Compact Unsupervised EEG Response Representation for Emotion Recognition

Xiaodan Zhuang*, Viktor Rozgić*, Michael Crystal

Abstract— **In this work, we propose a compact and unsupervised EEG response representation. Instead of directly extracting features from the whole response, as is commonly done for EEG signal processing, the proposed representation employs segment-level feature extraction and leverages a robust two-part unsupervised generative model to transform the segment-level features to a low-dimensional vector. The proposed method leads to rich and compact representation capability, and robust unsupervised estimation. While some previous work [1] based on segment-level features needs labeled training responses to transforms segment-level features to a response representation, the proposed method produces an EEG response representation in an unsupervised fashion, which can be directly used in various EEG response classification problems. We perform binary classification and regression of emotion dimensions on the DEAP dataset (Dataset for Emotion Analysis using electroencephalogram, Physiological and Video Signals) and demonstrate competitive performances.**

I. Introduction

Emotions provide salient cues to human psychological status and have been an active research area in human behavior analysis as well as human computer interaction. Various sensory data have been used for emotion recognition, including speech, text, facial expressions, physiological signals and electroencephalogram (EEG) [2], [3], [4]. As EEG becomes less intrusive and more affordable, its adoption in healthcare applications becomes more pervasive.

Common practices in EEG signal processing often extract features from the raw multi-channel signal on the response level [2], [3], [4], e.g. by calculating filter bank power coefficients on full response time series. Recently, segment-level feature extraction for EEG has been demonstrated to be effective for EEG emotion classification [1]: Each response is represented as multiple segment-level features extracted from overlapping temporal segments. This general approach enables new methods for response-level classification, in two broad categories: The first category performs classification directly from the set of segment-level features for each response. Particularly, Naive Bayesian Nearest Neighbor (NBNN) [5] can be used to infer the response-level label from the aggregation of segment-level nearest neighbor labels. The second category transforms the segment-level features into a fixed-length response-level representation, which enables more diverse classifier choices such as Support Vector Machines (SVM).

We propose a compact EEG response representation based on segment-level features and an unsupervised two-part generative model, inspired by recent audio modeling work [6]. In this model, the first part is a Gaussian Mixture Model (GMM) that generates the segment-level features. The second part, characterizing a low-dimensional factor, constrains the GMM parameters to a subspace with the most of the variation across responses preserved. This factor forms a compact and robust response-level representation.

The proposed representation has multiple properties.

- **Rich representation**: GMMs can represent arbitrary distributions of segment-level features with growing number of Gaussians, alleviating the loss-of-information drawback from traditional response-level features [2] and simple functional statistics (e.g., means and variances) of segment-level features in a response [1].
- **Robust estimation**: The GMM that generates the segment-level features is adapted from a Universal Background Model (UBM), leveraging a large set of response data, instead of trained from only one response. The constraints imposed on the generation of the GMM parameters, as in the second part of the generative model, enable robust estimation. The UBM and the constraints are trained jointly, enabling the UBM to represent the heterogeneous EEG data across responses from various subjects and with different emotions.
- **Compactness**: The final representation has low dimensionality and can be easily used with all classifiers that operate on fixed-length representations. The length of responses does not impact the computational cost of classifier training and testing, in contrast to classification directly based on segment-level features.
- **Unsupervised model**: The representation does not involve any modeling using annotated data, in contrast to [1], and can support different response-level classification problems. Please note that this refers to the unsupervised training of the two-part generative model used to construct the response representation. The resultant representation can be used in various EEG pattern recognition applications, such as unsupervised clustering and supervised classification or regression.

To validate the efficacy of the proposed method, we address single-trial binary classification and regression of

This material is based upon work supported by the Defense Advance Research Project Agency (DARPA) and Space and Naval Warfare Systems Center Pacific under Contract No. N66001-11-C-4094. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Defense Advance Research Project Agency (DARPA) and Space and Naval Warfare Systems Center Pacific.

