INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Introducing Weighted Kernel Classifiers for Handling Imbalanced Paralinguistic Corpora: Snoring, Addressee and Cold

Heysem Kaya[1], *Alexey A. Karpov*[2,3]

[1]Department of Computer Engineering, Namık Kemal University, Çorlu, Tekirdağ, Turkey
[2]St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences, Russia
[3] Department of Speech Information Systems, ITMO University, St. Petersburg, Russia

hkaya@nku.edu.tr, karpov@iias.spb.su

Abstract

The field of paralinguistics is growing rapidly with a wide range of applications that go beyond recognition of emotions, laughter and personality. The research flourishes in multiple directions such as signal representation and classification, addressing the issues of the domain. Apart from the noise robustness, an important issue with real life data is the imbalanced nature: some classes of states/traits are under-represented. Combined with the high dimensionality of the feature vectors used in the state-of-the-art analysis systems, this issue poses the threat of over-fitting. While the kernel trick can be employed to handle the dimensionality issue, regular classifiers inherently aim to minimize the misclassification error and hence are biased towards the majority class. A solution to this problem is oversampling of the minority class(es). However, this brings increased memory/computational costs, while not bringing any new information to the classifier. In this work, we propose a new weighting scheme on instances of the original dataset, employing Weighted Kernel Extreme Learning Machine, and inspired from that, introducing the Weighted Partial Least Squares Regression based classifier. The proposed methods are applied on all three INTERSPEECH ComParE 2017 challenge corpora, giving better or competitive results compared to the challenge baselines.

Index Terms: computational paralinguistics, imbalanced data, Snoring, Addressee, Fisher vector, Weighted PLS, ELM

1. Introduction

The flourishing field of Computational Paralinguistics, which focuses on the non-verbal aspects of speech such as speaker states and traits, grows rapidly with high success thanks to shared datasets with a common protocol. One of the factors that boosts the research in the field are common protocol challenges, that are led by the INTERSPEECH ComParE events. Since the first challenge in 2009 [1], the series introduced a variety of novel problems, including but not limited to emotion [1, 2], speaker traits [3], conflict and autism [2], Eating Condition [4], Sincerity and Native Language [5]. Over the years, both the quantity and quality of data have changed. The number of subjects/utterances and the extracted features increased from few hundreds to thousands. Moreover, the datasets became less lab-controlled, more "in-the-wild" reflecting the challenges of real-life conditions, which should be handled elaborately.

Increasing the number of potential acoustic features opens room for machine learning research, in robust high-dimensional classification and in feature selection. Past works on challenge corpora highlighted the importance of feature selection in handling high-dimensional paralinguistic datasets [6, 7, 8]. On the other hand, recent works have also shown that computer vision inspired high-dimensional representation methods for Low-Level Descriptors (LLDs), such as the Fisher Vector (FV) encoding [9, 10], can be successfully employed for paralinguistic analysis [11, 12]. Since labeled utterances in many paralinguistic corpora are still not enough to train a deep neural network, kernel machines (such as SVM) are popularly employed to cope with high-dimensionality. The kernel machines operate on the instance similarity matrix, dubbed *kernel*, which scales quadratically with the number of instances.

An important aspect of the real-life data is that since it is not controlled, the state/trait classes are typically imbalanced. This is mostly prevalent in biomedical datasets, where a small proportion of instances are positive (e.g. having anomaly). Since the regular classifiers aim to minimize the misclassification error, they are inherently biased towards the majority class. A typical approach to the problem is employing an instance upsampling strategy [5, 13, 14]. However, this increases memory and computational costs, particularly with kernel machines, while not providing additional information to the learner.

