CCA BASED FEATURE SELECTION WITH APPLICATION TO CONTINUOUS DEPRESSION RECOGNITION FROM ACOUSTIC SPEECH FEATURES

Heysem Kaya[†‡⋆]*, Florian Eyben*[‡]*, Albert Ali Salah*[†]*, Björn Schuller*[§‡]

[†]Bogazici University, Department of Computer Engineering, Istanbul / Turkey
[‡]Technische Universität München, Institute for Human-Machine Communication, München / Germany
[§]Imperial College London, Department of Computing, London / UK
heysem@boun.edu.tr, eyben@tum.de, salah@boun.edu.tr, schuller@ieee.org

ABSTRACT

In this study we make use of Canonical Correlation Analysis (CCA) based feature selection for continuous depression recognition from speech. Besides its common use in multimodal/multi-view feature extraction, CCA can be easily employed as a feature selector. We introduce several novel ways of CCA based filter (ranking) methods, showing their relations to previous work. We test the suitability of proposed methods on the AVEC 2013 dataset under the ACM MM 2013 Challenge protocol. Using 17% of features, we obtained a relative improvement of 30% on the challenge's test-set baseline Root Mean Square Error.

Index Terms— Canonical Correlation Analysis, feature selection, feature extraction, depression recognition, affect recognition, acoustic speech processing

1. INTRODUCTION

The state-of-the-art computational paralinguistics applications are built using suprasegmental features obtained from functionals operating on frame-level low level descriptors (LLD) [1]. Brute-force extraction of high-dimensional potent features is commonly encountered in competitive baseline feature sets of the most recent computational paralinguistics challenges [2, 3]. One such feature set is recently introduced for the ACM Multimedia 2013 Challenge Audio-visual Emotion Corpus (AVEC 2013) which is about prediction of continuous depression and affect labels from audio/visual data [4]. In this study, we present our work with AVEC 2013 focusing on prediction of depression level using acoustic features.

In the state-of-the-art pipeline of computational paralinguistics processing, high dimensional data are classified generally with Support Vector Machines (SVM) or Random Forests (RF), which are less vulnerable to vagaries of high dimensionality. Though effective in certain conditions, using high-dimensional feature vectors restricts the use of back-end classifiers as well as intermediate feature enhancement techniques due to the *curse of dimensionality*. Moreover, high-dimensional datasets are prone to contain many irrelevant and redundant features, which reduces the generalization power of any learner.

In this paper, we investigate new acoustic feature selection techniques to overcome the problem of over-fitting and to provide a compact set of high-quality features. In machine learning, feature selection aims at finding "a minimal yet predictive" subset of original features [5]. Feature selection techniques can broadly be categorized into filter/ranking based methods and wrapper methods. While wrapper methods use predictors (classifiers/regressors) to assess the suitability, filter based methods use a *heuristic merit* of a feature (or subset) to guide the ranking process.

We handle the problem of feature selection via Canonical Correlation Analysis (CCA), as it provides a general infrastructure for feature selection and extraction. Moreover, CCA based feature reduction can be employed both with categorical and continuous targets. This is important considering the research trend in paralinguistic analysis and prediction, which shifts partly towards continuous targets [6]. Yet, there is a growing interest in extending CCA with applications to acoustic speech processing [7, 8].

We present three feature selection methods using CCA. One of the methods is related to minimum Redundancy Maximum Relevance (mRMR) [9] feature selection and Correlation Based Feature Selection (CFS) [10], as it intends to minimize internal dependence/correlation of selected features and maximize the dependence between selected feature/set and the target variable. In mRMR, the heuristic merit uses Mutual Information (MI), which is a nonlinear measure of dependence [9]. MI quantifies the information shared by two variates in number of bits. In CFS, [10] refers to correlation as a general measure of dependence not restricting it to Pearson's correlation. We use CSF as one of our benchmark methods. We do not use mRMR as benchmark, as it requires

⋆Corresponding author. This work was performed while H. Kaya was at Technische Universität München, Institute for Human-Machine Communication on a visit funded by Turkish Higher Education Council (YÖK). This study is supported by Bogazici University BAP 6531 project.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

discretization to perform well and the sensitivity to discretization should be analyzed in a separate study. We also present results with a CCA based feature selection that does not require greedy selection. This method is computationally the simplest among presented CCA based feature selection methods and already has successful applications [11].