The authors are with the Speech, Language and Multimedia Business Unit, Raytheon BBN Technologies, Cambridge, MA 02138, USA {xzhuang,vrozgic,mcrystal}@bbn.com.

*The first and the second authors contributed equally to this work.

978-1-4799-2131-7/14/$31.00 ©2014 IEEE

emotion dimensions (arousal, valence, dominance and liking) using electroencephalogram (EEG) signals in the DEAP dataset [2]. We demonstrate competitive performances using the proposed EEG response representation.

II. Relationship with Previous Work

The proposed EEG representation is applicable to any response-level classification. The experiments in this paper focus on EEG-based emotion recognition, the same problem studied in [2], [3], [4]. The studies [7], [2] deal with the same dataset and the study [3] employs audio-visual stimuli, all suggesting feature extraction on the response level. [8] discusses grouping of segments, but focuses on simple combination of segment-level signals.

Like in [1], the proposed approach starts with segment-level EEG feature extraction. The proposed EEG response representation employs a two-part unsupervised generative model that emits segment-level features. This model does not require any training data labels. In contrast, [1] uses supervised transformation based on nearest neighbors to convert segment-level features into response-level representation. That approach untruly assumes that every individual segment has the same label as that of the response. Such supervised transformation also makes the response representation feasible only for fully supervised problems.

The proposed method is related to general “bag of words” approaches for data representation studied in other pattern recognition problems. Bag-of-words approaches [9] based on low-level features, extracted from local sliding windows, have been successful in large scale visual classification [10] and speech based speaker verification [11]. Like most bag-of-words approaches, the proposed method describes data by a set of local descriptors (in this case, segment-level EEG features). Many popular representations based on bag-of-words require high dimensionality, such as soft quantization histogram [10] and GMM supervector [11]. The proposed method employs a two-part generative model to model EEG segment-level features. Similar models have been successfully used in audio [6] and visual [13] modeling, but have not been studied for EEG signal processing to the best of our knowledge.

III. Compact EEG Response Representation

The proposed EEG representation shares the same design motivations and statistic modeling with the recent visual representation work by the first author [13]. Particularly, an ideal EEG response representation should be able to represent highly variant segment-level features. It should be robustly estimated from limited observations extracted from a response. Further, the representation should be compact for efficient access, and support different response classification tasks. To achieve these desirable properties, we leverage a two-part generative model for segment-level features (Figure 1). The segment-level features are generated by a unique Gaussian Mixture Model (GMM) for each response. The parameters of this GMM are generated as the summation of two components, confined to a low-dimensional subspace: a) a ”global mean component” m that captures the general distribution of segment-level features regardless of subjects or labels; b) a response-specific component characterized by a low-dimensional factor w and a subspace defined by T. w serves as a compact and robust representation response.

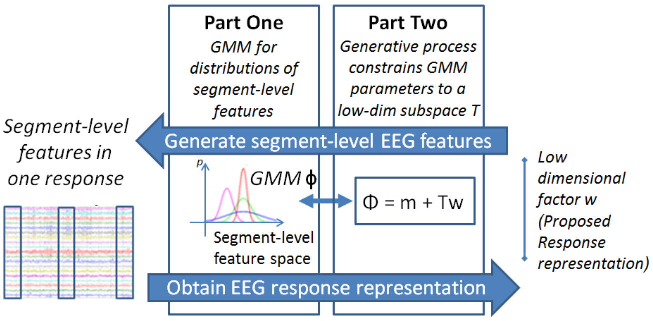
Part One
GMM for distributions of segment-level features
Part Two
Generative process constrains GMM parameters to a low-dim subspace T
Segment-level features in one response
Generate segment-level EEG features
P
GMM ϕ
Φ = m + Tw
Segment-level feature space
Obtain EEG response representation
Low dimensional factor w (Proposed Response representation)

Fig. 1. Two-part generative model for segment-level features

A. Gaussian Mixture Model

As the first part of the generative model, we consider a parametric Gaussian Mixture Model (GMM), in modeling the distribution of segment-level features in a response. GMMs meet the requirement of high representation capability as they can approximate any distribution with increased number of Gaussians. The parameters of the GMM encode information about the response, and have been used as a vector representation for classification [11]. In this subsection, we discuss this way of directly using GMM parameters as a vector representation and its limitations.

