Interspeech 2018
2-6 September 2018, Hyderabad

Computational Paralinguistics: Automatic Assessment of Emotions, Mood, and Behavioural State from Acoustics of Speech

Zafi Sherhan Syed[*], *Julien Schroeter*[†], *Kirill Sidorov, David Marshall*

Cardiff University, UK
{SyedMZ, SchroeterJ1, SidorovK, MarshallAD}@cardiff.ac.uk

Abstract

Paralinguistic analysis of speech remains a challenging task due to the many confounding factors which affect speech production. In this paper, we address the Interspeech 2018 Computational Paralinguistics Challenge (ComParE) which aims to push the boundaries of sensitivity to non-textual information that is conveyed in the acoustics of speech. We attack the problem on several fronts. We posit that a substantial amount of paralinguistic information is contained in spectral features alone. To this end, we use a large ensemble of Extreme Learning Machines for classification of spectral features. We further investigate the applicability of (an ensemble of) CNN-GRUs networks to model temporal variations therein. We report on the details of the experiments and the results for three ComParE sub-challenges: Atypical Affect, Self-Assessed Affect, and Crying. Our results compare favourably and in some cases exceed the published state-of-the-art performance.

Index Terms: social signal processing, speech analysis, computational paralinguistics, affective computing, deep learning, ensemble methods

1. Introduction

Paralinguistics is the study of non-verbal aspects of speech [1], and is increasingly becoming a mainstream topic within the domains of signal processing and machine learning, and is one of the hot research topics within Social Signal Processing [2]. In its tenth consecutive year, the Interspeech 2018 Computional Paralinguistics Challenge (ComParE) [3] offers three new corpora in the form of sub-challenges for recognition of Atypical Affect, Self-assessed affect, and Crying.

The objective of the Atypical Affect sub-challenge is to predict emotional state of individuals who have been diagnosed with mental, neurological, or physical disabilities. This sub-challenge is based on the Emotional Sensitivity Assistance System for People with Disabilities (EmotAsS) corpus, details of which were published at Interspeech 2017 [4]. While considerable research has focussed on emotion recognition from speech [5], the EmotAsS corpus is particularly challenging since it focuses on emotions of individuals with disabilities. Hante et al. [4] report that they observed very different emotional responses from individuals with disabilities for the same task. They state, for example, *"we found that very different emotional responses could be observed for the same task; for example, a woman laughed with joy when she had to describe a picture with a hurt dog because she simply liked the dog. Another woman wept over the picture because she had just lost her dog".*

The Self-Assessed Affect sub-challenge deals with the prediction of mood of individuals from their speech recordings, specifically, the valence of their emotions. As reported in [3], the objective of this sub-challenge is to lay foundations for applications that support individuals with affective disorders, as well as monitor rapport between therapists and patients.

The third sub-challenge, Crying, is somewhat unique in the sense that it focuses on infant vocalisations, instead of that of adults. The objective of this sub-challenge is to develop methods which can automatically recognise between vocalisations which represent neutral/positive mood of the infant, fussing, and just crying. This challenge is particularly important since automatic classification of infant vocalisation has many application for remote monitoring of children in intensive care units, as well as children wards.

2. Datasets

The Atypical Affect sub-challenge uses speech recordings which are part of the EmotAsS corpus [4], published at Interspeech 2017. There are however some significant changes. For this sub-challenge, organisers do not provide subject IDs, but instead provide provide partitions for training, development (validation), and test sets. The objective is to correctly predict between one of the four emotions which include angry, happy, neutral, and sad.

The Self-Assessed sub-challenge uses speech recordings from the Ulm State-of-Mind in Speech (USoMS) corpus, where the objective was to study core affect via valence in free speech. To the best of our knowledge, this corpus has not been published prior to Interspeech 2018, although the baseline paper [3] reports that self-assessed labels were collected on a 10-point Likert-scale. Later, these values were quantised to yield a three class classification task for this sub-challenge, with the following range: (i) low: 0-4, (ii) medium: 5-7, (iii) high: 8-10. Finally, the Crying sub-challenge uses vocalisation recorded by Marschik and his team as part of their work on early detection of neurodevelopmental disorders [6]. For further details on datasets used in these challenges, the reader is referred to the baseline paper [3].

