Fisher Vectors with Cascaded Normalization for Paralinguistic Analysis

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

[1]Department of Computer Engineering, Boğaziçi University, Bebek, İstanbul, Turkey
[2]St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences, Russia
[3] ITMO University, St. Petersburg, Russia
heysem@boun.edu.tr, karpov@iias.spb.su , salah@boun.edu.tr

Abstract

Computational Paralinguistics has several unresolved issues, one of which is coping with large variability due to speakers, spoken content and corpora. In this paper, we address the variability compensation issue by proposing a novel method composed of i) Fisher vector encoding of low level descriptors extracted from the signal, ii) speaker z-normalization applied after speaker clustering iii) non-linear normalization of features and iv) classification based on Kernel Extreme Learning Machines and Partial Least Squares regression. For experimental validation, we apply the proposed method on INTERSPEECH 2015 Computational Paralinguistics Challenge (ComParE 2015), Eating Condition sub-challenge, which is a seven-class classification task. In our preliminary experiments, the proposed method achieves an Unweighted Average Recall (UAR) score of 83.1%, outperforming the challenge test set baseline UAR (65.9%) by a large margin.

Index Terms: ComParE, computational paralinguistics, Eating Condition, Fisher vector, PLS, ELM, signal representation

1. Introduction

Paralinguistics is the study of non-verbal aspects of speech. It deals with how the words are spoken, rather than what is being spoken. A set of paralinguistic tasks, such as emotion [1, 2], depression [3] and personality [4] are popularly investigated and yet there is a plethora of other tasks to be discovered.

In this context, INTERSPEECH 2015 ComParE challenge [5] introduces a novel problem, which is to classify the eating condition (EC) of the speaker. There are seven different ECs (speech with no food plus six different types of food) to be classified using acoustic features. The challenge opens a new area of paralinguistic research that can be beneficial for existing studies e. g. by adapting speech and speaker recognizer systems to ECs. The problem is related to a "state" of the speaker, rather than a "trait", and therefore, individual differences should be minimized/compensated.

Modeling/compensating variability due to speakers is of interest in many speech related disciplines. In speaker recognition, state-of-the-art systems are built using i-Vector (i. e. total variability) modeling introduced by Dehak et al. [6], which has its roots in Joint Factor Analysis (JFA) approach [7, 8]. In this approach, the total variance is factorized, and it is postulated that some factors encode for idiosyncratic variations, whereas others are more general. i-Vectors are also used in other paralinguistic tasks [9, 10] for compensating variability due to speakers, rather than augmenting speaker related information for identification purposes.

We propose the use of Fisher vectors (FV) for encoding the low level descriptors (LLD) over utterances. This super vector modeling is introduced and popularly used in computer vision, especially in large scale image retrieval [11, 12]. The idea is to measure the amount of change induced by the utterance/video descriptors on a background probability model, which is typically a Gaussian Mixture Model (GMM). In other words, FV encodes the amount of change of model parameters to optimally fit the new-coming data. This requires the computation of the Fisher information matrix, which is the derivative of the log likelihood with respect to model parameters (hence the name "Fisher"). The encoding requires far less number of components in a GMM than the Bag of Words (BoW) approach [13].

In order to address the speaker variability issue in the EC sub-challenge by employing FV encoding, we first extract Mel Frequency Cepstral Coefficients (MFCC) and RASTA-style Perceptual Linear Prediction (PLP) Cepstrum to represent the signal properties. We show that the combination of RASTA-PLP and MFCC descriptors improve over their individual performances. Moreover, our experiments have shown that the FV encoding of extracted LLDs reaches the performance improvement obtained by the speaker based z-normalization of baseline feature set extracted via openSMILE tool [14]. The performance of the FV is further improved by applying speaker z-normalization. In order to apply this on the challenge test set, where the speaker labels are missing, we implemented Hierarchical Agglomerative Clustering (HAC), which is commonly used to identify speakers [15, 16, 10].

For modeling, we use Extreme Learning Machines (ELM) [17, 18] and Partial Least Squares (PLS) regression [19] based classifiers, motivated by their fast learning capability and outstanding performance in recent challenges [20, 21, 22].