A shared GMM, referred to as the Universal Background Model (UBM), is first trained using segment-level features across all training responses. Each GMM for a response is obtained by adapting the mean vectors of the UBM according to Maximum a Posteriori criterion based on the observed segment-level features. A linear approximation of the Kullback-Leibler distance $D(g_a||g_b)$ between two response-specific GMMs g_a and g_b, has the following form,

$$D(g_a||g_b) \approx \frac{1}{2}\sum_{k=1}^{K} w_k(\mu_k^a - \mu_k^b)^T \Sigma_k^{-1}(\mu_k^a - \mu_k^b),$$

where μ_k^a and μ_k^b denote the adapted means of the kth component in the GMMs g_a and g_b. w_k and Σ_k are the mixture weight and the covariance matrix. The corresponding kernel function can be considered a linear kernel between GMM supervectors ϕ in a high-dimensional feature space,

$$\phi(a) = [\sqrt{\frac{w_1}{2}}\Sigma_1^{-\frac{1}{2}}\mu_1^a; \cdots ; \sqrt{\frac{w_K}{2}}\Sigma_K^{-\frac{1}{2}}\mu_K^a]. \qquad (1)$$

The applicability of these vectors is hindered by extremely high dimensionality: For 140-dimensional segment-level features and a UBM with 128 Gaussians, the GMM supervector ϕ has 17,920 dimensions. High dimensionality also makes robust estimation difficult, as each GMM is adapted based on the segment-level features in a single response and the GMM supervectors have a low-rank covariance matrix.

B. Generative models constraining the GMM

In order to obtain a more compact and robust representation, we use the second part of the proposed generative model to constrain the parameters of the adapted response-specific GMM to a low-dimensional subspace. One way to achieve this is to have the GMM be subject to a parametric generative process. In recent speaker verification literature [6], the "i-vector" is proposed to confine the GMM parameters into a single "total variability" subspace T shown in Equation 2:

$$\phi = m + Tw. \tag{2}$$

m is the general mean vector in the high-dimensional GMM supervector space. The total variance matrix T is essentially a projection matrix from the supervector space to the low-dimensional subspace. Let the factor w observe a standard normal distribution $N(0; I)$, then ϕ observes $N(m, TT')$.

In this work, we adopt the i-vector approach as a general unsupervised modeling method. First, the UBM is trained on a randomly selected subset of the segment-level features from all responses. We obtain the general mean vector m and the overall covariance Σ, by concatenation of block diagonal matrices for the mean and covariance parameters across different Gaussians in the UBM. Second, we estimate T on all the segment-level features using unlabeled responses. T is estimated in the same way as the eigenvoice matrix in the speaker verification literature [14], except that each response carries a unique ID as its label (therefore unsupervised). Details of the EM algorithm can be found in [14].

Changing the factor w only changes response-specific GMM within the low-dimensional parameter space. This is more robustly estimated compared to GMM parameters directly estimated in the supervector method. The factor w serves as the final representation for the response, whose dimensionality is determined by the width of matrix T.

IV. Experiments

A. Dataset and setup

We tested the proposed EEG response representation for regression and binary classification on four emotion dimensions (arousal, valence, dominance and liking) using the DEAP dataset (Dataset for Emotion Analysis using electroencephalogram, Physiological and Video Signals) [2]. DEAP contains 32-channel EEG, multiple peripheral physiological signals, and frontal facial videos recorded for 32 participants while they were watching 40 one-minute long music videos. After the presentation of each stimulus, participants rated its content in terms of arousal, valence, likability, dominance (on scale from 1 to 9) and familiarity (on scale 1 to 5). In order to match experimental conditions for emotion dimension classification on DEAP in previously reported EEG studies [7], [2], [1], we applied the same set of EEG signal pre-processing steps as in [2] (down-sampling to 128Hz, removing eye-blinking artifacts, bandpass filtering each channel to 4-45Hz interval and averaged channels to a common reference). For binary classification experiments, we transformed ratings for arousal, valence, dominance and liking to two categories corresponding respectively to rating intervals [1,5) (class "0") and [5,9] (class "1"). Only features extracted from EEG signals are used.