The remainder of this paper is organized as follows. In the next section we provide background on CCA. We present CCA based feature selection methods in Section 3. In Section 4, we review mRMR and CFS showing their relation to our proposed feature selection methods. Then in Section 5 we introduce the corpus and features. Experimental work is presented in Section 6, while Section 7 concludes with future directions.

2. BACKGROUND: CANONICAL CORRELATION ANALYSIS

Proposed by Hotelling [12], CCA seeks to maximize the mutual correlation between two sets of variables by finding linear projections for each set. It is dened as a method to drive feature selection/extraction in order to elicit latent semantics by treating each view as a complex label for the other view [13]. Mathematically speaking, CCA seeks to maximize the mutual correlation between two *views* of the same semantic phenomenon (e. g. audio and video of a speech) denoted $A \in \mathbb{R}^{n \times d}$ and $B \in \mathbb{R}^{n \times p}$ via:

$$\rho(A, B) = \sup_{w,v} corr(w^T A, v^T B), \tag{1}$$

where "corr" corresponds to Pearson's correlation, w and v correspond to the projection vectors of A and B, respectively. Let C_{AB} denote the cross-set covariance between the sets A and B, and similarly let C_{AA} denote within set covariance for A. The problem given in eq. (1) can be re-formulated as:

$$\rho(A, B) = \sup_{w,v} \frac{w^T C_{AB} v}{\sqrt{w^T C_{AA} w \cdot v^T C_{BB} v}}. \tag{2}$$

The formulation in eq. (2) can be converted into a generalized eigenproblem once the terms in the denominator, i. e. $w^T C_{AA} w$ and $v^T C_{BB} v$, are constrained to unity, by observing the fact that the correlation does not depend on the scale of projection vectors. For both linear projections (i. e. w and v), the solution can be shown [13] to have the eigenform of:

$$C_{AA}^{-1} C_{AB} C_{BB}^{-1} C_{BA} w = \lambda w, \tag{3}$$

where the correlation appears to be the square root of eigenvalue:

$$\rho(A, B) = \sqrt{\lambda}. \tag{4}$$

To attain maximal correlation, the eigenvector corresponding to the largest eigenvalue in eq. (3) should be selected. Similarly, by restricting the new vectors to be uncorrelated with the previous ones, it can be shown that the projection matrices for each set are spanned by the k eigenvectors corresponding to the k largest eigenvalues.

There are also nonlinear extensions of CCA. Kernel CCA (KCCA), uses the *kernel trick* in the same vein with SVM [13] while Deep CCA (DCCA) is an efficient deep neural network alternative to KCCA [7].

3. CCA BASED FEATURE SELECTION

We present three different ways of CCA based feature selection. All three methods are valid for continuous and categorical targets, and can be used for feature extraction in the same way.

3.1. Samples versus Labels CCA

When CCA is used to find the canonical correlation between samples with high dimensionality and corresponding labels, the resulting projection vector for features can be directly used for feature selection. One can simply discard zero weight (or below a threshold) features and rank the absolute value of the remaining projection weights. This setting of CCA is called Samples versus Labels CCA (SLCCA) [14] and as a feature selection method it is successfully used for fMRI analysis [11]. To the best of our knowledge however, this method is not used in acoustic feature selection. We call this method the SLCCA-Filter.

3.2. Minimum Redundancy Maximum Relevance CCA

We propose the minimum Redundancy Maximum Relevance CCA (mRMR-CCA) first, which is related to CFS [10] and mRMR [9]. The difference from these methods is that instead of computing feature-wise internal correlations and averaging results, we directly compute canonical correlation of a candidate feature against the already selected subset:

$$\max_{x_j \in X - S_{k-1}} \left[\rho_{CCA}(x_j, t) - \rho_{CCA}(x_j, S_{k-1}) \right], \tag{5}$$

where S_{k-1} denotes the already selected subset with $k-1$ features, x_j is a candidate feature and t is the target variable (label). In eq. (5) the subtraction operator can be replaced with division, to account for the relative merit with respect to internal correlations. In our experiments, we used subtraction based measures.

