Canonical Correlation Analysis and Local Fisher Discriminant Analysis based Multi-View Acoustic Feature Reduction for Physical Load Prediction

Heysem Kaya[1*], *Tuğçe Özkaptan*[1,2], *Albert Ali Salah*[1], *Sadık Fikret Gürgen*[1]

[1]Department of Computer Engineering, Boğaziçi University, 34342, Bebek, İstanbul, Turkey
[2]Marmara Research Center, TÜBİTAK, 41470, Gebze, Kocaeli, Turkey

heysem@boun.edu.tr, tugce.ozkaptan@tubitak.gov.tr, salah@boun.edu.tr, gurgen@boun.edu.tr

Abstract

In this study we present our system for INTERSPEECH 2014 Computational Paralinguistics Challenge (ComParE 2014), Physical Load Sub-challenge (PLS). Our contribution is twofold. First, we propose using Low Level Descriptor (LLD) information as hints, so as to partition the feature space into meaningful subsets called *views*. We also show the virtue of commonly employed feature projections, such as Canonical Correlation Analysis (CCA) and Local Fisher Discriminant Analysis (LFDA) as ranking feature selectors. Results indicate the superiority of multi-view feature reduction approach to its single-view counterpart. Moreover, the discriminative projection matrices are observed to provide valuable information for feature selection, which generalize better than the projection itself. In our preliminary experiments we reached 75.35% Unweighted Average Recall (UAR) on PLS test set, using CCA based multi-view feature selection.

Index Terms: ComParE 2014, acoustic feature selection, Canonical Correlation Analysis, Local Discriminant Analysis, Physical Load

1. Introduction

The study of Computational Paralinguistics is about prediction of speakers' states and traits rather than the spoken content. The ComParE Challenge aims at bridging the gap between the state-of-the-art research on the field and comparability of the results. The first challenge in the series was held in 2009, where 384 baseline features were delivered by the organizers for the task of emotion classification [1]. While the tasks ranged from personality trait classification [2] to social signals detection [3], the baseline features grew in number and quality to reflect different aspects of the speech. ComParE 2014 presents two sub-challenges, namely the prediction of physical load (labeled as high pulse/low pulse) and the ternary level of cognive load. In this study, we focus on the binary classification of the physical load. Prediction of physical load is of interest with many applications [4]. Particularly, classification of the level of heart rate can be benefited in telemonitoring of the disabled/elderly, and in human behavior understanding as it correlates with affective arousal.

The last two years' baseline feature set, which is obtained by passing descriptive functionals (e. g. moments, extremes) over the Low Level Descriptor (LLD) contours (e. g. F0, shimmer, MFFC), contains 6 373 acoustic features [3, 5]. In machine learning literature, utilizing such a high number of features with a small amount of samples is known to reduce generalization power of the learner due to the *curse of dimensionality*. In order to overcome this problem, several feature reduction methods have been proposed in the literature [6, 7]. Fisher Discriminant Analysis (FDA) [8] and Canonical Correlation Analysis (CCA) [9] are two of the most commonly employed statistical methods. While FDA is used in classification problems to project the original features onto a discriminative lower dimensional space, the unsupervised CCA aims at maximizing the mutual correlation of two representations of the semantic object in the respective projected spaces [10].

In a recent study [11] CCA is employed as an acoustic feature selector for continuous depression level prediction. Here, the features are exposed to CCA against the continuous labels and are then ranked with respect to the absolute value of their weights in the projection vector (the eigenvector). In the case of classification, this setting is shown to reduce to FDA [12]. One problem in FDA is that it inherently assumes the classes to be unimodal. When classes are composed of several clusters, which is typical in acoustic speech processing, the within class scatter fails to reflect the structure and the projection does not generalize well for pattern recognition. To remedy this problem Sugiyama [13] proposed the incorporation of class-wise neighborhood information as in the Locality Preserving Projection (LPP) in a method called Local FDA (LFDA). LFDA considers the locality of instances of the same class, therefore aims at keeping the structural information in the discriminative embedding. We employ LFDA as an alternative to CCA for feature reduction.

