Hybrid Handcrafted and Learnable Audio Representation for Analysis of Speech Under Cognitive and Physical Load

Gasser Elbanna\textsuperscript{1,2*}, Alice Biryukov\textsuperscript{1,2}, Neil Scheidwasser-Clow\textsuperscript{1}, Lara Orlandic\textsuperscript{1}, Pablo Mainar\textsuperscript{2}, Mikolaj Kegler\textsuperscript{3}, Pierre Beckmann\textsuperscript{1}, Milos Cernak\textsuperscript{2}

\textsuperscript{1}École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
\textsuperscript{2}Logitech Europe S.A., Lausanne, Switzerland
\textsuperscript{3}Imperial College London, London, United Kingdom
\textsuperscript{4}Université de Lausanne, Lausanne, Switzerland

gasser.elbanna@epfl.ch, milos.cernak@ieee.org

Abstract

As a neurophysiological response to threat or adverse conditions, stress can affect cognition, emotion and behaviour with potentially detrimental effects on health in the case of sustained exposure. Since the affective content of speech is inherently modulated by an individual’s physical and mental state, a substantial body of research has been devoted to the study of paralinguistic correlates of stress-inducing task load. Historically, voice stress analysis has been conducted using conventional digital signal processing (DSP) techniques. Despite the development of modern methods based on deep neural networks (DNNs), accurately detecting stress in speech remains difficult due to the wide variety of stressors and considerable variability in individual stress perception. To that end, we introduce a set of five datasets for task load detection in speech. The voice recordings were collected as either cognitive or physical stress was induced in the cohort of volunteers, with a cumulative number of more than a hundred speakers. We used the datasets to design and evaluate a novel self-supervised audio representation that leverages the effectiveness of handcrafted features (DSP-based) and the complexity of data-driven DNN representations. Notably, the proposed approach outperformed both extensive handcrafted feature sets and novel DNN-based audio representation learning approaches.

Index Terms: computational paralinguistics, voice stress analysis, audio representation learning, DSP features, deep learning

1. Introduction

The emergence of machine learning methods for the paralinguistic analysis of speech signals has been of considerable influence for healthcare-related applications. This includes both the detection of long-term traits such as speech pathology \cite{1} and shorter-term mental states such as emotion, effort, and acute stress \cite{2,3}. For the latter, inter-individual differences in stress perception constitutes, among others, one of the major limitations of voice stress analysis (VSA) studies \cite{4}.

Among the main sources of acute stress \cite{5}, effort-induced stress can be elicited by increased physical or cognitive demands to complete the task (i.e., physical and cognitive load, respectively). Both behavioural states have been the subject of numerous paralinguistic studies \cite{6,7,8,9}. On the one hand, physical load is known to create notable differences in speech production, especially due to competition between breathing processes for speech generation and increased oxygen consumption to meet metabolic needs \cite{6}. Regarding voice-related low-level descriptors (LLDs), several studies have reported increases in mean fundamental frequency after intensive physical exercise \cite{7,8}. Similar effects on speech production have been observed in scenarios eliciting cognitive load \cite{7,9}. As a matter of fact, paralinguistic studies of stress have historically relied on handcrafted acoustic feature sets based on digital signal processing (DSP). Such feature sets could then be used for classification of cognitive or physical load levels using various classical machine learning approaches, including GMM supervisor systems \cite{10}, i-vector frameworks \cite{11}, and ensemble learning \cite{12}.

Alternatively, several deep learning frameworks have been proposed for cognitive and physical load detection \cite{13,14,15}. Yet, training such models directly on task load corpora remains a challenging issue due to their small sizes, as larger networks tend to overfit due to their higher model capacity. To overcome this issue, several studies have shown that pre-training models on large corpora of speech or audio can produce state-of-the-art performance in various data-constrained speech processing and paralinguistic tasks such as speech emotion recognition \cite{16}, speaker or language identification \cite{17,18,19}.

