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

Employing Bottleneck and Convolutional Features for Speech-Based Physical Load Detection on Limited Data Amounts

Olga Egorow[1], *Tarik Mrech*[2], *Norman Weißkirchen*[1], *Andreas Wendemuth*[1]

[1]Cognitive Systems Group, Otto von Guericke University, Magdeburg, Germany
[2]Virtual Engineering, Fraunhofer Institute for Factory Operation and Automation, Magdeburg, Germany

olga.egorow@ovgu.de

Abstract

The detection of different levels of physical load from speech has many applications: Besides telemedicine, non-contact detection of certain heart rate ranges can be useful for sports and other leisure time devices. Available approaches mainly use a high number of spectral and prosodic features. In this setting of typically small data sets, such as the Talk & Run data set and the Munich Biovoice Corpus, the high-dimensional feature spaces are only sparsely populated. Therefore, we aim at a reduction of the feature number using modern neural net inspired features: Bottleneck layer features, obtained from standard low-level descriptors via a feed-forward neural network, and activation map features, obtained from spectrograms via a convolutional neural network. We use these features for an SVM classification of high and low physical load and compare their performance. We also discuss the possibility of hyperparameter transfer of the extracting networks between different data sets. We show that even for limited amounts of data, deep learning based methods can bring a substantial improvement over "conventional" features.

Index Terms: deep learning, bottleneck features, convolutional neural networks, hyperparameter optimisation, physical stress detection, physical load detection, speech signal processing

1. Introduction

Physical stress or load can be defined as the exertion perceived by a subject and is therefore a subjective concept [1]. However, it is directly related to the heart rate since it correlates with the exercise intensity [2]: The physical load is often defined as high when the heart rate reaches around 80% of the maximum heart rate.

In contrast to physical stress, the heart rate can be directly measured – and, besides using ECG or other intrusive methods, it can also be estimated from other available data in a contactless way, for example from speech. Even when assuming that such a heart rate estimation would not conform to highest medical standards, it might be the only available data source for some applications such as telemedicine or emergency call centres. Furthermore, there are also ideas for non-medical use, e.g. physical exercise. Here, the heart rate monitoring can be used to detect whether the physical load is high enough for the training to be efficient and therefore to increase the user's motivation during exercise activities [3]. Another application is in speech-based advanced driver-assistance systems – for example, to detect stressful situations on the road by registering an unusually high heart rate [4].

The relationship between the heart rate and speech was addressed in several investigations. Orlikoff & Baken found that the heartbeat leads to perturbation of the fundamental frequency which can be measured on sustained vowels pronunciations [5]. This led to further investigations focussing on vowels, such as heart rate prediction from the 2D spectrum of human vowel speech [6] as well as from formant maximum peaks [7].

To work only on vowels is impractical for natural human-computer interaction systems – therefore the approaches must be extended to natural speech. The importance of this task was already acknowledged by addressing physical load detection in the Interspeech 2014 Computational Paralinguistics Challenge's Physical Load Sub-Challenge [8]. Here, the participants were given a baseline feature set consisting of Low-Level-Descriptors (LLDs) and their functionals as well as the Munich Biovoice Corpus (MBC), a data set containing not only vocals but also breathing sounds and read texts, that we also use in the present investigation. The winners of the challenge implemented a canonical correlation analysis and local Fisher discriminant analysis to rank the features and reduce their number, and employed Support Vector Machine (SVM) for the classification task, achieving over 75% Unweighted Average Recall (UAR) on the provided test set [9].

