Interspeech 2018
2-6 September 2018, Hyderabad

Implementing Fusion Techniques for the Classification of Paralinguistic Information

Bogdan Vlasenko[1], *Jilt Sebastian*[1,2,3], *D S Pavan Kumar*[1,3], *Mathew Magimai.-Doss*[1]

[1]Idiap Research Institute, CH-1920 Martigny, Switzerland
[2]Indian Institute of Technology Madras, India
[3]École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

bogdan.vlasenko@idiap.ch, jiltsebastian@gmail.com, dspavankumar@gmail.com, mathew.magimaidoss@idiap.ch

Abstract

This work tests several classification techniques and acoustic features and further combines them using late fusion to classify paralinguistic information for the ComParE 2018 challenge. We use Multiple Linear Regression (MLR) with Ordinary Least Squares (OLS) analysis to select the most informative features for Self-Assessed Affect (SSA) sub-Challenge. We also propose to use raw-waveform convolutional neural networks (CNN) in the context of three paralinguistic sub-challenges. By using combined evaluation split for estimating codebook, we obtain better representation for Bag-of-Audio-Words approach. We preprocess the speech to vocalized segments to improve classification performance. For fusion of our leading classification techniques, we use weighted late fusion approach applied for confidence scores. We use two mismatched evaluation phases by exchanging the training and development sets, and this estimates the optimal fusion weight. Weighted late fusion provides better performance on development sets in comparison with baseline techniques. Raw-waveform techniques perform comparable to the baseline.

Index Terms: fusion, feature selection, Multiple Linear Regression, raw-waveform CNN

1. Introduction

Computational paralinguistics covers various non-verbal information channels of speech and other types of vocalizations, and has developed rapidly over the last decade. The ComParE 2018 challenge comprises of four sub-challenges for predicting four basic emotions from the speech of handicapped subjects (*Atypical effect*), valence scores given by speakers themselves (*Self-Assessed Affect* (SAA)), three types of infant vocalizations (*Crying*) and three different types of heart beats (*Heart Beats*) [1]. The organizers of the challenge provide a baseline system composed of the state-of-the-art acoustic features and commonly used classification techniques. Feature extraction and machine learning techniques can be reproduced via freely available, open source tools [2,3]. It is difficult to outperform the baseline results owing to the complex fusion techniques used. A large variety of paralinguistic tasks are investigated in the recent years including but not limited to speaker traits [4], identifying child directed speech, cold and snoring identification [5], social signals, conflicts, emotions and autism [6].

As the paralinguistic applications tested varies from year to year, there has always been sufficient interest in feature selection and fusion techniques. Results presented in [7] show that combination of frame- and utterance level-analysis could significantly improve emotion recognition performance. Multi-modal decision level fusion is proposed in [8] for emotion recognition in the Wild [9]. Results presented in [10–12] show that the applied fusion methods improve the performance of the standalone detectors and provide systems capable of outperforming the baseline systems.

Feature selection for various models have been proposed in the past. This includes feature selection in Support Vector Machine (SVM) classifier [13] for emotion recognition, ensemble feature selection in [14] for model adaptation, and recent openSMILE [3]. Canonical correlation analysis (CCA) is employed in [15] for selecting apt features from a set of speech features for depression recognition. Acoustic, linguistic and psycholinguistic features are employed in [16] for learning the personality traits from the spoken data. Various feature selection methodologies in high-dimensional classification are presented in [17] where, speaker likability, intelligibility and personality traits are considered for classification. In [18], features extracted from a deep neural network are used for robust identification of child-directed speech, cold and snoring identification [5].

Year to year, the number of subjects/instances selected for the challenges has increased from few hundreds to several thousand. Increasing the number of acoustic features opens room for machine learning research, in optimizing feature set and improvement in robustness of classification techniques for high-dimensional fixed length turn-level features. Results presented during last ComParE challenges highlighted the importance of feature selection in handling high-dimensional paralinguistic datasets [19–21].

Recently, end-to-end systems are also used for classifying paralinguistic information. It avoids the use of hand-crafted features and allows the model-itself to learn most suitable feature representation for the given task [5], [22]. These applications are inspired by use of raw-waveform methods in ASR and speaker recognition tasks. We propose a raw-waveform CNN for three of the paralingustic sub challenges. We also compare the proposed approach with the corresponding end-to-end baseline. As the confidence scores are not available for the baseline system [23], it is not possible to use it for fusion techniques.

