

# SERAB: A MULTI-LINGUAL BENCHMARK FOR SPEECH EMOTION RECOGNITION

Neil Scheidwasser-Clow<sup>1,2\*</sup>, Mikolaj Kegler<sup>3</sup>, Pierre Beckmann<sup>1</sup>, Milos Cernak<sup>2</sup>

<sup>1</sup>École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

<sup>2</sup>Logitech Europe S.A., Lausanne, Switzerland

<sup>3</sup>Imperial College London, London, United Kingdom

## ABSTRACT

Recent developments in speech emotion recognition (SER) often leverage deep neural networks (DNNs). Comparing and benchmarking different DNN models can often be tedious due to the use of different datasets and evaluation protocols. To facilitate the process, here, we present the Speech Emotion Recognition Adaptation Benchmark (SERAB), a framework for evaluating the performance and generalization capacity of different approaches for utterance-level SER. The benchmark is composed of nine datasets for SER in six languages. Since the datasets have different sizes and numbers of emotional classes, the proposed setup is particularly suitable for estimating the generalization capacity of pre-trained DNN-based feature extractors. We used the proposed framework to evaluate a selection of standard hand-crafted feature sets and state-of-the-art DNN representations. The results highlight that using only a subset of the data included in SERAB can result in biased evaluation, while compliance with the proposed protocol can circumvent this issue.

**Index Terms**— emotion recognition, computational paralinguistics, deep neural networks, speech processing, transfer learning

## 1. INTRODUCTION

Speech emotion recognition (SER) is a cornerstone of computational paralinguistics, the analysis of non-verbal elements of speech [1]. Although a challenging task, being able to automatically and accurately classify emotions from voice could support a wide range of applications, including human-computer interaction [1], healthcare [2], and public safety [3]. At its inception, the development of hand-engineered features has proven effective in tackling various SER-related problems [1]. Such features are traditionally based on acoustic [4] or linguistic descriptors [5]. More recently, deep neural networks (DNNs) trained in self-supervised fashion were able to produce generalizable representations suitable for a range of audio and speech processing tasks [6, 7]. A notable advantage of such data-driven approaches over fixed hand-engineered feature sets is the ability to transfer learned knowledge from a large, unlabeled dataset to downstream tasks with less task-specific data available.

However, the estimated performance and generalization capacity of such self-supervised DNNs can strongly depend on the evaluation protocol. This makes open-source benchmarks, typically composed of fixed dataset(s) and evaluation pipelines, instrumental for informative, fair, and accessible comparisons of different methods. In visual object recognition, ImageNet [8] has established itself as the *de facto* image dataset and benchmark for deep learning models. In natural language processing (NLP), GLUE [9] is a widely used bench-

mark, with nine different tasks encompassing various characteristics of language understanding (e.g., sentiment analysis, paraphrase, and inference tasks). As one of the largest audio datasets available, AudioSet [10] is commonly used for self-supervised pre-training, as well as a benchmarking method for audio event classification [6, 7, 11]. A recently proposed HEAR challenge [12] focuses on evaluating *general-purpose* audio representations and extends the concept underlying AudioSet by including additional tasks. In speech representation learning, NOSS [6] was recently proposed as a platform for evaluating speech-specific feature extractors. It includes diverse non-semantic speech processing problems, such as speaker and language identification, as well as two SER tasks (CREMA-D [13] and SAVEE [14]).

In contrast to general audio and non-semantic speech representation learning, a standard, readily available multi-task SER benchmark is yet to be introduced. While recently [15] proposed a SER-specific benchmarking framework, it has two considerable shortcomings. First, it only includes a single dataset, implying the lack of diversity in terms of task difficulty, amount of task-specific data, or data acquisition setup (e.g., recording equipment and conditions). This effectively limits the estimation of generalization capacity, and thus the *real-life* impact of different methods. Second, the dataset is monolingual, with all speech material in English. As a paralinguistic cue, robust embeddings for speech emotion recognition should perform well across different languages.