Ignited by the challenge, the interest in physical load detection and heart rate estimation increased, leading to further investigations. Truong et al. presented a new data set in 2015, the Talk & Run Speech Database (TalkR), that we describe in detail in Section 2. Using this data set that contains speech of 21 subjects in two physical load intensity levels, Truong et al. performed a classification in high and low intensity exercise based on 224 acoustic features and an SVM with an Radial Basis Function (RBF) kernel. In a Leave-One-Speaker-Out (LOSO) setting, they achieved a 60% UAR for male and 74% UAR for female subjects [10]. Tsiartas et al. investigated the SRI BioFrustration Corpus and aimed at a prediction of heart rate changes, i.e. whether heart rate is increasing or decreasing. They used MFCCs, MFBs, energy contours, spectral tilt and intonation related features, as well as HNR in a random forest classifier, achieving an accuracy of 70% [11]. On the same data, Smith et al. used random forests and LASSO regression to predict the heart rate achieving around 10% better results than the baseline, which was defined as the standard deviation of speaker-normalised heart rate. They used MFCCs, MFBs, HNR and other spectral and prosodic features, as well as their functionals [12]. Aiming at heart rate estimation, Milton & Monsely used MFCCs and ARRCs as features for an SVM-based linear regression on the Tamil and English Speech Database they recorded, and achieved a minimum prediction error of $\pm$ 13 Beats Per Minute (BPM) [13].

So far, all of the mentioned approaches used relatively large feature sets with thousands of features based on hand-crafted

http://dx.doi.org/10.21437/Interspeech.2019-2502

acoustic characteristics of the speech signal. But the employed data sets are relatively small, therefore this results in sparsely populated high-dimensional feature spaces, leading to the curse of dimensionality. In our investigation, we want to test a possible solution to this problem and take a look on two small feature sets which we generate with state-of-the-art neural networks. In the first feature set, we feed the 3396 acoustic features used in the previously mentioned Interspeech 2014 Challenge to a Feed Forward Neural Network (FFNN) bottleneck architecture in order to reduce the dimensionality and extract the most important information condensed in 100 features. For the second feature set, we use spectrograms of the acoustic signal as input for a Convolutional Neural Network (CNN), resulting in a feature set containing 40 features. We compare the performance of these two feature sets for the recognition of low and high physical load on two previously mentioned data sets: TalkR and MBC. Furthermore, we evaluate the possibility of hyperparameter transfer of the extracting networks from one data set to another.

The paper is organised as follows: in Section 2 we describe the used data sets, Section 3 focuses on the experimental setup we implemented in order to obtain the two feature sets, Section 4 presents and discusses the achieved results, in Section 5 we summarise our work.

2. Data – TalkR & MBC

For our experiments, we use two data sets, the Talk & Run Speech Database (TalkR) and the Munich Biovoice Corpus (MBC).

The TalkR contains speech from 21 subjects (15 females and 6 males, 20 - 31 years old). The subjects were asked to read two short texts in English and Dutch during physical exercise. The corpus comprises 250 audio samples with a total duration of around 85 minutes. In the high physical load stage, the subjects experience heart rates between 172 and 198 BPM. Details on the experimental setup and data can be found in [10].

The MBC contains speech from 19 subjects (4 females and 15 males). The subjects were asked to utter the vocal /a/ in two frequencies and to read aloud one short text (one of the two texts used in TalkR) in English and German. Overall, the corpus comprises 74 text reading samples, 644 breath periods and 630 sustained vowel expressions. The recordings were made before physical exercise and during physical exercise with a heart rate of least 90 BPM. Further details on the experimental setup can be found in [14].

3. Experimental Setup

Our experimental setup consists of three stages: the feature extraction stage, the feature processing stage via FFNN and CNN, and the classification stage using SVM.

3.1. Feature Extraction

For the first of our two implemented systems, the FFNN system, we extracted 3396 features of the Interspeech 2011 Speaker State Challenge described in detail in [15]. These 3396 features are based on 54 LLDs (4 energy-related and 50 spectral features) and their functionals, and are the same features as used in the Interspeech 2014 Physical Load Sub-Challenge. The features are extracted using a 0.06 s window and a 0.01 s shift, the functionals are calculated over a window of 1 s.

For the second system, the CNN system, we used spectrograms of the audio signal as input. For this, we used STFT spectrograms and their first and second order deltas. We stack them to a three-dimensional matrix. The first layer of this matrix is a 40 x 40 spectrogram over a window of 20 ms with a shift of 10 ms. It should be noted that we removed frequencies above the 40th band, since they were negligible. The second layer contains the first order deltas of the spectrogram, and the third its second order deltas. Fig. 1 illustrates an exemplary feature map. This setup enables using temporal developments over longer periods of time compared to the "conventional" LLD-based features described above.

