LuÈders, HH Jasper, C Elger, et al.
The ten-twenty electrode system of the international federation.
Electroencephalogr Clin Neurophysiol, 52(3) :3–6, 1999.

[49] N Rowland, MJ Meile, S Nicolaidis, et al.
Eeg alpha activity reflects attentional demands, and beta activity reflects emotional and cognitive processes.
Science, 228(4700) :750–752, 1985.

[50] P.J. Lang, M.M. Bradley, and B.N. Cuthbert.
International affective picture system (iaps) : Affective ratings of pictures and instruction manual.
Technical report A-8, 2008.

[51] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang.
A survey of affect recognition methods : Audio, visual, and spontaneous expressions.
IEEE transactions on pattern analysis and machine intelligence, 31(1) :39–58, 2009.

[52] Min-Ki Kim, Miyoung Kim, Eunmi Oh, and Sung-Phil Kim.
A review on the computational methods for emotional state estimation from the human eeg.
Computational and mathematical methods in medicine, 2013, 2013.

[53] Bo Hjorth.
Eeg analysis based on time domain properties.
Electroencephalography and clinical neurophysiology, 29(3) :306–310, 1970.

[54] Kazuhiko Takahashi et al.
Remarks on emotion recognition from bio-potential signals.
In *2nd International conference on autonomous robots and agents*, volume 3, pages 1148–1153, 2004.

[55] Lindsay Brown, Bernard Grundlehner, and Julien Penders.
Towards wireless emotional valence detection from eeg.
In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 2188–2191. IEEE, 2011.

[56] Xiao-Wei Wang, Dan Nie, and Bao-Liang Lu.
Emotional state classification from eeg data using machine learning approach.
Neurocomputing, 129 :94–106, 2014.

[57] John Atkinson and Daniel Campos.

Improving bci-based emotion recognition by combining eeg feature selection and kernel classifiers.
Expert Systems with Applications, 47 :35–41, 2016.

[58] Wolfgang Klimesch, Michael Doppelmayr, H Russegger, Th Pachinger, and J Schwaiger.
Induced alpha band power changes in the human eeg and attention.
Neuroscience letters, 244(2) :73–76, 1998.

[59] Richard J Davidson.
Anterior cerebral asymmetry and the nature of emotion.
Brain and cognition, 20(1) :125–151, 1992.

[60] M Murugappan, M Rizon, R Nagarajan, S Yaacob, I Zunaidi, and D Hazry.
Eeg feature extraction for classifying emotions using fcm and fkm.
International journal of Computers and Communications, 1(2) :21–25, 2007.

[61] Reza Khosrowabadi, Hiok Chai Quek, Abdul Wahab, and Kai Keng Ang.
Eeg-based emotion recognition using self-organizing map for boundary detection.
In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 4242–4245. IEEE, 2010.

[62] Murugappan Murugappan, Nagarajan Ramachandran, and Yaacob Sazali.
Classification of human emotion from eeg using discrete wavelet transform.
Journal of Biomedical Science and Engineering, 3(04) :390, 2010.

[63] Dan Nie, Xiao-Wei Wang, Li-Chen Shi, and Bao-Liang Lu.
Eeg-based emotion recognition during watching movies.
In *Neural Engineering (NER), 2011 5th International IEEE/EMBS Conference on*, pages 667–670. IEEE, 2011.

[64] Kwang Shin Park, Hyun Choi, Kuem Ju Lee, Jae Yun Lee, Kwang Ok An, and Eun Ju Kim.
Emotion recognition based on the asymmetric left and right activation.
International Journal of Medicine and Medical Sciences, 3(6) :201–209, 2011.

[65] Mohammad Soleymani, Sander Koelstra, Ioannis Patras, and Thierry Pun.
Continuous emotion detection in response to music videos.
In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 803–808. IEEE, 2011.

[66] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu.
Differential entropy feature for eeg-based emotion classification.
In *Neural Engineering (NER), 2013 6th International IEEE/EMBS Conference on*, pages 81–84. IEEE, 2013.

[67] Viktor Rozgić, Shiv N Vitaladevuni, and Rohit Prasad.
Robust eeg emotion classification using segment level decision fusion.
In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 1286–1290. IEEE, 2013.

[68] Wei-Long Zheng and Bao-Liang Lu.
Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks.
IEEE Transactions on Autonomous Mental Development, 7(3) :162–175, 2015.

[69] Richard J Davidson and Nathan A Fox.
Asymmetrical brain activity discriminates between positive and negative affective stimuli in human infants.
Science, pages 1235–1237, 1982.

[70] John T Cacioppo.
Feelings and emotions : Roles for electrophysiological markers.
Biological psychology, 67(1) :235–243, 2004.

[71] Daniela Sammler, Maren Grigutsch, Thomas Fritz, and Stefan Koelsch.
Music and emotion : electrophysiological correlates of the processing of pleasant and unpleasant music.
Psychophysiology, 44(2) :293–304, 2007.

[72] Robert Jenke, Angelika Peer, and Martin Buss.
Feature extraction and selection for emotion recognition from eeg.
IEEE Transactions on Affective Computing, 5(3) :327–339, 2014.

[73] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui.
Eeg emotion recognition using dynamical graph convolutional neural networks.
IEEE Transactions on Affective Computing, 2018.

[74] Bruno A Olshausen and David J Field.
Sparse coding with an overcomplete basis set : A strategy employed by v1 ?
Vision research, 37(23) :3311–3325, 1997.