With that in mind, in this paper, we focus on studying the detection of cognitive and physical load from voice recording using data-driven audio representation models. We introduce four novel datasets related to cognitive load with varying task complexities, languages, and protocols, and one dataset related to physical load. We used the datasets to evaluate a novel hybrid training protocol for general-purpose speech representation (Figure 1). The proposed approach aims at taking advantage of both self-supervised representation learning and non-trainable DSP-based feature sets, which simply combine outputs from different mod-
els, our method incorporates the two models during the training phase to learn a new and unique representation better than the sum of its parts. We employed the proposed model to detect cognitive and physical stress from voice recordings in the above-outlined datasets. We showed that this hybrid speech representation generalizes across different datasets and produces better performance than existing DSP-based, as well as DNN-based representation learning methods. Importantly, the proposed hybrid model yielded better results than the ensemble of the methods used for its pre-training.

2. Materials & Methods

2.1. Tasks and Datasets

The speech datasets proposed in this study vary in language, duration, protocol, and more importantly task complexity. A brief description of the datasets is addressed in Table 1. Cognitive load was induced in the speakers through multitasking, based on the idea that cognitive load increases with the number of tasks to be performed simultaneously [20][21]. On the other hand, physical load was induced by recording speech samples before and after the speaker performed an intensive aerobic physical exercise (VO2 max running test). In each dataset, the recorded speech samples were labelled in accordance with the experimental condition (i.e., with load / without load).

**Cognitive Load 1 (Cog1)** - This dataset contains 21 speakers (3 female, 18 male) who participated in their native language (English or French). The experiment comprises six blocks. The first block consisted of a scripted speech task, in which participants read aloud at most three sentences from SIWIS, a multilingual speech and text database [22]. The other five blocks included the same scripted speech task, but this time participants were instructed to count spoken syllables mixed in additive Gaussian white noise (AGWN) while reading aloud. Syllables were selected from the Articulation Index dataset [23]. Each block contained five trials lasting 10 seconds each, producing approximately five minutes of speech material per speaker.

**Cognitive Load 2 (Cog2)** - 20 Chinese Mandarin speakers (10 female, 10 male) were recruited to perform a similar task to the Cog1 task. The experiment included 40 trials per participant: 20 of them included only reading scripted text from randomly selected articles across various media sources, while the other 20 trials comprised an additional working memory task, as participants had to listen to and memorise three numeric digits while reading. Each trial lasted 15 seconds, giving 10 min of speech material per participant.

**Cognitive Load 3 (Cog3)** - This task was identical to Cog2, except for the working memory task, as participants were asked to memorize alphanumeric digits (e.g., “A5” instead of “5” in Cog2). The dataset comprised 40 15-second trials from 20 speakers, producing 10 min of speech data per speaker.

**Cognitive Load 4 (Cog4)** - This dataset was collected from 22 subjects (7 female, 15 male). Audio recordings were collected as participants performed an experiment in which cognitive load was induced based on multitasking, with an experimental design inspired by [24]. The experiment was carried out in three blocks, a baseline block lasting four minutes, followed by two multitasking blocks lasting two minutes each. In the baseline block, participants were shown an image on a screen and were asked to describe it in as much detail as possible in their native language. The images shown on the screen were presented in colour and included cityscapes, landscapes, and famous paintings. Every 30 seconds, a new image appeared, and the participants began a new recording. In the multitasking blocks, a high level of cognitive load was applied by asking participants to perform four tasks simultaneously, in addition to describing the image appearing on the screen. The four tasks consisted of an auditory response task, a visual vigilance task, an arithmetic task, and a memory task. The participants were instructed to pay equal attention to all tasks and perform them simultaneously to the best of their ability, while at the same time describing the image on the screen in as much detail as possible. For these blocks, images were selected from commonly used images to study various mental and cognitive impairments through speech (e.g., Cookie-theft picture [25]). As in the baseline block, the presented image changed every 30 seconds. Each participant produced a total of eight minutes of speech data.

**Physical Load (Phys)** - 28 participants (13 female, 15 male) were recruited to perform a VO2 max running test [26]. Before and after the exercise, participants were instructed to speak spontaneously in their native language (English or French) for three minutes about personal topics, e.g., their current work, their favourite physical activities, or their daily routine. Subsequently, 10 utterances were extracted from each speaker, five for each condition (before and after exercise), where each utterance lasted between 5 and 15 seconds.