To that end, we introduce the Speech Emotion Recognition Adaptation Benchmark (SERAB), a collection of nine SER tasks spanning six languages, different dataset sizes and emotion categories. To streamline the comparison of different approaches, we set up a custom evaluation pipeline. We employed the framework to evaluate recent state-of-the-art pre-trained DNNs for speech/audio feature extraction [6, 7, 11, 16], as well as a classic set of hand-crafted features commonly used in computational paralinguistics [4]. Lastly, we also propose a novel Transformer-based model, which performs on par with state-of-the-art approaches. Results obtained for a range of baselines demonstrate apparent differences in performance achieved on single datasets, and illustrate the benefits of using the complete SERAB benchmarking framework.

## 2. SPEECH EMOTION RECOGNITION ADAPTATION BENCHMARK (SERAB)

### 2.1. Tasks & datasets

A summary of the tasks used in SERAB is presented in Table 1. The benchmark comprises nine speech emotion classification tasks in six languages: four in English (CREMA-D, IEMOCAP, RAVDESS & SAVEE), and one in French (CaFE), German (EmoDB), Greek (AESDD), Italian (EMOVO), and Persian (ShEMO). In each dataset,

\*NSC (neil.scheidwasser-clow@epfl.ch) performed this work as an intern at Logitech.

**Table 1.** SERAB tasks and datasets. IEM4: 4-class IEMOCAP [17]. Restricted-access datasets require additional registration on the data provider website. Open-access datasets can be downloaded without registration.

Dataset	Code	Access	Language	Classes	Utterances	Speakers	Average duration (s)	Total duration (h)
AESDD [18]	AES	Open	Greek	5	604	6	4.2	0.7
CaFE [19]	CAF	Open	French	7	864	12	4.5	1.1
CREMA-D [13]	CRE	Open	English	6	7,442	91	2.5	5.3
EmoDB [20]	EMB	Open	German	7	535	10	2.8	0.4
EMOVO [21]	EMV	Open	Italian	7	588	6	3.1	0.5
IEM4 [17]	IEM	Restricted	English	4	5,531	10	3.4	7.0
RAVDESS [22]	RAV	Open	English	8	1,440	24	3.7	1.5
SAVEE [14]	SAV	Restricted	English	7	480	4	3.8	0.5
ShEMO [23]	SHE	Open	Persian	6	3,000	87	4.0	3.3

speech samples have three attributes: audio data (i.e., the raw waveform, in mono), speaker identifier, and emotion label (e.g., angry, happy, sad). The datasets vary in size (i.e., number of utterances), number of speakers, class distribution, and number of classes. While anger, happiness, and sadness are found across all datasets, disgust, fear, neutral emotion, surprise, calm, and boredom appear in at least one dataset. On the other hand, all datasets have roughly the same average utterance duration (between 2.5 & 4.5 seconds).

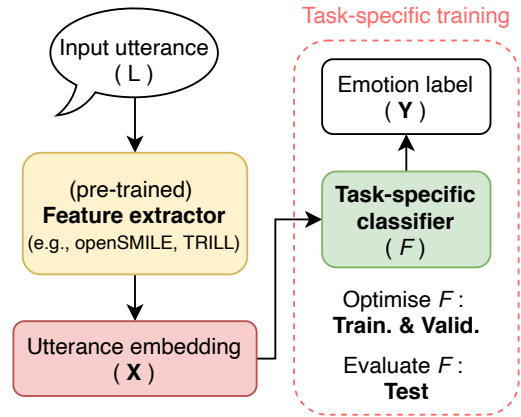
The benchmark was designed to balance dataset popularity, language diversity, and open access. In speech emotion recognition, EmoDB, IEMOCAP and RAVDESS are among the most widely used datasets [15, 23, 24]. In the same vein as [24], a 4-class subset of IEMOCAP (IEM4) was used to mitigate the severe class imbalance in the original dataset. For the other tasks, all samples and classes from the original datasets were used (Table 1). As already present in NOSS [6], CREMA-D and SAVEE were included in SERAB. To complete the benchmark, CaFE (French) and EMOVO (Italian) were chosen as Italic-language datasets, whereas AESDD (Greek) and ShEMO (Persian) represented the Hellenic and Indo-Iranian branches of the Indo-European family [25]. Overall, the benchmark mainly comprises scripted and acted speech, excepting IEM4 [17], RAVDESS [22] and ShEMO [23] which also feature spontaneous utterances.