3. The Proposed Approach

3.1. Spectral Modelling with Fisher Vectors

We posit that a substantial amount of paralinguistic information is contained in the speech spectra alone, and we demonstrate that by modelling the latter effectively, we can train classifiers to predict labels for the three sub-challenges for ComParE 2018 with reasonable accuracy. Spectral modelling has previously been shown to be useful for a variety of paralinguistic tasks such as those pertaining to emotion recognition [5] and screening of mental disorders [7] in addition of previous Interspeech ComParE challenges.

Spectral modelling typically involves computing spectral low-level descriptors (LLDs) over short segments of speech followed by feature aggregation. Feature aggregation is an approach

[*] Worked on GEWELMs
[†] Worked on CNN-GRU

10.21437/Interspeech.2018-2019

through which LLDs are summarised to create features which provide global information about the speech recordings. While several feature aggregation methods exist, such as functionals [8], GMM supervectors [9], Vectors of Locally Aggregated Descriptors (VLADs) [10], i-vectors [11] etc., we opt to use Fisher Vector encoding for aggregating spectral LLDs based on our previous experience: we found them effective for classifying between individuals with and without depression [12], as well as prediction of their depression severity [13].

While FV encoding was originally proposed by [14] for building visual vocabularies, it has become popular for a variety of applications in the field of social signal processing, such as depression recognition [15, 16, 12, 13], emotion recognition [17] as well as recent Interspeech Computational Paralinguistics (ComParE) challenges [18, 19, 20].

The recent popularity of Fisher Vector encoding, especially within the social signal processing community, is due to the fact that it combines the advantages of both generative and discriminative approaches for machine learning [21]. The process flow for Fisher Vector encoding starts with building a generative model (typically, using Gaussian Mixture Models, GMMs) of LLDs, and later computing the Fisher kernel from this generative model. Essentially, FV measures the deviation of the LLDs from the generative model. Fisher Vectors are quantified using first and second order statistics of the gradient of the sample log-likelihood with respect to the model parameters [14, 22].

As representations of speech spectra, we use three sets of low level spectral representations. These include the Mel Frequency Cepstral Coefficients (MFCCs), Perceptual Linear Prediction (PLP) coefficients, and the ComParE 2016 spectral LLDs. While MFCCs and PLPs are standard representations for spectra in speech processing, ComParE spectral LLDs have demonstrated impressive performance in the baseline paper for this challenge, and motivated us to use this feature set as well. MFCC, PLP and ComParE 2016 spectral LLDs were computed the using the `openSmile` toolkit version 2.3.0 [23] using `MFCC12_E_D_A.conf`, `PLP_E_D_A.conf`, and `ComParE_2016.conf` configuration files, respectively. For further details on these features, see [23].

The process of FV encoding for spectral modelling is summarised as follows: we concatenate spectral LLDs from each speech recording into a matrix and then build a background model for the spectral space using a GMM [24]. Next we compute Fisher Vectors using the `Matlab` API of `VLFeat` library [25] for both, estimating the vector means, covariance matrix and priors of the GMM, and implementing FV encoding. For FV encoding we compute both the vanilla Fisher Vectors as well as Perronnin's improved FVs [22].

3.2. Ensembles of Weighted Extreme Learning Machines

An Extreme Learning Machine (ELM) is essentially a single layer feed-forward neural network where the hidden layer is assigned randomly generated weights which are not updated during the training process. For classification, the output from the hidden layer can be mapped to the training labels using a least squares regression [26]. The idea is that even with random weights, the hidden layer can learn useful representation of input data which can be exploited by designing a suitable output layer. An outstanding advantage of ELMs is their very fast training time, which eases the process of tuning the hyper-parameters and experimentation.

We note that while ELMs were popularised by Huang et al. in 2004 [26], the fundamental concepts of ELMs have existed for much longer. Ping et al. proposed using least squares regression to compute weights of a neural network in [27]. The "random weights" concept of ELMs is analogous to the concept of random projections for feature mapping. If the number of neurons in the hidden layers is smaller than dimensionality of the input data, the ELM essentially implements dimensionality reduction. Conversely, when the the number of neurons are larger than the input dimensions than the ELM performs dimensionality expansion.