We evaluated classification and regression accuracies in a single-trial setup for each subject. We used leave-one-response-out cross validation scheme to obtain single subject accuracies, which are averaged over all subjects for four emotional categories: arousal (ARO), valence (VAL), dominance (DOM) and liking (LIK).

B. Compact EEG response representation

We segment EEG response signals into multiple overlapped segments with a 1 second window length and a 0.1 second step size. We extract the following features for each segment: spectral power in theta (4-8 Hz), slow alpha (8-10 Hz), alpha (8-12 Hz), beta (12-30Hz) and gamma (30+ Hz) bands for each channel and spectral power differences between 14 symmetric left-right channel pairs. This results in a 230 dimensional segment-level feature vector. In order to standardize input to the two-part generative model and remove some noise we have applied kernel principal component analysis reducing dimension to 140. For the two-part generative model, we use 128 Gaussians for the UBM and 100 dimensional factor. Note that the final response representation is therefore of 100 dimensions.

C. Results

In this section we present classification and regression performances obtained using the proposed response representations and compare them with the best reported performances on the DEAP dataset [7], [2], [1].

We tested two regression models: (1) linear ridge regression [15] in order to match the back-end setup used in [7] and demonstrate advantages of the proposed response-level representation; and (2) a support vector regression (SVR) [16] model to examine further performance variations. For the SVR we optimized C , the parameter that controls the trade-off between training errors and margin size, and kernel size γ using cross-validation on the training set for each cross-validation split, both on the grid of $\{2^{-6}, ..., 2^{6}\}$.

As shown in Table I, with a matched linear ridge regression model [7] our response-level representation outperformed the best reported results on the ARO, DOM and LIK tasks, and matched the VAL performance. [7] reports no benefits in using different back-end regression methods (Gaussian Process Regression [17] and Relevance Vector Machine regression [18]). Similarly, we observe that the linear ridge regression performs better than support vector regression based on the proposed representation.

On the binary classification task (Table II), we trained two different classifiers, naive Bayes matching [2] and RBF-SVM matching [1]. While naive Bayes has no tunable parameters, for the RBF-SVM we optimized C and γ parameters in the same manner and on the same grids as for SVR. Compared with [2] for the naive Bayes classifier we report higher accuracies on the VAL and LIK tasks and lower performance on the ARO task. Applying RBF-SVM, which doesn't have

TABLE I

RESPONSE EMOTION REGRESSION PERFORMANCE (MEAN ABSOLUTE ERROR) AVERAGED OVER SUBJECTS AND STANDARD DEVIATION

Method	Segment Feature	EEG Response Representation	Classifier	ARO[%]	VAL[%]	DOM[%]	LIK[%]
[7]	N/A	spectral power (unsupervised)	linear ridge	1.53(0.40)	1.59(0.39)	1.53(0.49)	1.78(0.51)
Proposed	spectral power	Two-part generative model (unsupervised)	linear ridge	1.50(0.42)	1.59(0.30)	1.42(0.52)	1.71(0.46)
Proposed	spectral power	Two-part generative model (unsupervised)	SVR	1.56(0.49)	1.61(0.39)	1.43(0.54)	1.74(0.56)

