INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Fusing Acoustic Feature Representations for Computational Paralinguistics Tasks

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

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

Abstract

The field of Computational Paralinguistics is rapidly growing and is of interest in various application domains ranging from biomedical engineering to forensics. The INTERSPEECH ComParE challenge series has a field-leading role, introducing novel problems with a common benchmark protocol for comparability. In this work, we tackle all three ComParE 2016 Challenge corpora (Native Language, Sincerity and Deception) benefiting from multi-level normalization on features followed by fast and robust kernel learning methods. Moreover, we employ computer vision inspired low level descriptor representation methods such as the Fisher vector encoding. After non-linear preprocessing, obtained Fisher vectors are kernelized and mapped to target variables by classifiers based on Kernel Extreme Learning Machines and Partial Least Squares regression. We finally combine predictions of models trained on popularly used functional based descriptor encoding (openSMILE features) with those obtained from the Fisher vector encoding. In the preliminary experiments, our approach has significantly outperformed the baseline systems for Native Language and Sincerity sub-challenges both in the development and test sets.
Index Terms: ComParE, computational paralinguistics, Native Language, Sincerity, Fisher vector, PLS, ELM

1. Introduction

Computational Paralinguistics, the study of non-verbal aspects of speech, has developed rapidly over the last decade. While speech based prediction of non-linguistic phenomena such as long term diseases and speaker identity had been under investigation earlier, the field flourished particularly around the subject of emotion recognition [1, 2]. The INTERSPEECH ComParE events played a prominent role in driving the study into a coherent field; by introducing novel problems, allowing comparability as well as repeatability of works. A large set of paralinguistic tasks, including but not limited to emotion [2, 3], speaker traits [4], conflict and autism [3] and Eating Condition [5] are investigated in past events. Yet, there is a plethora of other tasks to be discovered.

INTERSPEECH 2016 ComParE challenge [6] presents three sub-challenges for predicting Sincerity, Deception Condition and Native Language of the speaker, respectively. All three problems can be of important use in law enforcement community: e. g. for detecting lies and sincerity of confession regarding the first two; and for forensic purposes regarding the Native Language Sub-challenge. Moreover, studies in these directions can better enhance human-computer interaction (HCI) systems, e. g. the agent can adapt its English ASR depending on the Native Language of the speaker.

The organizers of the challenge provide a baseline system composed of a standard set of features and a commonly used classifier. Both of the system components can be reproduced via freely available, open source tools [7, 8]. The provided baseline system is sometimes very hard to outperform (see e. g. the results of Autism and Emotion Sub-Challenges of ComParE 2013 [3]), showing the success of brute-forced suprasegmental acoustic features extracted using the openSMILE tool [8]. Indeed, the tool gives a general purpose feature set that yields state-of-the-art results in a wide range of paralinguistic problems. On the other hand, there is also a need for alternative acoustic feature representations achieving state-of-the-art results on many paralinguistic tasks.

Motivated by the need to investigate different acoustic feature representations, in ComParE 2015 we proposed the use of Fisher vectors (FV) for encoding the low level descriptors (LLD) over utterances [9]. This super vector modeling is introduced and popularly used in computer vision, especially in large scale image retrieval [10, 11]. The super vector quantifies the amount of change induced by the utterance/video descriptors on a background probability model, which is typically a Gaussian Mixture Model (GMM). The advantage of FV is that it requires far less number of components in a GMM than the Bag of Words (BoW) approach [12] and does not require training on a very-large corpus, as in the case of Universal Background Model (UBM). FV is not only efficient but also accurate in encoding. In [9], the FV based method used to recognize the eating condition of speakers gave the best results, with a large margin compared to the first-runner up.

Inspired by the success of transferring FV encoding into speech domain, more specifically paralinguistics, here we seek to validate its efficiency and robustness in the new challenging tasks. In this work, we carry out extensive experiments using the baseline feature set and the FV encoding for comparison and fusion of their representations/predictions.