In order to further exploit the data, it is possible to make use of domain knowledge as hints. Some high dimensional data have natural feature partitions that are called *views* in the literature. As these views may be obtained from different modalities, it is also possible to divide the single modality feature set into views. This approach aims at bringing together the self sufficient subset of features to exploit the internal correlations while avoiding the curse of dimensionality. For example, in [14] bioinformatics domain knowledge is exploited to partition the features into views such as di/peptide composition and multi-view feature extraction is carried out. When the multi-view approach is used in filter feature selection, it also helps side-step the *irrelevant redundancy* (IR). IR is incurred when a potential feature that has unique information about the target is omitted due to a high dependency (e. g. correlation, mutual information) with already selected set of features [15]. Moreover, multi-view feature selection reduces space and time complexity, enables processing smaller chunks of data in parallel.

In this study, we propose the use of LLD information of features to partition the massive feature set. From acoustic speech processing, it is known that not all functionals work the same for every LLD. Therefore, our proposed system amounts to selection of functionals per LLD. We also compare and contrast

view-based feature selection and extraction using LFDA and CCA. As expected, we obtain better performance (UAR) using multi-view feature selection. We also observe that features selected by means of discriminative projections generalize to unseen data better than the projections themselves.

The layout of this paper is as follows. In the next section we provide the background on CCA and LFDA, then in Section 3 we introduce the feature selection/extraction scheme. The experimental results are given in Section 4. Finally Section 5 concludes with future directions.

2. Background

2.1. Canonical Correlation Analysis

Proposed by Hotelling [9], CCA seeks to maximize the mutual correlation between two sets of variables by finding linear projections for each set. Mathematically , CCA seeks to maximize the mutual correlation between two views of the same semantic phenomenon (e. g. audio and video of a speech) denoted $X \in \mathbb{R}^{n \times d}$ and $Y \in \mathbb{R}^{n \times p}$, where n denote the number of paired samples, via:

$$\rho(X,Y) = \max_{w,v} corr(w^T X, v^T Y), \tag{1}$$

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

$$\rho(X,Y) = \sup_{w,v} \frac{w^T C_{XY} v}{\sqrt{w^T C_{XX} w \cdot v^T C_{YY} v}}. \tag{2}$$

The formulation in Eq. (2) can be converted into a generalized eigenproblem for both projections (i. e. w and v), the solution can be shown [10] to have the form of:

$$C_{XX}^{-1} C_{XY} C_{YY}^{-1} C_{YX} w = \lambda w, \tag{3}$$

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

$$\rho(X,Y) = \sqrt{\lambda}. \tag{4}$$

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. In short, when CCA is applied between X and Y we get:

$$[W, V, r, U_X, U_Y] = CCA(X,Y), \tag{5}$$

where W and V are composed of (sorted) eigenvectors from the eigenproblem in Eq. (3), r is the m dimensional vector of canonical correlations given in Eq. (4) while U_X and U_Y are the covariates. In other words, $U_X = X \times W$, when features in X are mean removed. The relationship between the canonical correlation and the corresponding covariates is given by the the Pearson's Correlation Coefficient (PCC):

$$r^i = PCC(U_X^i, U_Y^i), \tag{6}$$

where i indexes the column. It is important to note that the maximum number of covariates m in U_X and U_Y are limited with the matrix rank of X and Y:

$$m = min(rank(X), rank(Y)) \tag{7}$$

The non-linear version of CCA using the *kernel trick* is known as KCCA [10]. Also, Deep CCA (DCCA) is an efficient deep neural network alternative to KCCA [16].