The contributions of this paper are threefold. First, we propose the use of a weighting scheme in kernel classifiers to handle the class imbalance problem. This strategy assigns each training instance an importance weight that is inversely proportional to the number of training set instances of its class. Thus, it favors the minority class(es) against the majority class and improves average recall instead of accuracy. We first implement Weighted Kernel Extreme Learning Machine (WKELM) [15, 16] for this purpose. Then, we transfer the same scheme to another fast and robust learner, namely Partial Least Squares (PLS) regression based kernel classifier, introducing Weighted Kernel PLS (WKPLS). Third, we apply the proposed methods on all three corpora from INTERSPEECH 2017 ComParE challenge [14] that presents three sub-challenges for classifying Addressee (Adult or Child directed speech), Cold and the source of Snoring, respectively. The challenge organizers provide a baseline system composed of a standard set of features and a commonly used classifier. This year the organizers provide three base systems: two feature representations (extracted using openSMILE [17] and openXBOW [18] tools) classified with SVM and one end-to-end trained neural network [19]. We benefit from the baseline sets and extract acoustic features using FV representation [9, 10].

The remainder of this paper is organized as follows. In Section 2, we introduce the proposed framework and give background on its major components. The experimental results are given in Section 3, Section 4 concludes with future directions.

http://dx.doi.org/10.21437/Interspeech.2017-653

2. Proposed Framework

In our approach to all three sub-challenges, we first try to exploit the given baseline feature sets using a Canonical Correlation Analysis (CCA) based feature selection method [8] and cascaded normalization strategies [11, 20]. We then employ FV encoding for feature extraction, comparing regular and weighted KELM and KPLS learners. We finally fuse the scores from best individual classifiers.

2.1. Feature Representation

For acoustic feature representation, we have recently introduced Fisher vectors (FV) to encode the LLDs over utterances, which rendered outstanding results in the recent ComParE challenges [11, 12]. Motivated from our past experiences, in this work we use both the FV encoding and the baseline feature sets extracted using openSMILE [17] and openXBOW [18].

As in our previous works, we extract Mel Frequency Cepstral Coefficients (MFCC) and RASTA-style Perceptual Linear Prediction (PLP) Cepstrum [21, 22] to represent the signal properties. Particularly, we extract MFCCs 0-24, and use a 12th order linear prediction filter giving 13 coefficients. Raw LLDs are augmented with their first and second order delta coefficients, resulting in 75 and 39 features for MFCC and RASTA-PLP, respectively. Following [11], we use combination of RASTA-PLP and MFCC descriptors as it is found to be better than their individual performances.

The Fisher vector (FV) provides a supra-frame encoding of the local descriptors, quantifying the gradient of the parameters of the background model with respect to the data [9, 10]. Given a probability model parametrized with θ, the expected Fisher information matrix $F(\theta)$ is the expectation of the second derivative of the log likelihood with respect to θ:

$$F(\theta) = -E[\frac{\partial^2 \log p(\mathcal{X}|\theta)}{\partial \theta^2}]. \quad (1)$$

The idea in FV in relation to $F(\theta)$ is taking the derivative of the model parameters and normalizing them with respect to the diagonal of $F(\theta)$ [9]. To make the computation feasible, a closed form approximation to the diagonal of $F(\theta)$ is proposed [9]. As a probability density model $p(\theta)$, GMMs with diagonal covariances are used. A K-component GMM is parametrized as $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$ where the parameters correspond to zeroth (mixture proportions), first (means) and second order (covariances) statistics, respectively. It has been shown that using the zeroth order statistics is equivalent to the Bag of Words (BoW) approach [23], and using only the first order statistics is equivalent to Vector of Locally Aggregated Descriptors (VLAD) encoding [24]. Note that in FV, we use far less number of clusters compared to BoW, and contribution of zeroth order statistics become negligible [9]. Therefore, only gradients of $\{\mu_k, \Sigma_k\}_{k=1}^K$ are used, giving a $2 \times d \times K$ dimensional super vector, where d is the LLD dimensionality.

2.2. Model Learning

To learn a classification model, we use Kernel ELM and PLS regression motivated from their fast and accurate learning capability and state-of-the-art results on recent paralinguistic/multimodal challenge corpora [11, 25, 26]. We obtain linear kernels from the dataset and use them in PLS and ELM, optimizing the hyper-parameters on the development set. For handling the imbalanced data, we employ a variant of ELM dubbed *Weighted ELM* [15]. Inspired from it, we applied this simple and efficient scheme to KPLS, introducing WKPLS. In the following, we briefly introduce the base classifiers and the *weighting trick*.

