INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Analysis and automatic prediction of exertion from speech: Contrasting objective and subjective measures collected while running

Andreas Triantafyllopoulos[1], Alexander Gebhard[1], Alexander Kathan[1], Maurice Gerczuk[1], Shahin Amiriparian[1], Björn W. Schuller[1,2]

[1]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2]GLAM – Group on Language, Audio, & Music, Imperial College London, UK

andreas.triantafyllopoulos@uni-a.de

Abstract

Monitoring runner exertion in real-time can provide unique insights that help improve training and reduce injuries. Most existing methods use heart rate (HR) as a physiological proxy of it, but this does not always correspond to self-perceived exertion. This is an additional factor in determining overall strain and is typically evaluated with the Borg rating of perceived exertion (RPE) scale. In recent years, speech has been one of the many modalities used to monitor exertion; however, mostly used to predict physiological measures using speech collected after a physical task. In this work, we contrast the manifestation of subjective vs objective exertion on speech signals obtained while running in real-life environments. We identify and interpret a set of prosodic and spectral features related to both markers, and proceed to train deep learning models that directly predict RPE and HR from speech, obtaining an average Pearson correlation of .341 and .418, respectively.

Index Terms: Perceived Exertion, Heart Rate, Speech Analysis, Computational Paralinguistics, Machine Learning

1. Introduction

Computer audition has become a cornerstone of recent advances in digital health, due to its noninvasive nature and ubiquitous presence in everyday devices such as smartphones and wearables, leading to a steady rise in associated applications [1]. Recently, it has been applied to the analysis and prediction of objective and subjective indicators of exertion in running applications [2–8]. Exertion during running is a key indicator of potential overuse injuries that plague up to 79% of runners on a yearly basis [9]. Promptly detecting the onset of exertion can help inform training and regulate training regimens to decrease the risk of injury [10], while additionally improving performance for experienced athletes and leading to the retainment of a new habit for amateurs [11]. Critically, most prior works focus on objective measures of exertion, such as heart rate or VO_2, both of which are known to impact the production of speech [2, 12, 13]. On the other hand, subjective, self-reported measures are relatively under-researched [14]. However, subjective exertion markers can be more informative with regards to determining dangerous strain and increasing overall activity enjoyment, as they account for signals from the peripheral muscles and joints, as well as the cardiovascular, respiratory, and central nervous systems [15], thus acting complementarily to heart rate (HR). The main focus of this work is to mitigate this gap and investigate the impact of subjective exertion on speech, and specifically whether speech signals can be used to predict these additional aspects of exertion.

The use of speech as a measure of physical activity has a long history, starting from the simple, yet informative, 'Talk Test' [16]. This has, in turn, led to the utilisation of speech as a biomarker of physical intensity [2–8]. These prior works have mostly investigated the impact of physical intensity *after* exercise, with subjects typically undergoing a fixed exercise protocol preceded and followed by a data recording – often of read speech. Analysing speech *during* exercise is relatively under-researched [3, 7], but with a considerably larger upside for providing the real-time feedback crucial for regulating training. The elicitation protocol in this case requires participants to talk while running, a setup we also follow in our work.

The basic premise is that out-of-breath speech exhibits changes to breathing patterns, namely a heavier inhalation and exhalation of air, which in turn impacts speech production [17]. A substantial part of this prior work has been dedicated to evaluating the impact of physical exertion on speech through feature analysis. It has been shown that physical load increases F0 [4, 18] and breathiness [4, 18], decreases voiced frames [8], leads to more irregular pauses [4, 18] and articulation [4], while lowering formant frequencies and widening their bandwidths [6] – though not consistently across all studies [18]. Crucially, previous authors often found the manifestation of exertion on speech parameters to be highly speaker dependent [6, 18, 19]. We follow up on these findings by investigating speech patterns related to heart rate and perceived exertion both on aggregate and on an individual basis using our data, while also evaluating the disaggregated performance of our models with respect to different groups.

Similarly, speech has been widely used to model affective states, like emotion [20]. This is motivated by the presence of affective cues in acoustic, prosodic, and tempo features [21], which can be used to predict emotion. More recently, these have been supplemented by automatically learnt features that can outperform previous, expert-driven approaches [22]. Such features can be used to predict both perceived [22] and subjective [23] emotional states. As exertion has both an objective component and a subjective one [15], we hypothesise these features to capture both.

