Speaker Embeddings as Individuality Proxy for Voice Stress Detection

Zihan Wu[1,2], *Neil Scheidwasser-Clow*[3], *Karl El Hajal*[1], *Milos Cernak*[2]

[1]École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[2]Logitech Europe S.A., Lausanne, Switzerland
[3]University of Copenhagen, Copenhagen, Denmark

`zihan.wu@epfl.ch, milos.cernak@ieee.org`

Abstract

Since the mental states of the speaker modulate speech, stress introduced by cognitive or physical loads could be detected in the voice. The existing voice stress detection benchmark has shown that the audio embeddings extracted from the Hybrid BYOL-S self-supervised model perform well. However, the benchmark only evaluates performance separately on each dataset, but does not evaluate performance across the different types of stress and different languages. Moreover, previous studies found strong individual differences in stress susceptibility. This paper presents the design and development of voice stress detection, trained on more than 100 speakers from 9 language groups and five different types of stress. We address individual variabilities in voice stress analysis by adding speaker embeddings to the hybrid BYOL-S features. The proposed method significantly improves voice stress detection performance with an input audio length of only 3–5 seconds.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

Recent developments in neural network-based models of speech have spearheaded progress in a variety of discriminative paralinguistic tasks, e.g., speaker count identification [1], verification [2], pitch estimation [3] and emotion recognition [4]. Beyond these tasks, robust representations of speech have garnered interest for their potential applications in healthcare, including speech enhancement of hearing aids [5, 6], speech pathology, and stress assessment [7–9].

Regarding stress assessment, a substantial body of literature has been dedicated to voice stress analysis. The categorized stressor types notably include cognitive load (i.e., task-induced mental effort) [7, 10–14], physical load (task-induced physical effort) [7, 11, 12], emotional load and alcohol/sleep-deprivation/hypoxia [14]. Nevertheless, inter-individual variability remains a common obstacle to producing robust and personalized evaluations [15]. There has been remarkably little scientific research to explore the voice as a potential stress indicator, caused by the high diversity of the different types of stress and previously mentioned significant individual differences in stress susceptibility [16].

In addition, several of the proposed models process medium to large utterances (10-30 s [7, 12, 17]) while real-time evaluation is a desired feature for medical or commercial usage. Furthermore, in a recent benchmark evaluation [7], each of the five datasets is evaluated by a separate classifier. However, different datasets contain different languages and different types of

cognitive or physical loads. As a result, the classifiers in the benchmark evaluations could be biased toward a specific dataset and thus not suitable for practical usage. It has been shown that cross-corpus testing performances deteriorate in many speech-related tasks such as speech emotion recognition [18, 19].

This paper also addresses individual variabilities in voice stress analysis. Recent work showed a significant relationship exists between a listener's perceived task load and their personality and frustration intolerance, which were measured for each individual through questionnaires [20]. Indeed, the effect of a stimulus on an individual's perception can be highly dependent on individuality factors such as their personality traits and current emotional states [21]. For instance, an easily frustrated listener will perceive tasks as more demanding than one with a higher tolerance for frustration. Consequently, extracting information about a speaker's individuality can enable a more accurate prediction of their perceived cognitive load. We, therefore, hypothesize that extracting speaker identity embeddings, which act as a proxy for individuality factors, can improve the prediction of a speaker's perceived cognitive load. If our hypothesis were correct, it would pave the way for more general personalized paralinguistic speech processing.

The goal of this paper is twofold:

1. Building a system that performs well with combined voice stress analysis datasets (using cross-dataset settings as confirmation of better generalization) and works well with short, 3–5 seconds long, input audio clips (some recordings used in [7] are more than five minutes long!).