The techinque of dimensionality reduction using random projections is supported by the 1984 Johnson-Lindenstrauss Lemma [28], according to which *"points in a vector space of sufficiently high dimension, may be projected into a suitable lower-dimensional space in a way which approximately preserves the distances between the points"*. Meanwhile, dimensionality expansion is supported by Cover's theorem [29], according to which *"a complex pattern-classification problem, cast in a high-dimensional space non-linearly, is more likely to be linearly separable than in a low-dimensional space, provided that the space is not densely populated"*.

In our work, we use ELMs as a method for dimensionality reduction followed by least squares regression towards class label prediction. As such, we do not use a non-linear activation function. Moreover, since the process of FV encoding expands the dimensionality to a factor of $2KD$, where K is the number of Gaussians in the GMM, and D is the dimensionality of the input vector, we use principal component analysis (PCA) to reduce the dimensionality such that 95% of variance is preserved, before using ELMs.

It is important to note that ELMs have previously been reported to provide good performance for tasks pertaining to emotion recognition [17], Interspeech ComParE [18, 19], and depression recognition [30]. Furthermore, to deal with class imbalance in datasets (which also exits in ComParE 2018), Zong et al. proposed Weighted Extreme Learning Machines (WELMs) in [31]. WELMs assign weights to each class according to the number of training examples available for that class. A typical WELM classifier has two hyper-parameters: (1) the number of neurons L in the hidden layer and (2) the regularisation parameter c required for the generalised Moore-Penrose inverse [31], which can be tuned. In our work, we fix $C = 1$, and use four values for L i.e. $L \in \{2, 5, 10, 50\}$. It is also worth mentioning that both us [12] and Kaya et al. [20] concurrently proposed, albeit at different conferences, the use of weighted extreme learning machines for tasks pertaining to social signal processing.

It is quite obvious that not all random projections will yield acceptable results in terms of UAR (which is used to measure accuracy for ComParE 2018) for the classification tasks at hand. Some random projection vectors may actually reduce the separability between classes, while others may increase the separability. Rather than manually sift for useful random projection vectors, in this work, we propose *Greedy Ensembles of Weighted Extreme Learning Machines* (GEWELMs).

The fundamental idea behind GEWELMs is to train a sufficiently large number of WELMs and then select those which have UAR above a certain threshold for the development partition. We arbitrarily fix the threshold as the value corresponding to 80th percentile of the UAR of all WELMs in the ensemble. We do appreciate the fact that GEWELMs can have a tendency to over-fit to the development partition, hence we train two sets of GEWELMs. The first regime is called T2D-GEWELMs, where is the conventional training on the training partition and testing on the development partition, and the second is called D2T-GEWELMs, where we train of the development partition

Table 1: *Performance (UAR, %) of an ensemble of 2000 WELMs on the Atypical Affect sub-challenge with various features vectors, as a function of number of Gaussians and neurons. Results are shown for the two regimes: training on the training set and testing on the development one* (T→D)*, and vice versa* (D→V)*;* Avg. *is the average of the two regimes.*

#Neur. →	2			5			10			50		
#Gauss. ↓	T→D	D→T	Avg	T→D	D→T	Avg	T→D	D→T	Avg	T→D	D→T	Avg
					ComParE Spectral Features							
8	48.5	44.4	*46.5*	49.2	47.4	***48.3***	49.0	43.5	*46.3*	41.7	37.2	*39.5*
16	44.4	43.0	*43.7*	47.2	45.8	*46.5*	45.4	44.8	*45.1*	40.0	36.6	*38.3*
32	43.9	42.0	*43.0*	50.0	45.3	*47.7*	45.8	42.5	*44.2*	42.7	36.3	*39.5*
64	43.6	40.8	*42.2*	46.0	40.8	*43.4*	47.0	41.9	*44.5*	38.3	36.4	*37.4*
					PLPs							
8	50.5	51.3	*50.9*	49.2	50.2	*49.7*	48.4	48.2	*48.3*	44.6	39.9	*42.2*
16	50.4	51.8	*51.1*	50.9	50.1	*50.5*	52.5	50.3	*51.4*	47.3	43.9	*45.6*
32	49.5	49.0	*49.2*	52.4	52.9	***52.6***	50.3	50.3	*50.3*	44.3	41.8	*43.1*
64	48.5	46.8	*47.6*	52.4	47.4	*49.9*	50.6	45.6	*48.1*	47.2	37.0	*42.1*
					MFCCs							
8	51.6	51.9	*51.8*	52.9	53.2	***53.0***	49.2	49.8	*49.5*	44.1	40.5	*42.3*
16	48.7	51.6	*50.1*	50.8	50.2	*50.5*	49.7	49.6	*49.7*	42.9	40.6	*41.7*
32	49.5	50.1	*49.8*	50.8	49.0	*49.9*	50.6	51.2	*50.9*	44.3	40.7	*42.5*
64	49.2	49.4	*49.3*	54.3	51.6	***53.0***	52.0	49.7	*50.9*	46.2	40.0	*43.1*