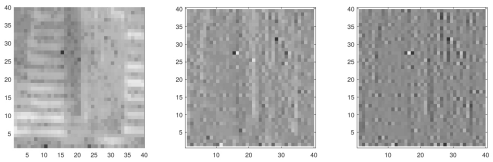

Figure 1: *Feature map consisting of the spectrogram and its first and second order deltas (from left to right).*

3.2. Feature processing

Our implemented FFNN system consists of 5 hidden layers and a bottleneck layer (with a linearly decreasing number of neurons per layer, from 1500 neurons in the first to 100 neurons in the last, bottleneck layer). Due to a very limited data amount, we pre-trained each layer in an unsupervised way using sparse autoencoders and scaled conjugate gradient. Following the standard autoencoder training procedure, we used the training data as input and output for the first autoencoder. The second autoencoder has the data encoded by the first autoencoder as input and so on. Before training, we scaled and centralised the training data. The network architecture is shown in Fig. 2.

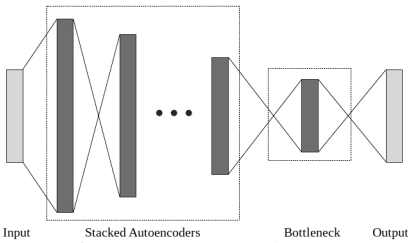

Figure 2: *The architecture of the FFNN for the bottleneck features processing.*

In order to ensure the generalisation ability for the later classification, we implemented a speaker-independent processing. For the FFNN hyperparameter optimisation, we used the 16 female speakers of the TalkR data (further referred to as *femTalkR* subset) in a LOSO manner: 12 speakers were used for training, two speakers as validation set for the adjustment of all the tested hyperparameters (number of hidden layers, number of neurons per layer etc.). The remaining two speakers were reserved for later testing. The whole procedure was repeated over all speakers of the *femTalkR* subset.

For the training of the autoencoders, we determined the number of training epochs depending on the number of weights of the layer to avoid overfitting. After the training of the autoencoders, the softmax layer is also trained – here the input was the

data trained by the last autoencoder, the target was given by the classes of the unencoded data. After this, we trained the whole network on the *femTalkR* subset and the two classes as target for 200 epochs. In order to avoid overfitting, we implemented early stopping using the validation set as explained above. In the end, we obtained 100 bottleneck features as the feature set and orthogonalised them by a Principal Component Analysis (PCA).

An overview of the CNN system architecture is shown in Fig. 3. We used two blocks of 4 convolutional layers, each followed by a max-pooling layer. The number of filters in the convolutional layers was 64 for all layers from the first block, with twice the amount in the second block. This number was determined by implementing Random Search [16]. The second max-pooling layer was followed by a fully connected layer (FC) with 40 neurons. This resulted in 40 features that were then orthogonalised by a PCA.

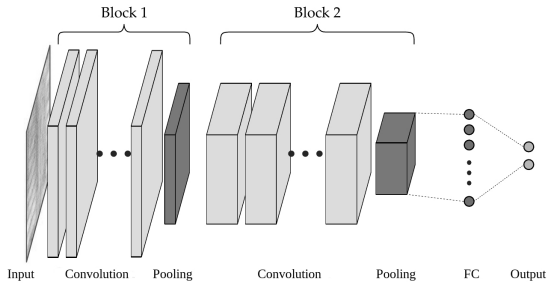

Figure 3: *The architecture of the CNN for the spectrogram processing.*

We used the architecture proposed in [17], and improved it by employing a variable network depth. For the training, we implemented a dropout layer after each block and a sparsity constraint. The network was trained for 25 epochs, starting with a learning rate of 0.0001 and decreasing this rate by a factor of 10^{-1} after every 5 epochs. Here, again, we implemented a LOSO procedure on the *femTalkR* subset for hyperparameter optimisation (number of blocks, number of layers, number of filters). Since the goal was not to use the network for classification but only for feature extraction, we aimed at a higher generalisation ability instead of classification performance and did not introduce further optimisation in order to avoid overfitting.