To summarise, our work introduces the following contributions: a) We analyse the effect of exertion on speech production from both a subjective and an objective perspective, using two different measures (perceived exertion and heart rate) obtained concurrently from speech *while* running. b) We expand beyond the usual indoor treadmill scenarios to also include naturalistic, outdoors recordings. c) We introduce an automated prediction algorithm for objective and subjective measures of exertion alongside traditional feature analysis. We present our methodology in Section 2, followed by our results in Section 3 and our concluding remarks in Section 4.

10.21437/Interspeech.2023-470

2. Methodology

This section outlines the data collected for this work, as well as our analysis and modelling methodology.

Dataset: For our experiments, we use a dataset of running speech [14], which includes multimodal recordings of biomechanical, HR, and speech data from 46 runners (f: 27, m: 19; mean age: 40 years; mean body-mass index (BMI): 23). The dataset was collected in 185 running sessions both indoors (treadmill) and outdoors (asphalt, forest roads, etc.). The HR data were recorded in beats per minute via a Polar H9 chest strap with a sample rate of 1 Hz, whereas the audio was recorded by the internal microphone of a smartphone strapped to the runners' arm at 16 kHz. Every participant conducted $1-5$ runs (median: 4) with an average duration of 33 minutes.

The runners were prompted at regular intervals (ca. 5 minutes) to give feedback on: a) their wellbeing level ($[-5, 5]$ scale), b) their self-perceived exertion level ($[6-20]$ rating of perceived exertion (RPE) scale), and c) the surface they are running on (free text; in the form of "I am running on..."). The answers were given *while* running. Our unit of analysis was their surface answers, as these are typically longer and, crucially, do not contain information about the variables we are trying to predict. As the surface answer comes mere seconds after the exertion one, it is reasonable to assume that the label remains the same. We also averaged the continuous HR values over the duration of the runners' answers; this resulted in a single label per utterance, as for the perceived exertion. In total, this resulted in 848 instances with corresponding RPE (mean: 11.9) and HR (mean: 154 BPMs) labels.

Features: From each audio recording, we use two standardised feature sets, the extended Geneva minimalistic acoustic parameter set (eGeMAPS) [21] and the Interspeech Computational Paralinguistics ChallengE feature set (ComParE) [24], and two contemporary, deep learning-based data representations, derived from *w2v2* [25]. eGeMAPS is an 88-dimensional, interpretable feature set comprising functionals of acoustic and prosodic low-level descriptors (LLDs) [21], while ComParE is a 6 373-dimensional brute-forced one. Both subsets were extracted using the *openSMILE* toolkit [26]. Moreover, we used the embeddings of *w2v2-l-r* [27] and *w2v2-l-emo* [22], two variants of *w2v2* [25], with the first trained on more data and the second additionally fine-tuned for dimensional emotion recognition. As *w2v2-l-emo* showed exceptional performance on the recognition of perceived emotional arousal [22], we expect it to contain useful information for the prediction of HR – an objective measure of runner intensity and, thus, of physiological arousal. Furthermore, it has been shown to be a robust predictor of self-perceived mood [23], and thus expected to show a correlation to perceived exertion as well. As features, we averaged the contextualised embeddings (output of the penultimate layer) over the time dimension.

Feature analysis: Feature analysis in prior work is usually conducted under a classification paradigm [4, 6], with recordings being categorised in either pre- vs post-exercise [4] or high vs low intensity [6]; the authors then proceed to analyse the differences between the two states. However, our recording procedure does not allow for a categorisation, as the RPE scale is continuous and reported at regular intervals, so we resorted to correlation analysis.

We are concerned with two research questions: a) whether there are differences in the manifestation of objective vs subjective exertion in speech, and b) whether there are individual trends in those manifestations, as reported in [6]. To answer those, we restricted our analysis to runners satisfying the following criteria: i) ones who have at least 10 recorded instances, thus lending stability to our results, and ii) ones who have a minimum reported RPE less than 10 and a maximum of more than 14, thus ensuring that the runners are recorded in both low and high levels of intensity. This resulted in 25 runners (14f, 11m; mean age: 36, mean BMI: 23).

