INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Using Fisher Vector and Bag-of-Audio-Words Representations to Identify Styrian Dialects, Sleepiness, Baby & Orca Sounds

Gábor Gosztolya

MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary
University of Szeged, Institute of Informatics, Szeged, Hungary
ggabor @ inf.u-szeged.hu

Abstract

The 2019 INTERSPEECH Computational Paralinguistics Challenge (ComParE) consists of four Sub-Challenges, where the tasks are to identify different German (Austrian) dialects, estimate how sleepy the speaker is, what type of sound a given baby uttered, and whether there is a sound of an orca (killer whale) present in the recording. Following our team's last year entry, we continue our research by looking for feature set types that might be employed on a wide variety of tasks without alteration. This year, besides the standard 6373-sized ComParE functionals, we experimented with the Fisher vector representation along with the Bag-of-Audio-Words technique. To adapt Fisher vectors from the field of image processing, we utilized them on standard MFCC features instead of the originally intended SIFT attributes (which describe local objects found in the image). Our results indicate that using these feature representation techniques was indeed beneficial, as we could outperform the baseline values in three of the four Sub-Challenges; the performance of our approach seems to be even higher if we consider that the baseline scores were obtained by combining different methods as well.

Index Terms: ComParE 2019 Challenge, Fisher vector representation, Bag-of-Audio-Words

1. Introduction

Besides linguistic information (meant in a strict sense), human speech incorporates a wide range of non-verbal content as well, encoding a huge variety of information about the physical and mental state of the speaker, and which enriches his message. The Interspeech Computational Paralinguistics Challenge (ComParE), held regularly at the Interspeech conference over the last decade, focuses on the automatic identification of this 'paralinguistic' (that is, 'beyond linguistic') aspect of human speech. The tasks set over the years covered dozens of different human speech aspects, ranging from emotion detection [1] through estimating blood alcohol level [2], and determining the speaker's age and gender [3].

A trend over the past few years among the challenge participants was to develop task-dependent features and/or techniques, such as extracting features from the middle of vowels [4], the amount of time when multiple subjects speak at the same time [5], intonation modeling and emotion-specific language models [6]. However, there is also a growing interest in developing general, task-independent approaches, which can be employed in a wide variety of tasks with slight or even no alteration (e.g. [7, 8]). The rise of such general approaches is also reflected in the challenge baselines: recently, the traditional, 6373-sized feature set ('ComParE functionals') was extended by utilizing Bag-of-Audio-Words feature representation (BoAW, [9, 10]) and end-to-end learning [11], and from 2018, sequence-to-sequence autoencoders were applied as well [12].

Following the latter trend, in our contribution to the 2019 ComParE Challenge [13], we experiment with three different types of feature representation. The first one is the classic, 6373-sized paralinguistic feature set developed by Schuller et al. and taking its final form in 2013 [14], which was proved to be a well-performing acoustic representation for a wide range of tasks over the years. As for the second feature set, we apply *Fisher vectors* [15]: this technique is well known in the image processing field (see e.g. [16, 17]), but it is used for audio processing tasks quite rarely (for some notable exceptions, see e.g. [18, 19, 20]). As the last feature extraction approach, we apply Bag-of-Audio-Words.

This year's ComParE Challenge [13] consists of four Sub-Challenges: in the Styrian Dialects Sub-Challenge, the task is to identify which of the three south-eastern Austrian dialects the speaker belongs to. In the Continuous Sleepiness Sub-Challenge, we have to automatically estimate the speaker's sleepiness on the Karolinska Sleepiness Scale (from 1 to 9). In the Baby Sounds Sub-Challenge, the vocalizations of children in the age range 2 to 33 months have to be categorized. Lastly, in the Orca Activity Sub-Challenge, the presence of orcas (or killer whales) have to be detected from digitized underwater sounds.

Following the Challenge guidelines (see [13]), we will omit the description of the tasks, datasets and the method of evaluation, and focus on the techniques we applied. We shall treat all four Sub-Challenges in the same way, except, of course, regarding evaluation metrics: the Styrian Dialects and the Baby Sounds sub-challenges are standard classification tasks, where the performance is measured via the Unweighted Average Recall (UAR) metric. However, as the Continuous Sleepiness Sub-Challenge is a regression task, there Spearman's Correlation Coefficient is used to rank the machine learning models; while for the Orca Activity Sub-Challenge, the Area Under the Curve (AUC) of the positive class is used for this purpose.