3.3. Classification

For classification, we used RBF-SVM. For training, we used all available audio from the training data set, including pauses and breathing. For a subject-independent evaluation, we again implemented a LOSO procedure. As mentioned above, we used the *femTalkR* subset for determining the hyperparameters of the neural networks, but we repeated the training process of the networks for all data sets in a LOSO manner without further changing the hyperparameters.

For both feature sets, we used the same procedure to optimise the parameters C and γ of the SVM by implementing Bayes optimisation. After fine-tuning C and γ, we performed the training and classification in a LOSO manner first on the female speakers of TalkR, then on the male speakers of TalkR. In the second step, we repeated the procedure on the female speakers of MBC, and on the male speakers of MBC. For both data sets, we trained n separate models for n speakers, using $n-2$ speakers for training, and 2 speakers for the final test. The final results are obtained on the n test speakers n times, and then averaged over all n speakers.

4. Results & Discussion

The main objective of our experiments was to find out how our deep learning inspired feature sets perform compared to the conventional approaches. For that, we first tested our features on the TalkR data, and then repeated the classification on the MBC data.

4.1. TalkR Data

The results achieved on the TalkR data are shown in Tab. 1, including the baseline system results as reported by Truong et al. [10], separated by the speakers' sex to enable a direct comparison. We also report the results achieved on all speakers – due to data dysbalance regarding the sex of the speakers, the results over all speakers are not the averages of the results on female and male speakers.

Table 1: *Classification results achieved on the TalkR data in terms of recall for both classes (High and Low) and their unweighted average (UAR), in %. Baseline refers to the results by Truong et al.*

System	Females	Males	Overall
Baseline UAR	73.5	60.0	70.1
FFNN			
– High	83.5	71.7	79.72
– Low	83.3	67.4	78.73
– UAR	83.4	69.7	79.22
CNN			
– High	81.7	74.3	79.94
– Low	82.0	72.5	79.09
– UAR	81.8	73.2	79.52

On the data of the female TalkR speakers, our FFNN system achieves a UAR of 83.4%, which is a clear performance gain compared to 73.5% achieved by the baseline. For the male speakers, our system also delivers a performance boost: It achieves 69.7% UAR compared to 60% of the baseline. Overall, the UAR of our system is around 9% abs. higher than the baseline.

The CNN system delivers comparable results, with a slightly lower performance on the female speakers of 81.8% UAR, and a remarkably higher performance on the male speakers of 73.2% UAR. Compared to the baseline, we can see that CNN features again outperform the "conventional" features by around 9% abs.

These results confirm that both, FFNNs and CNNs can be used to extract meaningful features even on limited amounts of data and generalise over different speaker groups – although the hyperparameters of the networks were determined in a time-consuming process only on the *femTalkR* subset, which does not seem to be a problem in this setting.

4.2. MBC Data

In the second step of our investigation, we take a look on the classification of the MBC data. Again, we report the results separately for female and male speakers as well as the average.

Table 2: *Classification results on the MBC data in terms of recall for both classes (High and Low) and their unweighted average (UAR), in %. Baseline refers to the results Kaya et al.*

System	Females	Males	Overall
Baseline UAR	–	–	75.35
FFNN			
– High	56.8	64.3	61.65
– Low	58.6	60.3	59.77
– UAR	57.7	62.3	60.71
CNN			
– High	56.9	55.9	56.20
– Low	55.0	55.3	55.19
– UAR	55.9	55.5	55.69

We can see that both, the FFNN features and the CNN features do not perform as well as for the TalkR data.

In the case of the FFNN features, the performance on female speakers delivers a UAR of only 57.7%. The overall UAR is just 60.71%, which is almost 15% abs. below the baseline.

Just as in the case of TalkR above, the CNN features deliver results similar to those of the FFNN features, with a UAR of only 55.7%, almost 20% abs. below the baseline.

