Continuous Emotion Recognition: Another Look at the Regression Problem

Pouria Fewzee, Fakhri Karray
Department of Electrical and Computer Engineering, University of Waterloo
Waterloo, ON, Canada N2L 3G1
Email: spfewzee@uwaterloo.ca

***Abstract*—Various regression models are used to predict the continuous emotional contents of social signals. The common trend to train those models is by minimizing a sense of prediction error or maximizing the likelihood of the training data. According to those optimization criteria, among two models, the one which results in a lower prediction error, or higher likelihood, should be favored. However, that might not be the case, since to compare the prediction quality of different models, the correlation coefficient of their prediction with the actual values is prevalently used. Hence, given the fact that a lower prediction error does not imply a higher correlation coefficient, we might need to reconsider the optimization criteria that we undertake in order to learn the regression coefficients, in order to synchronize it with the hypothesis testing criteria. Motivated by this reasoning, in this work we suggest to maximize a sense of correlation for learning regression coefficients. Two senses of correlation, namely Pearson's correlation coefficient and Hilbert-Schmidt independence criterion, are seen for this purpose. We have chosen the continuous audio/visual emotion challenge 2012 as the framework of our experiments. The numerical results of this study show that compared to support vector regression, the suggested learning algorithms offer higher correlation coefficient and lower prediction error.**

I. Introduction

Emotions can be recognized from social cues such as speech samples, facial expressions, or body movements [1], [2]. Depending on the available modalities to an emotion recognition system, different types of features can be employed. The objective of a statistical model is then to find patterns of variations between those features, or analogously explanatory variables, and the corresponding emotional contents, or response variables in the context of statistical learning. That is, to find a mapping $f : \mathbf{X} \rightarrow \mathbf{y}$ that reflects those patterns; explanatory and response variables are denoted by $\mathbf{X}$ and $\mathbf{y}$, respectively. Emotional states are represented using categorical and dimensional representations. According to the categorical view, emotional states can be described using discrete emotion categories such as anger or happiness [3], [4]. On the other hand, dimensional, or primitive-based, point of view suggests the use of some lower level continuous attributes (e.g., arousal and valence). Theories behind the dimensional representation claim that the space defined by those representations can subsume all the categorical emotional states [5], [6], [7]. Therefore, depending either the categorical or the dimensional point of view is of interest, the statistical learning task will be classification or regression. The focus of this work is on the dimensional modeling of emotion, hence we deal with solving a regression problem. Specifically, we are interested in linear regression, which dictates the function f to be of the form $w_0 + w_1\mathbf{x}_1 + \cdots + w_p\mathbf{x}_p$.

In principle, regression is an optimization problem that is used to set model parameters, in this case w_0 to w_P, so that the resulting model minimize the prediction error, or analogously maximize a sense of likelihood. However, if the criterion for the goodness of a regression model is other than the prediction error, we might accordingly need to modify the cost function of the corresponding optimization. Particularly, a commonly used measure for assessing the goodness of a prediction is the correlation coefficient of prediction values with the actual value of a response variable. In case a model can achieve perfect prediction, that is zero prediction error, that model also does maximize the correlation coefficient, however, otherwise, among two models, the one with a lower prediction error does not guarantee a higher correlation coefficient, i.e., it can be shown that in some subspaces of the response variable space $\mathcal{Y}$, the derivatives of correlation coefficient and an error measure, like root mean squared error, with respect to the prediction values happen to take similar signs. On the other hand, although various models are proved to asymptotically converge to their underlying distributions, achieving zero prediction error in real-world problems may not be a realistic target to set. Therefore, when the correlation coefficient is used as the measure of evaluating a regression model, it would be a fair decision to consequently adapt the optimization problem. However, the literature on the problem, in particular the works on the continuous audio/visual emotion challenge (AVEC 2012) [8], show that in spite of the fact that the correlation coefficient was set as the standard, the participants decided to minimize the prediction error [9], [10], [11], [12], [13], [14], [15].

To pursue our instinct, we hence decided to suggest two learning algorithms and perform some experiments, likewise based on the AVEC 2012. To do so, in addition to maximizing the Pearson correlation coefficient, we suggest a solution to the regression problem by maximizing the Hilbert-Schmidt independence criterion. Then, we compare the resulting correlation coefficient from regression by means of SVR (support vector regression) to those of the suggested techniques in

978-0-7695-5048-0/13 $26.00 © 2013 IEEE
DOI 10.1109/ACII.2013.39

this work. A major advantage of the suggested algorithms is that they have closed-form solutions, unlike the SVR. In addition, considering the simplicity of the considered optimization problems, the suggested algorithms are particularly faster than the support vector regression, in terms of the training and recall times. Moreover, the numerical results of the performed experiments show that depending on time granularity, the correlation coefficient of prediction using the suggested algorithms have a relative mean advantage of 46 to 64 percents with respect to SVR.