The contribution of this paper is threefold. First, we propose to use Voice Activity Detector (VAD) for pruning irrelevant information located in silent segments of challenge datasets. Second, we propose to combine splits for estimation of codebooks for Bag-of-Audio Words approach introduced in [24]. Third, we propose a raw-waveform CNN which has a comparable performance with the end-to-end baseline and also provide reliable confidence scores at frame-level and turn-level for late fusion. We postulate that combined training, development and test splits for specification codebook representatives should improve classification

10.21437/Interspeech.2018-2360

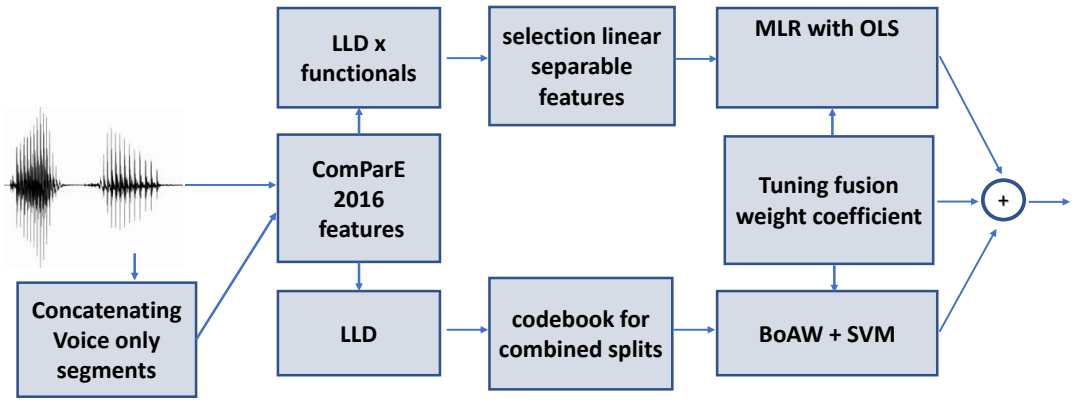

Figure 1: *Processing flow chart for Self-Assessed Affect sub-challenge.*

performance.

It is important to note that, similar to previous year, the baseline scores are obtained selectively from 23 test set evaluations (17 for individual systems and $3 + 3$ evaluations with fused systems): using n-best techniques for late fusion [1]. This makes the test baselines hard to outperform, as the challenge participants have a maximum of 5 submission options per sub-challenge. All presented baseline results were obtained with majority or confidence score based fusion. Also, results presented in Table 2 of [1] shows that in some cases we have a different optimal parameters for similar classification techniques evaluated on the development and test data. Hence, we decided to use late fusion for best performing classification techniques. For fusion we decided to use late pairwise fusion with weighted scores. Also, we decided to simulate direct and reverse evaluation scenarios by switching training and development sets.

2. General Framework

We present the general framework used for all the sub-challenges in this section. We remove the silence using a suitable VAD in the first step. Vocalized segments are used for acoustic feature selection using Low-Level-Descriptors (LLD) and functionals. We employed Multiple Linear regression with Ordinary Least Squares analysis, Bag-of-Audio-Words approach [24] with codebooks estimated on the combined splits, and CNN technique applied to raw waveforms for classification. Classification techniques applied varies for each sub-challenge. We finally fuse the scores from different best individual classifiers.

2.1. Speech Preprocessing

The preprocessing step consists of pruning irrelevant information using a suitable Voice Activity Detector (VAD). We used a Gaussian mixture model (GMM)-based VAD [25] which is more effective than simple energy-based counterparts when using with varying background noise levels. This converts the original signals to segments without silence. Finally, all non-silent segments were concatenated.

2.2. Acoustic Feature Extraction

For extraction of acoustic features, we used the openSMILE toolkit [26] with feature set configuration presented for Interspeech 2016 ComParE Challenge [27]. The same feature set has been used in the five previous editions of the Interspeech ComParE challenges as well. Original feature set contains 6373 static features resulting from the computation of various functionals over low-level descriptor (LLD) contours [28]. 65 LLDs and corresponding 65 delta coefficients are extracted for BoAW representation.

2.3. Modified Bag-of-Audio Words (BoAW) approach

In this work, we decided to used slightly modified BoAW approach for our experiments. Instead of using baseline configuration we decided to combine training, development and test split for better coverage of the codebook. Also, we skipped standardization step introduced in baseline system. We also used a concept of two codebooks: the first codebook is estimated for the 65 LLDs extracted from the ComParE feature set and the second codebook for the 65 deltas of these LLDs.