Table 2: *Summary of performance (UAR, %) of an ensemble of 2000 WELMs on the Self-Assessed sub-challenge with various features vectors, as a function of number of Gaussians and neurons, and a linear SVM classifier as a baseline*

Features	GEWELMs			SVM		
	UAR	#Gauss	#Neu	UAR	#Gauss	Cost
			Vanilla Fisher Vectors			
MFCCs	70.8	32	5	49.2	4	10^{-4}
PLPs	71.4	16	5	51.4	8	10^{-1}
ComParE Spec.	74.3	96	5	53.4	4	10^{-1}
eGeMAPS+	70.4	8	5	57.3	8	10^{-3}
			Improved Fisher Vectors			
MFCCs	69.6	8	10	52.0	4	10^{-1}
PLPs	67.8	4	2	52.0	32	10^{-2}
ComParE Spec.	72.8	4	10	58.2	4	10^{-1}
eGeMAPS+	72.4	8	10	56.8	16	10^{-2}

and test on the training partition. This serves to regularise the selection of WELMs in the ensemble by mandating that the set of random projections used for a particular WELM have acceptable performance for both T2D-GEWELMs and D2T-GEWELMs.

3.3. CNN-GRU model

Deep convolutional neural networks (CNNs) have recently demonstrated excellent performance on diverse tasks ranging from image classification, to speech recognition, to natural language processing [32, 33]. We have investigated the application of CNNs for the task at hand.

Structure The deep learning architecture proposed for this paper consists of a four-layer Convolutional Neural Network (CNN) followed by a Gated Recurrent Unit (GRU) [34]. The former acts as feature extractor, whereas the latter performs modelling of temporal features. More specifically, the four convolutional layers are each constituted by (3×3) convolutional filters and are interleaved with (2×3) pooling operations and ReLu activation functions. The first three layers contain 8, 16 and 16 filters respectively. In an attempt to avoid overfitting and to reduce the predictions variance, an ensemble of models approach has been implemented. To that end, the number of filters in the final convolutional layer was arbitrarily set to values ranging from 8 to 12 depending on the run. This last value is a key model hyper-parameter since it corresponds to the dimensionality of the features fed to the recurring unit.

For its part, the recurrent layer consists of a standard single-celled GRU. The specific number of units composing the cell is also considered a free parameter for the model ensembling procedure; it ranges from 12 to 16. Eventually, the respective prediction of each model is obtained by feeding the resulting recurring features to a unique fully-connected (dense) layer. Overall, the network architecture was intentionally kept small (< 7000 parameters) and simple in order to prevent overfitting. Taking this precautionary measure seemed necessary due to the limited amount of data available for certain classes.

Input data This multistep pipeline has been directly applied on centered and normalized Mel-spectrograms of the raw audio extracts. These were generated using short-time Fourier transform (STFT) with 1024 points, frame sizes 0.02s, frame strides 0.006s and 81 frequency bins. The resulting length of the spectrograms is quite diverse ranging from less than 100 to more than 3000 temporal bins. However, this is not an issue since the recurrent unit can dynamically handle variable length inputs.

