INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Identifying Conflict Escalation and Primates by Using Ensemble X-vectors and Fisher Vector Features

José Vicente Egas-López[1], *Mercedes Vetrab*[1], *László Tóth*[1], *Gábor Gosztolya*[1,2]

[1]Institute of Informatics, University of Szeged, Hungary
[2]MTA-SZTE Research Group on Artificial Intelligence, ELRN, Szeged, Hungary
{ egasj, vetrabm, tothl, ggabor } @ inf.u-szeged.hu

Abstract

Computational paralinguistics is concerned with the automatic identification of non-verbal information in human speech. The Interspeech ComParE challenge features new paralinguistic tasks each year; this time, among others, a cross-corpus conflict escalation task and the identification of primates based solely on audio are the actual problems set. In our entry to ComParE 2021, we utilize x-vectors and Fisher vectors as features. To improve the robustness of the predictions, we also experiment with building an ensemble of classifiers from the x-vectors. Lastly, we exploit the fact that the Escalation Sub-Challenge is a conflict detection task, and incorporate the SSPNet Conflict Corpus in our training workflow. Using these approaches, at the time of writing, we had already surpassed the official Challenge baselines on both tasks, which demonstrates the efficiency of the employed techniques.

Index Terms: human-computer interaction, computational paralinguistics, x-vectors, Fisher vectors, ensemble learning

1. Introduction

Speech is the primary communication channel of humans. Evidently, human speech not only encodes the actual words spoken, but it incorporates a wide range of non-verbal content as well, transmitting a variety of information about the physical and mental state of the speaker. In the past decade, a subfield has grown around the automatic identification of this 'paralinguistic' (that is, 'beyond linguistic') aspect of human speech. This area covers a wide variety of tasks from emotion detection [1, 2] to determining the alcohol intoxication of the speaker [3], estimating the sleepiness of the subject [4], screening mental or physical illnesses like depression [5], Parkinson's disease [6] and Alzheimer's disease [7].

Perhaps the most important task of computational paralinguistics is the choice of the features extracted from the audio utterances. In general, one might choose from two different approaches: to apply general feature extraction methods, which can be expected to work on a wide variety of tasks, or to incorporate external knowledge and develop task-specific features. A good example for the former, general feature type is the 'ComParE functionals', introduced by Schuller et al. [8], consisting of applying utterance-level statistical functions (e.g. mean, standard deviation, percentiles) for specific frame-level attributes. Other general feature sets include Bag-of-Audio-Words (BoAW, [9]) and Fisher vectors (FV, [10]), both of which were applied in several studies [2, 11, 12, 13].

The so-called i-vectors [14] were originally developed for speaker recognition, but they turned out to be suitable feature extractors for paralinguistic tasks as well (see e.g. [15, 16, 17, 18]). Recently, x-vectors [19] have become state-of-the-art in speaker recognition, and, similarly to i-vectors, they have been applied in a handful of studies like paralinguistic and medical feature extractors from voice [4, 20, 21, 22].

X-vectors are basically neural networks which map a variable-length speech utterance into a fixed-dimensional feature space; the extracted features (the x-vectors) are utterance-level embeddings, i.e. they are the activations of a specific layer of the DNN. But, similarly to other neural networks, they might prove to be sensitive to several training meta-parameters such as the number of hidden layers and neurons, to the learning rate or the number of training epochs. Furthermore, since when trained from scratch, the weights of a neural network are usually initialized randomly, they can even prove to be sensitive to the random seed of this weight initialization step. This might also hold for x-vectors, even if they are used only for feature extraction.

Our solution to this, on which we base our entry for the Escalation and the Primates sub-challenges of the 2021 Interspeech Computational Paralinguistic Challenge (ComParE, [23]), is the *ensemble x-vector* technique. That is, to reduce the stochasticity of the features extracted by the x-vectors, we repeat the x-vector DNN training step several times. Then, after extracting the x-vector features, we train independent classifier models on each of them, and simply average out the predictions (i.e. the posterior estimates). This process is supposed to reduce the stochasticity of the predictions themselves; that is, it should improve their robustness. Furthermore, as it turned out, it might even provide improvements in the classification performance over the individual x-vector-based models. Besides x-vectors, we also experiment with Fisher vectors (FV, [10]); and, similarly to our entries to the previous ComParE challenges (e.g. [24, 25]), we apply a fusion of the predictions. Then, in the last part of our study, we will focus on the Escalation sub-challenge by exploiting that it is essentially a conflict detection task, which allows us to utilize the recordings of the public SSPNet Conflict corpus [26] in our prediction workflow.