Clinging to the efficiency issues also in model learning, we use Extreme Learning Machines (ELM) [13, 14] and Partial Least Squares (PLS) regression [15] based classifiers, motivated by their fast learning capability and outstanding performance in recent challenges [9, 16, 17].

We present comparative analysis of each component of our framework. 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.2016-995

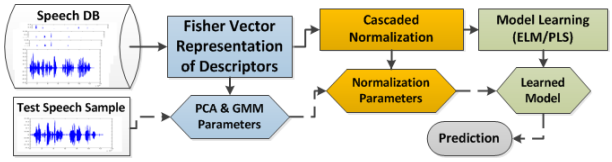

Figure 1: Proposed framework with alternative speech signal representation

2. Proposed Framework

In our approach to all three sub-challenges, we first try to exploit the given baseline feature set using recently proposed feature selection method [18] and cascaded normalization strategies. We then apply the proposed FV encoding based alternative framework, which is illustrated in Figure 1. Finally, we fuse the predictions of models trained on different feature representations to improve performance.

2.1. Extraction of Acoustic LLDs from Speech

MFCC and RASTA-PLP [19, 20] are the most popular descriptors used in a variety of speech technologies ranging from speaker identification to speech recognition, although they are initially designed to minimize the speaker dependent effects. They are also commonly employed in state-of-the-art paralinguistics studies, together with prosodic and voicing related features. Although the paralinguistic tasks at hand may well be enhanced with the use of linguistic model and prosody modeling, here we use only acoustic models in all 3 sub-challenges. In line with our previous work [9], 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. Although they are known to be alternative representations, in [9] RASTA-PLP and MFCC features were not found to be linearly dependent, therefore they have complementary rather than redundant information. In our preliminary experiments, we observed higher performance with the frame level combination of these two descriptors. The pipeline of FV feature extraction is given in Figure 2. Only in the Native Language Sub-Challenge, where the speech data is about 8GBs, we had to use only MFCCs due to memory limitations.

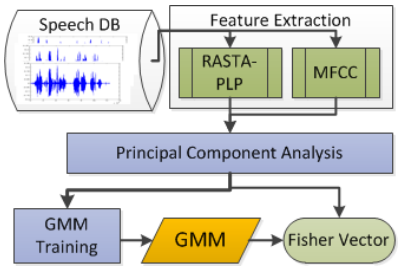

Figure 2: Fisher vector encoding based speech signal representation in the proposed framework

To distinguish the speech and non-speech frames, we use an energy based voice activity detector. In this approach, frames with lower energy than a threshold τ_E are considered to be non-speech. To smooth the decision boundary, we take the mean energy in a symmetric window of nine frames, centered at the frame of interest. As a measure of frame-level energy, we tried sum of RASTA-style auditory spectrum and MFCC 0 and observed that thresholding MFCC 0 gives more reliable results on speech signal segmentation.

2.2. Fisher Vector Encoding

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. 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 $F(\theta)$ [10]. To make the computation feasible, a closed form approximation to the diagonal of $F(\theta)$ is proposed [10]. 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 BoW model, however in FV, they have a negligible effect on performance [10]. 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.

In order to efficiently learn an Acoustic Background Model (ABM) using GMM with diagonal covariances, the data need to be decorrelated. Principal Component Analysis (PCA) is applied on the data for this purpose. To reduce the computational cost, we downsample LLDs prior to learning PCA and GMM. We take every second frame in the Sincerity and Deception SCs; every third frame in the Native Language SC, respectively.

2.3. Cascaded Normalization

Perronnin et al. proposed to improve the FV representation to be used in computationally efficient linear classifiers (e. g. Linear Kernel Support Vector Machines) with power normalization (POW), followed by instance level L_2 normalization [21]. This simple proposal is empirically verified to be very effective in a range of computer vision tasks, as well as our recent paralinguistic studies [9, 22]. The authors' argument is that power normalization helps "unsparsify" the distribution of feature values, thus improves discrimination:

$$f(x) = sign(x)|x|^\alpha, \quad (2)$$

where $0 \leq \alpha \leq 1$ is a parameter to optimize. In [21] the authors empirically choose $\alpha = 0.5$. Following [9], in this study we also investigate the suitability of sigmoid function (SIG):