2. The Feature Representations Tested

Next, we briefly describe the three different feature representation approaches we utilized in the ComParE 2019 Challenge.

2.1. 'ComParE functionals' Feature Set

Firstly, we used the 6373 ComParE functionals (see e.g. [14]), extracted by using the openSMILE tool [21]. The feature set includes energy, spectral, cepstral (MFCC) and voicing related frame-level attributes, from which specific functionals (like the mean, standard deviation, percentiles and peak statistics) are computed to provide utterance-level feature values.

http://dx.doi.org/10.21437/Interspeech.2019-1726

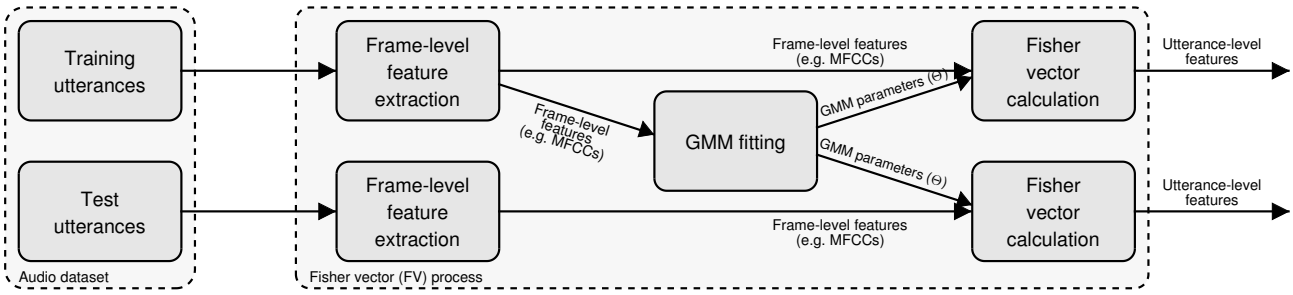

Figure 1: *Workflow of the Fisher vector representation used for audio processing.*

2.2. Fisher Vector Representation

The aim of the Fisher vector representation is to combine the generative and discriminative machine learning approaches by deriving a kernel from a generative model of the data [15]. First we describe the original version, developed for image representation; then we turn to the application of Fisher vectors to audio.

The main concept of the Fisher Vector (FV) representation is to take a set of low-level feature vectors, extracted from the image, and model them by their deviation from the distribution. That is, let $X = \{x_1, \ldots, x_T\}$ be d-dimensional low-level feature vectors extracted from an input sample, and let their distribution be modelled by a probability density function $p(X|\Theta)$, Θ being the parameter vector of the model. The Fisher score describes X by the gradient G_Θ^X of the log-likelihood function, i.e.

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

This gradient function describes the direction in which the model parameters (i.e. Θ) should be modified to best fit the data. Notice that the size of G_Θ^X is already independent of the number of low-level feature vectors (i.e. of T), and it depends only on the number of model parameters (i.e. on Θ). The Fisher kernel between the sequences X and Y is then defined as

$$K(X, Y) = G_\Theta^X F_\Theta^{-1} G_\Theta^Y, \quad (2)$$

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

$$F_\Theta = E_X[\nabla_\Theta \log p(X|\Theta) \nabla_\Theta \log p(X|\Theta)^T]. \quad (3)$$

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

$$\mathcal{G}_\Theta^X = L_\Theta G_\Theta^X = L_\Theta \nabla_\Theta \log p(X|\Theta). \quad (4)$$

In image processing, a varying number of low-level descriptors such as SIFT descriptors (describing occurrences of rotation- and scale-invariant primitives [22]) are extracted from the images as low-level features. The $p(X|\Theta)$ distributions are usually modelled by Gaussian Mixture Models [16]; hence, assuming a diagonal covariance matrix, the Fisher vector representation of an image has a length of twice the number of Gaussian components for each feature dimension (since each Gaussian component models each feature dimension with the help of two parameters: the mean and standard deviation).