Note that, following the Challenge guidelines (see [23]), we omit the detailed description of the tasks, datasets and the method of evaluation, and focus on the techniques we applied. We shall treat both sub-challenges (i.e. Escalation and Primates) in the same way, measuring the performance via the Unweighted Average Recall (UAR) metric.

2. X-vector Embeddings

The x-vector approach is a neural network-based feature extraction method that provides fixed-dimensional embeddings for variable-length utterances. Basically, it is a feed-forward Deep Neural Network (DNN) that computes such embeddings.

Fig 1 shows the structure of the DNN. The lower, *frame-level* layers have a time-delay architecture. After the frame-level layers, the *stats pooling* layer gets the frame-level acti-

http://dx.doi.org/10.21437/Interspeech.2021-1173

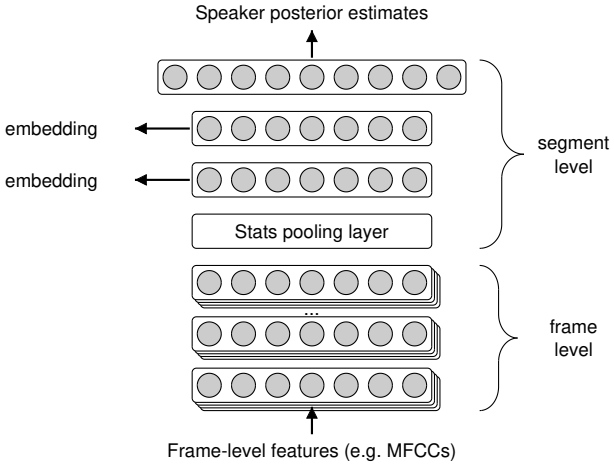

Figure 1: *The structure of the x-vector extractor neural network.*

vations of the last frame-level layer, aggregates over the input segment, and computes the mean and the standard deviation. These statistics are concatenated and used as input for the next, *segment-level* layer, which is followed by one (or possibly more) additional segment-level layers. The *x-vectors* embeddings can be extracted from any of *segment* layers [27], but from experience, embeddings from the *segment6* layer give a better performance than those from *segment7* [19]. Instead of predicting frames, the DNN is trained to predict speakers from variable-length utterances. Therefore, the output softmax layer has as many neurons as there are speakers in the training set. Notice that, to calculate the embeddings, this output layer is not required any more, so it can be discarded after training.

2.1. Ensemble X-vectors

The basic principle of ensemble learning is to train several different, but similar machine learning models, and combine their outputs in some way. In this study we build an ensemble based on the x-vector feature extractors. That is, we propose training several x-vector neural network models on the same data, but each time using a different random seed during random DNN weight initialization. By calculating the embeddings for each of them, we get a number of different representations of the same training data. Although in theory concatenating these feature vectors and training only one classifier model might lead to a more robust performance than relying on any of the individual representations, we would end up with an unfeasibly large feature vector. Therefore we chose to train separate machine learning (e.g. SVM) models on these x-vector representations in the next step. To make the predictions more robust (and thus, making hyperparameter selection more reliable), we suggest simply averaging out the prediction scores got after evaluation in an unweighted manner. Formally, we calculate the posterior estimate provided by the ensemble model as

$$P_e(c_i|X) = \frac{1}{m}\sum_{j=1}^{m} P_j(c_i|X) = \frac{1}{m}\sum_{j=1}^{m} P_j(c_i|H^j), \quad (1)$$

where c_i denotes the ith class ($1 \leq i \leq K$), X is the frame-level feature sequence of the actual utterance, H^j is the representation of X calculated by the jth x-vector extractor DNN, and the P_j value is the individual posterior estimate provided by the jth SVM model.