Data augmentation In addition, data augmentation has been implemented to extend the available data. This technique has shown its importance over the years in training deep learning models including audio classification models [35]. The benefits of an artificial increase in dataset size is even more noticeable for small datasets such as the ones provided for the challenge. In consequence several data augmentation techniques have been implemented for our submission: pitch shifting, time stretching and white noising. These transformations were not only applied for training, but also at inference time; the predictions are obtained by averaging the prediction of each individual augmentation fold. Overall, including such transformations has proven to lead to more stable training, less variability in the prediction and better out-of-sample performances.

Training regime In terms of training, a standard mini-batch cross-entropy minimizing procedure using Adam optimizer [36] was chosen. In order to overcome the great imbalance in the number of samples per class (see Section 2), stratified sampling was used to produce more balanced training batches. Additionally, a multi-task [37] learning approach was implemented to overcome the dataset size: the network was simultaneously trained to solve the task at hand as well as a voice recognition task. More precisely, a small subsample of the Librispeech corpus [38] was added to the training dataset; the voice extracts of four different people reading books were selected. Thus, the prediction space was increased by four new labels resulting in a more challenging task, but benefiting from a larger amount of data. This addition proved to be very useful in order to reduce overfitting.

In summary, we combine simple architecture, model ensembling, data augmentation and multi-task learning to overcome the challenge induced by the small size of the datasets.

4. Experimental Results

In order to evaluate the performance of the proposed approaches, we have conducted a series of experiments using the Interspeech 2018 ComParE challenge data (on three of the sub-challenges). The results are summarised in Table 3.

The best competition baseline results [3] were achieved using a fusion of multiple diverse classifiers. Since in this paper

Table 3: *Summary of the results. Unweighted average recall (UAR, %) is shown for the different methods and their aggregation on the three sub-challenges.*

Method	Atypical		Self-Ass.		Crying	
	Dev.	Test	Dev.	Test	LOSO	Test
GEWELM	56.2	36.3	69.0	49.5	72.7	–
CNN-GRU	38.3	40.4	–	–	77.4	70.7
Baseline (ComParE)	40.5	43.1	56.5	63.2	76.9	73.2
Baseline (CNN+LSTM)	41.8	28.0	49.7	46.6	–	63.5
Baseline (auDeep)	40.4	35.6	49.9	57.3	74.4	71.1
Baseline (Fusion)	–	43.4	–	66.0	–	74.6

we are interested in investigating the performance of individual methods, we omit the discussion of the comparison against fusion results (which are listed in Table 3 for completeness). Importantly, the best baseline results in Table 3 were not necessarily achieved in the same experiment, hence the variability between the *dev.* and *test* results is not completely reflected (see the complete breakdown in [3].

GEWELMS The performances of an ensemble of 2000 WELMs over a wide range of hyper-parameter setting are presented in Table 1. To that end, runs have been performed with a variable number of GMM Gaussians (ranging from 4 to 128) and a varying number of neurons (2,5,10 or 50) constituting the WELMs. Additionally, these experiments have not only been conducted using the *train* and the *dev.* set for training and testing purposes respectively, but the reverse has also been implemented. The average performance of the two different regimes is also displayed. We note that for the Atypical-Affect sub-challenge, MFCC based spectral features achieve an average T2D/D2T UAR = 53.0. This is better than the highest achieved UAR with ComParE spectral features, although PLP features achieve comparable performance with MFCCs. Furthermore, we note that performance generally decreases as the number of neurons are increased beyond $\#Neur. = 5$, however, we believe the grid search is not large enough to provide a definitive conclusion for this observation.

In Table 2, we provide a summary of experiments performed for the Self-Assessed Affect sub-challenge. Here, we note that ComParE spectral features with GEWELMS perform better than both MFCCs and PLPs, which is opposite to what was observed for the Atypical-Affect sub-challenge. In order to explore this further, we computed Fisher Vectors with spectral LLDs which are part of the extended Geneva Minimalistic Acoustic Parameter Set (GeMAPS) feature set [8]. We extended eGeMAPS spectral features by appending their velocity contours, and call this set of features as eGeMAPS+.

Our experiments show that with 'improved' Fisher Vectors [22], eGeMAPS achieved better performance than both MFCCs and PLPs, whereas for 'vanilla' Fisher Vectors [14] eGeMAPS+ features have a comparable performance to MFCCs and a slightly worse performance than PLP features. One can also note a trend where 'improved' Fisher Vector encoding [22] requires a smaller number of GMMs to achieve high classification accuracy with GEWELMS as compared to 'vanilla' Fisher Vector encoding [14].