3. Methodology for SSA sub challenge

The dataset for Self-Assessed Affect sub-challenge comprises four times five-minute sessions from around 150 individuals. They had to speak spontaneously twice about a negative, and twice about positive experiences in their life. Before and after, the speaker reports a self-assessment of their own state of mind (Arousal and Valence on a ten-point Likert scale). In this challenge, the task is to determine the emotion of individuals as was assessed by themselves. We proposed to use the pipeline shown in Figure 1 for this challenge.

The speech signals are preprocessed using VAD to select vocalized segments. Feature extraction is performed by selecting LLDs and functionals as mentioned in Section 2. We use two classification techniques in this challenge; BoAW with SVM (Section 2) and multiple linear regression with ordinary least squares analysis.

Classification performance on the development set for BoAW approach with applied modifications is presented in Table 1. Confidence scores obtained with modified BoAW technique are marked as $P_{II}(x_i)$ where x_i is one of the possible classes.

3.1. Multiple Linear regression

For the selected acoustic features we implement least square analysis and estimate Multiple Linear Regression (MLR) coefficients. 28 linear separable features, presented in Listing 1, were selected from 6373 turn-level features for SSA sub-challenge task. Selected features contains various spectral features.

Classification performance on development set for MLR approach with selected features is presented in Table 1. Confidence scores for obtained with MLR technique are marked with $P_I(x_i)$ where x_i is one of possible classes.

```
audSpec_Rfilt_sma[2]_percentile99.0
audSpec_Rfilt_sma[5]_lpc3
audSpec_Rfilt_sma[12]_lpc3
audSpec_Rfilt_sma[22]_upleveltime90
pcm_fftMag_spectralEntropy_sma_lpc4
pcm_fftMag_spectralVariance_sma_skewness
pcm_fftMag_spectralKurtosis_sma_quartile1
mfcc_sma[2]_pctlrange0-1
mfcc_sma[4]_kurtosis, mfcc_sma[8]_range
audSpec_Rfilt_sma_de[3]_lpc1
audSpec_Rfilt_sma_de[5]_pctlrange0-1
audSpec_Rfilt_sma_de[12]_kurtosis
audSpec_Rfilt_sma_de[20]_quartile2
mfcc_sma_de[6]_lpc0, mfcc_sma_de[11]_lpgain
mfcc_sma_de[14]_maxPos
shimmerLocal_sma_lpc4, jitterDDP_sma_de_amean
pcm_RMSenergy_sma_stddevRisingSlope
audSpec_Rfilt_sma[6]_peakMeanMeanDist
mfcc_sma[5]_meanPeakDist
mfcc_sma[10]_peakMeanRel
mfcc_sma[12]_stddevRisingSlope
mfcc_sma[13]_qregc2, mfcc_sma[3]_flatness
audSpec_Rfilt_sma_de[7]_minRangeRel
pcm_fftMag_spectralFlux_sma_de_peakRangeAbs
```

Listing 1: *Selected features for Self-Assest Affect*

3.2. Late fusion

Weighted late fusion is applied for merging the confidence scores obtained with MLR and modified BoAW approaches. Equation 1 represents weighted fusion approach applied for fusion MLR (I) and BoAW (II) approaches.

$$P_{fusion}(x_i) = \omega P_I(x_i) + (1-\omega)P_{II}(x_i) \quad \forall x_i \in \mathbf{X} \quad (1)$$

where $\mathbf{X}$ is a set of all possible classes, $0 < \omega < 1$. Class $\hat{x_k}$ with highest probability $P_{fusion}(\hat{x_k})$ is selected as recognized.

4. Raw-waveform CNN

Raw waveform methods have recently been exploited for speech processing applications with its inherent ability to extract features which are specific to the application. As a part of the challenge, we propose a raw-waveform CNN for Crying, Heartbeat and Self-assessed Assessment sub challenges and compare it with raw-waveform methods provided as the baselines. This is motivated by the use of raw-CNN for applications such as automatic speech recognition [29] and voice presentation attack detection [30].

The Crying sub challenge dataset comprises of more than 5000 vocalisations of 20 healthy infants (10 females). This is done under a study on postnatal neuro-functional and neurobehavioural changes and adaptations. The task is to automatically classify the three classes: (i) neutral/positive mood vocalisations, (ii) fussing vocalisations, and (iii) crying vocalisations. The heart-beat sub challenge dataset consists of heart sounds gathered from 170 (55 female, age ranges 21 – 88 years) subjects with various ages and health conditions. There are three classes to be recognised for the data: normal, mild, and moderate/severe as diagnosed by physicians specialized in heart disease. The task is to classify the sounds into: normal, mild, and moderate/severe.