3. Fisher Vector Representation

The basic idea of the Fisher Vector (FV) representation, adapted to audio processing (see e.g. [13, 2]), is to take the frame-level feature vectors of some corpus and model their distribution by a probability density function $p(X|\Theta)$, Θ being the parameter vector of the model. For example, when using Gaussian Mixture Models with a diagonal covariance matrix, Θ will correspond to the priors, and the mean and standard deviation vectors of the components. The Fisher score describes X by the gradient G_Θ^X of the log-likelihood function, i.e.

$$G_\Theta^X = \frac{1}{T}\nabla_\Theta \log p(X|\Theta). \quad (2)$$

This gradient function describes the direction in which the model parameters (i.e. Θ) should be modified to best fit the data. The Fisher kernel between the frame-level feature vector sequences (i.e. utterances) X and Y is then defined as

$$K(X,Y) = G_\Theta^X F_\Theta^{-1} G_\Theta^Y, \quad (3)$$

where F_Θ is the Fisher information matrix of $p(X|\Theta)$, defined as

$$F_\Theta = E_X[\nabla_\Theta \log p(X|\Theta)\nabla_\Theta \log p(X|\Theta)^T]. \quad (4)$$

Expressing F_Θ^{-1} as $F_\Theta^{-1} = L_\Theta^T L_\Theta$, we get the Fisher vectors as

$$\mathcal{G}_\Theta^X = L_\Theta G_\Theta^X = L_\Theta \nabla_\Theta \log p(X|\Theta). \quad (5)$$

4. Experimental Setup

4.1. X-vector DNN Training

Since speaker ID is required to train x-vectors, and it was not available for either sub-challenge corpus, we trained our x-vector extractor DNN models on an external dataset: on the combined training and development sets of the Dusseldorf Sleepy Language (SLEEP) corpus (11 hours and 39 mins) [28]. We employed the Kaldi framework [29] to do this; for this step, we did not add noise or reverberation to the training samples. (For the details, see [22].) We experimented with using 23-dimensional MFCCs, 40-dimensional FBANKs and spectrograms as features. The *segment6* layer of the DNN was used to compute the 512-dimensional x-vector embeddings. The number of models in the ensemble (m) was set to 10.

4.2. Fisher Vectors

We used the open-source VLFeat library [30] to fit GMMs and to extract the FV representation; we fitted Gaussian Mixture Models with 2, 4, 8, 16, 32, 64 and 128 components on the training sets of the sub-challenges. As the input frame-level feature vectors, we employed 40-dimensional FBANKs with energy as frame-level attributes; and following our previous experiments (e.g. [25]), we also experimented with adding the first and second order derivatives (i.e. Δ and $\Delta\Delta$).

4.3. Classification

Support Vector Machines (SVM) were utilized for classification; we relied on the LIBSVM implementation[31] with a linear kernel (nu-SVR method), and set the C complexity parameter in the range $10^{-5}, \ldots, 10^{1}$. All the features were standardized by removing the mean and scaling to unit variance before training the model. Since the datasets both in the Escalation and in the Primates sub-challenges were imbalanced, we handled this issue by employing downsampling: we randomly discarded training examples from the more frequent classes during

Table 1: *The results obtained for the Escalation Sub-Challenge*

Feature Set	**Dev**	**Test**
ComParE functionals	72.8%	—
Ensemble x-vectors (MFCC)	62.6%	—
Ensemble x-vectors (FBANK)	68.0%	—
Ensemble x-vectors (spectrogram)	72.5%	—
Fisher vectors (FBANK + Δ + $\Delta\Delta$)	74.3%	—
ComParE + x-vectors (spectr.)	74.5%	61.5%
ComParE + FV (FBANK + Δ + $\Delta\Delta$)	77.8%	63.2%
Official ComParE baseline	—	59.8%

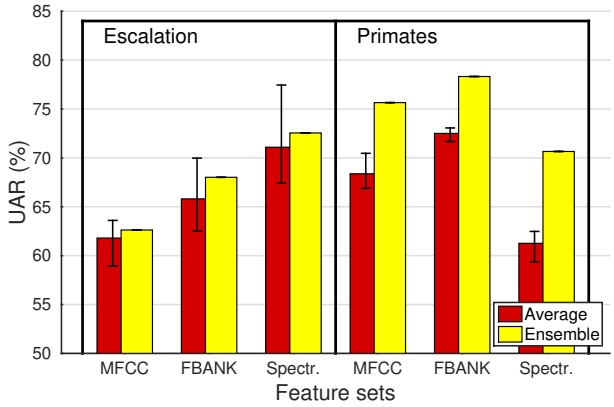

Figure 2: *The UAR scores of the individual and the ensemble x-vector approaches obtained on the development set; the error bars indicate minimum and maximum values.*

training. As downsampling introduces a further random factor into the training process, we decided to train several models and average out the resulting posterior values. For the Escalation sub-challenge, model training was repeated 100 times, while we trained 10 models in each case for the Primates sub-challenge. Test set predictions were obtained by training on the training and development sets together.