The ELM paradigm proposes unsupervised, even random generation of the hidden node output matrix $\mathbf{H} \in \mathbb{R}^{N \times h}$, where N and h denote the number of instances and the hidden neurons, respectively. The actual learning takes place in the second layer between $\mathbf{H}$ and the label matrix $\mathbf{T} \in \mathbb{R}^{N \times L}$, where L is the number of classes. $\mathbf{T}$ is composed of continuous annotations in case of regression, therefore is a vector. In the case of L-class classification, $\mathbf{T}$ is represented in one vs. all coding:

$$\mathbf{T}_{t,l} = \begin{cases} +1 & \text{if } y^t = l, \\ -1 & \text{if } y^t \neq l. \end{cases} \quad (2)$$

The second level weights $\beta \in \mathbb{R}^{h \times L}$ are learned by least squares solution to a set of linear equations $\mathbf{H}\beta = \mathbf{T}$. The output weights can be learned via:

$$\beta = \mathbf{H}^{\dagger}\mathbf{T}, \quad (3)$$

where $\mathbf{H}^{\dagger}$ is the Moore-Penrose generalized inverse [27] that gives the minimum L_2 norm solution to $\|\mathbf{H}\beta - \mathbf{T}\|$, simultaneously minimizing the norm of $\|\beta\|$. This extreme learning rule is generalized to use any kernel $\mathbf{K}$ with a regularization parameter C, without generating $\mathbf{H}$ [28], relating ELM to LSSVM [29]:

$$\beta = (\frac{\mathbf{I}}{C} + \mathbf{K})^{-1}\mathbf{T}, \quad (4)$$

where $\mathbf{I}$ is the $N \times N$ identity matrix. Weighted ELM [15] introduces a diagonal weight matrix $\mathbf{W_{t,t}} = \mathbf{1}/(\mathbf{N_l})$, where $y^t = l$ and N_l represents the total number of training set instances having class label l:

$$\beta = (\frac{\mathbf{I}}{C} + \mathbf{WK})^{-1}\mathbf{WT}. \quad (5)$$

This approach counter-balances the under-represented classes, while the weights of each class sum up to unity. Since total weight is equal for each class, models are forced to maximize average recall, instead of accuracy. In our experiments, we compare and fuse KELM learning rule given in eq. (4) and its weighted version WKELM given in eq. (5).

PLS regression between two sets of variables $\mathbf{X} \in \mathbb{R}^{N \times d}$ and $\mathbf{Y} \in \mathbb{R}^{N \times p}$ is based on decomposing the matrices as $\mathbf{X} = \mathbf{U}_x\mathbf{V}_x + r_x$, $\mathbf{Y} = \mathbf{U}_y\mathbf{V}_y + r_y$, where $\mathbf{U}$ denotes the latent factors, $\mathbf{V}$ denotes the loadings and r stands for the residuals. For further details of PLS regression, the reader is referred to [30]. PLS is applied to classification in one-versus-all setting between the kernel matrix $\mathbf{K} = \mathbf{X}\mathbf{X}'$ and the target matrix $\mathbf{T}$, then the class giving the highest regression score is taken as prediction. To realize the weighing trick, we input KPLS $\tilde{\mathbf{K}} = \mathbf{WK}$ and $\tilde{\mathbf{T}} = \mathbf{WT}$, thus practically obtaining *WKPLS*.

2.3. Fusion

We investigate two variants of score level fusion. The first is simple weighted fusion (SF) of scores, where the classifier confidence scores S^A and S^B are fused using weight $\gamma \in [0, 1]$:

$$S^{fusion} = \gamma * S^A + (1 - \gamma) * S^B. \quad (6)$$

Secondly, we apply weighted score fusion (WF) for each model and class. Let M and L denote the number of models and classes, respectively. The optimal fusion weights $W_{i,j}^{fusion} \in [0, 1], 1 \leq i \leq M, 1 \leq j \leq L, \sum_{i=1}^{M} W_{i,j}^{fusion} = 1$, are searched over a pool of randomly generated matrices.