[75] Philippe Schmid-Saugeon and Avideh Zakhor.
Dictionary design for matching pursuit and application to motion-compensated video coding.
IEEE Transactions on Circuits and Systems for Video Technology, 14(6) :880–886, 2004.

[76] Jing Su, Zuyuan Yang, Haiping Wang, and Wei Han.
Classification of motor imagery eeg based on sparsification and non-negative matrix factorization.
In *MATEC Web of Conferences*, volume 160, page 07007. EDP Sciences, 2018.

[77] Ayanendranath Basu, Ian R Harris, Nils L Hjort, and MC Jones.
Robust and efficient estimation by minimising a density power divergence.
Biometrika, 85(3) :549–559, 1998.

[78] Shinto Eguchi and Yutaka Kano.
Robustifing maximum likelihood estimation by psi-divergence.
ISM Research Memorandam, 802, 2001.

[79] Andrzej Cichocki, Rafal Zdunek, and Shun-ichi Amari.
Csiszar,Äôs divergences for non-negative matrix factorization : Family of new algorithms.
In *International Conference on Independent Component Analysis and Signal Separation*, pages 32–39. Springer, 2006.

[80] Cédric Févotte and Jérôme Idier.
Algorithms for nonnegative matrix factorization with the β-divergence.
Neural computation, 23(9) :2421–2456, 2011.

[81] D.D. Lee and H.S. Seung.
Algorithms for non-negative matrix factorization.
In *Advances in neural information processing systems*, pages 556–562, 2001.

[82] Julian Eggert and Edgar Korner.
Sparse coding and nmf.
In *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, volume 4, pages 2529–2533. IEEE, 2004.

[83] Bin Gao, Hong Zhang, Wai Lok Woo, Gui Yun Tian, Libing Bai, and Aijun Yin.

Smooth nonnegative matrix factorization for defect detection using microwave nondestructive testing and evaluation.
IEEE Transactions on Instrumentation and Measurement, 63(4) :923–934, 2014.

[84] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu.
Nonnegative matrix factorization with the itakura-saito divergence : With application to music analysis.
Neural computation, 21(3) :793–830, 2009.

[85] Jacob Cohen.
A coefficient of agreement for nominal scales.
Educational and psychological measurement, 20(1) :37–46, 1960.

[86] Jacob Cohen.
Weighted kappa : Nominal scale agreement provision for scaled disagreement or partial credit.
Psychological bulletin, 70(4) :213, 1968.

[87] Sofiane Boucenna, Salvatore Anzalone, Elodie Tilmont, David Cohen, and Mohamed Chetouani.
Learning of social signatures through imitation game between a robot and a human partner.
IEEE Transactions on Autonomous Mental Development, 6(3) :213–225, 2014.

[88] Agustin Petroni, Samantha Cohen, Nicolas Langer, Simon Henin, Tamara Vanderwal, Michael P Milham, and Lucas C Parra.
Age and sex affect intersubject correlation of eeg throughout development.
bioRxiv, page 089060, 2016.

[89] Lauri Nummenmaa, Enrico Glerean, Mikko Viinikainen, Iiro P Jääskeläinen, Riitta Hari, and Mikko Sams.
Emotions promote social interaction by synchronizing brain activity across individuals.
Proceedings of the National Academy of Sciences, 109(24) :9599–9604, 2012.

[90] Benjamin Blankertz, Steven Lemm, Matthias Treder, Stefan Haufe, and Klaus-Robert Müller.
Single-trial analysis and classification of erp components - a tutorial.
NeuroImage, 56(2) :814–825, 2011.

[91] H.P. Martinez, G.N. Yannakakis, and J. Hallam.
Don't classify ratings of affect ; rank them !
IEEE Transactions on Affective Computing, 5(3) :314–326, 2014.

[92] Jonathan Aigrain, Michel Spodenkiewicz, Severine Dubuisson, Marcin Detyniecki, David Cohen, and Mohamed Chetouani.
Multimodal stress detection from multiple assessments.
IEEE Transactions on Affective Computing, 2016.

[93] Alexander Yeh.
More accurate tests for the statistical significance of result differences.
In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 947–953. Association for Computational Linguistics, 2000.

[94] Eric W Noreen.
Computer-intensive methods for testing hypotheses.
Wiley New York, 1989.

[95] H. Kang and S. Choi.
Bayesian multi-task learning for common spatial patterns.
In *Pattern Recognition in NeuroImaging (PRNI), 2011 International Workshop on*, pages 61–64. IEEE, 2011.

[96] M. Kandemir, A. Vetek, M. Gönen, A. Klami, and S. Kaski.
Multi-task and multi-view learning of user state.
Neurocomputing, 139 :97–106, 2014.

[97] M.K. Abadi, A. Abad, R. Subramanian, N. Rostamzadeh, E. Ricci, J. Varadarajan, and N. Sebe.
A multi-task learning framework for time-continuous emotion estimation from crowd annotations.
In *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia*, pages 17–23. ACM, 2014.

[98] Peter Kuppens, Francis Tuerlinckx, Michelle Yik, Peter Koval, Joachim Coosemans, Kevin J Zeng, and James A Russell.
The relation between valence and arousal in subjective experience varies with personality and culture.
Journal of personality, 2016.

[99] F. Schweitzer and D. Garcia.
An agent-based model of collective emotions in online communities.
The European Physical Journal B, 77(4) :533–545, 2010.