Furthermore, we used a Support Vector Machine (SVM) classifier [39] with linear kernel to provide a baseline for comparing GEWELM classifier. The 'cost' parameter of linear SVM classifier was optimised over a logarithmically spaced grid with $C = \{10^2, 10^1, ...10^{-5}\}$. While it is clear from Table 2 that GEWELM classifiers perform much better that linear SVM, we also note that both ComParE spectral features and eGeMAPS+ features achieve better classification accuracy compared to MFCCs and PLPs. We intend to explore this observation further, beyond the ComParE 2018 sub-challenges.

In Table 3 we summarise results of our proposed approaches for ComParE 2018 sub-challenges. For the Atypical-affect sub-challenge, one can note that GEWELMs based classification of spectral features achieve better performance both deep learning based baseline approaches i.e. CNN+LSTM and auDeep, although it does not beat the baseline set by ComParE functionals.

GEWELMs also achieve better performance than the baseline CNN+LSTM approach for the Self-Assessed sub-challenge. However, we note that GEWELMs do have a tendency to overtrain, in spite of the T2D/D2T training/testing regularisation (see Section 3.2 for details).

CNN-GRU Our deep-learning based submission displays strong results on the two datasets it was applied on. In particular, our approach significantly outperforms all baseline deep-learning methods on the Atypical challenge, and on the Crying challenge outperforms all but one (auDeep in one of the configurations) of the baseline results. Our CNN-GRU architecture also favourably compares to the rest of the benchmark (non based on deep-learning). These results are even more relevant given that they are the results of single attempts on the test set, which is not the case for the benchmarks.

Therefore, we suggest that a combination of model ensembling, effective data augmentation, and multi-task learning is a viable angle of attack on the problem at hand, despite the problem of acute sample deficit. We further note, that the ensemble of CNN-GRUs served to alleviate the variability between the performance on the *dev.* vs *test* sets, a problem from which many of the baseline methods suffer.

5. Conclusions

In this paper, we have investigated two approaches to paralinguistic analysis in response to the Interspeech 2018 ComParE challenge. First, we proposed a novel technique, which we termed Greedy Ensemble of Weighted Extreme Learning Machines (GEWELM), that combines the well-known training efficiency of Extreme Learning Machines (ELM), with good classification performance. This combination of speed and accuracy, we speculate, will be especially important in real-time scenarios, such as screening.

Further, we have demonstrated that despite severe deficit of training data, a problem common to many datasets in social signal computing and paralinguistics, an effective deep-learning solution to the task at hand is viable. To this end, we proposed an effective combination of techniques (multi-task learning, model ensembling, and domain-specific data augmentation) that yielded very good performance (in many cases exceeding state-of-the-art).

Acknowledgements We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

6. References

[1] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language — State-of-the-art and the challenge," *Comput. Speech Lang.*, vol. 27, no. 1, pp. 4–39, 2013.

[2] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social Signals, Their Function, and Automatic Analysis: A Survey," in *ACM Int. Conf. Multimodal Interfaces*, 2008, pp. 61–68.

[3] B. W. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & Self-Assessed Affect, Crying & Heart Beats," in *INTERSPEECH*, 2018, pp. 1–5.

[4] B. S. Simone Hantke, Hesam Sagha, Nicholas Cummins, "Emotional Speech of Mentally and Physically Disabled Individuals: Introducing the EmotAsS Database and First Findings," in *INTERSPEECH*, 2017, pp. 3137–3141.

[5] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.

[6] P. B. Marschik, F. B. Pokorny, R. Peharz, D. Zhang, J. O'Muircheartaigh, H. Roeyers, S. Bölte, A. J. Spittle, B. Urlesberger, B. Schuller, L. Poustka, S. Ozonoff, F. Pernkopf, T. Pock, K. Tammimies, C. Enzinger, M. Krieber, I. Tomantschger, K. D. Bartl-Pokorny, J. Sigafoos, L. Roche, G. Esposito, M. Gugatschka, K. Nielsen-Saines, C. Einspieler, and W. E. Kaufmann, "A Novel Way to Measure and Predict Development: A Heuristic Approach to Facilitate the Early Detection of Neurodevelopmental Disorders," *Curr. Neurol. Neurosci. Rep.*, vol. 17, no. 5, 2017.