2.2. Local Fisher Discriminant Analysis

It is known that when classes are multimodal, FDA faces anomalies [17]. It is important to preserve the local structure in the embedded space while trying to maximize the class separability. To retain the multimodality in the target space without regarding the classes, Locality Preserving Projection (LPP) [18] is introduced as an alternative to Principal component Analysis. The approach uses the affinity matrix idea to weight (softly mask) the projections. This idea inspired Sugiyama to extend traditional FDA to Local FDA by first reformulating the scatter matrices [13]:

$$S^w = 1/2 \sum_{i,j}^{n} A_{i,j}^w (x_i - x_j)(x_i - x_j)', \tag{8}$$

$$S^b = 1/2 \sum_{i,j}^{n} A_{i,j}^b (x_i - x_j)(x_i - x_j)', \tag{9}$$

where $(')$ denotes transpose and

$$A_{i,j}^w = \begin{cases} 1/n_c & \text{if } y_i = y_j = c, \\ 0 & \text{if } y_i \neq y_j, \end{cases} \tag{10}$$

$$A_{i,j}^b = \begin{cases} 1/n - 1/n_c & \text{if } y_i = y_j = c, \\ 1/n & \text{if } y_i \neq y_j, \end{cases} \tag{11}$$

Here the affinity matrices do not contain locality information but class information. To obtain LFDA we have [13]:

$$\overline{S}^w = 1/2 \sum_{i,j}^{n} \overline{\mathcal{A}}_{i,j}^w (x_i - x_j)(x_i - x_j)', \tag{12}$$

$$\overline{S}^b = 1/2 \sum_{i,j}^{n} \overline{\mathcal{A}}_{i,j}^b (x_i - x_j)(x_i - x_j)', \tag{13}$$

and localized discriminative affinity matrices are defined as

$$\overline{\mathcal{A}}_{i,j}^w = \begin{cases} A_{i,j}/n_c & \text{if } y_i = y_j = c, \\ 0 & \text{if } y_i \neq y_j, \end{cases} \tag{14}$$

$$\overline{\mathcal{A}}_{i,j}^b = \begin{cases} A_{i,j}(1/n - 1/n_c) & \text{if } y_i = y_j = c, \\ 1/n & \text{if } y_i \neq y_j, \end{cases} \tag{15}$$

where $A_{i,j}$ is the $n \times n$ regular affinity matrix keeping the unsupervised locality information. $A_{i,j}$ can simply be composed of 1s for k-nearest neighbors for each instance and 0s for the rest. It is also possible to adopt a localized measure where the distance to the k-th nearest neighbor is used as bandwidth in Gaussian similarity. Let D denote the $n \times n$ Euclidean distance matrix of samples, d_k is the n dimensional vector keeping the square root of the Euclidean distance of each sample to its k-th neighbor, and $M./L$ denote the element-wise division, we can obtain a smoother affinity matrix A via:

$$L = d_k d_k', \tag{16}$$

$$A = \exp(-D./L). \tag{17}$$

Once the scatter matrices are computed, the regular FDA eigenproblem can be used to obtain the discriminative projection:

$$\overline{S}^b W = \Lambda \overline{S}^w W. \tag{18}$$

3. Proposed Feature Reduction System

3.1. Discriminative Projection Based Filters

Our proposed system extends the recent work of Kaya et al. [11] to classification and applies it to acoustic feature partitions based on LLDs. The main idea behind the CCA based filter in [11] is as follows. When all features on one view are subjected to CCA against the labels on the other view, the absolute value of the projection matrix W can be used to rank the features. The application to regression is straightforward since the resulting matrix is $n \times 1$, therefore a vector. It can be applied in the same way to 2-class classification where the classes can be denoted with 0 and 1 in the target vector. For $C > 2$, we can use the canonical correlation value (r^i) to weight the corresponding projection column (eigenvector W^i). In short, the Sample versus Labels CCA Filter (SLCCA-Filter) algorithm, which inputs the dataset $X \in \mathbb{R}^{n \times d}$ and label matrix $T \in \{0,1\}^{n \times (C)}$; and outputs feature ranking R is given as:

$$[W, V, r, U_X, U_Y] = CCA(X, T), \quad (19)$$

$$H = \sum_{i=1}^{m} abs(W^i) r^i, \quad (20)$$

$$R = \text{sort(H,'descend')}, \quad (21)$$

where as noted earlier $m = min(rank(X), rank(Y))$, and the 1-of-C coded label matrix T is defined as

$$T_{i,c} = \begin{cases} 1 & \text{if } y_i = c, \\ 0 & \text{if } y_i \neq c. \end{cases} \quad (22)$$

Since C classes have $C-1$ degrees of freedom, the rank of matrix T is $C-1$. Therefore it is possible to remove any of the columns from 1-of-C coded matrix. The filter can be applied to LFDA in a similar manner, where instead of the canonical correlation value square root of the corresponding eigenvalue λ^i is used.