Each dataset was split into training, validation, and testing sets to respectively train, optimize and evaluate task-specific speech emotion classifiers. Excepting CREMA-D, each dataset was split into 60% training, 20% validation, and 20% testing sets. For CREMA-D, we followed a 70/10/20% (training/validation/testing) split that was applied in NOSS [6]. Each data partition was speaker-independent, i.e., the sets of speakers included in each part were mutually disjoint. Since SERAB datasets vary in size, the fixed data split allows assessing how different methods cope with various amounts of task-specific data.

## 2.2. Evaluation pipeline

The SERAB evaluation pipeline is used to assess representations of speech-based emotion obtained using different feature extractors (Fig. 1). In particular, the workflow includes processing the input utterances through the pre-trained/non-trainable feature extractor and using the resulting embeddings together with a task-specific classifier to predict speech emotion. By using simple classifiers, the classification accuracy reflects the utility of the extracted features for the utterance-level SER tasks.

Importantly, the utterances included in SERAB vary in duration. The decision about how to integrate information across the utter-



**Fig. 1.** SERAB evaluation pipeline. The (pre-trained/non-trainable) feature extractor is used to obtain utterance-level embeddings ( $\mathbf{X}$ ) from the input.  $\mathbf{X}$  are used as input to the task-specific classifier  $F$  optimized for predicting the emotion  $\mathbf{Y}$  expressed in the input.

ance is a crucial design choice that may impact feature extractor performance. Thus, we left the decision to the authors of the feature extractor. Most of the current approaches tend to extract information from fixed-length frames and average the outputs across frames. However, other approaches may utilize temporal dependencies in the input utterance. As a result, we require the method to return one set of features for each input utterance with varying duration.

To assess a method's performance, all input utterances are processed through the feature extractor to obtain their embeddings ( $\mathbf{X}$ ). The resulting embeddings are then used as features for basic machine learning classifiers ( $F$ ) trained to predict emotion labels ( $\mathbf{Y}$ ). To assure the thorough evaluation, we considered several different classifiers: logistic regression (LR), support vector machine (SVM), linear and quadratic discriminant analysis (LDA/QDA), and random forests (RF). Classifier hyperparameters were optimized through grid-search using the training and validation portions of the data. The best-performing classifier from the grid-search procedure was evaluated on the set-aside test set. All classifier optimization and evaluation procedures were implemented using scikit-learn [26].

Test-set classification accuracy in each task was used as the performance metric. The resulting accuracies across the nine SERAB tasks were averaged to quantify the method's benchmark performance. In addition to the unweighted mean accuracy (**UM**) across the SERAB tasks, we also computed the weighted average derived from the test set size (**WM**), as well as the geometric mean (**GM**).

### 3. BASELINE APPROACHES

**openSMILE** - openSMILE [4] is a hand-engineered acoustic feature set based on functionals of low-level descriptor contours. Although not directly data-driven, openSMILE is capable of outperforming DNN-based feature extractors, e.g., in problems with little task-specific data [27]. Here, the most recent implementation of openSMILE<sup>1</sup> was used to extract features from each utterance in the SERAB tasks. Subsequently, for each task, the speech emotion classifier was optimized using the training and validation portions of the data and evaluated using the set-aside test set (Section 2.2).

**VGGish** - VGGish [11] is one of the first DNN-based feature extractors for audio, inspired by the VGG-16 convolutional DNN (CNN) [28]. Pre-trained model weights<sup>2</sup> were learned through supervised classification of audio events from the Youtube-8M dataset [29] ( $\geq 350,000$  hours of video, 3,000 classes). The model uses fixed-size input windows. To cope with variable-length audio clips, each input utterance was split into non-overlapping 960 ms-long frames. A log-mel magnitude spectrogram ( $N = 64$  mel frequency bins) was computed from a short-term Fourier transform with 25-ms windows in steps of 10 ms for each frame. The resulting frames were then fed to the pre-trained model for feature extraction. After processing  $M$  frames, the obtained  $M$  embeddings were averaged to obtain one feature set per utterance. The remaining evaluation followed the protocol outlined in Section 2.2.