[7] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Commun.*, vol. 71, pp. 10–49, 2015.

[8] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, 2016.

[9] X. Zhou, X. Zhuang, H. Tang, M. Hasegawa-Johnson, and T. S. Huang, "Novel Gaussianized vector representation for improved natural scene categorization," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 702–708, 2010.

[10] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3304–3311.

[11] M. Karafiát, L. Burget, P. Matějka, O. Glembek, and J. Černocký, "iVector-based discriminative adaptation for automatic speech recognition," in *IEEE Work. Autom. Speech Recognit. Underst.*, 2011, pp. 152–157.

[12] Z. S. Shah, K. Sidorov, and D. Marshall, "Psychomotor cues for depression screening," in *IEEE Int. Conf. Digit. Signal Process.*, 2017, pp. 1–5.

[13] Z. S. Syed, K. Sidorov, and D. Marshall, "Depression Severity Prediction Based on Biomarkers of Psychomotor Retardation," in *Audio/Visual Emot. Recognit. Chall.*, 2017.

[14] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2007.

[15] V. Jain, J. L. Crowley, A. K. Dey, and A. Lux, "Depression Estimation Using Audiovisual Features and Fisher Vector Encoding," in *Audio/Visual Emot. Recognit. Chall.*, 2014, pp. 87–91.

[16] A. Dhall and R. Goecke, "A temporally piece-wise fisher vector approach for depression analysis," in *Int. Conf. Affect. Comput. Intell. Interact.*, 2015, pp. 255–259.

[17] H. Kaya, F. Gürpinar, S. Afshar, and A. A. Salah, "Contrasting and Combining Least Squares Based Learners for Emotion Recognition in the Wild," in *ACMI*, ser. ICMI '15. New York, NY, USA: ACM, 2015, pp. 459–466.

[18] H. Kaya, A. A. Karpov, and A. A. Salah, "Fisher Vectors with Cascaded Normalization for Paralinguistic Analysis," in *INTERSPEECH 2015*, 2015, pp. 909–913.

[19] H. Kaya and A. A. Karpov, "Fusing Acoustic Feature Representations for Computational Paralinguistics Tasks," in *INTERSPEECH 2016*, 2016, pp. 2046–2050.

[20] ——, "Introducing weighted kernel classifiers for handling imbalanced paralinguistic corpora: Snoring, addressee and cold," in *INTERSPEECH*, vol. 2017-Augus, 2017, pp. 3527–3531.

[21] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, 2013.

[22] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Lect. Notes Comput. Sci.*, vol. 6314, no. Part 4, 2010, pp. 143–156.

[23] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *ACM MM 2013*, pp. 835–838.

[24] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, 1995.

[25] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," 2008.

[26] G.-b. Huang, Q.-y. Zhu, and C.-k. Siew, "Extreme Learning Machine : A New Learning Scheme of Feedforward Neural Networks," *IEEE Int. Jt. Conf. Neural Networks*, vol. 2, pp. 985–990, 2004.

[27] P. Guo, P. C. L. Chen, and Y. Sun, "An Exact Supervised Learning for a Three-Layer Supervised Neural Network," in *Int. Conf. neural Inf. Process.*, 1995, pp. 1041–1044.

[28] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," in *Conf. Mod. Anal. Probab. (New Haven, Conn., 1982) Contemp. Math. 26. Provid. RI Am. Math. Soc.*, 1984, pp. 189–206.

[29] T. M. T. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *Electron. Comput. IEEE Trans.*, pp. 326–334, 1965.

[30] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and Facial Biomarkers of Depression Based on Motor Incoordination and Timing," in *Int. Work. Audio/Visual Emot. Chall.*, 2014, pp. 65–72.

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

[32] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, and G. Wang, "Recent advances in convolutional neural networks," *CoRR*, vol. abs/1512.07108, 2015.

[33] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[34] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[35] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[37] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8599–8603.

[38] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.

[39] C. Cortes and V. Vapnik, "Support-Vector Networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.