To adapt Fisher vectors to audio processing, it is straightforward to use the frame-level features (e.g. MFCCs) of the utterances as the low-level features (i.e. X). Similar to image classification, the distribution of the frame-level components can be modelled by GMMs. For GMMs, using MFCCs is a plausible choice, since their components are quasi-orthogonal; therefore we can reasonably assume that the covariance matrix will be a diagonal one. A parameter of the method is N, the number of Gaussian components. For the workflow for using the FV representation for audio processing, see Figure 1.

We used the open-source VLFeat library [23] to fit GMMs and to extract the FV representation; we fitted Gaussian Mixture Models with $N = 2, 4, 8, 16, 32, 64$ and 128 components. As the input feature vectors, we utilized MFCCs, extracted by the HTK tool [24]. We experimented with using the 12 MFCC vectors along with energy as frame-level features, while also tried adding the first and second order derivatives (i.e. MFCC+Δ and MFCC+Δ+$\Delta\Delta$). We will also call these frame-level feature sets "MFCC13", "MFCC26" and "MFCC39".

2.3. Bag-of-Audio-Words Representation

The BoAW representation also seeks to extract a fixed-length feature vector from a varying-length utterance [9]. Its input is a set of frame-level feature vectors such as MFCCs. In the first step, clustering is performed on these vectors, the number of clusters (N) being a parameter of the method. The list of the resulting cluster centroids will form the *codebook*. Next, each original feature vector is replaced by a single index representing the nearest entry in the codebook (*vector quantization*). Then the feature vector for the given utterance is calculated by generating a histogram of these indices. To eliminate the influence of utterance length, it is common to use some kind of normalization such as L1 normalization (i.e. divide each cluster count by the number of frames in the given utterance).

To calculate the BoAW representations, we utilized the OpenXBOW package [10]. We tested codebook sizes of 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192 and 16384. We employed random sampling instead of kmeans++ clustering for codebook generation since it was reported that it allows a similar classification performance, and it is significantly faster [25]. We employed 5 parallel cluster assignments; otherwise, our setup followed the ComParE 2019 baseline paper (i.e. [13]): we used the 65 ComParE frame-level attributes as the input after standardization and taking the logarithm of the values, and a separate codebook was built for the first-order derivatives.

3. Experiments and Results

3.1. Utterance-level Classification

Our experiments followed standard paralinguistic protocols. After feature standardization (carried out on all the feature sets), we used nu-SVM with a linear kernel for utterance-level classification, using the LibSVM [26] library; the value of C was tested in the range $10^{\{-5,\ldots,1\}}$. The optimal meta-parameters (C for SVM and N for Fisher vectors and BoAW) were determined on the development sets. Final predictions on the test set

Table 1: *Results obtained for the Styrian Dialects Sub-Challenge; the performance is measured in terms of accuracy % and Unweighted Average Recall (UAR) %*

	Dev		Test	
Feature set	Acc.	UAR	Acc.	UAR
ComParE functionals	45.9	39.1	—	36.3
FV MFCC13	49.0	45.2	—	—
FV MFCC26	47.5	40.5	—	—
FV MFCC39	48.2	41.4	—	—
BoAW	46.1	41.4	—	30.4
ComParE + FV MFCC13	49.9	45.6	—	29.0
ComParE + FV MFCC26	47.5	40.6	—	—
ComParE + FV MFCC39	48.5	41.6	—	—
ComParE + BoAW	46.1	41.4	—	30.4
ComParE + FV + BoAW	49.9	45.6	—	29.0
ComParE baseline	—	—	—	47.0

Table 2: *Results obtained for the Continuous Sleepiness Sub-Challenge; the performance is measured in terms of Pearson's ("Pea") and Spearman's ("Spe") correlation coefficient*