Table 1: *Pearson correlation coefficient (PCC) between selected eGeMAPS features and RPE or HR computed over selected speakers. Correlation coefficients are statistically significant at the .05 level using a Wald test for both targets.*

Feature	RPE	HR
F0 (all - std)	0.141	0.259
Loudness (all - mean)	-0.131	-0.175
$MFCC_1$ (all - mean)	0.094	0.101
Alpha ratio (voiced - mean)	-0.129	-0.099
Hammarberg index (voiced - mean)	0.128	0.146
Slope [0-500 Hz] (voiced - mean)	-0.114	-0.212
Spectral flux (voiced - std)	0.117	0.116
Segment length (voiced - mean)	-0.101	-0.129

After selecting this reduced set of runners, we proceeded to compute the Pearson correlation coefficient (PCC) between all 88 features in eGeMAPS and both the RPE scale and HR, and subsequently tested for significance using the Wald test. We kept all features for which the null hypothesis is rejected at the .05 significance level for both RPE and HR; this yielded a total of 21 features. As most of those features are correlated with each other (different functionals of the same LLDs; same feature computed over different regions), we only present and interpret a selection to minimise redundancy, where we chose the most interpretable functionals (e. g., the mean or standard deviation) and the wider possible region (e. g., the entire signal over voiced regions only). Furthermore, to answer our second research question, we proceeded with computing the PCC for the selected features over the data of each runner separately, this time though only for the RPE scale.

Automatic modelling: For the automatic modelling of HR and RPE, we trained feed-forward deep neural networks (DNNs) on our features. The experiments were conducted via 5-fold speaker-independent cross-validation. Our DNNs comprise 4 fully-connected layers, each with a hidden size of 30, and are trained for 100 epochs. Each layer, except the last, is followed by a dropout of probability 50 % and a ReLU activation. The models are optimised with SGD with an initial learning rate of .001, which is reduced by 10 % after five epochs without improvement on the development set, a batch size of 16, and a Nesterov momentum of .9. We additionally used a weight decay of .0001. The loss function is the concordance correlation coefficient (CCC) loss, which has been shown to work well for exertion [14, 28, 29]. The model state from the best epoch was selected for evaluation on the test set.

3. Results

We begin by presenting the feature analysis results, first for the entire dataset in Table 1 and then for individual runners in Fig. 1. Our analysis and subsequent selection yielded the following features: the standard deviation of F0, the mean of loudness over all frames, the mean of $MFCC_1$ over all frames, the

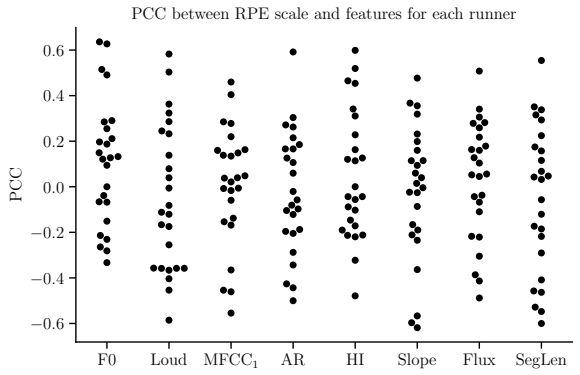

Figure 1: *Speaker-level PCC between exertion and the different features: F0, (Loud)ness, MFCC$_1$, (A)lpha (R)atio, (H)ammarberg (I)ndex, Spectral (Slope), Spectral (Flux), and (Seg)ment (Len)gth. Each point in the swarm plot represents an individual runner.*

mean of the alpha ratio over voiced frames, the mean of spectral slope in the range 0 Hz-500 Hz over voiced frames, the standard deviation of spectral flux over voiced frames, and the average duration of voiced segment length. Our first observation from Table 1 is that both the RPE scale and the HR showcase identical trends; however, the correlations are almost always more pronounced for the case of HR, which is consistent with previous observations showing that physiological indicators might manifest more clearly in speech [30]. We now proceed to interpret the trends presented in Table 1.

We observe an increase of F0 standard deviation with increasing exertion, which is counter to the previous observation that exertion increases sub-glottal pressure and thus increases F0 and limits its range [4]. However, this is contradicted by the decrease in loudness, which is heavily correlated with sub-glottal pressure; therefore, the observed increase in F0 standard deviation might simply be evidence of strain.