**YAMNet** - YAMNet [16] is another commonly used DNN-based feature extractor [6, 7]. This approach utilizes MobileNetv1 [30], an efficient CNN architecture optimized for mobile devices. Here, we used the weights<sup>3</sup> of the model pre-trained through supervised classification of events from AudioSet [10] ( $\approx 5,800$  hours of audio, 521 classes). Since the model operates using fixed-size windows, the input utterances were processed analogously to VGGish.

**TRILL** - While VGGish and YAMNet were trained on diverse audio sources (speech, music, environmental sounds, etc.), TRILL [6] was specifically developed as a non-semantic speech feature extractor. The DNN model adopted the architecture of ResNetish [11] and was pre-trained in self-supervised fashion using speech samples from AudioSet, which constitutes approximately 50% of the entire dataset ( $\approx 2,800$  hours of audio). The pre-trained model<sup>4</sup> used herein was obtained from triplet loss optimization, which aims at minimizing the embedding-space distance between an anchor and a positive sample (i.e., from the same clip) while maximizing the distance between the same anchor and a negative sample (i.e., from a different clip). In the context of audio, temporally neighboring audio segments will be closer in the representation space and vice versa. Once again, the model operates on fixed-size frames, so the input utterances were processed analogously to VGGish and YAMNet. Following [6], we used the embedding from the first 512-depth convolution layer (*layer 19*) which performed best on NOSS.

**BYOL-A** - As an alternative to contrastive learning setups such as TRILL, BYOL-A [7] proposes *bootstrapping your own latent* (BYOL) for audio representation learning, inspired by the success of BYOL [31] for self-supervised image classification. Pre-trained on the entire AudioSet, this approach achieved state-of-the-art results in various audio classification tasks, even outperforming TRILL [6] in speech processing problems. Instead of assessing the temporal proximity of two different audio segments, BYOL-A relies on comparing two augmented versions of a single sample. More specifically, each

**Table 2.** Baseline approaches evaluated on SERAB. The embedding size refers to the dimensionality of the utterance-level feature set (**X**) used to classify emotional labels (**Y**).

Model	Parameters (M)	Embedding size
openSMILE [4]	-	6,373
VGGish [11]	62.0	128
YAMNet [16]	4.2	1,024
TRILL (layer 19) [6]	9.0	12,288
BYOL-A [7]	0.6 / 1.6 / 5.3	512 / 1,024 / 2,048
<b>Proposed:</b>		
BYOL-S	0.6 / 1.6 / 5.3	512 / 1,024 / 2,048
BYOL-S/CvT, <i>small</i>	1.6	256
<i>CvT stages: 64/128/256</i>		
BYOL-S/CvT, <i>large</i>	5.0	512 / 2,048
<i>CvT stages: 64/256/512</i>		

version is respectively fed to an *online* network and a *target* network. While both are composed of an encoder and a projection block, the *online* network includes a *prediction* layer which aims at predicting the projected representation of the second augmented view. Thus, BYOL (and BYOL-A) learns a representation by negating the random data augmentations to capture the essential information about the input. Regarding BYOL-A, pre-trained weights for models of different sizes were released by the authors<sup>5</sup> and used in this work. Since the model accepts inputs of variable length, it returns a single embedding per input utterance. The resulting embeddings are used to train and evaluate the SER classifiers (Section 2.2).

**BYOL-S** - While BYOL-A can achieve state-of-the-art results on a range of audio classification tasks, its *general* audio representation might not be optimal for speech processing and especially paralinguistic problems. Thus, we re-trained BYOL-A using only speech samples of AudioSet, leading to the speech-specific BYOL-S (*S* denoting speech). The model architecture, pre-training routine, and usage remained the same as in the original version.