	Dev		Test	
Feature set	Pea	Spe	Pea	Spe
ComParE functionals	0.327	0.326	—	—
FV MFCC13	0.351	0.353	—	—
FV MFCC26	0.355	0.355	—	—
FV MFCC39	0.353	0.350	—	—
BoAW	0.300	0.309	—	—
ComParE + FV MFCC13	0.367	0.366	—	—
ComParE + FV MFCC26	0.366	0.365	—	0.382
ComParE + FV MFCC39	0.361	0.356	—	—
ComParE + BoAW	0.346	0.347	—	—
ComParE + FV + BoAW	0.368	0.367	—	0.383
ComParE baseline	—	—	—	0.343

were made by training an SVM model using these parameter values on the training and development sets combined. Following preliminary tests, we employed downsampling for the two corpora with the classification tasks (repeated 25 and 100 times, Styrian Dialects and Baby Sounds sub-challenges, respectively). In each downsampling iteration, training samples were chosen randomly, and the resulting posterior estimates were averaged out over all iterations. As the last step, we employed *late fusion* to combine the different feature representations: we took the weighted mean of the posterior estimates (classification) or predictions (regression); the weights were determined on the development set with 0.05 increments.

3.2. Results

Table 1 shows the accuracy and UAR scores we got for the **Styrian Dialects** Sub-Challenge; "—" denotes the scores which were not reported. We can see that, by using the Fisher vector representation, we outperformed the ComParE functionals on the development set; we got the best results when relying only on the 13 original feature dimensions. Combining this model

Table 3: *Results obtained for the Baby Sounds Sub-Challenge; the performance is measured in terms of accuracy % and Unweighted Average Recall (UAR) %*

	Dev		Test	
Feature set	Acc.	UAR	Acc.	UAR
ComParE functionals	49.5	56.5	—	—
FV MFCC13	37.8	43.3	—	—
FV MFCC26	44.8	49.3	—	—
FV MFCC39	45.9	51.5	—	—
BoAW	45.3	52.3	—	—
ComParE + FV MFCC13	50.2	57.1	—	—
ComParE + FV MFCC26	50.8	57.1	—	—
ComParE + FV MFCC39	50.3	58.0	—	59.5
ComParE + BoAW	49.6	57.8	—	—
ComParE + FV + BoAW	48.1	58.7	—	—
ComParE baseline	—	—	—	58.7

Table 4: *Results obtained for the Orca Activity Sub-Challenge; the performance in measured in terms of the AUC score of the "orca" class*

Feature set	Dev	Test
ComParE functionals	0.824	—
FV MFCC13	0.775	—
FV MFCC26	0.799	—
FV MFCC39	0.793	—
BoAW	0.804	—
ComParE + FV MFCC13	0.833	—
ComParE + FV MFCC26	0.836	—
ComParE + FV MFCC39	0.837	0.875
ComParE + BoAW	0.836	—
ComParE + FV + BoAW	0.843	0.879
ComParE baseline	—	0.866

with the one trained on the ComParE functionals led to a slight improvement, but on the test set we ended up with a UAR score of 29%. This is not only below the baseline, but it is also below the 33.3% UAR score achievable via simple random guessing.

On the **Continuous Sleepiness** Sub-Challenge (see Table 2) we can see similar trends on the development set. (Recall that, since this task is a regression task, the performance is measured via Spearman's correlation coefficient (CC).) FV encodings outperformed the ComParE functionals approach for each case, but now the measured CCs appeared to be quite similar. Late fusion of the two kinds of approaches brought a small improvement on both the development and on the test sets; adding the estimations obtained by using the BoAW features brought a further slight improvement. Note that the baseline score was also achieved via a combination of three approaches.

On the **Baby Sounds** Sub-Challenge (see Table 3), however, our results (at least on the development set) suggest that using the standard ComParE functionals is more efficient than employing Fisher vectors. This is not that surprising, however, if we examine the UAR values reported in the Challenge baseline paper [13]: there the ComParE functionals led to sig-