2.2. Audio Representation Models

All cognitive and physical load speech datasets were used as downstream tasks to benchmark different deep learning-based audio representations against well-established handcrafted acoustic feature sets. Importantly, the data-driven models were not fine-tuned directly on the task-specific data, but only pre-trained on independent datasets.

2.2.1. Handcrafted models

**openSMILE (OS)** - openSMILE [27] is an open-source feature extraction toolkit that generates handcrafted low-level descriptors (LLD) of audio inputs. This toolkit comprises three feature sets obtained by computing functionals on LLD contours. The most prominent of them is ComParE (OS/ComParE) [2], which contains 6373 static features. Another commonly used feature set is eGeMAPS (OS/eGeMAPS) [29] which includes 88 features. Both approaches are purely DSP-based and require no pre-training on auxiliary datasets.

2.2.2. Data-driven models

**VGGish** - VGGish [29] is a commonly used audio feature extractor derived from the convolutional neural network VGG-16 [30]. This model was pre-trained in a supervised fashion to classify audio events from the Youtube-8M dataset [31].

**YAMNet** - YAMNet [32] is a DNN pre-trained to classify events from AudioSet, a large collection of audio samples from YouTube videos [33]. Its feature extractor employs the MobileNetV1 convolutional architecture [34].

**TRILL, layer 19** - TRILL [17] is a model pre-trained in a self-supervised manner on speech samples from AudioSet. The CNN model based on ResNetish [29] was originally designed as a universal speech representation. It uses triplet-loss learning to create an embedding space where temporally adjacent samples are mapped close together. Here, we use the features generated from layer 19, as found by [17] to perform best on non-semantic speech tasks.

**BYOL-A** - Unlike contrastive learning frameworks, Bootstrap Your Own Latent for Audio (BYOL-A) [18] generates audio
representations using two augmented views of a single audio sample, inspired by the success of BYOL \[35\] for image representation. To obtain audio representations, the log-mel spectrogram (LMS) of an input audio sample is first fed to a data enhancement module, producing two randomly augmented copies of the input LMS (Figure 1). Subsequently, the augmented spectrograms are respectively passed to an online and a target network. Both networks share a similar architecture, with an encoder and a projector module. However, the online network comprises an additional module called predictor to avoid having collapsed representations \[35\]. The aim of this training paradigm is for the online network to predict the output generated by the target network by minimising a mean squared error (MSE) loss function. BYOL-A has achieved competitive performance in various speech processing tasks, e.g., language and speaker identification.

**BYOL-A Extensions** - Following the same implementation of BYOL-A \[16\], we proposed in a previous work BYOL-S \[35\]. BYOL, for speech, which was pre-trained on a speech subset of AudioSet. In addition, we modified the default CNN encoder architecture used in BYOL-A (a 5-layer CNN) using a lightweight version of the Convolutional vision Transformer (CvT) \[36\] to leverage the best of CNNs and Transformers. All proposed encoders produce an embedding of size 2048. The proposed models and associated methods are openly available online.

**2.2.3. Hybrid models**

Considering the effectiveness of handcrafted features for various paralinguistic challenges \[2,3\], we derived a new training protocol for BYOL-S by combining hand-made and data-driven features in a hybrid format. More specifically, we augmented the BYOL framework with a supervision network which generates DSP-based features extracted from the ComParE feature set in openSMILE, as shown in Figure 1. Therefore, in addition to the self-supervised BYOL loss ($L_{ss}$), we computed a supervised loss $L_{sup}$ between the outputs of the supervised and online networks, set as a MSE. The final loss function was defined as a weighted sum of the self-supervised and supervised losses:

$$L_{hybrid} = \alpha_{ss}L_{ss} + \alpha_{sup}L_{sup}$$

(1)

where $\alpha_{ss}$ and $\alpha_{sup}$ weigh the parts of the self-supervised and supervised losses, respectively.