4.4. Prediction Combination

From experience (e.g. [12]) we know that it might be beneficial to use multiple different (utterance-level) feature sets, as these might represent the individual utterances from a different aspect, and improve classification. We decided to opt for late fusion [12]: we trained independent SVM models for the different types of features, and combined the predictions in the second step. Following our previous studies, we took the weighted mean of the posterior estimates; the weights were determined on the development set with 0.05 increments.

5. Results

Fig. 2 shows the results obtained on the development sets with the individual x-vectors and the ensemble x-vectors. Notice that the ensemble approach always outperformed the average of the individual models. In the Escalation task, there was a large difference (4-10%) between the performance of the best model and the worst model, probably because of the limited amount of data. For the Primates sub-challenge, this variance was smaller (although still significant: between 1.4% and 3.6%); however, in this case, the ensemble model outperformed even the best of the 10 individual x-vector models. This, in our opinion, confirms that ensemble x-vectors is a viable approach.

Table 2: *The results obtained for the Primates Sub-Challenge*

Feature Set	**Dev**	**Test**
ComParE functionals	81.1%	—
Ensemble x-vectors (MFCC)	75.7%	—
Ensemble x-vectors (FBANK)	78.3%	—
Ensemble x-vectors (spectrogram)	70.7%	—
Fisher vectors (FBANK)	82.7%	—
ComParE + x-vectors (FBANK)	82.6%	83.3%
ComParE + FV (FBANK)	87.5%	88.8%
ComParE + FV (FBANK) + auDeep	88.2%	89.8%
Official ComParE baseline	—	87.5%

Table 1 shows the results obtained for the **Escalation** Sub-Challenge. We can see that the ensemble x-vector approach performed well, considering that it is a 3-class classification task: the UAR values are in the range 62.6 . . . 72.5%, the last being just as effective as ComParE functionals (72.8%). By combining the two feature types, we achieved a slight improvement (74.3%). Fisher vectors were slightly better (note that, due to the lack of space, we only reported the best FV configuration); in the end, we achieved the best results with the combination of ComParE functionals and FVs. Our two test set submissions achieved similar results to the scores on the development set: FVs slightly outperformed the ensemble x-vectors. However, both approaches scored above the official Challenge baseline (obtained via Bag-of-Audio-Words).

Table 2 lists our results obtained for the **Primates** Sub-Challenge. For this task, FBANK-based and MFCC-based (ensemble) x-vectors turned out to be better than the spectrogram-based one; and although even the best one, relying on FBANKs, performed below the standard ComParE functionals attribute set (78.3% and 81.1%, respectively), they could be combined effectively, as the UAR score on the development set improved to 82.6% in this case. Just like that for the Escalation corpus, we achieved even better scores with the Fisher vectors (although now Δs and $\Delta\Delta$s proved to be redundant); this UAR score of 82.7%, measured on the development set, could further be improved to 87.5% by a combination with the ComParE functionals. Regarding the test set scores, the combination of the ComParE feature set with ensemble x-vectors resulted in a test set UAR value below the Challenge baseline. However, we still managed to surpass the ComParE functionals score reported in the baseline paper (see [23]), while with the ComParE + FV method we even exceeded the official baseline score of 87.5%, which was a fusion of five(!) methods itself. This value was further exceeded by incorporating the auDeep features as well.