2. Evaluating the usage of speaker embeddings for addressing individual variabilities in voice stress analysis.

More specifically, in this paper, we re-evaluate the performance of Hybrid BYOL-S/CvT [22] as a general-purpose speech representation for *personalized* cognitive and physical load detection using the existing cognitive and physical load datasets. While [7] showed reasonable performance after training on each dataset individually, cross-training across combined datasets markedly deteriorated prediction performance. Further performance decrease is found by limiting the input length of recordings. To address the individual variabilities, we concatenate the pre-trained audio representation [22] with well-established speaker embeddings such as Resemblyzer [2] and ECAPA [23] and evaluated them on the 3-5-s long cross-dataset voices stress data. We conclude that personalized audio representation significantly improves voice stress detection.

The paper is structured as follows: Section 2 presents the proposed method of personalized model for voice stress analysis. Section 3 describes the experimental setup and discusses the obtained results. Finally, Section 4 concludes the paper and suggests future work.

---

*ZW performed this work as an intern at Logitech.

2. Methods

2.1. Related work

The hybrid BYOL-S/CvT [22], a self-supervised model pre-trained to extract features from speech audio input, was found as a state-of-the-art feature extractor for cognitive and physical task load detection [7]. The model is an extension of the BYOL-A model [24]. During pre-training, the model has two branches of neural networks, each receiving a different augmented version of the input log-mel spectrogram. The model aims to minimize the self-supervised loss: mean-squared error loss between the outputs of the two branches. In this way, the neural networks could capture the meaningful representations underlying the two differently augmented inputs. Both networks have an encoder and a projection module that share the same architectures across the networks, while one branch has an additional predictor to avoid collapsed representations.

In contrast to BYOL-A which is trained on AudioSet [25], BYOL-S is solely trained on the subset containing speech data. Besides, BYOL-S/CvT changed the convolutional neural network encoder in BYOL-S to a shallow convolutional vision transformer (CvT) [26]. The hybrid BYOL-S/CvT model adds a supervision loss that computes the difference between the predictor output and the openSMILE ComParE feature set [27], which comprises 6373 hand-engineered, DSP-based, audio features.

After pre-training, the CvT encoder of the hybrid BYOL-S/CvT model can be used to extract the embedding features from audio data. A log-mel spectrogram is first computed for each audio sample using a 25 ms window size and a 10 ms hop size. Subsequently, the encoder outputs one 2048-dimensional embedding per 160-ms frame. The utterance-level embeddings can be computed by adding a 'mean+max' temporal pooling. The embedding features extracted by the hybrid BYOL-S/CvT model have shown good performance on HEAR benchmark [28], which uses the extracted features in various downstream tasks, including speech, music, and environment-related tasks.

2.2. Utterance-level and framewise audio embeddings

Traditional methods for speech feature extraction mostly obtain a single feature vector for a given utterance [7, 27]. Such utterance-level embeddings condensed the temporal information stored in sequential data like speech. In contrast, framewise embeddings are obtained by using a model to extract a sequence of embeddings from sequential data. Such embeddings preserve the temporal information stored in the input data. Meanwhile, advancements in deep neural networks have proposed models like Transformers [29] that are capable of capturing temporal information in framewise embeddings. Framewise embeddings have been shown to outperform utterance-level embeddings in speech quality assessment tasks [30].

In this work, we extracted both types of embeddings and evaluated their performance in voice stress detection tasks. The utterance-level embeddings were extracted by using "mean+max" pooling on the Hybrid BYOL-S/CvT model output (Section 2.1), and the framewise embeddings were extracted by removing that 'mean+max' pooling layer.