The rest of this work is organized as follows. In Section 2 we go over the technicalities of correlation coefficient-maximizing regression algorithms. Section 3 is dedicated to the presentation of our experimental study, where we explain the details of the experimental setup and discuss the numerical results. This work ends with concluding remarks in Section 4.

II. Correlation-Maximizing Regression

Given a set of explanatory variables $\mathbf{X} \in \mathcal{X}$ ($\mathcal{X} \subset \mathbb{R}^p$) and a response variable $\mathbf{y} \in \mathcal{Y}$ ($\mathcal{Y} \subset \mathbb{R}$), our objective is to find a linear mapping of $\mathcal{X}$ onto $\mathcal{Y}$ that maximizes correlation with the response variable $\mathbf{y}$. That is, to solve the following optimization problem.

$$\underset{\mathbf{w}}{\operatorname{argmax}} \quad \text{correlation}(\mathbf{y}, \mathbf{X}\mathbf{w}) \tag{1}$$

where $\mathbf{y}$ is an $N \times 1$ vector, $\mathbf{X}$ an $N \times p$ matrix, and $\mathbf{w}$ a $p \times 1$ vector, with N and p being the number of instances and the number of explanatory variables, respectively.

To solve this problem, we consider two senses of correlation, namely Pearson's correlation coefficient and Hilber-Schmidt independence criterion. As for the notation that we use in this section, vectors and matrices are denoted by lower and uppercase boldface letters, respectively, and scalars by either lower or uppercase normal letters.

A. Pearson's Correlation Coefficient (CC)

For two vectors $\mathbf{y}$ and $\hat{\mathbf{y}}$, the definition of the Pearson's correlation coefficient is as follows.

$$r(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\mathbf{y}_c^\top \hat{\mathbf{y}}_c}{\sqrt{\mathbf{y}_c^\top \mathbf{y}_c}\sqrt{\hat{\mathbf{y}}_c^\top \hat{\mathbf{y}}_c}}, \tag{2}$$

where the sub-c notation indicates the centered variables. That is, $\mathbf{y}_c = \mathbf{y} - \overline{\mathbf{y}}$, where $\overline{\mathbf{y}}$ denotes the mean value of $\mathbf{y}$.

Our objective here is to choose a $\mathbf{w}$ that yields the maximum correlation coefficient of $\mathbf{X}\mathbf{w}$ with $\mathbf{y}$. However, since r varies between -1 and 1, and that the two extremes are equally as desired, one may instead maximize $r^2(\mathbf{y}, \mathbf{X}\mathbf{w})$.

$$\begin{aligned} r^2(\mathbf{y}, \mathbf{X}\mathbf{w}) &= \left(\frac{\mathbf{y}_c^\top \mathbf{X}_c \mathbf{w}}{\sqrt{\mathbf{y}_c^\top \mathbf{y}_c}\sqrt{\mathbf{w}^\top \mathbf{X}_c^\top \mathbf{X}_c \mathbf{w}}}\right)^2 \\ &= \frac{\mathbf{w}^\top \mathbf{X}_c^\top \mathbf{y}_c \mathbf{y}_c^\top \mathbf{X}_c \mathbf{w}}{(\mathbf{y}_c^\top \mathbf{y}_c)(\mathbf{w}^\top \mathbf{X}_c^\top \mathbf{X}_c \mathbf{w})}. \end{aligned}$$

Since $\mathbf{y}_c^\top \mathbf{y}_c$ is invariant with respect to $\mathbf{w}$, it does not impact a choice of $\mathbf{w}$. Therefore, we formulate the optimization problem as follows.

$$\underset{\mathbf{w}}{\operatorname{argmax}} \quad \frac{\mathbf{w}^\top \mathbf{A}\mathbf{w}}{\mathbf{w}^\top \mathbf{B}\mathbf{w}}.$$

Where $\mathbf{A} = \mathbf{X}_c^\top \mathbf{y}_c \mathbf{y}_c^\top \mathbf{X}_c$ and $\mathbf{B} = \mathbf{X}_c^\top \mathbf{X}_c$.