Of course, the performance of the ensemble x-vector approach (where the feature extractor neural networks used were identical) might be affected by the tasks themselves as well. That is, x-vectors were developed for speaker recognition, and the models were also trained on human speech (i.e. on the Dusseldorf Sleepy Language corpus). From the two sub-challenges, the recordings of Escalation indeed contained human speech, and the x-vectors proved to be quite efficient there. However, in the Primates task the "speakers" were different animals; there, DNNs trained solely on human speech might give a suboptimal performance (although, as we could see from the results,

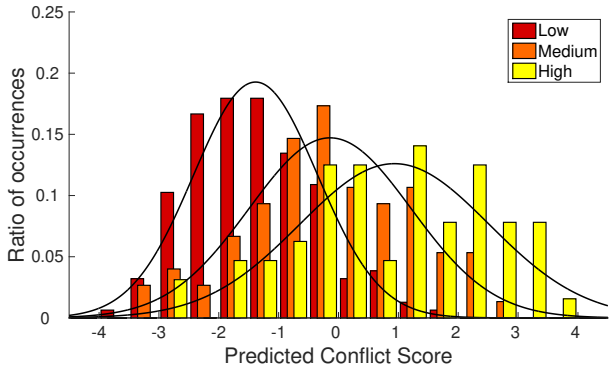

Figure 3: *The distribution of the conflict intensity predictions (based on the SSPNet Conflict Corpus) for the training set of the Primates Sub-Challenge.*

x-vectors were still relatively effective). On the other hand, the GMMs of the Fisher vector technique were trained on the training set of the given corpus, meaning a smaller mismatch; perhaps this also contributed to the difference between the performance of the two approaches.

6. Using the SSPNet Conflict Corpus

In our last experiment, we exploit the fact that the Escalation sub-challenge is a conflict detection task. As permitted by Challenge guidelines, we used another public dataset for model training purposes: the SSPNet Conflict Corpus [26], which contains recordings of Swiss French political TV debates. The total duration of this dataset is significantly larger than that of the Escalation Sub-Challenge: 11 hours and 55 minutes. However, in this corpus the task was to estimate the conflict intensity *scores* in the range -10 (no conflict at all) to 10 (very high level of conflict), while the Escalation sub-challenge was a classification task with three intensity labels. (For the better readability, we used the labels Low, Medium and High instead of the oversimplified 0 / 1 / 2 labels assigned by the Challenge organizers [23].) Next, we will describe our approach to handle this mismatch, and to utilize a corpus with continuous conflict level annotation to a categorical classification task.

Following the experimental setup of our former study (see [32]), we trained an SVR on the training set of the SSPNet Conflict corpus. It used the 130-sized frame-level feature set calculated by the OpenSMILE tool [33], and used Fisher vectors (with 32 Gaussian components) to extract utterance-level attributes. The SVR was trained using a linear kernel, with $C = 10^{-4}$; then we evaluated this model on all the utterances of the Escalation task (i.e. on its training, development and test sets). The distribution of these predictions (on the training set) can be seen in Fig. 3; it is obvious that the predictions do correlate with the class labels, although (as expected) it is not possible to perfectly separate the Low, Medium and High categories based on the intensity predictions alone.

As we sought to combine the predictions of the Escalation test set with the other approaches tested, we had to employ an approach that provided posterior estimates (or values satisfying the formal requirements of posteriors, i.e. falling in the interval $[0, 1]$ and adding up to one). To do this, we modeled the distribution of the SSPNet-based predictions using a normal distribution for each class; these can also be seen in Fig. 3. To obtain 'posterior estimates' for a predicted conflict intensity value, we just calculated the probability density of each class, and normalized these values to add up to one.

Table 3: *The results obtained for the Escalation Sub-Challenge with the SSPNet Conflict Corpus-based approaches*

Feature Set	**Dev**	**Test**
ComParE functionals	72.8%	—
SSPNet Conflict Corpus-based	62.6%	—
ComParE + SSPNet Conflict	73.8%	62.4%
ComParE + x-vectors + FV + SSPNet	79.8%	63.9%
Official ComParE baseline	—	59.8%

Table 3 shows the UAR values obtained via this approach. Although the UAR score of 62.8% on the development set might seem low compared to the ComParE functionals case, for a 3-class (and cross-corpus, as the test set of the Escalation sub-challenge is comes a different dataset than its training and development sets) task it is realistic, as it significantly exceeds the 33.3% value achievable via random guessing. Furthermore, we did not want to utilize this approach on its own, but we sought to use it to aid the other classification methods; and by combination, we achieved an UAR value of 73.8%. On the test set we reached 62.4% with this approach, which was improved to 63.9% by combining all four methods. Both values exceed the official baseline of the Escalation sub-challenge, which, in our opinion, indicates the usefulness of this cross-corpus method.