2.3. Combined audio and speaker embeddings

We propose to use speaker embeddings as a proxy for the individuality factors like personality and frustration intolerance, which are important for cognitive load perception [20]. In this
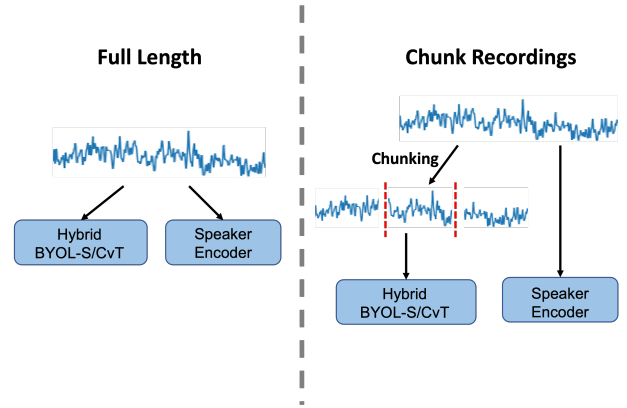

Figure 1: *Left: Full-length audio embedding extraction. Right: Chunked audio embedding extraction.*
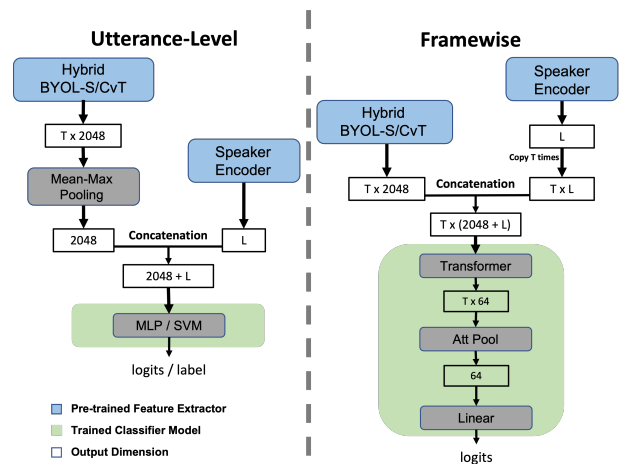

Figure 2: *Concatenation of speaker embeddings and model embeddings. Left: Utterance-level embedding concatenation; Right: Framewise embedding concatenation.*

work, we considered two speaker encoders to extract speaker embeddings: Resemblyzer [2] and ECAPA [23]. Both encoders generate a speaker embedding vector for each speech utterance. Resemblyzer embeddings have 256 features, and ECAPA embeddings have 196 features.

Fig. 1 shows full-length and chunked audio speaker embedding computation. When using full-length data, the speaker embeddings and the audio embeddings were concatenated. When using chunked (3-5 seconds long) data, the audio embeddings were computed using only the chunked audio clips. Each chunk has the same target label as its original recording. The audio embeddings of the chunked audio clip and the speaker embeddings of the full-length utterance were then concatenated and used for training downstream classifiers. We did not compute separate speaker embeddings using each chunked clip because the whole utterance better represents the speaker's identity. Clips chunked from the same utterance share the same speaker embeddings.

Fig. 2, left, shows the concatenation of the speaker and audio utterance-level embeddings. After concatenation, we used either Support Vector Machine (SVM) or Multi-Layer Perceptrons (MLP) for binary classification. Fig. 2, right, shows the

Table 1: *Unweighted Average Recall (UAR) on the test set. Score: mean ± 95% bootstrapped confidence interval (CI)*

	full length utterance-level	5-s chunk utterance-level	5-s chunk framewise	3-s chunk utterance-level	3-s chunk framewise
Only audio embeddings:					
Hybrid BYOL-S/CvT	74.99 ± 2.88	70.43 ± 1.27	70.66 ± 1.22	67.00 ± 1.10	67.65 ± 1.06
Hybrid BYOL-S/CvT ($\alpha = 2$)	**77.07** ± 2.82	70.20 ± 1.27	68.54 ± 1.30	70.18 ± 1.04	70.73 ± 1.06
Only speaker embeddings:					
ECAPA	63.16 ± 3.24	75.73 ± 1.26	-*	75.49 ± 1.04	-
Resemblyzer	66.45 ± 3.31	70.57 ± 1.32	-	69.94 ± 1.01	-
ECAPA + Resemblyzer[1]	62.41 ± 3.12	74.23 ± 1.28	-	74.55 ± 0.98	-
Concatenate ECAPA:					
Hybrid BYOL-S/CvT	72.93 ± 2.95	**80.21** ± 1.16	**79.32** ± 1.10	**80.62** ± 0.96	79.59 ± 0.86
Hybrid BYOL-S/CvT ($\alpha = 2$)	76.20 ± 2.74	79.24 ± 1.10	78.25 ± 1.12	79.38 ± 0.94	**80.89** ± 0.90
Concatenate Resemblyzer:					
Hybrid BYOL-S/CvT	76.55 ± 2.89	72.08 ± 1.26	68.52 ± 1.34	71.67 ± 0.98	70.32 ± 0.96
Hybrid BYOL-S/CvT ($\alpha = 2$)	75.87 ± 2.86	74.50 ± 1.27	71.26 ± 1.32	74.68 ± 1.02	72.83 ± 1.02