Since the objective function is invariant with respect to $\mathbf{w}$ by a scaling factor, therefore there is no unique solution to this problem. To pin down the magnitude of the $\mathbf{w}$ and make the optimization problem well-defined, we can assume that the denominator is an arbitrary fixed scalar, i.e. $\mathbf{w}^\top \mathbf{B}\mathbf{w} = 1$. Therefore, the problem will change to the following:

$$\begin{aligned} \underset{\mathbf{w}}{\operatorname{argmin}} \quad & -\mathbf{w}^\top \mathbf{A}\mathbf{w} \\ \text{subject to} \quad & \mathbf{w}^\top \mathbf{B}\mathbf{w} = 1. \end{aligned} \tag{3}$$

Using the lagrangian, we have

$$\mathcal{L} = -\mathbf{w}^\top \mathbf{A}\mathbf{w} + \lambda(\mathbf{w}^\top \mathbf{B}\mathbf{w} - 1).$$

According to the KKT conditions and properties of eigenvalues and eigenvectors, it can be shown that the solution to this problem is the eigenvector that corresponds to the greatest eigenvalue of the following generalized eigenvalue problem:

$$\mathbf{A}\mathbf{w} = \lambda \mathbf{B}\mathbf{w}. \tag{4}$$

B. Hilbert Schmidt Independence Criterion (HSIC)

Assuming $\mathcal{F}$ and $\mathcal{G}$ to be two separable reproducing kernel Hilbert spaces [16] and $\mathcal{Z} := \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\} \subseteq \mathcal{X} \times \mathcal{Y}$, the empirical estimate of HSIC [17] is defined as

$$\text{HSIC}(\mathcal{Z}, \mathcal{F}, \mathcal{G}) = (N-1)^{-2}\mathbf{tr}(\mathbf{KHLH}). \tag{5}$$

Where $\mathbf{K}$,$\mathbf{L}$,$\mathbf{H} \in \mathbb{R}^{N \times N}$, $K_{ij} := k(\mathbf{x}_i, \mathbf{x}_j)$, $L_{ij} := l(\mathbf{y}_i, \mathbf{y}_j)$, and $H := \mathbf{I} - N^{-1}\mathbf{e}\mathbf{e}^T$, $\mathbf{e}$ is a vector of all ones. It can be shown that the HSIC of two independent kernels $\mathbf{K}$ and $\mathbf{L}$ is zero. Therefore, in order to maximize the correlation between the two kernels, we need to maximize $\mathbf{tr}(\mathbf{KHLH})$ [18]. We assume $\mathbf{K}$ to represent a kernel of the linear mapping, that is $\mathbf{X}\mathbf{w}$, and $\mathbf{L}$ a kernel of the response variable $\mathbf{y}$. By further assuming that the two kernels are linear, i.e., $\mathbf{K} = \mathbf{X}\mathbf{w}\mathbf{w}^\top \mathbf{X}^\top$ and $\mathbf{L} = \mathbf{y}\mathbf{y}^\top$, as a result we will have:

$$\begin{aligned} \mathbf{tr}(\mathbf{KHLH}) &= \mathbf{tr}(\mathbf{X}\mathbf{w}\mathbf{w}^\top \mathbf{X}^\top \mathbf{H}\mathbf{y}\mathbf{y}^\top \mathbf{H}) \\ &= \mathbf{tr}(\mathbf{w}^\top \mathbf{X}^\top \mathbf{H}\mathbf{y}\mathbf{y}^\top \mathbf{H}\mathbf{X}\mathbf{w}) \end{aligned}$$

Hence, we are interested in the solution of the following optimization problem.

$$\begin{aligned} \underset{\mathbf{w}}{\operatorname{argmax}} \quad & \mathbf{tr}(\mathbf{w}^\top \mathbf{X}^\top \mathbf{H}\mathbf{L}\mathbf{H}\mathbf{X}\mathbf{w}) \\ \text{subject to} \quad & \mathbf{w}^\top \mathbf{w} = 1. \end{aligned} \tag{6}$$

Without the constraint on the magnitude of $\mathbf{w}$, there wouldn't be any upper bound to the objective function. Through a set of algebraic manipulations, it can be shown that the solution to this optimization problem is the eigenvector of $\mathbf{X}^\top \mathbf{H}\mathbf{L}\mathbf{H}\mathbf{X}$ that corresponds to its greatest eigenvalue.