3. Experimental Results

All challenge corpora are divided into speaker-independent training, development (validation) and test sets. The labels of the test set are not known to competitors. In ComParE series, the challenge measure for classification based tasks is Unweighted Average Recall (UAR, mean recall of all classes). For L-class classification, this measure has a constant chance level (worst case) score of $1/L$. In our experiments, hyper-parameters of the classifiers and score fusion weights are tuned to maximize the validation set UAR. The baseline systems employ an upsampling strategy to handle the class-imbalance issue. We use exactly the same upsampling to showcase the effect of our proposed weighting scheme. Due to space limitations, we refer the reader to the challenge paper [14] for further details.

For ease of reproducibility, we use open source tools in our experiments. For MFCC and RASTA-PLPC feature extraction we use RASTAMAT library [31], for GMM training and FV encoding we use MATLAB API of VLFeat library [32]. In all 3 tasks, Fisher vectors are tested with 110 PCA dimensions and $K_{GMM} = \{64, 128\}$ components for GMM, since these were found to render the best results in our former paralinguistic studies [11, 12]. The regularization parameter in KELM variants is optimized in the set $10^{-6,-5,...,3}$ with exponential steps. The number of latent factors for KPLS variants is searched in the range [2, 20] with steps of two.

It is important to note that, unlike former challenges, this year the baseline scores are obtained selectively from among 20 test set probes: using 3 classifier systems and their decision level fusion schemes [14]. This makes the test baselines hard to surpass, as the competitors have a maximum of 5 submission options per SC.

3.1. Experiments on the Addressee Sub-Challenge

The task of Addressee SC is to discriminate adult speakers' Adult Directed Speech (ADS) and Child Directed Speech (CDS) on HOMEBANK CHILD/ADULT ADDRESSEE CORPUS (HB-CHAAC) [14]. The recordings are carried out in natural conditions, thus contain background noise and in some cases overlapped speech (child and adult). Out of 10 866 instances, which are partitioned into the training, validation (development) and test sets, about 40% is adult directed speech.

As a preliminary work, we analyze the openSMILE functional features with four classifiers, namely KPLS, WKPLS, KELM and WKELM, respectively to assess the effectiveness of weighting versus upsampling strategies. From Table 1, we observe that concerning UAR scores i) weighed classifiers outperform their regular versions ii) upsampling may be helpful for regular classifiers, while for weighted learners effect may be worsening. Note that since the class proportions are approximately 40% to 60%, the boost due to upsampling/weighting is not markedly high for this data.

Table 1: *Comparison of development set UAR performance (%) scores for original and upsampled data in regular and weighted classifiers using z-normalization*

Data	KPLS	WKPLS	KELM	WKELM
Original	59.91	62.82	61.40	62.24
Upsampled	61.44	62.29	61.28	61.74

Applying a CCA based feature selection [8], it is possible to improve the performance of baseline openSMILE functional set (OS) to 63.26%, using WKPLS with top 1300 features. As feature reduction does not give a marked increase with other classifiers on this data, in the subsequent experiments we use the original baseline set to treat all classifiers evenly. We next extract Fisher Vectors (FV) as described in Section 2.1, and compare their performance against the baseline openSMILE functional and Bag-of-Audio-Words (BOAW) representations. In BOAW representation, we use 4000/4000 words for raw/delta LLDs, based on their reported performance in the challenge paper [14]. The best results of each feature type are summarized in Table 2.

Table 2: *Comparison of development set UAR performance (%) for three acoustic feature representations using z-normalization. Best results on each column are shown in* ***bold***.

Feature	KPLS	WKPLS	KELM	WKELM
OS	59.91	62.82	61.40	62.24
BOAW	61.76	64.12	63.65	63.47
FV	**62.36**	**65.62**	**64.34**	**64.37**

In our test set prediction submissions for this SC, we fused the scores of the best classifiers. Since this task is binary classification (so the class confidences are linearly dependent), we use the simple weighted fusion approach that gives weights to classifier models, rather than the version that gives weights per model and class. The development and test set scores of the top three submitted systems are shown in Table 3. We observe that the test set scores are higher compared to the development set, however they still remain below the challenge baseline scores. This can be attributed to the competitive baseline scores set out this year.