3.2. Feature Partitioning

While it is possible to obtain features statistically or randomly, in this study we used the 65 LLDs along with their first derivatives to form the 130 views. Details of the corpus and feature set can be found in [5, 19]. The canonical correlations obtained from applying CCA between the LLD based views and the target labels are shown in Fig. 1. The canonical correlation values range from 0.29 to 0.53.

4. Experimental Results

In our experiments we utilized the CCA implementation in MATLAB, author's own implementation for LFDA [1]. We used Weka Data Mining tool [20] in classification with Support Vector Machines, and also in our preliminary studies with Correlation Based Feature Selection (CFS).

4.1. System Development

To show the superiority of multi-view versus single view (i. e. full feature set) feature selection independently from the proposed CCA and LFDA based Filter, we used CFS. CFS measures the merit between a feature set S and target t via [21]:

$$r_{S,t} = \frac{k\overline{r_{ti}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}}, \quad (23)$$

[1] Available from `http://sugiyama-www.cs.titech.ac.jp/~sugi/software/LFDA/index.html`

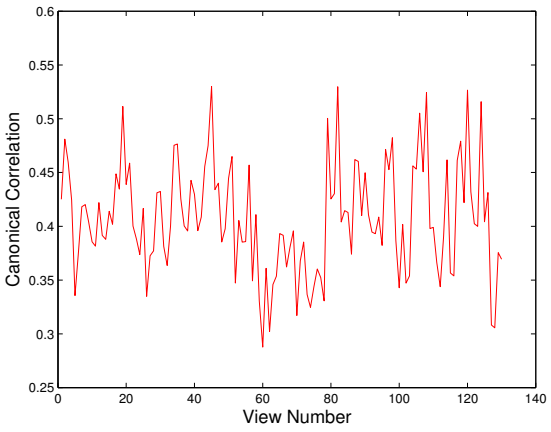

Figure 1: Canonical Correlation of LLD based views against the physical load labels (Low/High). The views are sorted lexicographically with respect to the LLD names

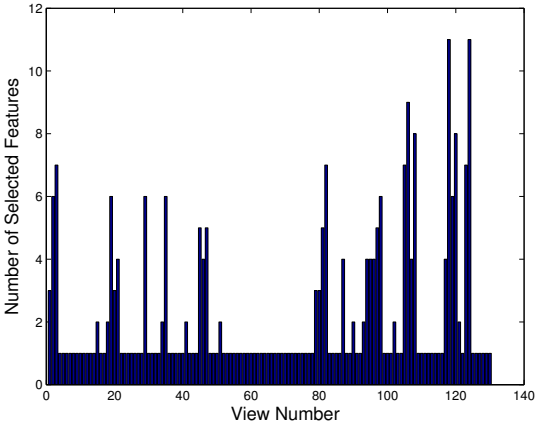

Figure 2: The number of selected features with CFS per view. The views are sorted lexicographically with respect to the LLD names

where k is number of features, $\overline{r_{ti}}$ 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. Therefore, Eq. (23) tries to reduce internal correlation (redundancy) and favors higher average feature-target correlations (relevance).