C. Regression Error

To this point, we have discussed two ways for solving a regression problem with the intention of maximizing the correlation rather than minimizing the regression error. In addition, we would like to achieve the best that we can in terms of the regression error. From the properties of the covariance matrix and the definition of Pearson's correlation coefficient, we have:

$$r(\mathbf{y}, \theta\hat{\mathbf{y}} + \theta_0) = r(\mathbf{y}, \hat{\mathbf{y}}). \tag{7}$$

Therefore, without disturbing the result of correlation-maximizing regression, we can use θ and θ_0 to introduce two degrees of freedom, by use of which we can minimize the regression error. A natural solution to this problem is the least squares. Hence, the solution that we suggest for the regression problem consists of two optimization problems. The primary optimization guarantees maximum correlation and the secondary optimization minimizes the regression error.

III. Numerical Experiments

To investigate the applicability of the suggested correlation-maximizing regression for continuous emotion recognition, we have performed some experiments that we present in this section. The experiments are designed in the framework of the continuous audio/visual emotion challenge (AVEC) 2012 [8]. Set up in two different time granularities, the challenge targets fully continuous and word-level emotion recognition. In the fully continuous setting, the emotional states of the speakers in every 50 mSec window are estimated, whereas in the word-level setting the emotional content of the experssion of each word is of interest. As for the emotional primitives, arousal, expectation, power, and valence are considered. Since the purpose of this study is to propose two algorithms and compare those with a state of the art algorithm like support vector regression, the choise of modalities and features may not impact the objectives of the study, as long as similar conditions are maintained across the experiments. In this work, the experiments are performed based on the audio signal.

In the remainder of this section, we briefly talk about the database that is used in the experiments and our choice of speech features. We then present the result of the study, and discuss our findings.

A. Dataset

For this experiment, we have used the part of the SEMAINE corpus [19], [20] that was used for the AVEC 2012 [8]. SEMAINE is a database recorded based on the sensitive artificial listener (SAL) interaction scenario [21]. The aim of SAL is to evoke strong emotional responses in a listener by controlling the statements of an operator (the script is predefined in this scenario). For this purpose, four agents are introduced, and a user can decide at any time to which operator she/he would like to talk; each of those agents tries to simulate a different personality: Poppy tries to evoke happiness, Obadiah tries to evoke sadness, Spike tries to evoke anger, and Prudence tries to make people sensible. Therefore, a combination of those decisions is claimed to result in a highly emotional conversation. Solid SAL is a similar scenario to SAL, for which there is no predefined script given to the operators. Instead, they are free to act as one of the four SAL agents at any time. This is done for the sake of a more natural face-to-face conversation, as in the SAL scenario reading the script or recalling it (in case operators have memorized the script) may not allow such non-verbal interactions.

Selected from the recordings of the SEMAINE, AVEC 2012 provides three sets of data, labeled as training, development, and testing sets. Since the response variables for the testing set is not made available to the public, we use the training and development sets for training and hypothesis testing purposes. The number of sessions in the training and development set are 31 and 32, respectively.

B. Features

As for the audio features, we use the spectral energy distribution (SED) in this work [22], [10]. SED is comprised of a set of components, where each component represents the relative energy of the signal in a specific band of the spectrum. For the speech signal $s[n]$, the definition of the component i is as follows.

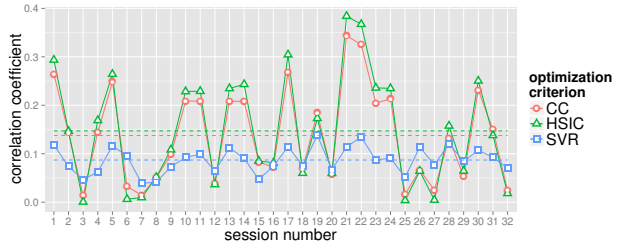
$$\text{SED}_i = \sum_{k=1}^{N} [H[k - U_i] - H[k - L_i]] g(S[k])^2, \tag{8}$$

where $S[k]$ is the discrete Fourier transform of $s[n]$; $H[k]$ is the unit step function (a.k.a. the Heaviside step function); L_i and U_i indicate the lower and upper bounds of the component in the spectrum; and $g(.)$ is a normalizing function. In this equation, N denotes the number of samples of the signal, or similarly of a window of the signal, which by principle equals the length of the signal times its sampling frequency.