We explain the effect of each component of our framework separately and in a combined fashion. The remainder of this paper is organized as follows. In Section 2 we introduce the proposed method and give background on its major components. The experimental results are given in Section 3, Section 4 concludes with future directions.

2. Proposed Method

The overview of the proposed speech signal representation method is given in Figure 1, where speaker IDs are used for speaker z-normalization as the first stage of cascaded normalization.

2.1. Speech Signal Processing

MFCC and RASTA-PLP [23, 24] are the most popular descriptors used in a variety of speech technologies ranging from speaker identification to speech recognition, although they are

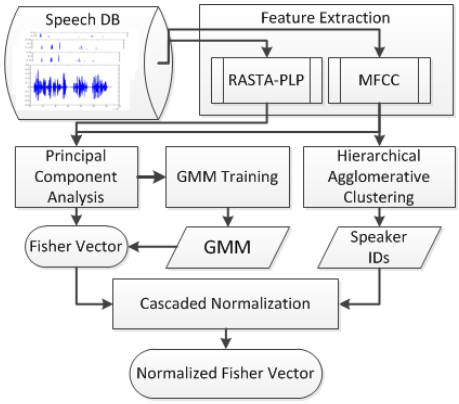

Figure 1: Overview of the proposed method: speech signal representation using Fisher vectors with cascaded normalization

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. Since the task at hand is to recognize the EC, which can be identified with the existence of specific acoustic characteristics (e. g. due to the sound of crunching or chewing), we implement an acoustic model, ignoring the prosody that can be biased towards the speaker identity.

For the purpose of speech signal representation, 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. After a preliminary analysis, we have found that MFCC bands 22-24 are linearly dependent on the first 21 bands; nonetheless, their removal decreased the performance. Moreover, although they are known to be alternative representations, RASTA-PLP and MFCC features are not found to be linearly dependent, therefore they have complementary rather than redundant information.

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 $F(\theta)$ [11]. To make the computation feasible, a closed form approximation to the diagonal of $F(\theta)$ is proposed [11]. 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 [11]. 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, we first need to decorrelate the data gathered from all utterances. Principal Component Analysis (PCA) is applied on the data for this purpose. To reduce the computational cost, we take LLDs from every second frame to learn PCA and GMM. In our preliminary tests, this sub-sampling did not decrease the performance. Once the parameters of PCA projection and GMM are learned, we use all speech frames from each utterance without sub-sampling to represent them as a FV.

2.3. Speaker Clustering

Despite the fact that FV encoding aims to compensate the speaker dependent variability, there may still be bias in this representation towards speakers. To further enhance the features in eliminating the speaker dependent information, we use speaker based z-normalization. Since the speaker IDs of utterances are given only for the training/validation set, we need to employ a clustering method to obtain speaker ID information on the challenge test set.

A literature review on speaker clustering reveals that GMM or K-Means clustering on LLDs do not give desirable results. Moreover, the *must-link condition* for LLDs of an utterance is not met with these partitional clustering methods. The most popular method employed for this purpose is single Gaussian based bottom up Hierarchical Agglomerative Clustering with Generalized Likelihood Ratio (GLR) as distance measure [10, 15, 16]. In this method, initialization is done by modeling LLDs of each utterance with a full covariance Gaussian. Then the GLR is computed for each pair of components, and the pair with minimum GLR distance is merged into a single Gaussian component. This continues until one component is left. If the optimal number of components K^* is known, the clustering with K^* components can be taken. Otherwise, one needs to use automatic model selection criteria, such as Bayesian Information Criterion (BIC) [25], or Minimum Description Length (MDL) [26]. In our problem, the number of speakers in the test set is given in the challenge [5].

In HAC, we use MFCC 1-12 as in [16], instead of 75 dimensional MFCCs used in the ABM. We also use a higher energy threshold compared to the one used in ABM, since here we are interested in clean speech rather than "eating noise" that is useful in discrimination of the EC.

2.4. Feature Normalization

Perronnin et al. further improve the FV representation to be used in linear classifiers (e. g. Linear Kernel Support Vector Machines) with power normalization, followed by instance level L_2 normalization [27]. The authors argue 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 [27] the authors empirically choose $\alpha = 0.5$. In this study we investigate the suitability of sigmoid function, which is commonly used as hidden layer activation function of Neural Networks:

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

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 2. 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 does not necessitate this step.