Given the size of the ComParE feature set (6373), we projected the BYOL-S model outputs, from both online and target networks, to produce 6373-dimensional embeddings to compute the hybrid loss ($L_{hybrid}$). However, the output size of the pre-trained encoder module remained unchanged (2048).

2.3. Evaluation Pipeline

All models were evaluated on the five cognitive and physical load tasks mentioned above. Audio samples were processed to produce their representation vectors per model. Each downstream dataset was divided into training and test sets with ratios of 70% and 30%, respectively. To ensure a fair and robust evaluation, the partitions were constructed to be speaker-independent (i.e., all utterances from a given speaker belonged to either the training or the test set) and to have a similar gender distribution. Each partition was also standardised (to zero mean and unit variance) to alleviate inter-speaker variability. Subsequently, classification was performed by training and validating linear support-vector machines (SVMs) using five-fold cross-validation. A grid search was applied to optimise the penalty hyperparameter, with values ranging from $10^{-5}$ to $10^5$. Using such a simple classifier with linear kernel allows us to place more emphasis on the predictive power of the extracted features. Lastly, model performance was evaluated by computing the unweighted average recall on the test set to compensate for any imbalanced data distributions. All procedures in the evaluation pipeline were implemented using scikit-learn \[37\].

3. Results

Table 2 presents the performance of each model in all five tasks by reporting the unweighted average recall (UAR, in percentages) on an unseen and speaker-independent test set. Importantly, using the ComParE handcrafted features exhibited better performance than most deep learning-based models, including recent self-supervised models for speech and audio representation (TRILL, BYOL-A and BYOL-S). Only BYOL-S/CvT achieved better performance, although only slightly better (1.1% average improvement). This result highlights the effectiveness of DSP-based features in paralinguistic tasks.

On the other hand, hybrid representation models, which aim to take advantage of both data-driven and handcrafted features, yielded significantly better performance. The two proposed hybrid models, one using BYOL-A’s CNN encoder \[18\] and one with CvT encoding \[36\], consistently outperformed their corresponding BYOL-S models. This result suggests that adding the DSP-based supervision to the self-supervised representation learning framework helped to improve its generalization capacity in cognitive/physical load detection tasks.

As the proposed hybrid loss consists of a weighted sum of its self-supervised and supervised counterparts (Eq. 1), we reported model performance on the five datasets for different $\alpha_{ss} : \alpha_{sup}$ ratios. Here, we selected the Hybrid BYOL-S/CvT variant, which yielded the best overall performance. The results shown in Table 2 indicate a steady increase in model performance with increasing weight assigned to the supervised loss ($\alpha_{sup}$). The 1:2 loss ratio ($\alpha_{ss} : \alpha_{sup}$) yielded the best model for detecting cognitive and physical load in the considered tasks. Increasing the weight of supervised loss further led to a consistent decrease in performance across all tasks.

### Table 1: Cognitive/physical load speech datasets summary. ENG: English, FRA: French, CMN: Mandarin Chinese.

| Dataset | Language | Utterances | Speakers | Duration (h) | Speech data acquisition protocol |
|---------|----------|------------|----------|--------------|---------------------------------|
| Cog1    | ENG/FRA  | 586        | 21       | 1.75         | Count audio stimuli (syllables) while reading scripted text |
| Cog2    | CMN      | 800        | 20       | 3.33         | Memorise three numeric digits while reading scripted text |
| Cog3    | CMN      | 800        | 20       | 3.33         | Memorise three alphanumeric digits while reading scripted text |
| Cog4    | 9 languages | 349       | 22       | 2.93         | Describe an image while performing four cognitive tasks |
| Phys    | ENG/FRA  | 280        | 28       | 0.70         | Spontaneous speech before/after VO2 max running test |

[https://github.com/nttcslab/byol-a](https://github.com/nttcslab/byol-a)  
[https://github.com/lucidrains/vit-pytorch](https://github.com/lucidrains/vit-pytorch)  
[https://github.com/Neclow/SERAB](https://github.com/Neclow/SERAB)  
[https://github.com/nttcslab/byol-a](https://github.com/nttcslab/byol-a)
Finally, we further validated the proposed hybrid framework by examining the results obtained by concatenating ComParE features with either BYOL-S or BYOL-S/CvT embeddings ($\alpha = 1$). To leverage their efficacy, we proposed a novel hybrid approach obtained by combining data-driven features from a BYOL-derived model (BYOL-S/CvT) with handcrafted features extracted from openSMILE during pre-training. This hybrid training protocol outperformed all data-driven representations, likely reduces any potential overfitting to the pre-training dataset. As a matter of fact, such a hybrid model also produced highly competitive results within the HEAR benchmark, outperforming several state-of-the-art approaches.