In our preliminary experiments, we used a single view feature selection and compared its performance with the features selected from LLD based views. In WEKA CFS implementation, we used BestFirst Forward search option with a backtracking limit of 5 steps. In single view setting the algorithm found 75 features, while in multi view setting 283 features were attained. The distribution of the number of selected features to views is given in Fig. 2. When the algorithm does not find a good merit in any subset, it generally outputs a single feature. For classification we used SVMs with Linear and RBF kernels. The precision parameter γ in RBF kernel is set to 0.0005 using cross validation on development set. In both kernels, min-max normalization (min-max Norm) and z-normalization (z-Norm) were tested in the preliminary studies. The set of SVM complexity parameter ranged roughly in double increments, for Lin-

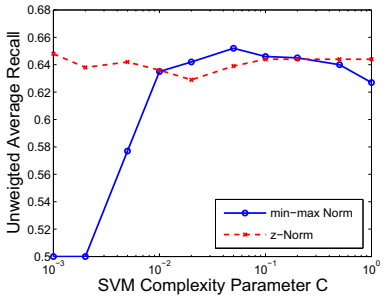

Figure 3: Performance of SVM with Linear Kernel on multi view LFDA selected features.

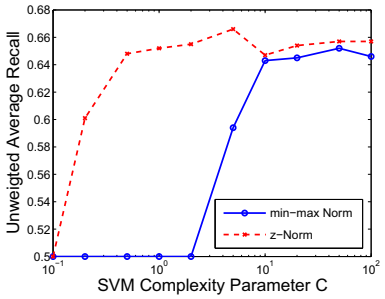

Figure 4: Performance of SVM with RBF Kernel on multi view LFDA selected features.

ear Kernel we used $\{0.0001, 0.0002, 0.0005, 0.001..., 1\}$ and for RBF kernel a set in the range $0.1 - 50$ is used. The best single view CFS performance is obtained as 61.7 UAR, using z-normalization with RBF kernel with $C = 2$. The best multi view CFS performance is given in Table 1. As can be seen from the table, all multi view settings provide better results than single view setting, z-Norm with RBF kernel yielding relatively better. While these UAR results are lower than baseline development set performance, they motivate further work in multi view approach.

Table 1: *Best SVM performance with multi view CFS features per normalization type and Kernel*

Norm	Kernel	C	UAR (%)
Min-max	Linear	0.02	64.6
Min-max	RBF	50	64.6
Z-norm	Linear	0.001	65.6
Z-norm	RBF	1	66.2

4.2. CCA and LFDA for Feature Selection

Since SLCCA and LFDA provide discriminative projections, which are popularly used in pattern recognition, we also compared the performance of feature transformation against feature selection using the same projection matrix. Encouraged from the preliminary experiments, we use a multi view setting to obtain 130 discriminative features, one from each LLD based view. The best SVM performance (z-Norm, RBF kernel, C=10) on LFDA features was found as 62.9%, the best SLCCA counterpart obtained was 63.5 (z-Norm, RBF kernel, C=0.2).

Next we issued another set of experiments using 5 highest ranking features from LLD based views with LFDA (5x130=650 features in total). The aim of this set of experiments was to choose an appropriate kernel and normalization method. Similar to preliminary experiments with CFS, we observed that z-norm with RBF kernel worked better than other combinations. See Fig. 3 and Fig. 4 for the effect of normalization on Linear and RBF kernels, respectively. We finally chose z-Norm along with RBF kernel, where the C parameter ranged from 0.1 to 50 with roughly double increments and $\gamma = 0.0005$ as stated before. The number of selected features per view ranged from 10 to 30 with a step of 5 (max features per view is 54). The best results of the two methods are given in Table 2. The best overall results are achieved using SLCCA-Filter with 15 features per view (C=0.1). We observe that most of the results are either better or on par with the challenge baseline on development set (67.2% UAR).

Table 2: Best RBF SVM Performance of multi view SLCCA-Filter and LFDA-Filter for varying number of features per view

Method	SLCCA		LFDA	
#Feats	C	UAR(%)	C	UAR (%)
10	2	67.1	1	67.7
15	0.1	69.4	1	68.4
20	0.1	68.5	1	68.8
25	2	68.9	1	68.0
30	2	66.2	2	67.0

Finally, we used the best hyper-parameter/method setting to predict the test set. We reached a UAR of 74.16%, beating the test set baseline 71.9%. We further ranked groups of multi-view selected features using minimum Redundancy Maximum Relevance Filter from [11]. On development set, we obtained the best performance using 1845 features of first 123 views. These relatively slightly refined set of features increased the UAR performance to 75.35% on challenge test set.