The proposed network consists of two-to-three convolutional layers depending on the application and a hidden layer with relu activation function. This architecture of convolution layers followed by a multilayer perceptron is motivated by the success on various tasks, such as speech recognition [29, 31], presentation attack detection [30] and speaker recognition [32]. The output layer performs a softmax operation to obtain frame-based posteriors.

Table 1: *UAR [%] rates for best baseline system and our approach. Results obtained on development set for Self Assessed Affect sub-challenge.*

Case	System	UAR
Direct	Baseline	56.7
Direct	MLR	63.3
Reverse	MLR	60.1
Direct	BoAW(GLOB)+SVM	62.1
Reverse	BoAW(GLOB)+SVM	53.3

We provide the architecture of Baby Crying sub challenge for general understanding of the proposed approach. We use 400ms window length ($\mathbf{W_{len}}$) with a 40 ms shift ($\mathbf{W_{shift}}$) for making frame level decisions about the underlying classes. First convolutional layer has 80 filters ($\mathbf{N_{filters1}}$) with a filter width of 30 samples ($\mathbf{N_{seq1}}$) and is shifted by 5 samples ($\mathbf{N_{shift1}}$) to have 75 such chunks over the window length. We use a max pooling size of 3 ($\mathbf{mp_i}, \mathbf{i} = \mathbf{1}..\mathbf{N}$) for all convolutional layers($\mathbf{N}$). Pooling is followed by non-linear activation and we choose relu based on its applications in ASR and ASV Spoof Detection tasks. Second convolutional layer has 60 filters ($N_{filters2}$), filter width of 7 (N_{seq2}) and a filter shift of 1 (N_{shift2}). Hence, the total number of samples actually considered in filtering is $7 * 75 = 525$. Third convolutional layer (if exists) also has the same configurations as the second one. The parameters such as window length, number of filters, filter width and stride are selected based on heuristics and assumptions about the task. For example, a very small $\mathbf{N_{seq1}}$ and ($\mathbf{N_{shift1}}$) in Crying task is justified by the need of analysis at micro-level, so as to cover at least two pitch cycles of a baby, which is typically of $1-2$ ms. The pitch plays a very important role in their voice as their vocal tract is not fully developed. A large $\mathbf{W_{len}}$ captures the contextual information and it avoids the need of recurrent units.

In the challenge paper [1], development data is used for evaluating the top performing model and the testing is also performed with the same data. This could lead to a bias towards the development data and increase the performance gap with the testing data. As compared to the experiments reported in the challenge paper, we follow a different procedure to train the proposed and baseline networks. 90% of the training examples are used for training and rest is used for validation. Training instances are repeated in the training set so as to have an even distribution of instances among classes. The development data provided is used for testing the models. As the development data is unseen during the training, testing the model with this data is justified. The experiments are performed with 2 LSTM layers as it has better unweighted average recall (UAR) in the development set [1]. The baseline model is trained for 10 epochs and top performing model is evaluated to obtain the performance. The CNN model is trained with an initial learning rate (LR) of 0.1. The LR is halved whenever the validation loss stagnates between successive epochs. Training is terminated when the LR drops below 10^{-5} and the final model is used for evaluation.

5. Results

During first experiment phase, we used development data for tuning classification parameters of MLR and modified BoAW techniques. During the second phase, we used development data for the estimation of late fusion weight. Afterwards, optimized architecture was used for training and fusing models on combined training and development sets. Finally, fused models were evaluated on test data.

5.1. Evaluation of development data

We used two different evaluation setups for SSA sub-challenge: we trained models on the training set and evaluated on the development set (direct case), and trained models on the development set and evaluated on the the training set (reverse case).

One could see, that proposed MLR and BoAW with global codebooks outperformed results obtained with baseline system on development set

Afterwards, by using technique described in Section 3.2 we were estimating optimal ω value for late fusion of MLR and BoAW. Results obtained with direct and reverse evaluation cases were used. UAR as a function of ω are presented in Figure 2.