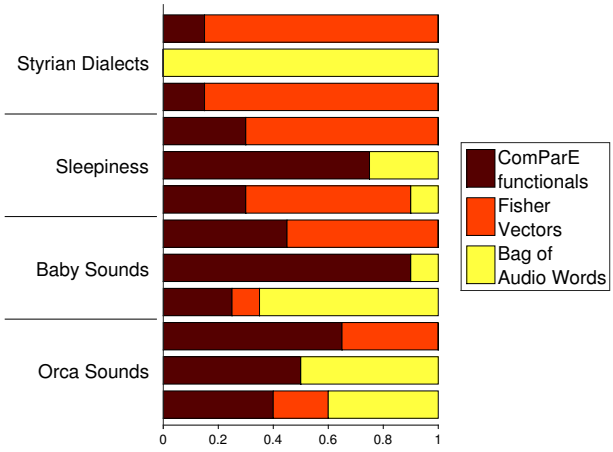

Figure 2: *Optimal weights found in late fusion.*

nificantly higher UAR scores than either Bag-of-Audio-Words or sequence-to-sequence autoencoders did, and combining the three approaches improved this score by only a little bit. In our case, the combined UAR values on the development set exceeded the 'ComParE functionals' case only by 0.6-1.5% absolute; the best-performing combination, however, was slightly above the baseline 58.7% UAR score on the test set, giving a 0.8% absolute improvement (2% relative). For this particular task, in our experience, using the BoAW representation as well was not beneficial, as it had a late fusion weight of 0.00 when combining all three approaches tested.

As for the **Orca Activity** Sub-Challenge, the Fisher vectors again produced similar AUC scores regardless of whether we used the first and second order MFCC derivatives, although these values slightly lagged behind those of the ComParE functionals. This similarity persisted after combination as well: we measured AUC values in the range 0.833-0.837. This hybrid model also outperformed the baseline score on the test set, although the difference is probably statistically not significant. Incorporating the BoAW predictions brought a further slight improvement, leading to AUC values of 0.843 and 0.879 for the development and test sets, respectively.

3.3. Late Fusion Weights

We can gain an insight into the utility of Fisher vectors and Bag-of-Audio-Words by examining the late fusion weights (see Fig. 2). In the **Styrian dialects** Sub-Challenge, due to the overconfidence (overfitting?) of the classifiers trained on the Fisher vectors, the ComParE functionals had a weight of only 0.05-0.10 for the three FV models. In the **Continuous Sleepiness** Sub-Challenge, when we used the first and/or the second-order derivatives of the MFCCs (i.e. models "FV26" and "FV39"), they had a weight of 0.70. This, in our opinion, means that although Fisher vectors turned out to be the more descriptive feature types, the ComParE functionals also represent a valuable ingredient in the final, combined classifier for this particular task. In the **Baby Sounds** Sub-Challenge the two feature types had similar weights (in the range 0.45-0.55), while for the **Orca Activity** Sub-Challenge, ComParE functionals had a weight value of 0.65, suggesting that it was found to be the (slightly) more valuable feature set. Interestingly, these hypotheses mirror the tendency of the UAR values given in the ComParE challenge baseline paper [13].

Regarding the BoAW features, on the Continuous Sleepines Sub-Challenge it was assigned a much smaller weight than Fisher Vectors did, while for the Orca Sounds it had a somewhat larger one. Lastly, on the Baby Sounds Sub-Challenge, when we just combined it with the ComParE functionals, BoAW was assigned a weight of just 0.1. Surprisingly, when we combined all three methods, the very same predictions produced a weight of 0.65, which we cannot regard as anything else but overfitting. Clearly, even the reported results of the Challenge baseline paper indicate that the Bag-of-Audio-Words approach suffers from its stochastic nature: the codebook size which gave an optimal performance on the development set led to a metric value on the test set which was significantly below the optimal one (see [13]). Based on our experimental results, we consider Fisher vectors a much more robust approach in this aspect.

3.4. Building an Ensemble of Classifiers

By our recent experience, the Bag-of-Audio-Words approach is inherently stochastic due to the randomness being present in its clustering step (for our initial findings, see [7]). Our hypothesis was that this also holds for the Fisher vector representation, since fitting GMMs also involves clustering. Therefore, to increase the robustness of our predictions, we extracted the FV and the BoAW features using ten different random seed values, trained our SVMs on all of them, and simply averaged out the predictions we got (Continuous Sleepiness Sub-Challenge) or posterior estimates (Baby Sounds and Orca Activity sub-challenges). Unfortunately, due to time limitations, we were unable to finish these experiments for the Styrian Dialects Sub-Challenge. In the other three tasks, however, we obtained further slight improvements over the single-seed case (when combining all three feature sets): for the Continuous Sleepiness Sub-Challenge, Spearman's CC increased to 0.387 (from 0.383) on the test set; on the Baby Sounds Sub-Challenge, we obtained an UAR value of 59.9%, while for the Orca Activity Sub-Challenge, the AUC value of the "orca" class rose from 0.879 to 0.884.