This is a startling result – the only difference compared to the results on the TalkR data is that the hyperparameters of both networks are not fine-tuned to the MBC data. Even taking into account that the MBC data contains only 4 female speakers and 15 male speakers as opposed to 15 females and 6 males of TalkR, the experiments show that the features do not perform in a stable way over different data sets.

4.3. Discussion

The results show that there are remarkable differences regarding the performance on the two data sets – the choice of the data seems to play a bigger role than the choice of the features, since both feature sets achieve almost identical results of around 79% UAR on TalkR and around 55–60% UAR on MBC.

Regarding the performance on TalkR, this is an improvement of around 9% abs. compared to the results achieved by Truong et al. with LLD-based features. As far as we know, we performed the evaluation in the same manner, therefore the results are directly comparable.

But regarding the MBC, both feature sets deliver a suboptimal performance,with the results at least 15% abs. below those of the baseline system of Kaya et al. – although the evaluation procedure employed by the original authors is based on one development and one test set instead of a true LOSO setting, so the results are not directly comparable.

As already explained above, the hyperparameters of the neural networks generating the feature sets were determined only on the *femTalkR* subset, assuming that the subsequent training process, which was performed on the appropriate data sets, would be sufficient to fit the networks to the data during the training process. But our results suggest the opposite: Apparently, the hyperparameters have a great influence on the generated features. One possible explanation for this are the differences in the data sets. We already mentioned the different speaker sex distribution: This is especially interesting since female voices generally have higher frequencies, and the speakers of MBC had to utter vowels in two pre-defined frequencies, which is also the main difference to the TalkR data containing only speech. Another point is the difference in the definitions of high and low physical load: For TalkR, the heart rate during the high load condition was between 172 and 198 BPM, for MBC everything above 90 BPM was defined as high load.

Comparing the performance of the two employed feature sets, we can state that both feature sets perform fairly well, with the FFNN features achieving an overall slightly better performance than the CNN features. One explanation for this is surely the low amount of data, leading to overfitting especially for the CNN features: In the case of FFNN features, this is partly mitigated by employing autoencoders to initialise the network. Also, the FFNN features are based on LLDs, which are known to perform well for speech-related recognition tasks. In contrast to this, the CNN features are based on pure spectrograms, where redundancy might obscure the relevant information. For CNNs, this issue could be addressed by implementing data augmentation techniques known from image classification, where additional data is created by rotating, cropping and flipping input images [18, 19], i.e. the spectrograms in our case.

We are aware of the fact that our features are based on a deep learning procedure, and therefore obviously more difficult and time-consuming to obtain compared to conventional LLD-based features. Nevertheless, as deep learning is gaining importance in more research fields, we strongly argue for usage of such systems, since we have shown that it is possible to employ such systems even on limited amounts of data, such as the TalkR and MBC data sets.

5. Conclusion

In this paper, we have shown that both, FFNN and CNN-based features are suitable for the task of speech-based classification of high and low physical load. Both feature sets perform fairly well on the TalkR data, achieving around 79% UAR in a LOSO setting – this is remarkable, since the data set has a rather limited length of only 85 minutes in 250 audio samples. But both the FFNN and CNN feature sets are able to extract the relevant information – with only 40 features in the CNN setting and 100 features in the FFNN setting, instead of almost 4000 features contained in conventional LLD-based feature sets.

Nevertheless, the results also show that the hyperparameters of such networks need to be chosen carefully, since the same architectures for FFNN and CNN did not perform well on the MBC data. Here, more insight into the data is necessary, as leaving it to the mechanics of deep learning does not achieve the desirable results. Since both employed data sets are available to the community, we hope that this issue will gain more attention in the future.

6. Acknowledgements

This work was sponsored by the German Federal Ministry of Education and Research (BMBF) in the research alliance 3Dsensation (www.3d-sensation.de) under grant number 03ZZ0414. Furthermore, the authors acknowledge support by the project Intention-based Anticipatory Interactive Systems (IAIS) funded by the European Funds for Regional Development (EFRE) and by the Federal State of Sachsen-Anhalt, Germany (grant number ZS/2017/10/88785). The authors also thank K. Truong and B. Schuller for providing the data sets.