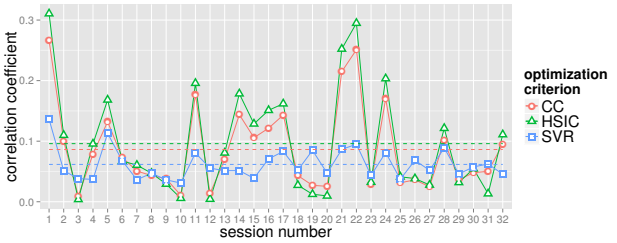
Features are extracted from 100 mSec-long windows of speech signal. The length of the spectral intervals is set to 100 Hz, where two consecutive intervals do not intersect, and collectively they cover 0 to 8 kHz of the spectrum. To set each of these parameters, which are the window size in time domain and the spectral bandwidth, a line search is performed. As for statistics, we use the min, max, median, mean, and standard deviation of the features over windows of a speech signal. This makes a feature vector of 400 dimensions.

C. Results

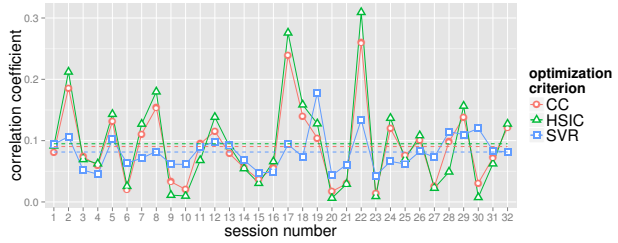
The objective of these experiments is to investigate the quality of a model, in terms of correlation coefficient and mean absolute error of the model's predictions. Although all the models used in this work are of the family of linear models, we have used three different criteria to learn the linear coefficients. Those are as follows: maximizing Pearson's correlation coefficient (CC), maximizing Hilbert-Schmidt independence criterion (HSIC), and support vector regression (SVR). For each of the two time granularities, which are fully continuous and word-level, and for each of the four emotion dimension, which are arousal, expectation, power, and valence, we have trained three models, using the three criteria. For



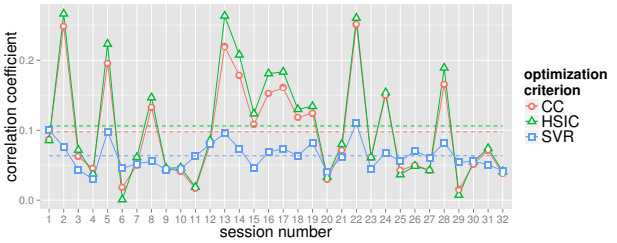
(a) Fully Continuous Emotion Recognition – Arousal



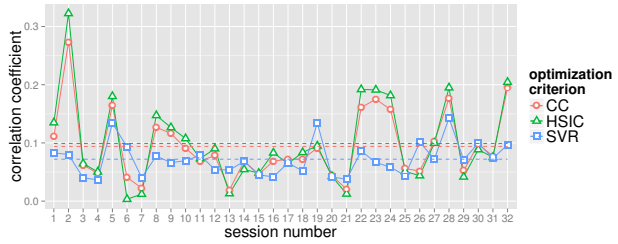
(b) Word Level Emotion Recognition – Arousal



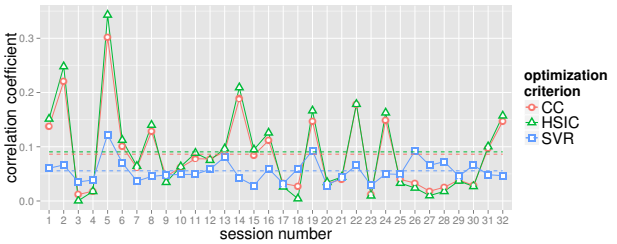
(c) Fully Continuous Emotion Recognition – Expectation



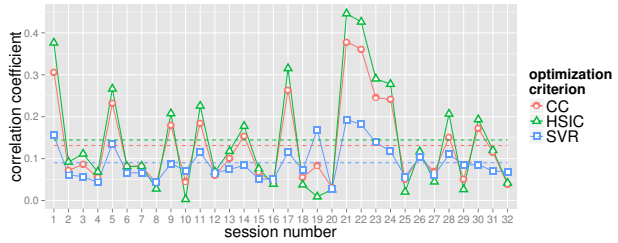
(d) Word Level Emotion Recognition – Expectation