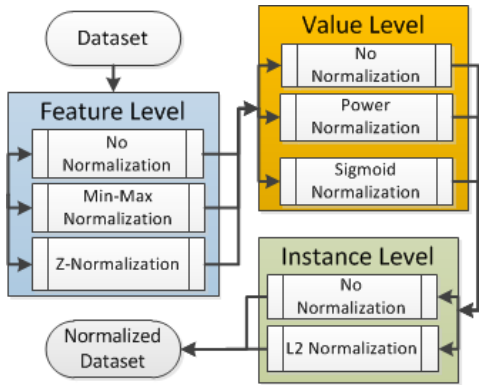

Figure 2: Cascaded feature normalization pipeline

2.5. 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:

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

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:

$$\beta = \mathbf{H}^{\dagger}\mathbf{T}, \quad (5)$$

where $\mathbf{H}^{\dagger}$ is the Moore-Penrose generalized inverse [28] 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 $\mathbf{H}$ [18], relating ELM to Least Square SVM [29]:

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

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. The decomposition is done by finding projection weights $\mathbf{W}_x, \mathbf{W}_y$ that jointly maximize the covariance of corresponding columns of $\mathbf{U}_x = \mathbf{X}\mathbf{W}_x$ and $\mathbf{U}_y = \mathbf{Y}\mathbf{W}_y$. For further details of PLS regression, the reader is referred to [19]. 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.

3. Experimental Results

The corpus of the EC sub-challenge is from [30], which features a total of 1 414 speech utterances produced by 30 speakers with either no food or eating one of the following six food types: apple, nectarine, banana, crisp, biscuit, gummy bear. Each speaker is expected to give data for 7 utterances for each class. However, since some speakers refuse to eat some food, some classes are missing for these speakers. The challenge measure is Unweighted Average Recall (i. e. mean recall of all seven classes), used as challenge measure since INTERSPEECH 2009 challenge [1]. The challenge organizers provide a baseline feature set consisting of 6 373 suprasegmental features extracted via the latest version of openSMILE tool [14].

The data are segmented into a training set with 20 speakers, and a test set with 10 disjoint speakers. Model optimization is done via 20-fold leave-one-speaker-out (LOSO) cross-validation (CV). The test set labels and speaker IDs are not known by the competitors. For further details on challenge protocol, reader is referred to the challenge paper [5].

For ease of reproducibility, we use open source tools in our experiments. For MFCC and RASTA-PLP feature extraction we use RASTAMAT library [31], for GMM training and FV encoding we use MATLAB API of VLFeat library [32]. Prior to the experiments with FV, we analyze the baseline features with cascaded normalization strategies.

In all our experiments we generated linear kernels from the preprocessed data and used these kernels in ELM and PLS. The regularization parameter in ELM is optimized in the set $10^{-6,-5,...,5}$ with exponential steps. The number of latent factors for PLS is searched in [2,24] range with steps of two.

3.1. Experiments with the Baseline Feature Set

As mentioned earlier, the baseline UAR for the training/validation set is computed via LOSO CV by combining the predictions on each fold to get an overall performance. The baseline UAR scores are 61.3% and 65.9%, for the training/validation and the test sets, respectively.

We first analyzed the features using the combination of normalization strategies described in Section 2.4. Combination of z-norm + logistic sigmoid + L_2-norm reached the highest LOSO UAR score (63.2%) among other alternatives.

We analyzed the effect of feature selection separately using z-normalization and min-max normalization with two feature filters: multi-view discriminative projection based feature selection [33] that generated the best performance in INTERSPEECH 2014 Physical challenge [34] and a randomized version of this filter [35]. To our surprise, the highest improvement

over the baseline was less than one percent, remaining below the individual contribution of the cascaded normalization.

Using the ground truth speaker IDs for speaker z-normalization, it was possible to dramatically increase the UAR performance to 70.1% with PLS and to 70.8% with ELM. When speaker z-norm is augmented with logistic sigmoid + L_2-norm combination, UAR reaches 71.6% with ELM. The results indicate that the features are highly biased towards speakers and a marked improvement can be obtained by minimizing speaker variability.