The increase of MFCC$_1$ along with exertion indicates a relative shift of energy towards the lower frequencies; this is corroborated by the concurrent hike of the Hammarberg index (ratio of the strongest peak in [0 kHz-2 kHz] vs the strongest peak in [2 kHz-5 kHz]) and the decrease in spectral slope. This trend is contradicted by the decrease in alpha ratio (ratio of summed energy from [50 Hz-1000 Hz] over [1k Hz-5 kHz]), which shows instead a shift towards higher frequencies – a trend which would be consistent with the expected increase in breathiness due to higher exertion. We hypothesise that a higher exertion might correlate with a higher running speed, which in turn results in more (and perhaps louder) steps being picked up by the microphone; this effect also showed in a preliminary manual inspection of the audio files. As the sound produced by those steps is concentrated in lower energies, this could introduce some unwanted noise in our interpretation. In particular, the maximum in the Hammarberg index might be more susceptible to it than the means or sums of the other functionals, whereas the ranges used by spectral slope and emphasised by the cosine weights in MFCC$_1$ would also add more weights to those frequencies. This interpretation is consistent with previous results showing that exertion can be predicted from the surrounding, non-speech audio [14], and it is something we intend to follow up in future work.

Table 2: *Pearson correlation coefficient (PCC) for the automatic prediction of RPE and HR using different speech features. For RPE, we additionally include a linear regression baseline using the ground truth HR as the predictor, as well as the late fusion of the best-performing speech model with this baseline. Mean and standard deviation results over 5 folds.*

Features	**RPE**	**HR**
HR baseline	.553 (.090)	N/A
eGeMAPS	.158 (.144)	.298 (.028)
ComParE	.229 (.077)	.380 (.136)
w2v2-l-r	.259 (.112)	.418 (.044)
w2v2-l-emo	.341 (.062)	.416 (.037)
w2v2-l-emo + HR	.501 (.053)	N/A

The last two features, spectral flux and length of voiced segments, are also related to articulation: the former indicates a higher rate of articulation while the latter points to the presence of more pauses [18]. Both patterns are consistent with our expectation of increased breathing cycles and thus a need for more pauses, which could accordingly increase the articulation rate to convey the same amount of information [4]. Additionally, voiced segment length decreases with increased exertion and HR, as was found in previous work [4, 18], and is an indication of more noise-like components in the signal due to increased breath emission levels [8]. However, we note that the same caveat as before applies to spectral flux, namely that it might be affected by an increase in running speed.

Finally, we examine the correlations computed over the data of individual runners in Fig. 1. There, we show the correlation between each runner and each feature as a single point in a swarm plot. All correlations between features and the RPE scale span a range between $[-0.6, 0.6]$, which lends evidence to previous work claiming that the manifestation of exertion in speech is heavily individualised [6, 18]. This interesting observation serves as preliminary evidence that the prediction of subjective exertion will greatly benefit from personalised models – a modelling approach which has already been explored for other modalities [29]. This needs to be further elucidated in a follow-up analysis which also checks for potential confounders, such as the underground surface, which has been shown to influence the audio [28].

After interpreting the impact of exertion on individual features, we proceed with modelling both RPE and HR using the experimental setup discussed in Section 2. Our results are presented in Table 2. We additionally include a simple, linear regression baseline using HR as the single predictor for RPE – this constitutes the state-of-the-art which utilises physiological measures as a proxy for subjective exertion. We note that our correlation is low compared to established knowledge; in fact, the design of the RPE scale is supposed to correspond exactly with the HR [15]. However, we are only able to obtain a PCC of .553 on our data; this moderate correlation is potentially explained by the naturalistic conditions under which we collected our dataset, and has been observed in other studies as well [31], highlighting the importance of predicting perceived exertion separately.