Table 3: *Development and test set UAR (%) scores of submitted fusion systems*

System	Devel	Test
Baseline	66.40	70.20
FV_{WPLS}, OS_{WPLS}	65.98	68.53
FV_{WPLS}, OS_{WPLS}, $BOAW_{WPLS}$	66.45	68.63
FV_{WPLS}, OS_{WPLS}, $BOAW_{WPLS}$, FV_{PLS}	66.61	68.66

3.2. Experiments on the Cold Sub-Challenge

The Cold SC is also a binary task to predict the state of having a *Cold*. Since the sickness affects the speech production system, it changes the acoustic characteristics, which can be used to model a prediction system. There are a total of 28 652 utterances in the corpus, where Cold (C) class comprises about 10% of them. Due to high class imbalance, regular classifiers have a poor UAR performance on original data.

We follow the same steps as in the Addressee SC and observe higher improvement due to feature selection (on OS functionals) and weighting strategy. When performance of four classifiers with respect to top features ranked with SLCCA-RAND method [8] are analyzed, we found that the UAR performance of baseline openSMILE set can be improved from 65% to 68% via feature selection. We also observe that weighting remarkably improves the UAR performance of PLS from 53% to 68%. We use the reduced feature set in the subsequent score fusion experiments.

Extracting FVs with 110 PCA eigenvectors and 128 GMM components gives a development set UAR of 69.5%, using combination of z-normalization and L_2 normalization with KELM.

This is an interesting result, as the best weighted classifier performance is (68.9% UAR with WKELM) remains below this performance. Since we have limited test submission options, we could not probe the performance of all sub-systems. Therefore, we used the fusion of best OS and best FV systems in our all test set submissions.

The score fusion of two feature types (OS and FV) resulted in a development set UAR of 71.4%, interestingly rendering a lower test set score of 65.2%. When FV and OS systems are evaluated separately, they gave test set UAR performances of 65.3% and 65.6%, respectively. These scores remain below the test set baseline score of 71.0%. In both Cold and Addressee SCs, the classifier hyper-parameters are optimized on the development set, which are used to train models on the combination of training and development sets for final prediction on the test set. A probable reason for the performance drop is the shift in optimal hyper-parameters due to change in distribution of the combined set. Furthermore, feature-target variable correlations may also be different in the test set, which hampers the performance. Note also that the baseline system uses BOAW features and end-to-end trained neural network that were not investigated in our systems.

3.3. Experiments on the Snoring Sub-Challenge

Snoring is a biological effect that is generated during inspiration in sleep, by vibrating soft tissue in the upper airways. This dramatically degrades the sleep quality of the bed/room partner(s) and may cause social problems among the partners. An important issue in treating the problem with a surgical operation is locating the source of snoring. Therefore, the task in Snore SC is to classify the excitation location of snore sounds [14]. The patients are audio-video recorded during Drug Induced Sleep Endoscopy (DISE) in three medical centers. The snore recordings are annotated by ear, nose, throat (ENT) experts from video for commonly used VOTE scheme (**V**elum, **O**ropharyngeal lateral walls, **T**ongue and **E**piglottis) [33, 34]. The four classes are highly imbalanced, where T and E utterances comprise 5% and 11% of the data, respectively.

Since this corpus is small in terms of total number of utterances (843), we used 64 GMM components in FV feature extraction. Moreover, here we applied a 2-fold cross-validation (CV) using training and development sets in their entirety.

Table 4: *Comparative UAR (%) scores for two feature types and two folds. Best results on each column are shown in* ***bold***.