$$h(\mathbf{x}) = \frac{1}{1 + exp(-\mathbf{x})}. \quad (3)$$

This way we avoid a hyper-parameter to optimize, while providing a non-linear normalization into [0,1] range. The flowchart of the normalization steps we applied on the baseline openSMILE features and extracted FVs is given in Figure 3. We use the combination of feature, value (applied to each value of the data

matrix separately) and instance level normalization strategies. Without using feature level normalization, the performance is poor for the baseline set, while FV encoding may also work without normalization.

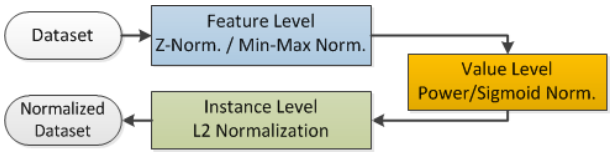

Figure 3: Cascaded feature normalization pipeline

2.4. Model Learning

To learn a classification model, we use Kernel ELM and PLS regression due to their fast and accurate learning capability.

ELM 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:

$$\mathbf{T}_{t,l} = \begin{cases} +1 & \text{if } y^t = l, \\ -1 & \text{if } y^t \neq l. \end{cases} \tag{4}$$

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:

$$\beta = \mathbf{H}^{\dagger}\mathbf{T}, \tag{5}$$

where $\mathbf{H}^{\dagger}$ is the Moore-Penrose generalized inverse [23] 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 $\mathbf{H}$ [14], relating ELM to Least Square SVM [24]:

$$\beta = (\frac{\mathbf{I}}{C} + \mathbf{K})^{-1}\mathbf{T}, \tag{6}$$

where $\mathbf{I}$ is the $N \times N$ identity matrix. In our experiments, we use Kernel ELM learning rule given in eq. (6).

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 [15]. PLS is applied to classification in one-versus-all setting between the feature matrix $\mathbf{X}$ and the binary label vector $\mathbf{Y}$, then the class giving the highest regression score is taken as prediction. The number of latent factors is a hyper-parameter to tune via cross-validation.

2.5. Fusion

We investigate feature level fusion, weighted score level fusion and their multi-level combination. Score level fusion is more appropriate when different feature representations necessitate distinct normalization pipelines. On the other hand, when the preprocessing is similar, combining features is likely to improve the performance.

In weighted score level fusion, the classifier confidence scores S_A and S_B are fused with a weight $0 \leq \gamma \leq 1$, searched with steps of 0.05:

$$S_{fusion} = \gamma * S_A + (1 - \gamma) * S_B. \tag{7}$$

The fusion parameter is optimized on the development set, as it is the case with other hyper-parameters.

3. Experimental Results

The challenge measure for classification based tasks is Unweighted Average Recall (i. e. mean recall of all classes). Sincerity SC is a regression task, where the measure is Spearman correlation.

For ease of reproducibility, we use open source tools in our experiments. For MFCC and RASTA-PLP feature extraction we use RASTAMAT library [25], for GMM training and FV encoding we use MATLAB API of VLFeat library [26]. In all 3 tasks, Fisher vectors are tested with PCA dimensions that explain 99.9% of the total variability, and $K_{GMM} = \{64, 128\}$ components for GMM. Prior to the experiments with FV, we analyze the baseline features with cascaded normalization strategies.

3.1. Experiments on the Native Language Sub-Challenge

The task in the Native Language SC[1] is to predict L1 from L2 (English speech). It is known that L2 speakers tend to transfer the linguistic and prosodic patterns existing in their native language. The task of this SC is to classify non-native English speakers from eleven different native languages: Arabic (ARA), Chinese (CHI), French (FRE), German (GER), Hindi (HIN), Italian (ITA), Japanese (JPN), Korean (KOR), Spanish (SPA), Telugu (TEL), and Turkish (TUR). The baseline development and test UAR scores for this SC are 45.1% and 47.5%, respectively. For other details, please refer to the paper on challenge [6].