*The speaker embeddings are utterance-level embeddings, so there is no framewise performance for speaker embedding only.
[1]The two speaker embeddings are concatenated

concatenation of the speaker and audio framewise embeddings. The dimension of audio framewise embeddings is $T \times 2048$, where T is the total audio length divided by 160ms. We concatenated the same utterance-level speaker embedding at each timestamp. As a result, the framewise embeddings had the shape $T \times (2048 + L)$ where L is the length of speaker embedding. After concatenation, we used a lightweight Transformer followed by an attention-pooling layer and a linear projection layer for binary classification.

3. Experimental setup and results

3.1. Data

We used the existing cognitive and physical load dataset corpus as presented in [7]. The corpus consists of five datasets, with a total number of 111 speakers, nine languages, and an overall duration of 12 hours. For each data sample, the task is to predict whether the recorded speech was produced while the speaker was simultaneously performing a cognitive or physical load-inducing task or not.

In each one of the five datasets, the audio samples were split based on speaker identities into 60% for training, 15% for validation, and 25 % for testing. As a result, the testing data only consisted of speakers never seen by the model during training.

3.2. Baseline systems

To be consistent with the benchmark, we first used the hybrid BYOL-S/CvT model to extract the audio embeddings. In our experiments, we picked two best-performing models from the benchmark paper: the hybrid BYOL-S/CvT model and a variation of the hybrid BYOL-S/CvT model in which the weight of supervision loss to self-supervised loss ($\alpha = \alpha_{sup} : \alpha_{sl} = 2$) is 2:1.

The utterance-level embeddings were extracted by the hybrid BYOL-S/CvT models with the 'mean+max' pooling layer to compress the temporal dimension of the model output. Then, we used an SVM as the downstream classifier and trained an SVM on each dataset separately. During training, we searched the optimal penalty weight from 10^{-5} to 10^5, and we used the 5-fold cross-validated recall score to determine the opti-

Table 2: *Test UAR after combining datasets. Score: mean ± 95% bootstrapped CI*

	Average*	Combined
Hybrid BYOL-S/CvT	80.36	73.68 ± 3.10[1]
Hybrid BYOL-S/CvT ($\alpha = 2$)	82.43	73.72 ± 2.93

*CI not provided because it is an average performance on different datasets

mal weight. Table 2 ("average" column) shows the reproduced results of [7] as an average of the unweighted average recall (UAR) of the five datasets.

We also see that the UAR performance significantly dropped ("combined" column) after we combined the five datasets. We used the same SVM model and hyperparameter search methods.

3.3. Neural network downstream classifier

To improve performance, we changed the SVM classifier to a shallow MLP, with one hidden layer of 512 neurons. During the training, the classification accuracy of the validation dataset is used for early stopping, with 30 epochs of patience. We also used the validation performance to search the optimal learning rate from $3.2 \times 10^{-3}, 10^{-3}, 3.2 \times 10^{-4}, 10^{-4}, 3.2 \times 10^{-5}$.

As shown in Table 1 (column "full length" and rows "only audio embedding"), MLP-based classification improved performance on the combined dataset.