5. Conclusions and Outlook

In this study we examined the filter capability of the discriminative projections specifically Local Fisher Discriminant Analysis and Samples Versus Labels Canonical Correlation Analysis. We also proposed the use of domain knowledge to partition the acoustic feature set into views so as to divide and conquer the data. Using a multi-view approach in feature selection helps avoid the so called irrelevant redundancy, hence allows higher generalization. We show that multi-view setting provides superior scores against its single-view counterpart. Moreover, utilizing the projection matrices for feature selection is found to generalize better to unseen data than the projection itself. Combining the multi-view approach and the proposed feature selection method we obtain 75.35% UAR on the Physical Sub-challenge. The study can be extended using kernel trick for both statistical methods.

6. Acknowledgements

Work of the first and second authors are supported by The Scientific and Technological Council of Turkey (TÜBİTAK).

7. References

[1] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 Emotion Challenge," in *Proc. INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association*, ISCA. Brighton, UK: ISCA, September 2009, pp. 312–315.

[2] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 Speaker Trait Challenge," in *Proc. INTERSPEECH 2012*. Portland, OR, USA: ISCA, September 2012.

[3] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. INTERSPEECH 2013*, ISCA. Lyon, France: ISCA, August 2013, pp. 148–152.

[4] S. Harada, J. Lester, K. Patel, T. S. Saponas, J. Fogarty, J. A. Landay, and J. O. Wobbrock, "Voicelabel: Using speech to label mobile sensor data," in *in Proc. 10th International Conference on Multimodal Interfaces*, ser. ICMI '08. ACM, 2008, pp. 69–76.

[5] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load," in *Proc. INTERSPEECH 2014*, ISCA. Singapore, Singapore: ISCA, September 2014.

[6] C. M. Bishop, *Pattern Recognition and Machine Learning*, ser. Information Science and Statistics. Springer, 2006.

[7] E. Alpaydin, *Introduction to Machine Learning*, 2nd ed. The MIT Press, 2010.

[8] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

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

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

[11] H. Kaya, F. Eyben, A. A. Salah, and B. W. Schuller, "CCA Based Feature Selection with Application to Continuous Depression Recognition from Acoustic Speech Features," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, Florence, Italy, 2014.

[12] M. S. Bartlett, "Further aspects of the theory of multiple regression," in *Proc. of the Cambridge Philosophical Society*, vol. 34, no. 1, 1938, pp. 33–40.

[13] M. Sugiyama, "Local fisher discriminant analysis for supervised dimensionality reduction," in *Proc. of the 23rd International Conference on Machine Learning*, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 905–912.

[14] H. Kaya, O. Kurşun, and H. Şeker, "Stacking class probabilities obtained from view-based cluster ensembles," in *Artificial Intelligence and Soft Computing*, ser. Lecture Notes in Computer Science, L. Rutkowski, R. Scherer, R. Tadeusiewicz, L. Zadeh, and J. Zurada, Eds. Springer Berlin Heidelberg, 2010, vol. 6113, pp. 397–404.

[15] C. O. Sakar, O. Kursun, and F. 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] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. of the 30th International Conference on Machine Learning*, Atlanta ,Georgia, USA, 2013, pp. 1247–1255.

[17] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. Boston: Academic Press, 1990.

[18] X. He and P. Niyogi, "Locality Preserving Projections," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. Saul, and B. Scholkopf, Eds. Cambridge, MA: MIT Press, 2004.

[19] B. Schuller, F. Friedmann, and F. Eyben, "The Munich Bio Voice Corpus: Effects of Physical Exercising, Heart Rate, and Skin Conductance on Human Speech Production," in *Proc. 9th Language Resources and Evaluation Conference (LREC 2014)*. Reykjavik, Iceland: ELRA, September 2014.

[20] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[21] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.