Feat./Fold	KPLS	WKPLS	KELM	WKELM
OS-F1	39.11	47.58	43.31	43.06
FV-F1	28.56	38.95	45.03	**50.12**
OS-F2	**52.65**	**50.65**	**49.47**	**50.64**
FV-F2	29.40	43.07	41.40	46.25

Best results (over normalization alternatives and hyper-parameters) of models trained with baseline openSMILE (OS) and FV features on 2-fold CV are summarized in Table 4. We see dramatic performance differences between the two folds and between the feature types. On Fold-1, the challenge paper reports 40.6% UAR using baseline OS features, upsampling strategy and SVM classifier [14]. Using the same features with WKPLS, it was possible to obtain 47.6% UAR. Using FV features and WKELM on the same fold, the performance further improves to 50.1%. On the other hand, surprising results are observed in Fold-2 (i. e. when trained on development and tested on training set). Baseline OS features give on the average higher than 50% UAR with this fold. Although, KPLS scores are low in Fold-1, in Fold-2 they are found to give the highest performance. We should note that the number of minority class (T) instances are 8 and 15 for training and development sets, respectively. The number of instances in other classes remains similar across these two sets. Thus a better performance and shifted optimal hyper-parameters are observed under Fold-2 with baseline OS features. The results with FV features do not exhibit this tendency. For our test submissions, we combined results from the best models of each classifier type using class and model weighted fusion scheme. In our first two test submissions, we combined only systems trained with OS features, reaching 62.34% and 61.08% UAR scores, respectively. In our third submission, we fused confidence scores of three classifiers (KPLS, WKPLS and KELM) trained on OS features with WKELM model trained with FVs. This improved the test set UAR to 64.23%, outperforming the challenge baseline (58.5%) about 9.8%, relatively. The corresponding confusion matrix is illustrated in Figure 1, where we see a moderate recall (43.75%) for the minority class (T), but a high recall (85.2%) for the second smallest class (E).

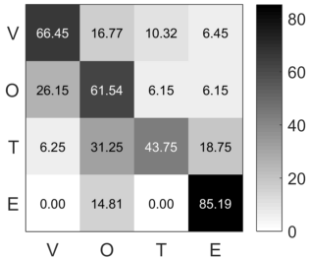

Figure 1: Confusion matrix for test set evaluation with fused approach (KPLS, WKPLS, KELM and WKELM)

4. Conclusion

In this work, we propose the use of weighted kernel classifiers to handle the imbalanced data. Inspired from Weighted Kernel ELM, we introduce the weighting trick into another fast and robust learner, namely Kernel PLS regression based classifier. The processing pipeline also combines popular suprasegmental acoustic features with computer vision inspired FV encoding and applies multi-level normalization. The preliminary results on the challenge validation and test sets indicate that the presented system is effective in handling the class imbalance and efficient as it does not necessitate over-sampling. Noting that the challenge baselines are selected from among 20 test set probes, and hence are highly competitive; the initial results with the proposed system are promising as they outperform the challenge test set UAR baseline in Snoring SC and are on par with those in the Addressee SC. Ongoing works focus on fusion strategies of regular and weighted classifiers, that inherently favor accuracy and UAR, respectively.

5. Acknowledgments

This work is partially supported by RFBR (project № 16-37-60100), grant of the President of Russia (№ MD-254.2017.8) and by the Government of Russia (grant № 074-U01).

6. References

[1] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 Emotion Challenge," in *INTERSPEECH*, Brighton, UK, Proceedings, 2009, pp. 312–315.

[2] 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 *INTERSPEECH*, Lyon, France, Proceedings, 2013, pp. 148–152.

[3] 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 *INTERSPEECH*, Portland, OR, USA, Proceedings, 2012, pp. 254–257.

[4] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The INTERSPEECH 2015 Computational Paralinguistics Challenge: Nativeness, Parkinson's & Eating Condition," in *INTERSPEECH*, Dresden, Germany, Proceedings, 2015, pp. 478–482.

[5] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language," in *INTERSPEECH*, San Francisco, USA, Proceedings, 2016, pp. 2001–2005.

[6] A. Ivanov and X. Chen, "Modulation spectrum analysis for speaker personality trait recognition," in *INTERSPEECH*, Portland, OR, USA, Proceedings, 2012, pp. 278–281.

[7] H. Kaya, T. Özkaptan, A. A. Salah, and S. F. Gürgen, "Canonical Correlation Analysis and Local Fisher Discriminant Analysis based Multi-View Acoustic Feature Reduction for Physical Load Prediction," in *INTERSPEECH*, Singapore, Proceedings, 2014, pp. 442–446.