7. Conclusions

This study describes the techniques we based our entry on the Escalation and Primates sub-challenges of the Interspeech 2021 Computational Paralinguistic Challenge. Our main contribution was to employ x-vectors as features; to improve both the robustness and the performance of the x-vectors, we built an ensemble x-vector classifier by training 10 independent x-vector extractor neural networks on the same data. Our UAR scores on the development set demonstrated the superiority of the ensemble classifiers over the independent x-vector-based ones.

Since Challenge guidelines allow only five submissions for each task, we were unable to extensively verify the performance of all our approaches on the test set. Therefore, we can only extrapolate from the test UAR scores of specific combinations; but based on these values, ensemble x-vectors seem to be an effective approach. Our other, perhaps more traditional feature extractors, Fisher vectors, were even more successful. Our last technique, which used the SSPNet Conflict Corpus in the Escalation sub-challenge, also led to promising UAR values. Overall we managed to exceed the official Challenge baselines for both tasks, which, our opinion, supports the efficacy of the applied techniques.

8. Acknowledgements

This research was supported by the Hungarian Ministry of Innovation and Technology NRDI Office with grant FK-124413 and within the framework of the Artificial Intelligence National Laboratory (MILAB). This research was also partially supported by grant NKFIH-1279-2/2020 of the Hungarian Ministry for Innovation and Technology. G. Gosztolya was also funded by the János Bolyai Scholarship of the Hungarian Academy of Sciences and by the Hungarian Ministry of Innovation and Technology New National Excellence Program ÚNKP-20-5.

9. References

[1] H. Kaya and A. A. Karpov, "Efficient and effective strategies for cross-corpus acoustic emotion recognition," *Neurocomputing*, vol. 275, pp. 1028–1034, 2018.

[2] G. Gosztolya, "Using the fisher vector representation for audio-based emotion recognition," *Acta Polytechnica Hungarica*, vol. 17, no. 6, pp. 7–23, 2020.

[3] C. Montacié and M.-J. Caraty, "Combining multiple phoneme-based classifiers with audio feature-based classifier for the detection of alcohol intoxication," in *Proceedings of Interspeech*, 2011, pp. 3205–3208.

[4] M. Huckvale, A. Beke, and M. Ikushima, "Prediction of sleepiness ratings from voice by man and machine," in *Proceedings of Interspeech*, Shanghai, China, Oct 2020, pp. 4571–4575.

[5] G. Kiss, M. G. Tulics, D. Sztahó, and K. Vicsi, "Language independent detection possibilities of depression by speech," in *Proceedings of NoLISP*, 2016, pp. 103–114.

[6] J.-R. Orozco-Arroyave, J. Arias-Londono, J. Vargas-Bonilla, and E. Nöth, "Analysis of speech from people with Parkinson's disease through nonlinear dynamics," in *Proceedings of NoLISP*, 2013, pp. 112–119.

[7] I. Hoffmann, D. Németh, C. Dye, M. Pákáski, T. Irinyi, and J. Kálmán, "Temporal parameters of spontaneous speech in Alzheimer's disease," *International Journal of Speech-Language Pathology*, vol. 12, no. 1, pp. 29–34, 2010.

[8] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The Interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Proceedings of Interspeech*, San Francisco, CA, USA, 2016, pp. 2001–2005.

[9] S. Pancoast and M. Akbacak, "Bag-of-Audio-Words approach for multimedia event classification," in *Proceedings of Interspeech*, Portland, OR, USA, Sep 2012, pp. 2105–2108.

[10] T. S. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *Proceedings of NIPS*, Denver, CO, USA, 1998, pp. 487–493.

[11] M. Schmitt, F. Ringeval, and B. Schuller, "At the border of acoustics and linguistics: Bag-of-Audio-Words for the recognition of emotions in speech," in *Proceedings of Interspeech*, San Francisco, CA, USA, 2016, pp. 495–499.

[12] G. Gosztolya and R. Busa-Fekete, "Ensemble Bag-of-Audio-Words representation improves paralinguistic classification accuracy," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 29, pp. 477–488, 2021.