3.3. Maximum Collective Relevance CCA

Our second proposed method, Maximum Collective Relevance CCA (MCR-CCA) focuses on maximizing the joint correlation of the selected subset and the candidate feature against the target. The redundancy in the ranked subset can further be reduced using feature extraction. The formulation

is similar to wrapper based forward selection, but we do not employ a classifier:

$$\max_{x_j \in X - S_{k-1}} \left[\rho_{CCA}(S_{k-1} \cup x_j, t) \right]. \quad (6)$$

4. RELATION TO PREVIOUS WORK

One of the novel methods we introduce, namely mRMR-CCA, is inspired from CFS [10] and mRMR [9]. CFS measures the heuristic merit between a feature set S and target t via [10]:

$$r_{S,t} = \frac{k\overline{r_{zi}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}}, \quad (7)$$

where k is number of features, $\overline{r_{zi}}$ denote average correlation between the features in the subset and the target variable, and the term $\overline{r_{ii}}$ denote average inter-correlation between features. In short, eq. (7) punishes internal correlation and favors higher average feature-target correlations. Hall (1999) proposes several measures of dependence to compute feature-feature and feature-target merits of a subset. When the target variable is continuous, Pearson's correlation coefficient is used. In our approach we simplify eq. (7), keeping the notion of high relevance low redundancy. Similarly, mRMR drives the feature selection in a set X, at step k maximizing the difference or ratio between relevance and redundancy terms [9]:

$$\max_{x_j \in X - S_{k-1}} \left[MI(x_j, t) - \frac{1}{k-1} \sum_{x_i \in S_{k-1}} MI(x_j, x_i) \right], \quad (8)$$

where $MI(x, y)$ is mutual information between random variables x and y. In KCCAmRMR, Sakar et al. [15] improved mRMR feature selection using correlated functions of variables (i. e. projections attained by CCA) weighted with corresponding correlations with the target variable. In our work, we completely replace MI with CCA. Moreover, CCA not only eliminates discretization for continuous targets, but also is capable of handling multiple targets in the feature reduction process.

We next introduce the data for experimental validation.

5. AVEC 2013 DATABASE

5.1. Depression Corpus

AVEC 2013[4] uses a subset of the audio-visual depressive language corpus (AVDLC), which includes 340 video clips of subjects performing a Human-Computer Interaction task while being recorded by a webcam and a microphone. In AVDLC, the total number of subjects is 292 and only one person appears per clip, i. e. some subjects feature in more than one clip. The speakers were recorded between one and four times, with a period of two weeks between the measurements. Table 1 summarizes basic statistics of the corpus [4].

Table 1. Statistics of the AVDLC [4]

Property	Statistic
# of Clips	340
# of Subjects	292
Range of Clip Length	20-50 min.
Mean Clip Length	25 min.
Total Duration	240 hours
Age Range of Subjects	18-63 years
Mean±Std of Age of Subjects	31.5±12.3 years
BDI-II Score Range	0-45

Recorded behavior includes speaking out loud while solving a task, counting from 1 to 10, read speech (excerpts of a novel and a fable), singing in German, telling a story from the subjects' own past (the best event and a sad event from childhood). The depression levels were labeled per clip using Beck Depression Inventory-II (BDI-II) [16], a subjective self-reported 21 item multiple-choice inventory.

For the AVEC 2013 challenge, the recordings were split into three partitions: training, development, and test sets of 50 recordings each, respectively.

5.2. Baseline Acoustic Feature Sets

The AVEC 2013 audio baseline feature set, which is an extended set of features with respect to AVEC 2012 [2], consists of 2 268 features. Due to space limitations, the reader is referred to the challenge paper [4] for the details of the LLDs and functionals.

The audio features are computed on short episodes of audio data. Since the Challenge dataset contains long continuous recordings, three segmentations have been performed: 1) voice activity detection (VAD) based 2) overlapping short fixed length segments (3 seconds) and, 3) overlapping long fixed length segments (20 seconds). For VAD segmentation, pauses of more than 200 ms are used to split speech activity segments. In short and long segmentation, the windows are shifted forward at a rate of one second. Functionals are then computed over each segment. Together with the per instance computation of functionals, the baseline feature set is provided in 4 versions to grasp relatively short-long acoustic characteristics of speech intended for depression and affect tasks. See Table 2 for the distribution of instances.