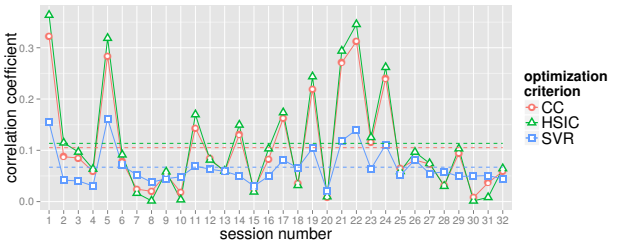
(e) Fully Continuous Emotion Recognition – Power



(f) Word Level Emotion Recognition – Power



(g) Fully Continuous Emotion Recognition – Valence

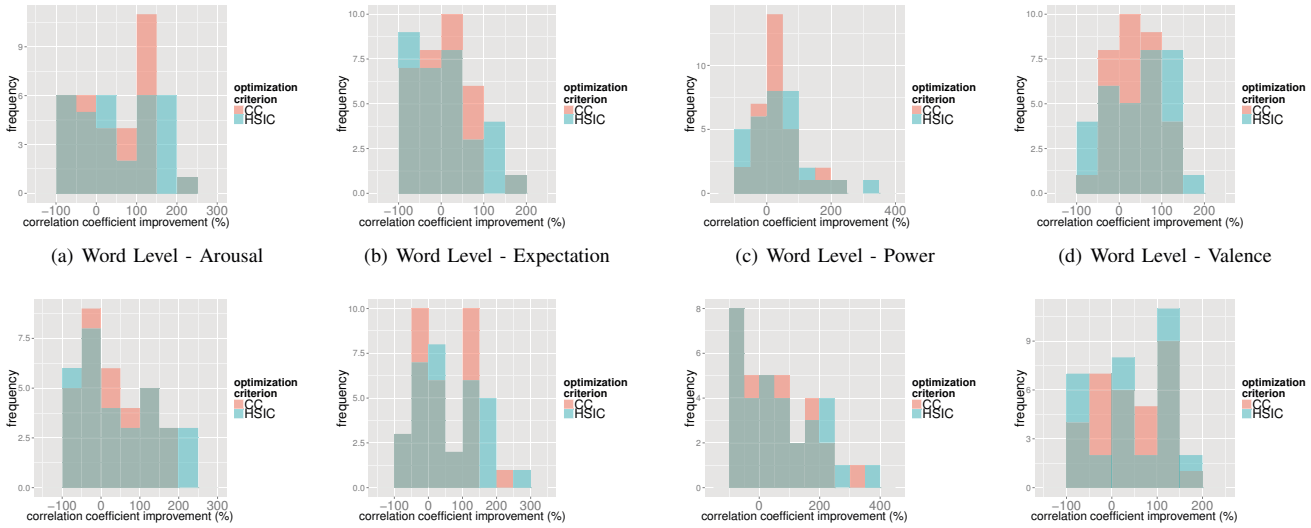


(h) Word Level Emotion Recognition – Valence

Fig. 1. Correlation coefficient of the predictions with the actual values, per session of the development set. The dashed lines indicate the average correlation coefficient of predicted values with the actual response values, for each method and over all the sessions.

the experiments in this work, all the response variables were normalized to vary between zero and one. For each of these tasks, while there is no need to set any parameters for the two correlation-maximizing algorithms, we choose the complexity parameter of the SVR from $C \in 10^{\{-2,-1,\ldots,2\}}$. To train SVR, explanatory variables were normalized.

We have used each of the models to predict the emotional content of each of the 32 sessions of the development set. Correlation coefficient of the predictions with the actual emotional content of each of the sessions is shown in Figure 1. The dashed lines in these figures show the average correlation coefficient of the prediction over all the sessions. For the arousal dimension, for both time granularities, we notice that the HSIC and CC regression result in higher average correlation coefficients than that of the SVR. Similarly, for the expectation dimension, the two correlation maximizing regression algorithms surpass the SVR with respect to the correlation coefficient of their prediction, however in the case of the fully continuous recognition, the differences are as noticeable. Again, for the power and valence emotion dimensions, for both time granularities, HSIC regression results in a higher correlation than those of the CC regression and SVR.



(a) Word Level - Arousal (b) Word Level - Expectation (c) Word Level - Power (d) Word Level - Valence

(e) Fully Continuous - Arousal (f) Fully Continuous - Expectation (g) Fully Continuous - Power (h) Fully Continuous - Valence

Fig. 2. Distribution of the development sessions population in terms of the relative correlation coefficient advantage of the CC and HSIC regressions with respect to SVR.