3.4. Decreasing input audio length

To train a model capable of making predictions based on shorter audio inputs, we first chunked all audio data to five seconds. Then, we used the hybrid BYOL-S/CvT models to extract the utterance-level embeddings on these shorter inputs and searched for the best downstream classifier among the SVM and MLP. The training process for SVM and MLP was the same as before chunking. The performance decreased significantly after chunking the data (Table 1, 5-s chunk utterance-level column and only audio embedding rows).

3.5. Utterance-level speaker embeddings

3.5.1. Combined with audio embeddings from 5-s chunked data

Our next step was to address individual variabilities in voice stress analysis by using speaker embeddings. We chose two speaker embeddings encoders: ECAPA [23] (implemented by speechbrain [31]) and Resemblyzer [2] (implemented by resemble-ai[1]). After concatenation of ECAPA speaker embeddings and utterance-level audio embeddings, performance increases significantly, with ECAPA and hybrid BYOL-S/CvT embeddings reaching a test UAR of 80.21% (Table 1, 5-s chunk utterance-level column and concatenate ECAPA rows). Interestingly, we found Resemblyzer embeddings give much less improvement on the performance than ECAPA does (Table 1, 5-s chunk utterance-level column and concatenate Resemblyzer rows).

Besides, we tried to add speaker embeddings to the audio embeddings of the original full-length dataset. The speaker embeddings, however, do not always help improve the model performance (Table 1, full-length utterance-level column).

3.5.2. Speaker embeddings only

To investigate the differences between the two speaker encoders, we evaluated the performance using only speaker embeddings. For each chunked data, the speaker embeddings were still computed from the whole utterance before chunking. We found ECAPA has a better performance than Resemblyzer (Table 1, "only speaker embeddings" rows). This matches the performance gap between concatenating ECAPA embeddings and concatenating Resemblyzer embeddings. Interestingly, for chunked data, we found that ECAPA embeddings themselves even outperformed the embeddings from the two hybrid BYOL-S models. We speculated that the speaker embedding models themselves contained not only information on speaker identities, but also some paralinguistic features that help the downstream model to detect the cognitive or physical loads. The additional paralinguistic information stored in the ECAPA embeddings helped to improve the classification performance. We also observed significant improvement when using speaker embeddings only on 5-s chunks instead of full-length utterance-level data. We believe it was due to having more balanced chunked data – in the full-length version, for example, several minutes long audio had a single label, whereas with chunked data the label distribution became more balanced.

Furthermore, we examined whether the improvement brought by ECAPA concatenation is simply due to the additional paralinguistics information in the ECAPA embeddings. Thus, we evaluated the concatenation of ECAPA and Resemblyzer embeddings. The obtained performance does not exceed the better performance of ECAPA-only on chunked data (Table 1, only speaker embedding rows). In this case, adding additional information does not lead to performance incease. Consequently, we argue that the speaker identity and paralinguistics information encoded by ECAPA really complements the audio features extracted by the hybrid BYOL-S/CvT in voice stress detection tasks.

3.5.3. Combined with audio embeddings from 3-s chunked data

Finally, we examined the sensitivity of the different models to shorter chunk sizes (3 instead of 5 s). Regarding utterance-level embeddings, performance of hybrid BYOL-S/CvT ($\alpha = 2$)

showed little variation, while the hybrid BYOL-S/CvT showed worse performance on shorter audio chunks. After concatenating the 3-s chunked utterance-level embeddings with the speaker embeddings, both models showed a performance similar to 5-s chunking length. Concatenation of speaker embeddings helped the models to be more robust to different chunking lengths.

3.6. Framewise audio embeddings

We have reported so far only the results for the utterance-level embeddings, whereas framewise embeddings have also been used to improve the performance of many speech-related tasks like speech quality assessment [30]. With chunked data, we were able to get framewise embeddings with shorter and less varied temporal lengths. Therefore, we also extracted framewise embeddings and trained a two-layer, 64-dimensional, Transformer-based classifier on the embeddings to detect the voice stress.