3.2. Experiments with the Proposed Method

We test the effect of RASTA-PLP and MFCC separately to evaluate their individual and combined performance. We then apply speaker z-normalization using ground truth and predicted speaker IDs for our final system.

Fisher vectors for RASTA-PLP are tested with a range of PCA dimensions, and $K = \{64, 128, 256\}$ components for GMM. The best UAR performance (62.7%) is obtained with 30 PCA dimensions, 128 GMM components, preprocessed using power-normalization + L_2-norm and PLS based classifier. Note that this performance is slightly higher than the baseline.

In the remaining experiments, we use $K = 128$ to train GMMs, as it gives a good compromise between computational complexity and UAR performance. Using MFCC features, the performance is increased to 66.9% with 70 PCA dimensions, with logistic sigmoid + L_2 norm. In results not reported here, we observed that the classifiers react differently to non-linear preprocessing alternatives. We also noticed a jump in UAR performance (64.3%→66.9%) from 60 to 70 PCA dimensions. This implies that the "devil is in the details", as the variability due to eating noise that is useful for discrimination might be contained in these eigenvectors.

When the two descriptors are combined, the best overall UAR is obtained as 70.4%, with 110 PCA dimensions and power-norm + L_2 norm combination. Further dramatic improvement is attained when speaker z-normalization is applied. We reach 76.1% and 77.0% UAR using speaker clustering and real speaker IDs, respectively (see Table 1). Score fusion of the best two systems gave 77.5% UAR both with predicted and ground truth speaker IDs.

Table 1: UAR scores of RASTA-PLP + MFCC combination

UAR (%)	Power-L_2		Logsig-L_2		No-norm	
Preprocess	PLS	ELM	PLS	ELM	PLS	ELM
PCA 80	**66.1**	64.2	65.5	64.8	65.7	64.0
PCA 100	**67.5**	63.9	65.7	**67.4**	66.8	65.8
PCA 110	69.4	**70.4**	66.8	67.9	66.0	68.7
PCA 110 with speaker z-normalization						
Real ID	75.3	75.6	**77.0**	**77.3**	76.3	76.5
Pred. ID	74.2	73.9	75.5	74.4	**76.1**	74.1

For the challenge test set, we have submitted predictions of five systems. The first is score fusion of the best two systems with predicted speaker IDs (see the last row of Table 1). This resulted in a test set UAR of 81.4%. For the second submission, we re-trained GMM based ABM using the descriptors from the training and test sets. This increased the best training set UAR to 78.9% using logistic sigmoid + L_2-norm combination with PLS. The test set UAR increased slightly to 81.6%. This result is motivating as it shows that using only training set for acoustic background modeling generalizes as good as combination of the training and test sets. We observed that the predicted labels for the first two submissions differ in 72 instances, therefore we fused their scores for the third submission. This combination reached a test set UAR of 83.1%. The corresponding confusion matrix is given Figure 3, where we see a perfect recall of "No Food" class. We also observe high recall for "Crisp" and "Biscuit" classes, where we may expect high confusion. The lowest recall is observed with "Nectarine", which is confused generally with "Apple" (25%) and "Banana" (13%).

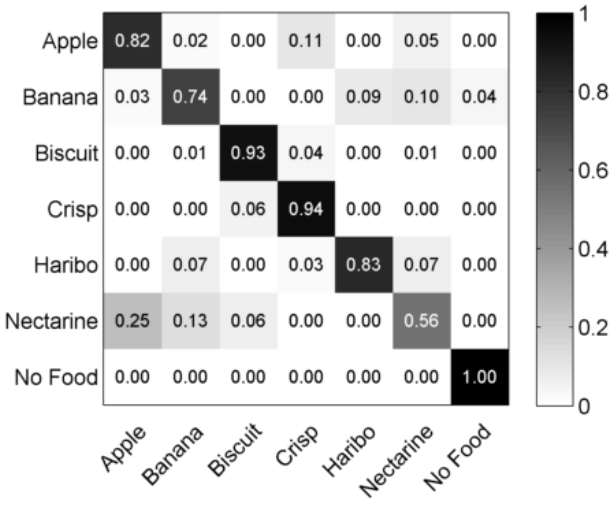

Figure 3: Confusion matrix of submission 3 (UAR 83.1%)

We finally employed a weighted fusion scheme on the best performing four base classifiers, reaching test set UAR scores of 74.5% and 82.9%, for our fourth and fifth submissions, respectively.