4. Conclusions

In our contribution to the INTERSPEECH 2019 Computational Paralinguistic Challenge (ComParE), we tested general-purpose feature representations on all four sub-challenges. Besides the now standard ComParE functionals, we employed Bag-of-Audio-Words, and utilized the Fisher vector representation to construct fixed-size utterance-level feature vectors from recordings of varying lengths. In the end, we managed to outperform the Challenge baselines in three tasks out of four. Besides showing that using Fisher Vectors is an efficient way of representing utterances in paralinguistic tasks, we also found that building an ensemble of classifiers (based on either FV or BoAW feature sets) could improve the robustness of predictions.

5. Acknowledgements

This research was partially funded by the Ministry of Human Capacities, Hungary (grants 20391-3/2018/FEKUSTRAT and TUDFO/47138-1/2019-ITM), and by the National Research, Development and Innovation Office of Hungary via contract NKFIH FK-124413. 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-19-4.

6. References

[1] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proceedings of Interspeech*, Brighton, United Kingdom, Sep 2009, pp. 312–315.

[2] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 speaker state challenge," in *Proceedings of Interspeech*, Florence, Italy, Aug 2011, pp. 3201–3204.

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

[4] D. Sztahó, G. Kiss, and K. Vicsi, "Estimating the severity of parkinson's disease from speech using linear regression and database partitioning," in *Proceedings of Interspeech*, 2015, pp. 498–502.

[5] F. Grèzes, J. Richards, and A. Rosenberg, "Let me finish: Automatic conflict detection using speaker overlap," in *Proceedings of Interspeech*, 2013, pp. 200–204.

[6] C. Montacié and M.-J. Caraty, "Vocalic, lexical and prosodic cues for the INTERSPEECH 2018 self-assessed affect challenge," in *Proceedings of Interspeech*, 2018, pp. 541–545.

[7] 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.

[8] B. Vlasenko, J. Sebastian, D. P. Kumar, and M. Magimai-Doss, "Implementing fusion techniques for the classification of paralinguistic information," in *Proceedings of Interspeech*, Hyderabad, India, Sep 2018, pp. 526–530.

[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] M. Schmitt and B. Schuller, "openXBOW – introducing the Passau open-source crossmodal Bag-of-Words toolkit," *The Journal of Machine Learning Research*, vol. 18, pp. 1–5, 2017.

[11] B. Schuller, S. Steidl, A. Batliner, S. Hantke, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A. S. Warlaumont, G. Hidalgo, S. Schnieder, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2017 computational paralinguistics challenge: Addressee, Cold & Snoring," in *Proceedings of Interspeech*, Stockholm, Sweden, Aug 2017, pp. 3442–3446.

[12] 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 *Proceedings of Interspeech*, Hyderabad, India, Sep 2018, pp. 122–126.

[13] 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. 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.

[14] B. W. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social signals, conflict, emotion, autism," in *Proceedings of Interspeech*, Lyon, France, Sep 2013, pp. 148–152.

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

[16] G. Csurka and F. Perronnin, "Fisher vectors: Beyond Bag-of-Visual-Words image representations," in *Proceedings of VISIGRAPP*, Angers, France, May 2010, pp. 28–42.

[17] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International Journal of Computer Vision*, vol. 105, no. 3, pp. 222–245, 2013.

[18] P. J. Moreno and R. Rifkin, "Using the Fisher kernel method for web audio classification," in *Proceedings of ICASSP*, Dallas, TX, USA, 2010, pp. 2417–2420.

[19] 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.

[20] Z. Zajíc and M. Hrúz, "Fisher vectors in PLDA speaker verification system," in *Proceedings of ICSP*, Chengdu, China, 2016, pp. 1338–1341.

[21] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proceedings of ACM Multimedia*, 2010, pp. 1459–1462.

[22] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

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

[24] S. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[25] 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.

[26] 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.