The proposed hybrid representation learning approach could be easily incorporated into any existing data-driven audio representation learning methods. Moreover, both components of the proposed framework could be substituted with other (potentially larger/complex) models. In particular, the fixed DSP-based feature extractor could be replaced with another pre-trained DNN-based representation. In a similar vein to self-distillation approaches, the two learnable representations could be iteratively updated in the pre-training process to produce a potentially more robust representation. Alternatively, the DSP-based feature extractor could also generate a specific family of acoustic features (e.g., prosodic features) which could be incorporated into any existing data-driven audio representation learning methods. Moreover, both components of the proposed framework could be substituted with other (potentially larger/complex) models. In particular, the fixed DSP-based feature extractor could be replaced with another pre-trained DNN-based representation. In a similar vein to self-distillation approaches, the two learnable representations could be iteratively updated in the pre-training process to produce a potentially more robust representation. Alternatively, the DSP-based feature extractor could also generate a specific family of acoustic features (e.g., prosodic features) which could be beneficial for building more specialized audio representations.

Table 2: Test UAR (%) on cognitive and physical load datasets. $\{\}$ indicates the concatenation of feature vectors. OS: openSMILE.

| Model                          | Cog1 | Cog2 | Cog3 | Cog4 | Phys | Average |
|-------------------------------|------|------|------|------|------|---------|
| VGGish [29]                   | 64.0 | 69.6 | 90.0 | 77.1 | 55.0 | 71.1    |
| TRILL, layer 19 [17]          | 58.8 | 72.9 | 93.3 | 74.0 | 62.5 | 72.3    |
| OS/GeMAPS [25]                | 62.9 | 65.4 | 93.8 | 72.9 | 67.5 | 72.5    |
| BYOL-A [18]                   | 68.0 | 70.4 | 90.0 | 81.3 | 61.3 | 74.2    |
| YAMNet [32]                   | 58.1 | 71.7 | 99.2 | 83.3 | 62.5 | 75.0    |
| BYOL-S [16]                   | 65.3 | 72.9 | 91.7 | 80.2 | 66.3 | 75.3    |
| OS/ComParE [27]               | 78.7 | 74.6 | 100.0| 79.2 | 52.5 | 77.0    |
| BYOL-S/CvT [16]               | 79.2 | 80.0 | 100.0| 74.0 | 57.5 | 78.1    |

Table 3: Impact of hybrid loss weighting (Eq. 1) in Hybrid BYOL-S/CvT pre-training on test UAR (%).

| $\alpha_{\text{ss}} : \alpha_{\text{sup}}$ (Eq. 1) | Cog1 | Cog2 | Cog3 | Cog4 | Phys | Average |
|-------------------------------------------------|------|------|------|------|------|---------|
| $4 : 1$                                         | 73.5 | 71.3 | 97.5 | 87.5 | 60.0 | 77.9    |
| $2 : 1$                                         | 72.2 | 72.5 | 99.2 | 92.7 | 55.0 | 78.3    |
| $4 : 3$                                         | 76.7 | 70.0 | 98.3 | 88.5 | 62.5 | 79.2    |
| $1 : 1$                                         | 78.4 | 74.6 | 99.2 | 83.3 | 66.3 | 80.3    |
| $3 : 4$                                         | 80.2 | 75.0 | 99.2 | 86.5 | 65.0 | 81.2    |
| $1 : 2$                                         | 83.8 | 72.5 | 98.3 | 87.5 | 70.0 | 82.4    |
| $1 : 4$                                         | 81.4 | 72.5 | 97.9 | 86.5 | 63.8 | 80.4    |