Moving to our speech-based results, we observe consistently higher performance when predicting HR than when predicting the RPE scale – as also seen in the preceding correla-

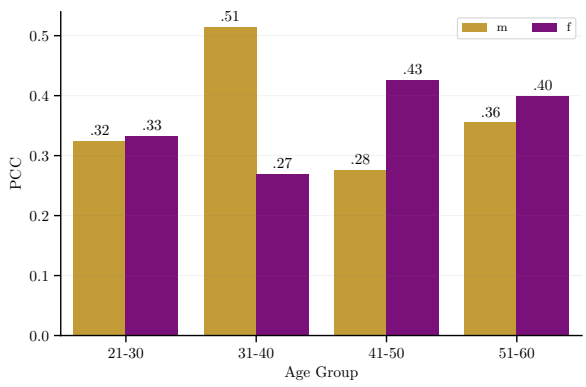

Figure 2: *Intersectional model performance of w2v2-l-emo for different age groups and genders on subjective exertion.*

tion analysis. This once again showcases that the former manifests more strongly in speech than the latter. The best performance is shown by *w2v2* variants. Of those, *w2v2-l-emo* shows a considerably higher, and more stable, correlation than *w2v2-l-r* on RPE (.341 vs .359), while both features are almost equivalent for HR prediction (.416 vs .418). The fact that the gains are observed only for perceived exertion, but not for the objective measure of HR, points to the fact that emotional information might be beneficial only for this task; this is exploited by *w2v2-l-emo*, which is just a derivative of *w2v2-l-r*, but additionally fine-tuned for dimensional emotion recognition [22]. This is followed by ComParE (.229/.380), which improves over eGeMAPS (.158/.298) but still lags behind the two learnt feature sets.

We additionally evaluated the disaggregated performance of our model with respect to age and gender, where we cluster the age in decades and use a self-reported binary gender categorisation. Results are presented in Fig. 2, where we show the PCC per gender-age group for exertion prediction using the best-performing *w2v2-l-emo* model, averaged over all 5 folds. The model shows no consistent bias in favour of one gender group or another, as male performance is higher for some age groups and female for others, even though there are more females in the dataset. There is a trend towards better performance for subjects aged in the middle of the scale.

The causes for this discrepancy in performance are not readily apparent. We found no significant differences in the ground-truth exertion rates for the two genders (two-sided Mann-Whitney U test). We did find small negative correlations of exertion with age ($-.165$) and years of running experience ($-.096$), which were statistically significant with a Wald test. As age and years of experience are highly correlated (.495), we think the correlation with age is caused because more experienced runners are a) either better able to regulate their speed and remain at lower exertion levels, or b) more "used" to running at extreme levels and so rate their exertion lower. a) seems to be better backed by data as there is a stronger negative correlation between age and hr ($-.309$), though this is confounded by other factors [32]. We hypothesise that these differences make for differing sub-populations in our group of runners. Naturally, our results are impacted by the relatively limited size of some sub-populations and should be verified in other, larger datasets.

Finally, even our best-performing method for predicting subjective exertion falls short of the HR regression baseline, which achieves a PCC of .533 An attempt to further improve results via a late fusion (averaging) between the predictions of our best-performing speech model, *w2v2-l-emo*, and those of the HR baseline also fails to improve upon the baseline. This illustrates that, for the time being, speech lags far behind the industry standard of HR as a proxy for exertion – which makes sense given that the RPE scale was explicitly designed to closely follow HR [15]. To shed further light into this, we compute the correlation of *w2v2-l-emo* predictions trained to predict exertion with HR, resulting in an average correlation of .372 over the five folds – higher than the correlation of the model with the target it was trained to predict. This illustrates how models trained to predict subjective measures of exertion are primarily picking up on cues informed by objective measures, and thus fail to uncover any additional, subjective information. A by-product of this is that the late fusion of a speech-based and an HR-based predictor fails to improve performance. Therefore, further work is required to disentangle the two variables and uncover these more subjective aspects of exertion.

However, it remains the case that obtaining reliable measurements of HR requires specialised hardware which might not be available to the average runner, whereas our approach can be deployed on a standard smartphone; thus, the more widespread availability of our solution partially counterbalances the loss in performance. Overall, we expect to see big gains when fine-tuning these models in-domain and coupling them with a speech enhancement frontend to remove unwanted interference by other sources (e. g., steps, wind, etc.), both of which we intend to pursue in future work. This should allow us to bridge the remaining gap between speech-based exertion prediction and the HR baseline which represents the current state-of-the-art.