4. Conclusion

In this study, we proposed a novel method combining the FV encoding with cascaded normalization that is composed of speaker z-normalization and non-linear normalization. The results indicate superior performance of the proposed method over the challenge baselines obtained with openSMILE features. The experiments with the baseline feature set reveal the importance of compensating speaker variability, which is handled partly by the FV approach and partly by speaker z-normalization employed after Hierarchical Agglomerative Clustering. The best overall test set performance is obtained with score fusion of systems trained on combination of RASTA-PLP and MFCC descriptors. The results on both baseline and extracted features indicated that proposed sigmoid normalization is a good alternative to power-normalization used to enhance non-linear discrimination capability of linear classifiers. Application of the proposed method on the challenging task of cross-corpus acoustic emotion recognition constitutes our nearest future work. Cascaded normalization can be improved using other speaker adaptation transforms such as the one proposed in [36].

5. Acknowledgments

This research is partially supported by the Council for grants of the President of Russia (project № MD-3035.2015.8) and 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] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge," in *4th ACM Intl. Workshop on Audio/Visual Emotion Challenge, Orlando, Florida, USA, Proceedings*, 2014, pp. 3–10.

[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.

[6] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[7] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, 2005.

[8] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[9] J. M. K. Kua, V. Sethu, P. Le, and E. Ambikairajah, "The UNSW Submission to Interspeech 2014 ComParE Cognitive Load Challenge," in *INTERSPEECH, Singapore, Proceedings*, 2014, pp. 746–750.

[10] M. Van Segbroeck, R. Travadi, C. Vaz, J. Kim, M. P. Black, A. Potamianos, and S. S. Narayanan, "Classification of cognitive load from speech using an i-vector framework," in *INTERSPEECH, Singapore, Proceedings*, 2014, pp. 751–755.

[11] 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.

[12] 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, San Francisco, CA, USA, Proceedings*, 2010, pp. 3384–3391.

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

[14] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent Developments in openSMILE, the Munich open-source Multimedia Feature Extractor," in *ACM MM '13 – 21st ACM International Conference on Multimedia, Barcelona, Spain, Proceedings*, 2013, pp. 835–838.

[15] W. Wang, P. Lu, and Y. Yan, "An improved hierarchical speaker clustering," *ACTA ACUSTICA*, vol. 33, no. 1, p. 9, 2008.

[16] K. J. Han, S. Kim, and S. S. Narayanan, "Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1590–1601, 2008.

[17] 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.

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

[19] H. Wold, "Partial least squares," *Encyclopedia of Statistical Sciences*, 1985.

[20] H. Kaya and A. A. Salah, "Combining modality-specific extreme learning machines for emotion recognition in the wild," in *ICMI '14 – 16th International Conference on Multimodal Interaction, Istanbul, Turkey, Proceedings*, 2014, pp. 487–493.

[21] G. Varol and A. A. Salah, "Extreme learning machine for large-scale action recognition," in *ECCV Workshop on Action Recognition with a Large Number of Classes, Zürich, Switzerland, Proceedings*, 2014.

[22] 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 *ICMI '14 – 16th International Conference on Multimodal Interaction, Istanbul, Turkey, Proceedings*, 2014, pp. 494–501.

[23] 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.

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

[25] G. Schwarz, "Estimating the Dimension of a Model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1979.

[26] J. Rissanen, "A Universal Prior for Integers and Estimation by MDL," *The Annals of Statistics*, vol. 11, no. 2, pp. 416–431, 1983.

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

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

[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] S. Hantke, F. Weninger, R. Kurle, A. Batliner, and B. Schuller, "I hear you eat and speak: automatic recognition of eating condition and food type," *(to appear)*.

[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] H. Kaya, T. Özkaptan, A. A. Salah, and 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.

[34] 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 *INTERSPEECH, Singapore, Proceedings*, 2014, pp. 427–431.

[35] 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.

[36] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, "MLLR transforms as features in speaker recognition," in *Eurospeech, Lisbon, Portugal, Proceedings*, 2005, pp. 2425–2428.