Table 2. Instance Distribution per Partition and Segmentation

#	Train	Dev	Test
Per Clip	50	50	50
VAD Seg	6015	5763	5946
Short Seg	23863	23513	23824
Long Seg	23439	23087	23399

6. EXPERIMENTAL WORK

In our experiments we used the AVEC 2013 [4] challenge baseline feature set focusing on the depression sub-challenge. We used the WEKA [17] implementation of CFS with "Best First" search and Bagging-REPTree (BRep) from the same package as classifier. The hyper-parameters of both methods are left as default. As detailed before, we followed the training, development and testing protocol of the challenge. Therefore, we optimized the investigated feature selection methods on the development set and finally used the optimal setting for predicting labels on the sequestered test set.

6.1. Results and Discussion

For developing candidate hypotheses and selecting the best features for challenge test set, we utilized the training and development set. Baseline acoustic features using Support Vector Machine Regressor (SVR) with linear kernel gives Mean Absolute Error (MAE) of 8.66 and Root Mean Square Error (RMSE) of 10.75 for the development set [4].

We first used SVR (Linear Kernel, C=0.0001) and Bagging REPTree (V = 0.001) as classifiers on VAD segmented features. In all our experiments we tested five feature settings: 1) All Baseline Features (denoted All) together with selected sets using 2) SLCCA-Filter 3) mRMR-CCA 4) MCR-CCA and as independent benchmark 5) Correlation Based Feature Selection (CFS). Over five feature settings, we obtained better results with BRep (mean RMSE 11.15) against SVR (mean RMSE 12.24) with an order-of-magnitude less training time. So, we choose BRep as regressor.

We next experimented with five feature settings in all four segmentations. Considering the computational complexity of CCA (whose bottleneck is inversion of covariance matrix of samples, which scales cubicly with the number of selected features) we used the first 100 ranked features for MCR-CCA and mRMR-CCA. Once the threshold is determined, the number of features for SLCCA-Filter is automatically determined after a single application of CCA between the whole feature set versus the continuous depression labels. To probe the SLCCA-filter performance, we tested a set of thresholds $10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}$. The best development set results were obtained with a threshold of 10^{-5}. A summary of experiments is given in Table 3. In accordance with the results reported in [4], we observe that the long segmentation provides the best results for the depression task. Moreover, the simplest CCA based selection method, namely SLCCA-Filter, yields the best RMSE results in all segmentations. The number of selected features with SLCCA-Filter ranges from 387 (long seg) to 467 (short seg) with segmented sets. However, due to increased nullity of covariance (with 50 samples as opposed to 2 268 dimensions) in per instance set, the number of features contributing to covariate is found to be 49.

Focusing on long segmentation, we tested the first 400 ranked features for MCR-CCA and mRMR-CCA. The results were not found to improve considerably over the first 100 features: 8.13 MAE, 10.3 RMSE for MCR-CCA and 8.40 MAE and 10.97 RMSE for mRMR-CCA. Moreover, union and intersection of the selected feature sets pairwise did not improve over the performance of the SLCCA-Filter, individually.

Table 3. Development Set Performances per Feature Setting and Segmentation

	VAD SEG		SHORT SEG	
	MAE	**RMSE**	**MAE**	**RMSE**
All	9.01	11.25	8.37	10.42
SLCCA-Filter	9.13	11.15	7.99	10.36
MCR-CCA	9.31	11.54	8.73	10.86
mRMR-CCA	9.47	11.48	8.93	10.83
CFS	9.31	11.42	8.55	10.57

	LONG SEG		PER CLIP	
	MAE	**RMSE**	**MAE**	**RMSE**
All	7.93	10.24	9.75	11.89
SLCCA-Filter	7.84	10.22	8.92	11.00
MCR-CCA	8.27	10.72	9.81	11.55
mRMR-CCA	8.80	10.98	9.11	11.01
CFS	8.24	10.22	9.30	11.46

We therefore used SLCCA-Filter method with long segmentation to train a model for challenge test set. We obtained 7.83 MAE and 9.78 RMSE, improving challenge baseline test set RMSE performance (14.12) 30%, relative. These results also compare favorably to the best test set result of Meng et al. [18] (10.96 RMSE using audio-visual fusion) and Cummins et al. [19] (10.17 RMSE by using only audio information). Interestingly, unlike these recent studies that report better development set results using more complex systems, the development set performance of our computationally efficient SLCCA-Filter system is highly indicative of test set performance. Thus, SLCCA-Filter is thought to achieve the intended goal of avoiding over-fitting.