TABLE II

RESPONSE EMOTION CLASSIFICATION ACCURACY AVERAGED OVER SUBJECTS AND STANDARD DEVIATION

Method	Segment Feature	EEG Response Representation	Classifier	ARO[%]	VAL[%]	DOM[%]	LIK[%]
[2]	N/A	spectral power (unsupervised)	Naive Bayes	62.0(N/A)	57.6(N/A)	N/A(N/A)	55.4(N/A)
Proposed	spectral power	Two-part generative model (unsupervised)	Naive Bayes	56.3(11.1)	65.1(15.8)	66.0(13.2)	63.0(16.4)
Proposed	spectral power	Two-part generative model (unsupervised)	RBF-SVM	67.1(14.2)	70.9(11.4)	70.9(12.8)	70.5(17.1)
[1]	spectral power	NN-based (supervised)	RBF-SVM	68.4(12.1)	76.9(6.4)	73.9(11.1)	75.3(10.6)

the false assumption of independence between dimensions in naive Bayes, we observe further improvements on all the four tasks. For the ARO task, the proposed representation achieves similar performance compared to [1]. Note that [1] needs access to labeled training data in constructing the Nearest Neighbor (NN) based representation and assumes that the response labels apply to all segments, therefore applicable only to fully supervised problems.

V. CONCLUSIONS AND DISCUSSION

We propose a compact EEG response representation modeled via unsupervised training. Segment-level features are first extracted from uniformly sampled sliding windows. Then, we leverage a robust unsupervised two-part generative model to transform the segment-level features to a low-dimensional response representation. Since training this generative model does not require annotated training data, the proposed response representation can be directly used in many different EEG response classification problems. Using the proposed EEG response representation, we demonstrate competitive results in binary classification and regression tasks of emotion recognition on the DEAP dataset.

Direct response-level representation in [7], [2] is constructed directly from the whole response, which involves inherent averaging across all parts of the response. Nearest Neighbor based response level representation [1] was based on nearest neighbor distances between segments within a response and all training segments in different classes. That representation favored most representative segments and captured class information in a small number of descriptive isolated intervals. The representation proposed in this paper is based on segment-to-response transformation method that uses an unsupervised generative model. This model captures both isolated and averaged segment characteristics, with less emphasis on extreme local properties when compared to the representations in [1]. As these different types of representations may capture complementary information about the EEG responses, we plan to investigate effective fusing for further improved EEG response recognition performance.

REFERENCES

[1] V. Rozgic, S. Vitaladevuni, and R. Prasad, "Robust eeg emotion classification using segment level decision fusion," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 1286–1290.

[2] 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," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 18–31, 2012.

[3] D. Nie, X.-W. Wang, L.-C. Shi, and B.-L. Lu, "Eeg-based emotion recognition during watching movies," in *Neural Engineering (NER), 2011 5th International IEEE/EMBS Conference on*. IEEE, 2011, pp. 667–670.

[4] Y. Liu, O. Sourina, and M. K. Nguyen, "Real-time eeg-based human emotion recognition and visualization," in *Cyberworlds (CW), 2010 International Conference on*. IEEE, 2010, pp. 262–269.

[5] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[6] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788 –798, May 2011.

[7] M. Soleymani, S. Koelstra, I. Patras, and T. Pun, "Continuous emotion detection in response to music videos," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 803–808.

[8] L. Hausfeld, F. De Martino, M. Bonte, and E. Formisano, "Pattern analysis of eeg responses to speech and voice: Influence of feature grouping," *NeuroImage*, vol. 59, no. 4, pp. 3641–3651, 2012.

[9] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *In Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, pp. 1–22.

[10] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek, "Visual word ambiguity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1271–1283, 2010.

[11] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *Signal Processing Letters, IEEE*, vol. 13, no. 5, pp. 308–311, 2006.

[12] X. Zhuang, S. Wu, and P. Natarajan, "Compact bag-of-words visual representation for effective linear classification," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 521–524.

[13] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 345 – 354, May 2005.

[14] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, "Numerical recipes: the art of scientific computing," 1987.

[15] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[16] C. E. Rasmussen, "Gaussian processes for machine learning," 2006.

[17] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *The Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.