[8] H. Kaya, T. Özkaptan, A. A. Salah, and F. Gürgen, "Random Discriminative Projection based Feature Selection with Application to Conflict Recognition," *Signal Processing Letters, IEEE*, vol. 22, no. 6, pp. 671–675, 2015.

[9] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, Minnesota, USA, Proceedings, 2007, pp. 1–8.

[10] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, "Large-scale Image Retrieval with Compressed Fisher Vectors," in *23rd IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3384–3391.

[11] H. Kaya, A. A. Karpov, and A. A. Salah, "Fisher vectors with cascaded normalization for paralinguistic analysis," in *INTERSPEECH*, Dresden, Germany, Proceedings, 2015, pp. 909–913.

[12] H. Kaya and A. A. Karpov, "Fusing acoustic feature representations for computational paralinguistics tasks," in *INTERSPEECH*, San Francisco, USA, Proceedings, 2016, pp. 2046–2050.

[13] G. Gosztolya, T. Grósz, R. Busa-Fekete, and L. Tóth, "Determining native language and deception using phonetic features and classifier combination," in *Proceedings of Interspeech*, San Francisco, CA, USA, Sep 2016, pp. 2418–2422.

[14] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A. Warlaumont, G. Hidalgo, S. Schneider, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring," in *INTERSPEECH*, Stockholm, Sweden, Proceedings, 2017.

[15] W. Zong, G.-B. Huang, and Y. Chen, "Weighted extreme learning machine for imbalance learning," *Neurocomputing*, vol. 101, pp. 229 – 242, 2013.

[16] K. Li, X. Kong, Z. Lu, L. Wenyin, and J. Yin, "Boosting weighted ELM for imbalanced learning," *Neurocomputing*, vol. 128, pp. 15 – 21, 2014.

[17] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent Developments in openSMILE, the Munich open-source Multimedia Feature Extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*. ACM, 2013, pp. 835–838.

[18] M. Schmitt and B. W. Schuller, "openXBOW-introducing the Passau open-source crossmodal bag-of-words toolkit," *preprint arXiv:1605.06778*, 2016.

[19] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5200–5204.

[20] H. Kaya, A. A. Karpov, and A. A. Salah, "Robust acoustic emotion recognition based on cascaded normalization and extreme learning machines," in *13th International Symposium on Neural Networks - ISNN'16, LNCS 9719*, St. Petersburg, Russia, pp. 115–123. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-40663-3_14

[21] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[22] H. Hermansky and N. Morgan, "RASTA Processing of Speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 4, pp. 578–589, 1994.

[23] J. Sivic and A. Zisserman, "Efficient visual search of videos cast as text retrieval," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 4, pp. 591–606, 2009.

[24] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3304–3311.

[25] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining Multiple Kernel Methods on Riemannian Manifold for Emotion Recognition in the Wild," in *Proceedings of the 16th International Conference on Multimodal Interaction*, ser. ICMI '14. New York, NY, USA: ACM, 2014, pp. 494–501.

[26] H. Kaya, F. Gürpınar, and A. A. Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion," *Image and Vision Computing*, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885617300367

[27] C. R. Rao and S. K. Mitra, *Generalized inverse of matrices and its applications*. Wiley New York, 1971, vol. 7.

[28] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme Learning Machine for Regression and Multiclass Classification," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 42, no. 2, pp. 513–529, 2012.

[29] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[30] H. Wold, "Partial least squares," in *Encyclopedia of Statistical Sciences*, S. Kotz and N. L. Johnson, Eds. Wiley New York, 1985, pp. 581–591.

[31] D. P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, online web resource. [Online]. Available: http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

[32] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," 2008. [Online]. Available: http://www.vlfeat.org/

[33] E. J. Kezirian, W. Hohenhorst, and N. de Vries, "Drug-induced sleep endoscopy: the VOTE classification," *European Archives of Oto-Rhino-Laryngology*, vol. 268, no. 8, pp. 1233–1236, 2011.

[34] N. Charakorn and E. J. Kezirian, "Drug-induced sleep endoscopy," *Otolaryngologic Clinics of North America*, vol. 49, no. 6, pp. 1359–1372, 2016.