Looking closer at Figure 1, we notice that for most of the sessions HSIC regression works the best among the three, however, there are some sessions and emotional dimensions for which SVR or CC result in better correlation coefficients. To show these differences more clearly, we use the distribution of the relative advantage of the CC and HSIC regressions with respect to the SVR. What we mean by relative advantage is the correlation difference between each of the CC and HSIC regression with the SVR, normalized by the correlation of the SVR. Those distributions are shown in Figure 2. In Table I, the average relative advantage of the CC and HSIC regression to SVR is recorded. According to this information, except the expectation and power dimensions in the fully continuous recognition task, for the other six cases, which are arousal and valence for both time granularities and expectation and power for word-level recognition, HSIC regression is more than 50% advantageous to the SVR.

A valid question here is how the two correlation-maximizing models perform in terms of the prediction error. To answer to this question, we have included the average root mean squared error (RMSE) of the prediction over all the sessions along with the average correlation coefficients of the predictions in Table II. There, one can notice that compared to the SVR, both the CC and HSIC regressions result in higher correlation coefficient and lower prediction error. Moreover, in the same table the average training and recall times (T_T and T_R) are included. According to these numbers, both CC and HSIC regressions are, by an order of 4 to 23 time for training and 60 to 87 times for recall, faster than the SVR.

We are convinced that the advantage of the CC and HSIC regressions to SVR is due to the different nature of the objective functions that they optimize. Regarding the advantage of the HSIC to CC regression, since the matrix $\mathbf{B}$ in the equation 4 can be numerically near singular, calculating the inverse of the matrix may result in approximate solution for the regression coefficients. Whereas, there is need for the use of the inverse operation in the calculation of the maximum HSIC regression coefficients. Due to the same reason, the training time of the HSIC regression is less than that of the CC regression. About the processing time of these three algorithms, CC and HSIC regressions take considerably less time to train than SVR, since unlike the latter, the formers have closed-form solutions. Moreover, those two algorithms are non-parametric, therefore there are no parameters to be optimized in their learning process.

TABLE I
THE RELATIVE ADVANTAGE OF THE CORRELATION COEFFICIENT OF PREDICTION USING HSIC AND CC REGRESSION, WITH RESPECT TO THAT OF PREDICTED BY SVR.

	Arousal	Expectation	Power	Valence
Fully Continuous Emotion Recognition				
CC	58%	11%	31%	46%
HSIC	69%	17%	38%	60%
Word-Level Emotion Recognition				
CC	40%	54%	55%	57%
HSIC	56%	67%	63%	69%

Regarding the prediction error that we get, we suspect that the reason why the two correlation-maximizing regressions result in lower error than that of the SVR can be explained by the fact that to minimize the error, we introduced two new parameters to the model in Section II-C, therefore the resulting model has one more degree of freedom than that of the SVR.

IV. CONCLUSION

Fitting a regression model is an inevitable part of a solution to the continuous recognition of emotions. The classic training algorithms minimize the prediction error of the regression model, however in some cases minimizing the error might

TABLE II

COMPARISON OF THE PERFORMANCE OF DIFFERENT REGRESSION MODELS, IN TERMS OF THE CORRELATION COEFFICIENT (CC) AND THE ROOT MEAN SQUARED ERROR (RMSE) OF THE PREDICTED VALUES, AS WELL AS THE TRAINING AND RECALL TIME (T_T AND T_R) IN MILLISECOND.

Optimization	Arousal		Expectation		Power		Valence		Average			
Criterion	CC	RMSE	CC	RMSE	CC	RMSE	CC	RMSE	CC	RMSE	T_T	T_R
Fully Continuous Emotion Recognition												
CC	0.138	0.193	0.090	0.132	0.094	0.038	0.131	0.180	0.113	0.136	5368	1
HSIC	**0.147**	0.193	**0.095**	0.132	**0.099**	0.038	**0.144**	0.180	**0.121**	0.136	**1416**	1
SVR	0.087	0.200	0.081	0.138	0.072	0.039	0.090	0.185	0.082	0.140	23968	60
Word-Level Emotion Recognition												
CC	0.086	0.189	0.098	0.157	0.086	0.118	0.105	0.180	0.094	0.161	5615	2
HSIC	**0.096**	0.189	**0.106**	0.157	**0.091**	0.118	**0.114**	0.181	**0.102**	0.161	**3304**	2
SVR	0.062	0.194	0.064	0.164	0.056	0.123	0.067	0.185	0.062	0.166	74893	174