7. References

[1] G. A. Borg, "Psychophysical bases of perceived exertion," *Medicine & Science in Sports & Exercise*, vol. 14, no. 5, pp. 377–381, 1982.

[2] J. Karvonen and T. Vuorimaa, "Heart rate and exercise intensity during sports activities," *Sports Medicine*, vol. 5, no. 5, pp. 303–311, 1988.

[3] T. Han, X. Xiao, L. Shi, J. Canny, and J. Wang, "Balancing accuracy and fun: designing camera based mobile games for implicit heart rate monitoring," in *Proc. of the the 33rd ACM Conference on Human Factors in Computing Systems*. ACM, 2015, pp. 847–856.

[4] R. R. Singh, S. Conjeti, and R. Banerjee, "An approach for real-time stress-trend detection using physiological signals in wearable computing systems for automotive drivers," in *Proc. of the 14th Int. IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2011, pp. 1477–1482.

[5] R. F. Orlikoff and R. J. Baken, "The effect of the heartbeat on vocal fundamental frequency perturbation," *Journal of Speech, Language, and Hearing Research*, vol. 32, no. 3, pp. 576–582, 1989.

[6] D. Skopin and S. Baglikov, "Heartbeat feature extraction from vowel speech signal using 2d spectrum representation," in *Proc. the 4th Int. Conference on Information Technology (ICIT)*, 2009.

[7] A. Mesleh, D. Skopin, S. Baglikov, and A. Quteishat, "Heart rate extraction from vowel speech signals," *Journal of computer science and technology*, vol. 27, no. 6, pp. 1243–1251, 2012.

[8] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The interspeech 2014 computational paralinguistics challenge: Cognitive & physical load," in *Proc. of the INTERSPEECH 2014*, 2014, pp. 427–431.

[9] 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 *Proc. of the INTERSPEECH 2014*, 2014, pp. 442–446.

[10] K. P. Truong, A. Nieuwenhuys, P. Beek, and V. Evers, "A database for analysis of speech under physical stress: detection of exercise intensity while running and talking," in *Proc. of the INTERSPEECH 2015*, 2015, pp. 3705–3709.

[11] A. Tsiartas, A. Kathol, E. Shriberg, M. d. Zambotti, and A. Willoughby, "Prediction of heart rate changes from speech features during interaction with a misbehaving dialog system," in *Proc. of the INTERSPEECH 2015*, 2015, pp. 3715–3719.

[12] J. Smith, A. Tsiartas, E. Shriberg, A. Kathol, A. Willoughby, and M. de Zambotti, "Analysis and prediction of heart rate using speech features from natural speech," in *Proc. of the IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP) 2017*. IEEE, 2017, pp. 989–993.

[13] A. Milton and K. A. Monsely, "Tamil and english speech database for heartbeat estimation," *International Journal of Speech Technology*, pp. 1–7, 2018.

[14] B. W. Schuller, F. Friedmann, and F. Eyben, "The Munich Biovoice Corpus: Effects of Physical Exercising, Heart Rate, and Skin Conductance on Human Speech Production," in *Proc. of the Int. Conference on Language REsources and Evaluation (LREC) 2014*, 2014, pp. 1506–1510.

[15] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The interspeech 2011 speaker state challenge," in *Proc. of the INTERSPEECH 2011*, 2011, pp. 3201–3204.

[16] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. Feb, pp. 281–305, 2012.

[17] N. Kurpukdee, T. Koriyama, T. Kobayashi, S. Kasuriya, C. Wutiwiwatchai, and P. Lamsrichan, "Speech emotion recognition using convolutional long short-term memory neural network and support vector machines," in *Proc. of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) 2017*. IEEE, 2017, pp. 1744–1749.

[18] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.

[19] L. Taylor and G. Nitschke, "Improving deep learning with generic data augmentation," in *Proc. of the 2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2018, pp. 1542–1547.