We also used validation accuracy with 30 epochs of patience for early stopping. The optimal learning rate was again searched from $3.2 \times 10^{-3}, 10^{-3}, 3.2 \times 10^{-4}, 10^{-4}, 3.2 \times 10^{-5}$.

On 5-s chunked data, using framewise embeddings did not improve the recall performance, regardless of speaker embedding concatenation (Table 1, 5-s chunk utterance-level columns and concatenate ECAPA or Resemblyzer rows). When we chunked the data to 3 s, the performance improves slightly for the hybrid BYOL-S/CvT ($\alpha = 2$) model, but not for the hybrid BYOL-S/CvT model. The effect of using framewise embeddings is thus model-dependent. We obtained the best performance by concatenating the hybrid BYOL-S/CvT ($\alpha = 2$) model embeddings and ECAPA embeddings. However, the slight improvement was not significant when compared with the best utterance-level embeddings. Overall, we found framewise embedding as being not helpful in voice stress detection tasks.

4. Conclusion

In this work, we have constructed a pipeline to use the embeddings of a pre-trained audio feature extraction model in a cross-dataset, lower latency voice stress detection system. We re-evaluated the state-of-the-art pre-trained model from the previous benchmark paper [7] and found a worse performance in the new pipeline. We improved the cross-dataset performance by replacing the downstream SVM classifier with a shallow MLP or Transformer. We also confirmed that individual variabilities in voice stress analysis can be achieved by concatenating ECAPA speaker embeddings, which significantly improved the performance after we chunked the data into shorter segments.

In our future work, we aim to include other types of stressors, such as emotional load, to train a more general voice stress analysis model. Besides, we wish to apply the speaker embeddings in other speech-related downstream tasks, including emotion recognition and quality assessments, to examine the role of individuality factors in these tasks. Finally, it is also worthy of investigating the relationship between speaker identity and paralinguistic features to help explain the good performance of speaker embeddings on this voice stress detection task.

5. Acknowledgements

---

[1]https://github.com/resemble-ai/Resemblyzer

6. References

[1] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked Spectrogram Modeling using Masked Autoencoders for Learning General-purpose Audio Representation," in *HEAR: Holistic Evaluation of Audio Representations (NeurIPS 2021 Competition)*, vol. 166, 2022, pp. 1–24.

[2] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *ICASSP*, 2018, p. 4879–4883.

[3] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *ICASSP*, 2018, pp. 161–165.

[4] N. Scheidwasser-Clow, M. Kegler, P. Beckmann, and M. Cernak, "Serab: A multi-lingual benchmark for speech emotion recognition," in *ICASSP*, 2022, pp. 7697–7701.

[5] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, R. Viveros Munoz *et al.*, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in *Proc. Interspeech*, 2021, pp. 686–690.

[6] N. A. Lesica, N. Mehta, J. G. Manjaly, L. Deng, B. S. Wilson, and F.-G. Zeng, "Harnessing the power of artificial intelligence to transform hearing healthcare and research," *Nat. Mach. Intell.*, vol. 3, no. 10, pp. 840–849, 2021.

[7] G. Elbanna, A. Biryukov, N. Scheidwasser-Clow, L. Orlandic, P. Mainar, M. Kegler, P. Beckmann, and M. Cernak, "Hybrid Handcrafted and Learnable Audio Representation for Analysis of Speech Under Cognitive and Physical Load," in *Proc. Interspeech*, 2022, pp. 386–390.

[8] P. Janbakhshi and I. Kodrasi, "Experimental investigation on stft phase representations for deep learning-based dysarthric speech detection," in *ICASSP*, 2022, pp. 6477–6481.

[9] L. P. Violeta, W. C. Huang, and T. Toda, "Investigating Self-supervised Pretraining Frameworks for Pathological Speech Recognition," in *Proc. Interspeech*, 2022, pp. 41–45.