We first carried out tests using combinations of cascaded normalization with the baseline feature set and PLS/ELM based classifiers. The best development set results using the baseline set are obtained using combination of z-normalization (ZN), sigmoid-normalization (SIG) and L_2 normalization, with Kernel ELM as classifier. This scheme gave a development set UAR score of 51.6% and a test set score of 53.4%. We then applied Canonical Correlation Analysis based Randomized feature selection proposed in [18]. The resulting ranking are given to classifiers in [100, 6300] range with steps of 100 features. The best results are obtained with 5300 features, which means an elimination of 1K features. This feature eliminated system gave 53.7% and 55.3% UAR for development and test sets, respectively.

We then proceed with our proposed FV based framework. Since the data in this corpus is very large compared to a moderate personal computer, we carried out our preliminary experiments only using 75 dimensional MFCC based LLDs. After application of PCA, 50 features are retained. Comparative results with respect to $K_{GMM} = \{64, 128\}$ GMM components, two classifiers and best normalization cascaded steps are given in Table 1. In this table, results of commonly used feature normalization approaches are given in the first part. The second part summarizes the best four combinations of cascaded normalization approach. Best UAR scores per column are shown

[1] **Source:** Derived from data provided by ETS. Copyright © 2016 ETS. www.ets.org. The opinions set forth in this publication are those of the authors not ETS.

in **bold**. Here, we observe that i) employed cascaded normalization boosts the performance as in the case of openSMILE features and ii) 128 GMM components give better results compared to 64 in FV encoding. In the remaining experiments, we carry out tests with $K_{GMM} = 128$.

Table 1: Native Language SC: comparative results using FV encoding with 50 PCA dimensions on MFCC descriptors.

UAR (%)	K_{GMM}**=64**		K_{GMM}**=128**	
Preprocessing	**PLS**	**ELM**	**PLS**	**ELM**
ZN	60.9	61.3	61.9	62.3
MM	60.6	59.1	61.6	60.8
POW+L_2	**64.9**	**64.8**	**66.8**	65.9
ZN+POW+L_2	63.1	62.9	65.5	**66.8**
ZN+SIG	63.2	62.0	65.3	65.0
ZN+SIG+L_2	62.5	62.4	64.6	66.4

Weighted score fusion of PLS and ELM using POW+L_2 combination of FV encoding (50 PCA dimensions, K_{GMM}=128) yielded an UAR score of 67.4%. When this FV representation is combined with the reduced openSMILE feature set, the development UAR scores rose to 67.1% and 66.6% for PLS and ELM, respectively. Further, the score fusion performance increased to 67.6%. This multi-level fusion approach attained a test set UAR of 71.5%, which outperforms baseline UAR by 50%, relatively.

The corresponding confusion matrix is given in Figure 4, where we observe the highest confusion between Hindi and Telugu followed by Italian and Spanish.

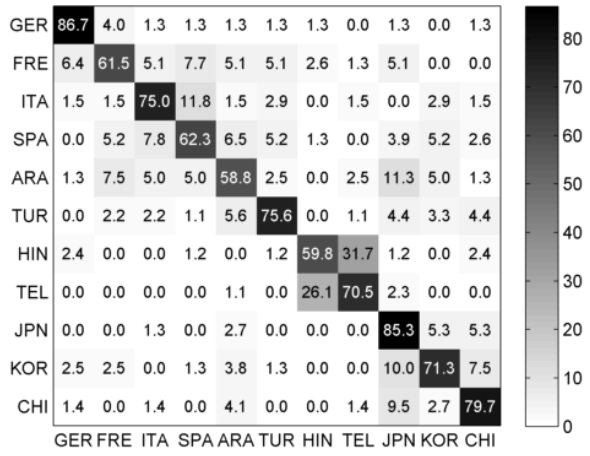

Figure 4: Test set confusion matrix in percent (UAR 71.5%)

3.2. Experiments on the Sincerity Sub-Challenge

The Sincerity SC is about predicting the perceived sincerity. Each recording is rated by 13 to 19 annotators and the average z-normalized individual scores are taken as ground truth.