**BYOL-S/CvT** - In this model, we propose an extension of BYOL-S with a Transformer representation. More specifically, we replaced the convolution blocks in BYOL-S with Convolutional Transformer<sup>6</sup> (CvT) [32]. CvT notably extends self-attention with depthwise convolution to project the queries, keys, and value embeddings. Between the attention modules, traditional convolution layers are added to decompose the input as in most CNNs. Consequently, CvT combines the qualities of CNNs (e.g., translation invariance) and Transformers (e.g., capturing long-range dependencies and generalization abilities). Here, each CvT stage included only one self-attention layer to allow fair comparisons with BYOL-S, both in terms of model architecture and the number of parameters. We experimented with three different configurations of the model. To explore the impact of model size, the number of filters in CvT stages was manipulated to reduce the number of parameters (Table 2), analogously to BYOL-A [7]. In addition, the model was tested with three different embedding dimensions: 256, 512 and 2048. The latter used *mean + max* temporal aggregation in the last layer instead of global average pooling, in the same vein as [7]. Like BYOL-S, the pre-training and application to SERAB tasks was analogous to BYOL-A.

<sup>1</sup><https://audeering.github.io/opensmile-python/>

<sup>2</sup><https://tfhub.dev/google/vggish/1>

<sup>3</sup><https://tfhub.dev/google/yamnet/1>

<sup>4</sup><https://tfhub.dev/google/nonsemantic-speech-benchmark/trill/3>

<sup>5</sup><https://github.com/nttclab/byol-a>

<sup>6</sup><https://github.com/lucidrains/vit-pytorch>

**Table 3.** Test accuracy (%) on the different downstream tasks in SERAB, referred to by their code from Table 1. **UM**: unweighted mean, **WM**: weighted mean (by the number of utterances in the test set), **GM**: geometric mean. Models are sorted by their **UM** across all tasks. The best performing approaches for each task and metric are denoted in bold.

Model	AES	CAF	CRE	EMB	EMV	IEM	RAV	SAV	SHE	UM	WM	GM
YAMNet	53.6	49.6	53.9	60.7	35.7	56.1	52.3	54.2	81.7	55.3	55.9	54.2
VGGish	46.4	50.0	55.5	73.8	36.2	60.1	53.0	53.3	83.6	56.9	57.7	55.4
TRILL, layer 19	66.7	69.2	73.3	81.0	36.7	57.7	73.7	76.7	86.8	69.1	68.3	67.3
openSMILE	70.0	67.1	72.8	<b>90.5</b>	37.2	62.1	71.3	72.5	84.9	69.8	69.1	68.1
BYOL-A, 512	71.5	72.6	70.2	84.5	39.3	62.5	74.7	76.7	90.1	71.3	69.5	69.7
BYOL-A, 2048	72.0	72.6	73.7	88.1	38.3	62.8	<b>77.7</b>	78.3	89.0	72.5	71.0	70.7
BYOL-A, 1024	75.4	<b>76.1</b>	71.3	88.1	44.4	62.1	76.0	80.8	89.5	73.7	70.6	72.4
<i>Proposed:</i>												
BYOL-S, 512	74.9	73.5	74.4	86.9	34.2	63.3	77.3	79.2	90.6	72.7	71.6	70.4
BYOL-S/CvT, 256	72.9	70.9	72.9	85.7	47.4	64.8	76.0	75.8	89.0	72.8	71.4	71.8
BYOL-S, 1024	75.4	71.8	75.3	84.5	39.3	63.8	74.0	<b>82.5</b>	90.9	73.1	72.0	71.4
BYOL-S/CvT, 512	71.0	74.4	74.0	88.1	46.9	65.0	76.3	80.0	87.9	73.7	72.1	72.6
BYOL-S/CvT, 2048	75.8	70.1	<b>76.9</b>	84.5	<b>48.5</b>	<b>65.1</b>	76.3	76.7	<b>93.0</b>	74.1	73.5	73.1
BYOL-S, 2048	<b>77.3</b>	75.6	<b>76.9</b>	88.1	44.4	64.8	76.7	81.7	91.1	<b>75.2</b>	<b>73.7</b>	<b>73.8</b>

#### 4. EVALUATION RESULTS & DISCUSSION

Table 2 presents configurations of different baseline approaches outlined in Section 3. For BYOL-like models (BYOL-A, -S, -S/CvT), we explored different model sizes and sizes of the output embedding fed to the task-specific classifiers predicting emotion labels. Benchmark performance for all baseline methods is presented in Table 3.