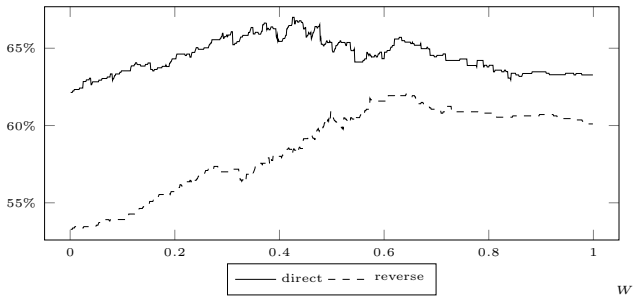

Figure 2: *UAR as a function of linear regression weight ω. Direct case: training on training data, evaluation on development. Reverse order - training on development set and evaluation on training set.*

As one could see from Figure 2 curves for the direct and the reverse case have different maximum values. With $\omega = 0.426$ we can improve classification performance for the direct case from $UAR_{MLR} = 63.3\%$ and $UAR_{BoAW} = 62.1\%$ to $UAR_{fusion} = 67.0\%$. In the case of reverse case, on optimal fusion result was obtained with $\omega = 0.642$.

5.2. Evaluation of test data

By averaging UAR curves for the direct and the reverse case we found that an optimal performance for fused technique can be obtained $\omega = 0.63$. Hence, for late fusion of MLR and BoAW techniques trained on combined training and development sets we used $\omega = 0.63$. Results obtained on test sets with proposed technique is presented in Table 2 The corresponding confusion matrix is given in Table 3, where we observe the highest confusion between the low (class l) and middle level (class m) of valence. As one could see recall rates for m and h valence classes significantly outperform UAR reported for best baseline system. For final test evaluations we are planing to improve the results for valence class l.

Table 2: *UAR [%] rates for best baseline system (Majority vote for 3 best) and our approach with tunned fusion weight. Results obtained of test set for Self Assessed Affect sub-challenge.*

System	UAR
Baseline [1]	66.0
Proposed	63.5

Table 3: *Confusion Matrix for evaluation on test set for Self Assessed Affect Sub Challenge.*

System	**l [%]**	**m [%]**	**h [%]**
l	21.3	58.7	20.0
m	7.4	81.6	11.0
h	0.7	11.8	87.5

5.3. Results on raw-waveform CNN

We compare the performance of proposed approach with raw waveform-based END2YOU network [23] for three sub challenges. In terms of UAR, proposed raw-CNN approach has a comparable performance with baseline for Crying task, whereas it outperforms the baseline for other two tasks, see Table 4. This is achieved in spite of being a less complex network. For example, baseline system has $2,115,035$ parameters in total as compared to $1,668,539$ parameters of the proposed model for crying sub challenge. This is reflected in both training and testing times. The LSTM layers of the baseline causes delay in training the network and the models validation performance need not necessarily have a convergence with respect to the epochs.

Table 4: *UAR [%] rates for End2End methods trained and evaluated on raw-waveform. Evaluated on development sets.*

System	**Crying**	**Heart beat**	**Self-Assessed**
Baseline [1]	73.7	25.6	43.0
Proposed	71.1	44.8	49.2

6. Conclusions and Future work

In this work, we propose to use MLR for selecting acoustic features, and we offer to use advanced technology for BoAW codebook creation. We showed that implementing weighted late fusion for MLR and BoAW could significantly improve performance on the development set. We simulated direct and reverse evaluation cases for tuning the late fusion weight parameter for evaluating on the test set. The preliminary results on SSA development and test sets indicate that presented techniques are effective in the classification of Self-Assessed Affect. We also proposed a raw-waveform CNN for three sub-challenges. Raw-waveform techniques perform comparable to the baseline for Crying sub challenge and outperform the baseline results for SSA and Heart Beats sub challenge.

7. Acknowledgment

This work was partly supported by the IdeARK funded project COBALT, the Swiss Government Excellence Scholarship Project with ESKAS No: 2017.0575 and the HASLER Foundation project FLOSS.

8. References

[1] B. W. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. Pokorny, F. 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 *Proceedings of the Interspeech 2018*, Hyderabad, India, 2018.

[2] 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.

[3] 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*. Barcelona, Spain: ACM, 2013, pp. 835–838.

[4] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet *et al.*, "A survey on perceived speaker traits: personality, likability, pathology, and the first challenge," *Computer Speech & Language*, vol. 29, no. 1, pp. 100–131, 2015.

[5] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom *et al.*, "The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Proceedings of the Interspeech 2017*, Stockholm, Sweeden, 2017, pp. 3442–3446.

[6] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings of the Interspeech 2013*, Lyon, France, 2013.

[7] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Combining Frame and Turn-Level Information for Robust Recognition of Emotions within Speech," in *Proceedings of the Interspeech 2007*, Antwerp, Belgium, August 2007, pp. 2249–2252.