4. Conclusion

We presented an empirical analysis on the impact of both objective (heart rate) and subjective exertion on speech collected while running in indoor and outdoor environments. Our findings show that heart rate, as a physiological measure of exertion, is more clearly manifested in speech than self-assessment using the Borg RPE scale, with a best average PCC of .418 vs one of .341. These results were obtained with embeddings generated from pre-trained transformer models that have been fine-tuned for dimensional emotion recognition. Our accompanying exploratory data analysis partially matches our expectations of shorter breathing cycles, which lead to more pauses, higher breathiness and articulation rate, and sub-glottal strain manifesting in an increase of F0 variability – but the potentially confounding effects of background noise need to be further elucidated in future work. Moreover, we identified individualised trends in the manifestation of exertion.

Follow-up work is needed to disentangle the effects of background noise on speech features using denoising, which would have the added benefit of improving overall performance. Furthermore, multimodal, personalised approaches could be employed to provide a more holistic characterisation of exertion by integrating data from all available sensors. Finally, we plan to expand our dataset to include more runners in order to increase the robustness and validity of our findings.

5. Acknowledgment

This work was funded from the DFG project No. 442218748 (AUDI0NOMOUS) and the Zentrales Innovationsprogramm Mittelstand (ZIM) grant agreement No. 16KN069455 (KIRun).

6. References

[1] A. Triantafyllopoulos, A. Kathan, A. Baird, L. Christ, A. Gebhard, M. Gerczuk, V. Karas, T. Hübner, X. Jing, S. Liu, *et al.*, "Hear4health: A blueprint for making computer audition a staple of modern healthcare," *arXiv preprint arXiv:2301.10477*, 2023.

[2] S. Baker, J. Hipp, and H. Alessio, "Ventilation and speech characteristics during submaximal aerobic exercise," *Journal of Speech, Language, and Hearing Research*, vol. 51, no. 5, pp. 1203–1214, 2008.

[3] 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. INTERSPEECH*, ISCA, Dresden, Germany, 2015, pp. 3705–3709.

[4] J. Trouvain and K. P. Truong, "Prosodic characteristics of read speech before and after treadmill running," in *Proc. INTERSPEECH*, Dresden, Germany, 2015.

[5] S. Deb and S. Dandapat, "Fourier model based features for analysis and classification of out-of-breath speech," *Speech Communication*, vol. 90, pp. 1–14, 2017.

[6] S. Sahoo and S. Dandapat, "Analyzing the vocal tract characteristics for out-of-breath speech," *The Journal of the Acoustical Society of America*, vol. 150, no. 2, pp. 1524–1533, 2021.

[7] A. Gebhard, S. Amiriparian, A. Triantafyllopoulos, A. Kathan, S. Ottl, M. Gerczuk, M. Jaumann, D. Hildner, V. Dieter, P. Schneeweiss, I. Rösel, I. Krauss, and B. W. Schuller, "Towards heart rate categorisation from speech in outdoor running conditions," in *Proc. E-Health and Bioengineering Conference (EHB)*, to appear, IEEE, Iasi, Romania, 2022, pp. 1–6.

[8] S. Deb and S. Dandapat, "Analysis of out-of-breath speech for assessment of person's physical fitness," *Computer Speech & Language*, vol. 76, p. 101 391, 2022.

[9] R. Van Gent, D. Siem, M. van Middelkoop, A. Van Os, S. Bierma-Zeinstra, and B. Koes, "Incidence and determinants of lower extremity running injuries in long distance runners: A systematic review," *British Journal of Sports Medicine*, vol. 41, no. 8, pp. 469–480, 2007.

[10] B. Saragiotto, T. Yamato, L. Junior, M. Rainbow, I. Davis, and A. Lopes, "What are the main risk factors for running-related injuries?" *Sports Medicine*, vol. 44, no. 8, pp. 1153–1163, 2014.

[11] S. D. Anton, M. G. Perri, J. Riley, W. F. Kanasky, J. R. Rodrigue, S. F. Sears, and A. D. Martin, "Differential predictors of adherence in exercise programs with moderate versus higher levels of intensity and frequency," *Journal of Sport and Exercise Psychology*, vol. 27, no. 2, pp. 171–187, 2005.

[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. ICASSP*, IEEE, New Orleans, LA, USA, 2017, pp. 989–993.

[13] A. Jati, P. G. Williams, B. Baucom, and P. Georgiou, "Towards predicting physiology from speech during stressful conversations: Heart rate and respiratory sinus arrhythmia," in *Proc. ICASSP*, IEEE, Calgary, AB, Canada, 2018, pp. 4944–4948.