The large BYOL-S, with a 2048 embedding size, emerged as the best model across all considered performance metrics and yields the best individual accuracy on two out of the nine SERAB tasks. Importantly, model ranks were generally similar across all benchmark-wide metrics. Thus, we chose to sort model performance by **UM** in accordance with previous benchmarks for computer vision systems [33]. Although reaching slightly lower scores across the entire benchmark, the largest BYOL-S/CvT remained competitive by providing the best results in four out of nine tasks. Moreover, the increase in the embedding size and the overall model size tend to consistently improve the proposed approaches’ performance.

More generally, all BYOL-inspired models, even with small sizes, achieved significantly higher scores (up to a 5% absolute difference in **UM**) than TRILL, VGGish, YAMNet, and openSMILE. This considerable difference in performance most likely originates from the vastly different pre-training strategies. Another reason for the supremacy of BYOL-derived models might be the fact that they are designed to process variable-length inputs rather than fixed-length frames as openSMILE, VGGish, YAMNet, and TRILL do. This, in turn, suggests that aggregation of the temporal context could improve utterance-level SER performance.

Interestingly, BYOL-S models performed consistently better than original BYOL-A approaches. This indicates that specializing the pre-training task by focusing only on speech excerpts resulted in more suitable embeddings for SER. In such a speech-specific pre-training, the model presumably developed better capacity for representing speech, including language-independent paralinguistic cues such as speech-based emotion.

On the other hand, enriching BYOL-S with self-attention mechanisms via CvT did not yield a notable performance increase we anticipated, but the model was lighter with 0.3M fewer parameters than BYOL-S. The slight difference in terms of performance might be due to the minimal inductive biases implied in Transformer-like

models, in contrast to CNNs [34]. While advantageous when training large models on large datasets [35], such biases become critical in smaller network and smaller dataset setups [34]. Thus, increasing the pre-training dataset’s size could help develop the generalization ability of Transformers and thus improve the overall performance of BYOL-S/CvT.

While SERAB allows comparing different models across a diverse range of tasks, more importantly, it provides a streamlined benchmarking platform for the comparison of different approaches. In particular, some tasks exhibit significant variations between models (e.g., EMOVO, SAVEE, EmoDB) such that the overall poorer-performing approaches may appear better than they really are. Some of these differences might be dataset-specific, introducing even larger bias that is not trivial to overcome. The inclusion of multiple tasks across different languages provides robust performance estimates, as shown by our evaluation of the baseline approaches.

#### 5. CONCLUSIONS

We introduce SERAB, a multi-lingual benchmark for speech emotion recognition. With the rapid emergence of DNN-based representations of speech and speech-based emotion, the benchmark provides a universal platform for comparing different methods. Due to the inclusion of diverse tasks spanning across different languages, dataset sizes, and emotional categories, SERAB produces robust estimates of performance and generalization capacity. We used SERAB to evaluate a range of recent baselines. Among the tested frameworks, BYOL-based approaches yielded superior performance across all considered metrics. Interestingly, pre-training BYOL-A models on only speech samples of AudioSet (BYOL-S) led to an almost 3% accuracy improvement compared to the original method. Presented evaluation results can be used as baselines for developing novel approaches, such as CvT-based methods explored here. Future work should focus on incorporating more datasets in even more languages into SERAB, as well as extending the task range to include regression problems such as valence or arousal estimation. To facilitate the usage of SERAB, the framework, including setup instructions, evaluation pipelines & examples, is freely available online<sup>7</sup>.