7. CONCLUSIONS AND FUTURE WORK

In this study, we presented two novel CCA based feature selection methods to reduce the massive dimensionality observed with the state-of-the-art acoustic feature sets in affective computing. We experimented on the recently published AVEC 2013 set to predict depression level. Results revealed that the computationally simple CCA based feature selection method worked the best on the development set. Using only 17% of original features, the SLCCA-Filter system yielded 30% decrease of RMSE over the baseline on challenge test set, advancing the state-of-the-art on this data. Extending the work with Kernel CCA, and testing the introduced methods on other emotion datasets comprise our nearest future work.

8. REFERENCES

[1] Björn Shuller, *In Salah, A. A. and Gevers, T. (eds) Computer Analysis of Human Behavior*, chapter Voice and Speech Analysis in Search of States and Traits, pp. 227–253, Springer, 2011.

[2] Björn Schuller, Michel Valstar, Florian Eyben, Roddy Cowie, and Maja Pantic, "AVEC 2012 – The Continuous Audio/Visual Emotion Challenge," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction, ICMI*, Louis-Philippe Morency, Dan Bohus, Hamid K. Aghajan, Justine Cassell, Anton Nijholt, and Julien Epps, Eds., Santa Monica, CA, October 2012, ACM, pp. 449–456.

[3] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, Marcello Mortillaro, Hugues Salamin, Anna Polychroniou, Fabio Valente, and Samuel Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proceedings INTERSPEECH 2013*, Lyon, France, August 2013, pp. 148–152, ISCA.

[4] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic, "AVEC 2013–The Continuous Audio/Visual Emotion and Depression Recognition Challenge," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 2013, AVEC '13, pp. 3–10.

[5] Ethem Alpaydin, *Introduction to Machine Learning*, The MIT Press, 2nd edition, 2010.

[6] Hatice Gunes and Björn Schuller, "Categorical and Dimensional Affect Analysis in Continuous Input: Current Trends and Future Directions," *Image and Vision Computing, Special Issue on Affect Analysis in Continuous Input*, vol. 31, no. 2, pp. 120–136, February 2013.

[7] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu, "Deep canonical correlation analysis," in *Proceedings of the 30th Int. Conf. on Machine Learning*, Atlanta ,Georgia, USA, 2013, pp. 1247–1255.

[8] Raman Arora and Karen Livescu, "Multi-view cca-based acoustic features for phonetic recognition across speakers and domains," in *Proceedings 38th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2013*, Vancouver, Canada, 2013, IEEE, pp. 7135–7139.

[9] Hanchuan Peng, Fuhui Long, and Chris Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[10] Mark A Hall, *Correlation-based feature selection for machine learning*, Ph.D. thesis, The University of Waikato, 1999.

[11] David R Hardoon, John Shawe-Taylor, and Ola Friman, "KCCA Feature Selection for fMRI Analysis," Technical Report TR_soton_04_03, University of Southampton, 2004.

[12] Harold Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[13] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[14] Olcay Kursun, Ethem Alpaydin, and Oleg V Favorov, "Canonical correlation analysis using within-class coupling," *Pattern Recognition Letters*, vol. 32, no. 2, pp. 134–144, 2011.

[15] Cemal Okan Sakar, Olcay Kursun, and Fikret Gürgen, "A feature selection method based on kernel canonical correlation analysis and the minimum redundancy maximum relevance filter method," *Expert Systems with Applications*, vol. 39, no. 3, pp. 3432–3437, 2012.

[16] Aaron Beck, Robert Steer, Roberta Ball, and William Ranieri, "Comparison of beck depression inventories -ia and -ii in psychiatric outpatients," *Journal of Personality Assessment*, vol. 67, no. 3, pp. 588–597, 1996.

[17] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[18] Hongying Meng, Di Huang, Heng Wang, Hongyu Yang, Mohammed AI-Shuraifi, and Yunhong Wang, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 2013, AVEC '13, pp. 21–30.

[19] Nicholas Cummins, Jyoti Joshi, Abhinav Dhall, Vidhyasaharan Sethu, Roland Goecke, and Julien Epps, "Diagnosis of depression by behavioural signals: a multimodal approach," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 2013, AVEC '13, pp. 11–20.