[13] H. Kaya, A. A. Karpov, and A. A. Salah, "Fisher Vectors with cascaded normalization for paralinguistic analysis," in *Proceedings of Interspeech*, 2015, pp. 909–913.

[14] N. Dehak, P. J. 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.

[15] M. V. 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 *Proceedings of Interspeech*, Singapore, Sep 2014, pp. 751–755.

[16] J. Grzybowska and S. Kacprzak, "Speaker age classification and regression using i-vectors," in *Proceedings of Interspeech*, San Francisco, CA, Sep 2016, pp. 1402–1406.

[17] J. Weiner and T. Schultz, "Selecting features for automatic screening for dementia based on speech," in *Proceedings of SPECOM*, Leipzig, Germany, Sep 2018, pp. 747–756.

[18] J. V. Egas López, L. Tóth, I. Hoffmann, J. Kálmán, M. Pákáski, and G. Gosztolya, "Assessing Alzheimer's Disease from speech using the i-vector approach," in *Proceedings of SPECOM*, Istanbul, Turkey, Aug 2019, pp. 289–298.

[19] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker verification," in *Proceedings of ICASSP*, Calgary, Canada, Sep 2018, pp. 5329–5333.

[20] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "X-vectors meet emotions: A study on dependencies between emotion and speaker verification," in *Proceedings of ICASSP*, Barcelona, Spain, May 2020, pp. 7169–7173.

[21] S. Zargarbashi and B. Babaali, "A multi-modal feature embedding approach to diagnose Alzheimer's disease from spoken language," *arXiv preprint arXiv:1910.00330*, 2019.

[22] J. V. Egas-López and G. Gosztolya, "Deep Neural Network embeddings for the estimation of the degree of sleepiness," in *Proceedings of ICASSP*, Toronto, Canada, Jun 2021, p. accepted.

[23] B. W. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen, S. Ottl, M. Gerczuk, P. Tzirakis, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, M. R. Leon J. J. Zwerts, J. Treep, and C. Kaandorp, "The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 cough, COVID-19 speech, escalation & primates," in *Proceedings of Interspeech*, Brno, Czech Republic, Sep 2021, to appear.

[24] G. Gosztolya, T. Grósz, and L. Tóth, "General utterance-level feature extraction for classifying crying sounds, atypical & self-assessed affect and heart beats," in *Proceedings of Interspeech*, Hyderabad, India, Sep 2018, pp. 531–535.

[25] G. Gosztolya, "Using Fisher Vector and Bag-of-Audio-Words representations to identify Styrian dialects, sleepiness, baby & orca sounds," in *Proceedings of Interspeech*, Graz, Austria, Sep 2019, pp. 2413–2417.

[26] S. Kim, F. Valente, M. Filippone, and A. Vinciarelli, "Predicting continuous conflict perception with Bayesian Gaussian Processes," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 187–200, 2014.

[27] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep Neural Network embeddings for text-independent speaker verification," in *Proceedings of Interspeech*, Stockholm, Sweden, Aug 2017, pp. 999–1003.

[28] B. W. Schuller, A. Batliner, C. Bergler, F. B. Pokorny, J. Krajewski, M. Cychosz, R. Vollmann, S.-D. Roelen, S. Schnieder, E. Bergelson, A. Cristia, A. Seidl, A. S. Warlaumont, L. Yankowitz, E. Nöth, S. Amiriparian, S. Hantke, and M. Schmitt, "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity," in *Proceedings of Interspeech*, Graz, Austria, Sep 2019, pp. 2378–2382.

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Proceedings of ASRU*, Big Island, HI, USA, Dec 2011.

[30] A. Vedaldi and B. Fulkerson, "Vlfeat: an open and portable library of computer vision algorithms," in *Proceedings of ACM Multimedia*, 2010, pp. 1469–1472.

[31] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1–27, 2011.

[32] G. Gosztolya, "Very short-term conflict intensity estimation using Fisher Vectors," in *Proceedings of Interspeech*, Shanghai, China (online), Oct 2020, pp. 3127–3131.

[33] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proceedings of ACM Multimedia*, Barcelona, Catalonia, Spain, Oct 2013, pp. 835–838.