[14] A. Triantafyllopoulos, S. Ottl, A. Gebhard, E. Rituerto-Gonzalez, M. Jaumann, S. Hüttner, V. Dieter, P. Schneeweiss, I. Krauss, M. Gerczuk, S. Amiriparian, and B. Schuller, "Fatigue prediction in outdoor running conditions using audio data," in *Proc. EMBC*, IEEE, Glasgow, UK: IEEE, 2022, pp. 2623–2626.

[15] G. A. Borg, "Psychophysical bases of perceived exertion.," *Medicine and Science in Sports and Exercise*, 1982.

[16] C. Foster, J. P. Porcari, J. Anderson, M. Paulson, D. Smaczny, H. Webber, S. T. Doberstein, and B. Udermann, "The talk test as a marker of exercise training intensity," *Journal of Cardiopulmonary Rehabilitation and Prevention*, vol. 28, no. 1, pp. 24–30, 2008.

[17] J. H. Hansen and S. Patil, "Speech under stress: Analysis, modeling and recognition," *Speaker classification I: Fundamentals, features, and methods*, pp. 108–137, 2007.

[18] K. W. Godin and J. H. Hansen, "Analysis and perception of speech under physical task stress," in *Proc. INTERSPEECH*, Brisbane, QLD, Australia, 2008, pp. 1674–1677.

[19] ——, "Physical task stress and speaker variability in voice quality," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–13, 2015.

[20] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[21] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[22] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Eyben, and B. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–13, 2022.

[23] M. Song, A. Triantafyllopoulos, Z. Yang, H. Takeuchi, T. Nakamura, A. Kishi, T. Ishizawa, K. Yoshiuchi, X. Jing, Z. Zhao, V. Karas, K. Qian, B. W. Schuller, and Y. Yamamoto, "Daily mental health monitoring from speech: A real-world Japanese dataset and multitask learning analysis," in *Accepted in ICASSP 2023*, Rhodos, Greece, 2023.

[24] 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 *Proc. INTERSPEECH*, Lyon, France, 2013.

[25] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proc. NeurIPS*, vol. 33, pp. 12 449–12 460, 2020.

[26] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proc. ACMMM*, Florence, Italy, 2010, pp. 1459–1462.

[27] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," *arXiv preprint arXiv:2104.01027*, 2021.

[28] A. Gebhard, A. Triantafyllopoulos, S. Amiriparian, S. Ottl, V. Dieter, M. Gerczuk, M. Jaumann, D. Hildner, P. Schneeweiß, I. Rösel, I. Krauß, and B. W. Schuller, "Improving exertion and wellbeing prediction in outdoor running conditions using audio-based surface recognition," in *Proc. International Workshop on Multimedia Content Analysis in Sports (MMSports) at ACMMM*, 9 pages, ACM, Lisbon, Portugal: ACM, Oct. 2022.

[29] A. Kathan, A. Triantafyllopoulos, S. Amiriparian, A. Gebhard, S. Ottl, M. Gerczuk, M. Jaumann, D. Hildner, V. Dieter, P. Schneeweiss, I. Rösel, I. Krauss, and B. W. Schuller, "Investigating individual- and group-level model adaptation for self-reported runner exertion prediction from biomechanics," in *Proc. E-Health and Bioengineering Conference (EHB)*, to appear, IEEE, Iasi, Romania, 2022, pp. 1–4.

[30] A. Triantafyllopoulos, S. Zänkert, A. Baird, J. Konzok, B. M. Kudielka, and B. W. Schuller, "Insights on modelling physiological, appraisal, and affective indicators of stress using audio features," in *Proc. EMBC*, IEEE, Glasgow, UK: IEEE, 2022, pp. 2619–2622.

[31] S. G. Karavatas and K. Tavakol, "Concurrent validity of borg's rating of perceived exertion in african-american young adults, employing heart rate as the standard," *Internet Journal of Allied Health Sciences and Practice*, vol. 3, no. 1, p. 5, 2005.

[32] I. Antelmi, R. S. De Paula, A. R. Shinzato, C. A. Peres, A. J. Mansur, and C. J. Grupi, "Influence of age, gender, body mass index, and functional capacity on heart rate variability in a cohort of subjects without heart disease," *The American Journal of Cardiology*, vol. 93, no. 3, pp. 381–385, 2004.