[10] 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*, 2013, pp. 148–152.

[11] H. Jing, T.-Y. Hu, H.-S. Lee, W.-C. Chen, C.-C. Lee, Y. Tsao, and H.-M. Wang, "Ensemble of Machine Learning Algorithms for Cognitive and Physical Speaker Load Detection," in *Proc. Interspeech 2014*, 2014, pp. 447–451.

[12] 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. Interspeech 2014*, 2014, pp. 427–431.

[13] M. Van Segbroeck, R. Travadi, C. Vaz, J. Kim, M. P. Black, A. Potamianos, and S. S. Narayanan, "Classification of Cognitive Load from Speech using an i-vector Framework," in *Proc. Interspeech 2014*, 2014, pp. 751–755.

[14] M. Van Puyvelde, X. Neyt, F. McGlone, and N. Pattyn, "Voice stress analysis: A new framework for voice and effort in human performance," *Front. Psychol.*, vol. 9, 2018.

[15] C. L. Giddens, K. W. Barron, J. Byrd-Craven, K. F. Clark, and A. S. Winter, "Vocal Indices of Stress: A Review," *J. Voice*, vol. 27, no. 3, p. 390.e21–390.e29, 2013.

[16] K. R. Scherer, "Acoustic Patterning of Emotion Vocalizations," in *The Oxford Handbook of Voice Perception*. Oxford University Press, 2018.

[17] A. Gallardo-Antolín and J. M. Montero, "A Saliency-Based Attention LSTM Model for Cognitive Load Classification from Speech," in *Proc. Interspeech*, 2019, pp. 216–220.

[18] S. Zhang, R. Liu, X. Tao, and X. Zhao, "Deep cross-corpus speech emotion recognition: Recent advances and perspectives," *Front. Neurorobotics*, p. 162, 2021.

[19] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[20] R. P. Spang, K. E. Hajal, S. Möller, and M. Cernak, "Personalized Task Load Prediction in Speech Communication," in *ICASSP*, 2023, pp. 1–5.

[21] R. Principi, C. Palmero, J. Junior, and S. Escalera, "On the effect of observed subject biases in apparent personality analysis from audio-visual signals," *IEEE Trans. Affect. Comput.*, vol. 12, no. 03, pp. 607–621, 2021.

[22] G. Elbanna, N. Scheidwasser-Clow, M. Kegler, P. Beckmann, K. El Hajal, and M. Cernak, "BYOL-S: Learning Self-supervised Speech Representations by Bootstrapping," in *HEAR: Holistic Evaluation of Audio Representations (NeurIPS 2021 Competition)*, vol. 166, 2022, pp. 25–47.

[23] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[24] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "BYOL for Audio: Self-supervised learning for general-purpose audio representation," in *IJCNN*, 2021, pp. 1–8.

[25] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017, pp. 776–780.

[26] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing convolutions to vision transformers," in *Proc. ICCV*, 2021, pp. 22–31.

[27] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. ACM Multimedia (MM)*, 2010, pp. 1459–1462.

[28] J. Turian, J. Shier, H. R. Khan, B. Raj, B. W. Schuller, C. J. Steinmetz, C. Malloy, G. Tzanetakis, G. Velarde, K. McNally, M. Henry, N. Pinto, C. Noufi, C. Clough, D. Herremans, E. Fonseca, J. Engel, J. Salamon, P. Esling, P. Manocha, S. Watanabe, Z. Jin, and Y. Bisk, "HEAR: Holistic Evaluation of Audio Representations," in *NeurIPS Competitions and Demonstrations Track*, 2022, pp. 125–145.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, vol. 30, 2017.

[30] K. El Hajal, Z. Wu, N. Scheidwasser-Clow, G. Elbanna, and M. Cernak, "Efficient speech quality assessment using self-supervised framewise embeddings," in *ICASSP*, 2023, pp. 1–5.

[31] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.