fall at a second level of importance. In this work, with the objective of maximizing the correlation of the prediction values with the actual value of the response variable, we suggested two algorithms, one based on the Pearson's correlation coefficient and the other based on the Hilbert-Schmidt independence criterion (HSIC). The numerical experiments of this study, which were performed in the framework of the continuous audio/visual emotion challenge 2012, show that compared to the support vector regression, the two correlation-maximizing algorithms improve the prediction accuracy, from both correlation coefficient and prediction error points of view. As a future work, we would like to see the generalization of the proposed algorithms to other modalities, e.g. facial expressions. Moreover, as discussed, the HSIC regression is defined based on the kernel spaces of the explanatory and response variables, however, in this work we only considered the linear kernels of those variables. Therefore, as another potential future direction of this work, we would like to study the effect of deviation from linear kernel.

REFERENCES

[1] H. G. Wallbott and K. R. Scherer, "Cues and channels in emotion recognition," *Journal of personality and social psychology*, vol. 51, no. 4, pp. 690–699, 1986.

[2] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39 –58, jan. 2009.

[3] P. Ekman, *Basic Emotions*. Sussex, U.K.: John Wiley & Sons, Ltd, 1999.

[4] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32 –80, jan 2001.

[5] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 1980.

[6] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, pp. 261–292, 1996.

[7] J. Fontaine, K. Scherer, E. Roesch, and P. Ellsworth, "The world of emotions is not two-dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, 2007.

[8] B. Schuller, M. Valstar, F. Eyben, R. Cowie, and M. Pantic, "Avec 2012 – the continuous audio/visual emotion challenge," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, ser. ICMI '12. ACM, 2012, pp. 449–456.

[9] A. C. Cruz, B. Bhanu, and N. Thakoor, "Facial emotion recognition with expression energy," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, ser. ICMI '12. New York, NY, USA: ACM, 2012, pp. 457–464.

[10] P. Fewzee and F. Karray, "Elastic net for paralinguistic speech recognition," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, ser. ICMI '12. New York, NY, USA: ACM, 2012, pp. 509–516.

[11] M. Glodek, M. Schels, G. Palm, and F. Schwenker, "Multiple classifier combination using reject options and markov fusion networks," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, ser. ICMI '12. ACM, 2012, pp. 465–472.

[12] L. van der Maaten, "Audio-visual emotion challenge 2012: a simple approach," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, ser. ICMI '12. New York, NY, USA: ACM, 2012, pp. 473–476.

[13] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, "Robust continuous prediction of human emotions using multiscale dynamic cues," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, ser. ICMI '12. ACM, 2012, pp. 501–508.

[14] D. Ozkan, S. Scherer, and L.-P. Morency, "Step-wise emotion recognition using concatenated-hmm," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, ser. ICMI '12. ACM, 2012, pp. 477–484.

[15] C. Soladié, H. Salam, C. Pelachaud, N. Stoiber, and R. Séguier, "A multimodal fuzzy inference system using a continuous facial expression representation for emotion detection," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, ser. ICMI '12. ACM, 2012, pp. 493–500.

[16] M. Hein and O. Bousquet, "Kernels, associated structures and generalizations," Max Planck Institute for Biological Cybernetics, Tech. Rep. 127, 2004.

[17] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with hilbert-schmidt norms," in *Algorithmic Learning Theory*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2005, vol. 3734, pp. 63–77.

[18] E. Barshan, A. Ghodsi, Z. Azimifar, and M. Z. Jahromi, "Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds," *Pattern Recognition*, vol. 44, no. 7, pp. 1357 – 1371, 2011.

[19] G. McKeown, M. Valstar, R. Cowie, and M. Pantic, "The semaine corpus of emotionally coloured character interactions," in *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, july 2010, pp. 1079 –1084.

[20] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 5 –17, jan.-march 2012.

[21] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amier, and D. K. J. Heylen, "The sensitive artificial listner: an induction technique for generating emotionally coloured conversation," in *LREC Workshop on Corpora for Research on Emotion and Affect, Marrakech, Marokko*. Paris, France: ELRA, 2008, pp. 1–4.

[22] P. Fewzee and F. Karray, "Emotional speech: A spectral analysis," in *Proceedings of Interspeech 2012, Portland, OR, USA*, September 2012.