[8] F. Ringeval, S. Amiriparian, F. Eyben, K. Scherer, and B. Schuller, "Emotion recognition in the wild: Incorporating voice and lip activity in multimodal decision-level fusion," in *Proceedings of the 16th International Conference on Multimodal Interaction*. Orlando, USA: ACM, 2014, pp. 473–480.

[9] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon, "Emotion recognition in the wild challenge 2013," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. Barcelona, Spain: ACM, 2013, pp. 509–516.

[10] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing," in *Proceedings of the ACII 2007*, A. Paiva, R. W. Picard, and R. Prada, Eds. Berlin/Heidelberg: Springer, 2007, vol. 4738/2007, pp. 139–147.

[11] D. Tavarez, X. Sarasola, A. Alonso, J. Sanchez, L. Serrano, E. Navas, and I. Hernáez, "Exploring fusion methods and feature space for the classification of paralinguistic information," in *Proceedings of the Interspeech 2017*, Stockholm, Sweeden, 2017, pp. 3517–3521.

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

[13] I. Trabelsi and M. S. Bouhlel, "Feature selection for gumi kernel-based svm in speech emotion recognition," in *Artificial intelligence: concepts, methodologies, tools, and applications*. IGI Global, 2017, pp. 941–953.

[14] M. Abdelwahab and C. Busso, "Ensemble feature selection for domain adaptation in speech emotion recognition," in *Proceedings of the ICASSP 2017*. New Orleans, USA: IEEE, 2017, pp. 5000–5004.

[15] H. Kaya, F. Eyben, A. A. Salah, and B. Schuller, "CCA based feature selection with application to continuous depression recognition from acoustic speech features," in *Proceedings of the ICASSP 2014*. Florence, Italy: IEEE, 2014, pp. 3729–3733.

[16] F. Alam and G. Riccardi, "Fusion of acoustic, linguistic and psycholinguistic features for speaker personality traits recognition," in *Proceedings of the ICASSP 2014*. Florence, Italy: IEEE, 2014, pp. 955–959.

[17] J. Pohjalainen, O. Räsänen, and S. Kadioglu, "Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits," *Computer Speech & Language*, vol. 29, no. 1, pp. 145–171, 2015.

[18] G. Gosztolya, R. Busa-Fekete, T. Grósz, and L. Tóth, "DNN-based Feature Extraction and Classifier Combination for Child-Directed Speech, Cold and Snoring Identification," in *Proceedings of the Interspeech 2017*, Stockholm, Sweeden, 2017, pp. 3522–3526.

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

[20] H. Kaya, T. Özkaptan, A. A. Salah, and F. Gürgen, "Random discriminative projection based feature selection with application to conflict recognition," *IEEE Signal Processing Letters*, vol. 22, no. 6, pp. 671–675, 2015.

[21] H. Kaya, T. Özkaptan, A. A. Salah, and S. F. Gürgen, "Canonical correlation analysis and local fisher discriminant analysis based multi-view acoustic feature reduction for physical load prediction," in *Proceedings of the Interspeech 2014*, Singapore, 2014.

[22] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proccedings of the ICASSP 2016*. Shanghai, China: IEEE, 2016, pp. 5200–5204.

[23] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[24] M. Schmitt and B. Schuller, "openXBOW—Introducing the Passau open-source crossmodal bag-of-words toolkit," *Journal of Machine Learning Research*, vol. 18, no. 96, pp. 1–5, 2017.

[25] [Online]. Available: https://webrtc.org

[26] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. Florence, Italy: ACM, 2010, pp. 1459–1462.

[27] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings of the Interspeech 2013*, Lyon, France, 2013.

[28] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: what speech, music, and sound have in common," *Frontiers in psychology*, vol. 4, p. 292, 2013.

[29] D. Palaz, R. Collobert, and M. M. Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," *arXiv preprint arXiv:1304.1018*, 2013.

[30] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, "End-to-end convolutional neural network-based voice presentation attack detection," in *IEEE IAPR International Joint Conference on Biometrics (IJCB)*, 2017.

[31] D. Palaz, M. Magimai.-Doss, and R. Collobert, "End-to-End Acoustic Modeling using Convolutional Neural Networks for Automatic Speech Recognition," in *Idiap*, no. EPFL-REPORT-219847, 2016.

[32] H. Muckenhirn, M. Magimai.-Doss, and S. Marcel, "Towards directly modeling raw speech signal for speaker verification using CNNs," in *Proceedings of the ICASSP 2018*, Calgary, Canada, 2018.