In this SC, we applied the same pipeline as mentioned in the Native Language SC. We found that feature selection does not generalize well on this corpus, partly due to vagueness of the target variable. The first test set submission with selected 1280 features rendered a Spearman Correlation of 0.573, falling behind the baseline score of 0.602. The summary results of other three submissions that used cascaded normalization on the full baseline set and score level combination of baseline and FV based systems are listed in Table 2. FV encoding in this SC uses MFCC and RASTA-PLP descriptor combination reduced to 80 PCA dimensions. Our current best test set score (.640), outperforms the test set 7%, relatively.

Table 2: Development and test set results of proposed systems for Sincerity SC

Normalization	Features	Regr.	Devel.	Test
ZN+SIG+L_2	Baseline [6]	PLS	0.501	0.613
MM+POW+L_2	Baseline+FV: K_{GMM}=64	ELM	0.569	0.636
MM+POW+L_2	Baseline+FV: K_{GMM}=128	ELM	0.588	0.640

3.3. Experiments on the Deception Sub-Challenge

The Deception SC is a binary classification task. The data is collected in an empirical study at the University of Arizona. In the setup, some participants were asked to retrieve an exam key from a computer in department office, while some others retrieved a leaflet. Those who stole the exam key were asked to tell the truth in one session and to lie in another session, which provided the deception case. The development and test set baseline UAR scores are 61.9% and 68.3%, respectively.

Due to the imbalanced nature of the data, baseline system uses instance upsampling strategy. We implemented the same strategy in our experiments to avoid bias towards the majority class. Applying feature selection [18] on the upsampled training set and cross-validating on the development set yielded much higher performance compared to the baseline. Therefore, we dedicated 4 of 5 submissions to feature selected systems. A score fusion of models trained on FV encoding and on the baseline features was also probed. The best development UAR score obtained was 75.2% using selected 45 features from INTERSPEECH ComParE 2010 baseline set [27]. However, this rendered a test set UAR score of 66.6%. Apart from possibility of over-fitting, there may be a shift in the optimal hyperparameters for the combined training and development set.

4. Conclusion

In this work, we propose a framework that combines popular suprasegmental acoustic features with computer vision inspired FV encoding and applies multi-level normalization. For the Native Language SC, proposed framework achieves dramatically higher performance compared to that of the baseline system. Using feature and decision level fusion, we attain 50% increase relative to the baseline test set performance. In the Sincerity SC, we also get better results with the proposed approach, although not as dramatic as in the former SC. Given the short time and high difficulty level of the challenge conditions, using efficient and robust system components for signal processing and machine learning is of high importance. The proposed framework can be further enhanced with linguistic and prosodic modeling, which constitute our future directions. Furthermore, the proposed framework will be investigated for unsupervised domain adaptation in cross-corpus acoustic emotion recognition tasks.

5. Acknowledgments

This research is financially supported by the Russian Foundation for Basic Research (project № 16-37-60100).

6. References

[1] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *IEEE International Conference on Multimedia and Expo*, vol. 1. Los Alamitos, CA, USA: IEEE Computer Society, 2003, pp. 401–404.

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

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

[4] 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.

[5] 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.

[6] 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.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.

[14] 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.

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

[16] H. Kaya and A. A. Salah, "Combining modality-specific extreme learning machines for emotion recognition in the wild," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 139–149, 2016. [Online]. Available: http://dx.doi.org/10.1007/s12193-015-0175-6

[17] F. Gürpınar, H. Kaya, H. Dibeklioğlu, and A. A. Salah, "Kernel ELM and CNN Based Facial Age Estimation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016, pp. 80–86.

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

[19] 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.

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

[21] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *11th European Conference on Computer Vision*, 2010, pp. 143–156.

[22] 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

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

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

[25] 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/

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

[27] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, and S. S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge." in *INTERSPEECH*, Makuhari, Japan, Proceedings, 2010, pp. 2794–2797.