<sup>7</sup><https://github.com/Neclow/serab/>

## 6. REFERENCES

- [1] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*, Wiley, Chichester, UK, 2013.
- [2] 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, pp. 1994, 2018.
- [3] I. Lefter, L. J. Rothkrantz, D. A. Van Leeuwen, and P. Wiggers, "Automatic stress detection in emergency (telephone) calls," *Int. J. Intell. Defence Support Syst.*, vol. 4, no. 2, pp. 148–168, 2011.
- [4] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: the Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia*, 2010, pp. 1459–1462.
- [5] M. Schmitt and B. Schuller, "openXBOW: Introducing the Passau open-source crossmodal bag-of-words toolkit," *J. Mach. Learn. Res.*, vol. 18, no. 96, pp. 1–5, 2017.
- [6] J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. C. Quitry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. Haviv, "Towards Learning a Universal Non-Semantic Representation of Speech," in *Proc. Interspeech*, 2020, pp. 140–144.
- [7] 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.
- [8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR. IEEE*, 2009, pp. 248–255.
- [9] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *ICLR*, 2018.
- [10] 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. IEEE*, 2017, pp. 776–780.
- [11] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al., "CNN architectures for large-scale audio classification," in *ICASSP. IEEE*, 2017, pp. 131–135.
- [12] Neural Audio AI, "HEAR 2021 NeurIPS Challenge Holistic Evaluation of Audio Representations," <https://neuralaudio.ai/hear2021-holistic-evaluation-of-audio-representations.html>.
- [13] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, 2014.
- [14] S. Haq, P. J. Jackson, and J. Edge, "Speaker-dependent audio-visual emotion recognition," in *AVSP*, 2009, pp. 53–58.
- [15] W. Fan, X. Xu, X. Xing, W. Chen, and D. Huang, "LSSSED: a large-scale dataset and benchmark for speech emotion recognition," in *ICASSP. IEEE*, 2021, pp. 641–645.
- [16] M. Plakal and D. Ellis, "YAMNet," <https://github.com/tensorflow/models/tree/master/research/audioset/yamnet>, 2020.
- [17] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.
- [18] N. Vryzas, R. Kotsakis, A. Liatsou, C. A. Dimoulas, and G. Kalliris, "Speech emotion recognition for performance interaction," *J. Audio Eng. Soc.*, vol. 66, no. 6, pp. 457–467, 2018.
- [19] P. Gournay, O. Lahaie, and R. Lefebvre, "A Canadian French emotional speech dataset," in *Proc. ACM Multimedia Systems*, 2018, pp. 399–402.
- [20] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *EU-ROSPEECH*, 2005.
- [21] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "EMOVO corpus: an Italian emotional speech database," in *LREC. ELRA*, 2014, pp. 3501–3504.
- [22] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS ONE*, vol. 13, no. 5, 2018.
- [23] O. M. Nezami, P. J. Lou, and M. Karami, "ShEMO: a large-scale validated database for persian speech emotion detection," *Lang. Resour. Eval.*, vol. 53, no. 1, pp. 1–16, 2019.
- [24] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2D continuous space," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 3–14, 2015.
- [25] L. Campbell, *Historical Linguistics: An Introduction*, Edinburgh University Press, Edinburgh, UK, 2013.
- [26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [27] B. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen, et al., "The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 cough, COVID-19 speech, escalation & primates," *arXiv preprint arXiv:2102.13468*, 2021.
- [28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.
- [29] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8M: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.
- [30] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
- [31] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, et al., "Bootstrap your own latent - a new approach to self-supervised learning," in *NeurIPS*, 2020, pp. 21271–21284.
- [32] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing convolutions to vision transformers," *arXiv preprint arXiv:2103.15808*, 2021.
- [33] X. Zhai, J. Puigcerver, A. Kolesnikov, P. Ruysen, C. Riquelme, M. Lucic, J. Djolonga, A. S. Pinto, M. Neumann, A. Dosovitskiy, et al., "A large-scale study of representation learning with the visual task adaptation benchmark," *arXiv preprint arXiv:1910.04867*, 2019.
- [34] G. Cazenavette and S. Lucey, "On the bias against inductive biases," *arXiv preprint arXiv:2105.14077*, 2021.
- [35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.